School of Civil, Environmental and Land Management Engineering
Master of science in civil engineering for risk mitigation

Urban Crash Hotspot Identification Using Network-Based Analysis: A Case Study of NoLo, Milan

Author: Zeinab Awada

Student ID: 10967603
Advisor: Prof. Lorenzo Mussone
Co-advisor: Prof. Roberto Notari
Academic Year: 2025-2026

Abstract

Urban environments face significant challenges in ensuring the safety of pedestrians, cyclists, and motor vehicle users due to increasing traffic demand, complex road networks, and the coexistence of multiple transportation modes. The NoLo (North of Loreto) district in Milan represents a particularly relevant case study because of its high levels of pedestrian activity, mixed traffic conditions, and dense urban road network. Understanding the spatial distribution of crashes and identifying future high-risk locations are essential for supporting effective road safety policies and targeted interventions.

This thesis is composed of two complementary parts aimed at analyzing and predicting urban traffic crashes involving pedestrians, bicycles, and other vehicles within the NoLo district.

The first part presents a comprehensive descriptive analysis of five years of crash data (2017–2021). The analysis investigates the spatial, temporal, environmental, and behavioral characteristics of crashes, highlighting differences among pedestrian, bicycle, and vehicle incidents. Particular attention is given to crash distribution patterns, road characteristics, user behavior, vehicle involvement, and the relationship between crashes and pedestrian crossing facilities. The results reveal clear spatial concentrations of crashes along the main urban corridors of the district and confirm the importance of localized interaction points such as intersections and crosswalks in shaping crash occurrence.

The second part develops a network-based framework for crash prediction and hotspot identification. The road network is represented as a graph and several node-weighting approaches are evaluated, including Raw Data indicators and the ICENTR centrality indicators. Hierarchical clustering techniques are then applied to partition the network into homogeneous clusters and generate probability-based crash prediction models for pedestrians, bicycles, and vehicles separately. The models are evaluated using prediction accuracy, Crash Prediction Ratio (CPR), Network Coverage Ratio (NCR), Crash Prediction Efficiency (CPE), and Homogeneity Ratio (HR).

The results demonstrate that crash-based node-weighting approaches provide the most reliable prediction performance within the NoLo district. The final selected models are RN–11 for pedestrian crashes, RE–5 for bicycle crashes, and RN–10 for vehicle crashes. These models successfully identify the most critical crash-prone areas and provide a balanced combination of prediction accuracy, hotspot concentration, and spatial consistency. The findings indicate that localized exposure conditions and recurring crash concentrations play a greater role in determining crash occurrence than network hierarchy alone.

The proposed methodology provides a practical decision-support tool for urban planners and transportation authorities by identifying priority locations for safety interventions. More broadly, the framework offers a transferable approach for crash analysis, hotspot identification, and risk-based urban safety planning in complex urban environments.

Acknowledgment

This thesis marks the completion of an important chapter in my academic and personal development, one that would not have been possible without the guidance, encouragement, and support of many people.

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Lorenzo Mussone, for his continuous guidance, insightful feedback, and invaluable support throughout this research. His expertise, humility, and dedication were a constant source of motivation. I am especially grateful for his prompt feedback, availability, and genuine interest in helping me overcome challenges and improve the quality of this work.

I would also like to extend my appreciation to Professor Roberto Notari for his valuable observations and continuous support throughout the development of this work, which contributed significantly to improving the quality of this thesis.

I would also like to thank the professors and academic staff of Politecnico di Milano, whose teaching, support, and commitment to excellence have greatly enriched my educational experience and contributed to my growth both academically and professionally.

To my family, I owe my deepest gratitude. To my mother, who has always been my greatest source of strength and unconditional love, this achievement belongs to you as much as it belongs to me. Your sacrifices, resilience, and unwavering belief in me have guided me through every challenge and encouraged me to pursue my goals with confidence.

To my fiancé, thank you for walking beside me throughout this journey with patience, understanding, and endless kindness. Your presence brought comfort during stressful moments and perspective during difficult times. Whether through your words of reassurance or simply being there when I needed support, you helped make this experience more manageable and meaningful.

I would also like to thank my friends and colleagues, especially Eng. Youssef Beik, whose support, encouragement, and friendship made this academic journey smoother and more memorable.

Finally, I would like to express my gratitude to everyone who has contributed, directly or indirectly, to my academic and personal journey. This accomplishment is not mine alone, but a reflection of all those who stood beside me with guidance, support, and love. To each of you, I offer my heartfelt thanks.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1. Road Traffic Safety: Global Perspective and Local Context

Road traffic crashes represent one of the most pressing global public health challenges, causing significant human suffering and substantial social and economic losses each year. The World Health Organization (2023) reported that approximately 1.19 million people die on the world's roads annually, placing road traffic injuries among the leading causes of death across all age groups. Urban environments are particularly critical, accounting for an estimated 54–60% of all fatal road crashes, largely due to high traffic volumes, complex road layouts, and the frequent and often unpredictable interactions between different categories of road users.

Some road users face a greater risk of death in road crashes. They are commonly referred to as Vulnerable Road Users (VRUs) and include pedestrians, cyclists, micro-mobility users, roadway workers, and other road users who are not protected by a vehicle. The World Health Organization (2023) reported that VRUs account for nearly 50% of all road deaths, a proportion that rises even further in urban settings where exposure to motorized traffic is greatest. The European Commission (2022) reported that, in the European Union alone, pedestrians accounted for 3,715 fatalities in 2022, with car–pedestrian collisions responsible for 2,409 deaths, while cyclists recorded 2,009 fatalities, of which 921 resulted from collisions with passenger cars (Figure 1.1). This elevated risk is largely attributable to the absence of physical protection for these users, which significantly increases their vulnerability to severe or fatal injuries in collisions with motorized vehicles.

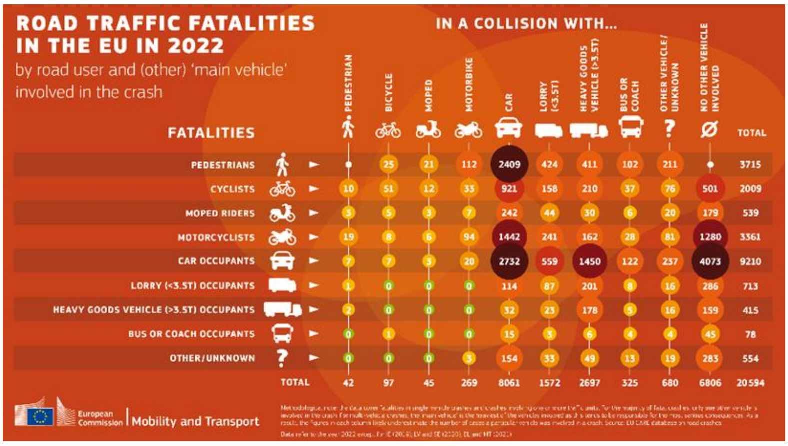

Figure 1. 1 EU road traffic fatalities by road user type, 2022

ISTAT (2022) reported that approximately 73–75% of all road crashes in Italy occur on urban roads, where interactions between vehicles and VRUs are most frequent. ISTAT (2022) also reported that pedestrians and cyclists account for around 35–40% of total road fatalities nationwide, with even higher proportions recorded in large metropolitan areas. Such statistics underline the urgent need for more targeted, network-aware safety approaches in urban environments.

As urbanization continues to intensify globally, improving road safety in cities has become a central objective of transport and urban planning policies. At the European level, this is reflected in the Vision Zero initiative, which aims to eliminate all road fatalities and serious injuries. Milan exemplifies both the scale of these challenges and the ambition to address them, promoting sustainable mobility through the Piano Urbano della Mobilità Sostenibile (PUMS) (Comune di Milano, 2020), which aims to reduce traffic-related fatalities while encouraging walking, cycling, and public transport use. While these policies have increased active mobility, they have also intensified interactions between different road users within compact urban districts characterized by high demand and limited space, making a detailed, network-based understanding of crash risk within individual urban districts both necessary and urgent.

Within this context, the NoLo (Nord di Loreto) district in Milan represents a particularly suitable case study for network-based urban crash analysis (Figure 1.2). Located north of Piazzale Loreto and shaped by major urban corridors such as Viale Monza and Via Padova, NoLo combines a dense local street network with mixed residential and commercial land uses and sustained pedestrian and cyclist activity.

Figure 1. 2 Location of the NoLo district within the city of Milan

The district is home to approximately 27,000 residents, with high population density and strong social diversity, with foreign residents accounting for around one-third of the total population. ISTAT (2023) reported that the demographic profile is dominated by young and working-age residents alongside long-established migrant communities. In recent years, rapid gentrification has increased residential attractiveness while further intensifying daily mobility demand. The urban structure, characterized by compact residential blocks, narrow local streets, high-capacity arterial roads, and fragmented cycling infrastructure, generates frequent interactions between vulnerable road users and motorized traffic, particularly at intersections and near transit access points (Comune di Milano, 2020).

The two main corridors passing through the district, Via Padova and Viale Monza (Figure 1.3), rank among the most dangerous roads in Milan for pedestrians and cyclists. According to the AMAT analysis (2017–2021), Via Padova recorded 479 crashes involving 105 pedestrians and 90 cyclists, while Viale Monza recorded 520 crashes with 48 pedestrians and 79 cyclists involved. During the period 2017–2021, 495 crashes were recorded in the NoLo district, including 93 pedestrian and 84 bicycle crashes. This confirms the high exposure of VRUs in the district and justifies the focus of this research on network-based crash analysis at the micro-scale.

Figure 1.3 Mixed traffic conditions on Viale Monza, Milan [15]

1.2. Thesis Approach

Pedestrian and cyclist safety in urban environments remains a critical public health concern, requiring advanced analytical frameworks capable of identifying high-risk locations with greater precision than traditional methods allow. This thesis proposes a network-based methodology that integrates spatial analysis with advanced centrality indices to improve the identification and prediction of crash concentrations through a robust predictive framework for crashes involving pedestrians, ciyclist, and motor vehicles within the urban road network of the NoLo district in Milan.

The study utilizes a detailed dataset of geo-localized crashes collected by the local municipality police over a five-year observation period (2017–2021). These crashes are mapped onto the underlying road network, and specific weights are assigned to each road segment and intersection based on crash occurrence through a dedicated methodology that integrates road network topology with crash data. A recently developed centrality index, ICENTR (Mussone et al., 2021), is employed to provide a more detailed understanding of network connectivity and the spatial relationships influencing crash occurrence

Through hierarchical clustering techniques, the study addresses the over-dispersed nature of crash data by grouping network elements into spatially coherent clusters that facilitate more reliable analysis and prediction. Probability density functions (PDFs) are then generated from these clusters to simulate and predict crash occurrence patterns, providing quantitative insights to support urban planners and traffic safety authorities in prioritizing safety interventions.

MATLAB was used to process the network and crash datasets, implement the centrality-based methodologies, perform hierarchical clustering analyses, and evaluate prediction performance through validation against observed crash data.

Overall, the thesis is structured around two main parts. The first part involves a multivariate analysis of five years of crash data, systematically examining the relationships between crash variables and crash occurrence within the NoLo road network. By identifying the spatial, structural, and temporal factors most strongly associated with crash frequency and severity, this component aims to develop a deeper understanding of why, when, and how crashes concentrate across the district.

The second part focuses on the development of predictive models for pedestrian, bicycle, and vehicle crashes separately. These models are designed to estimate crash concentration at the network level, enabling the identification of high-risk locations.

This research aims to contribute to urban traffic safety and transportation planning within the NoLo district by supporting targeted interventions to reduce crash occurrence and improve the safety of vulnerable road users. The proposed framework demonstrates the effectiveness of combining spatial analysis, graph theory, and clustering techniques for crash prediction and hotspot identification and can be adapted for similar studies in other urban areas.

Nevertheless, some limitations should be acknowledged, including the absence of traffic flow and exposure data, the lack of detailed infrastructure characteristics, and possible geolocation inaccuracies within the crash database that may affect crash-to-network assignment precision.

2 Review of Related Literature

This chapter reviews the literature relevant to the objectives of this thesis, focusing on the factors influencing urban traffic crashes and the analytical methods used to investigate them. The review is organized into two main parts. The first examines the categorization and analysis of crash data, with particular emphasis on the spatial, temporal, environmental, and behavioral factors associated with pedestrian, bicycle, and vehicle crashes. The second reviews predictive models and network-based approaches used in transportation safety research, including probability models, crash weighting methods, centrality measures, clustering techniques, and model validation approaches for crash analysis and prediction.

2.1. Urban Crash Patterns and Risk Factors

2.1.1. Spatial and Topological Aspects

Spatial and topological characteristics are fundamental for understanding the distribution of pedestrian and bicycle crashes within urban road networks. While spatial characteristics describe where crashes occur, topological characteristics explain how the configuration and connectivity of the road network influence crash patterns. Together, these factors support the identification of high-risk locations and the development of effective road safety strategies.

ZIAKOPOULOS & YANNIS (2020), through a review of spatial road safety studies, found that traffic crashes are unevenly distributed across urban environments and tend to concentrate in locations with intense interactions between vehicles, pedestrians, and cyclists. Their findings indicate that urban centers, major intersections, commercial districts, and public transport hubs consistently exhibit higher crash frequencies because these locations concentrate multiple traffic movements and interactions among different road users. Infrastructure deficiencies, including poorly designed intersections, inadequate pedestrian crossings, and insufficient signal timing, further increase crash risk by creating additional conflict points (KIM ET AL., 2010; MA ET AL., 2022).

ZIAKOPOULOS & YANNIS (2020) reviewed the application of GIS-based methods, including hotspot analysis, kernel density estimation, and spatial autocorrelation techniques, for identifying crash hotspots and supporting evidence-based road safety planning. The review identified crash concentrations near commercial areas, shopping centers, bus stops, and public transport hubs, particularly where complex road geometries or inadequate pedestrian facilities exist. For cyclists, the availability and continuity of dedicated bicycle infrastructure are critical determinants of safety. Discontinuous or poorly integrated bicycle lanes increase interactions with motor vehicles and consequently the likelihood of crashes (MA ET AL., 2022; THEOFILATOS & YANNIS, 2014).

Beyond the spatial distribution of crashes, the characteristics and connectivity of the road network also influence crash frequency and severity. MUSSONE ET AL. (2017), using Principal Component Analysis (PCA) and Self-Organizing Maps (SOM), highlighted the influence of roadway characteristics on crash severity at urban intersections, demonstrating the importance of infrastructure characteristics in road safety analysis. Their findings indicate that roads with poor pedestrian integration, limited crossing opportunities, and inadequate safety facilities increase the exposure of vulnerable road users to vehicular traffic. Similarly, discontinuous or poorly separated bicycle lanes force cyclists to travel within mixed traffic, increasing conflicts with motor vehicles (MA ET AL., 2022). Road hierarchy is another important factor, as arterial roads with multiple lanes and higher operating speeds generally present greater risks than local streets because of their limited pedestrian and cyclist facilities (AGUERO-VALVERDE & JOVANIS, 2006). Improving network connectivity through dedicated cycling infrastructure, safe pedestrian crossings, and appropriate traffic management measures can therefore contribute to reducing conflicts and enhancing road safety (ELVIK ET AL., 2009).

BASILE ET AL. (2010), through the application of the Analytic Hierarchy Process (AHP) to pedestrian crossings, identified intersections as critical locations where inadequate traffic control substantially increases crash risk. Their study demonstrated that properly designed signalized intersections with visible crosswalks, pedestrian refuges, and effective pedestrian signals can significantly improve pedestrian safety. HUANG & ABDEL-ATY (2010), using Bayesian hierarchical modelling, further demonstrated how multilevel statistical models can account for spatial heterogeneity and roadway characteristics when evaluating crash risk at intersections. MA ET AL. (2022) also highlighted that improving pedestrian crossing facilities remains one of the principal strategies reported in the literature for reducing pedestrian crashes. In contrast, non-intersection locations, particularly midblock crossings, present different safety challenges. Where formal crossing facilities are unavailable, pedestrians often cross unpredictably, increasing crash risk, especially on roads with higher traffic speeds (ZIAKOPOULOS & YANNIS, 2020). These findings demonstrate that effective safety strategies should consider the distinct characteristics of both intersections and non-intersections to better protect all road users (THEOFILATOS & YANNIS, 2014).

2.1.2. Temporal Patterns

Temporal characteristics, including the time of day and day of the week, provide valuable insights into crash occurrence and help identify periods associated with elevated crash risk. Understanding temporal variations supports the development of targeted road safety measures for different groups of road users.

KIM ET AL. (2010), using mixed logit models to analyse pedestrian injury severity, and THEOFILATOS & YANNIS (2014), through a review of the relationship between traffic flow and crash occurrence, demonstrated that both crash frequency and severity vary considerably throughout the day. Their findings indicate that the highest crash frequencies generally occur during peak traffic periods, when increased traffic volumes, pedestrian activity, and reduced visibility create complex traffic conditions. HUANG & ABDEL-ATY (2010), applying Bayesian hierarchical modelling, further showed that nighttime conditions are associated with more severe crashes because limited visibility and higher vehicle speeds increase the risk for vulnerable road users. The transition from daylight to darkness during late afternoon and early evening also represents a critical period, as heavy traffic demand coincides with deteriorating lighting conditions. These findings support the implementation of time-dependent safety measures, including improved street lighting, enhanced road markings, and adaptive traffic control systems to reduce crash risk during high-risk periods (KIM ET AL., 2010; THEOFILATOS & YANNIS, 2014).

Temporal variations are also evident across different days of the week. MA ET AL. (2022), through a scientometric review of pedestrian safety research, highlighted temporal patterns as one of the principal themes investigated in road safety studies, while YU & ABDEL-ATY (2014) analysed crash data to examine the influence of traffic conditions on crash severity. Their findings indicate that weekday crashes occur more frequently during commuting hours, particularly in business districts, school zones, and public transport hubs. Although these crashes are generally less severe because congestion reduces operating speeds, their frequency remains relatively high. In contrast, crashes occurring during weekends are generally more severe because of higher travel speeds, alcohol consumption, and distracted driving, especially in recreational and nightlife areas (YU & ABDEL-ATY, 2014). These temporal variations indicate that traffic management and enforcement strategies should be adapted to the characteristics of different periods of the week, combining increased enforcement and public awareness campaigns during weekends with traffic flow improvements during weekday peak periods.

2.1.3. Weather Conditions

Weather conditions are important environmental factors that influence both the frequency and severity of traffic crashes. Variations in visibility, pavement conditions, and road user behavior under different weather conditions substantially affect the safety of motorists, pedestrians, and cyclists.

ZHAI ET AL. (2019), through an analysis of weather-related crash data, and MA ET AL. (2022), through a scientometric review of pedestrian safety research, demonstrated that adverse weather conditions, including rain, snow, and fog, increase crash risk by reducing visibility and deteriorating pavement conditions. Their findings indicate that wet surfaces reduce tire-road friction and braking efficiency, while fog and heavy rainfall impair drivers' visibility, particularly at intersections and complex road layouts where rapid decision-making is required. These conditions are especially hazardous in urban environments because poor visibility and reduced pavement friction increase the vulnerability of pedestrians and cyclists (ZIAKOPOULOS & YANNIS, 2020).

JUNG ET AL. (2010), using polychotomous response models to analyse the effects of rainfall on crash severity, and BERGEL-HAYAT ET AL. (2013), using statistical analyses of crash records under different weather conditions, showed that the influence of adverse weather varies across regions. Their findings indicate that sudden rainfall following prolonged dry periods can substantially increase crash occurrence because drivers often fail to adapt to rapidly changing road conditions. In addition to motorists, pedestrians and cyclists face increased risks as snow and ice create slippery walking and cycling surfaces, while longer stopping distances reduce drivers' ability to avoid collisions (THEOFILATOS & YANNIS, 2014; MA ET AL., 2022). ZIAKOPOULOS & YANNIS (2020) further highlighted that integrating weather information with spatial analyses enables the identification of weather-related crash hotspots and supports proactive road safety planning.

MA ET AL. (2022), through a scientometric review of pedestrian safety research, and ANTONIOU ET AL. (2013), through analyses of seasonal traffic patterns, demonstrated that seasonal changes also influence traffic crash occurrence by altering both environmental conditions and travel behavior. Their findings indicate that winter is generally associated with higher overall crash rates because snow, ice, and reduced daylight hours create hazardous driving and walking conditions. In contrast, although total crash frequencies often decline during summer, crashes involving pedestrians and cyclists may increase because warmer weather encourages greater outdoor activity and active transportation.

BERGEL-HAYAT ET AL. (2013) further showed that transitions between seasons increase crash risk because changes from dry to wet pavement conditions often catch road users unprepared. Consequently, seasonal road safety strategies should combine infrastructure maintenance, such as timely snow removal and road treatment during winter, with public awareness campaigns and targeted traffic management measures during periods of increased pedestrian and cyclist activity. ANDREY ET AL. (2003) also demonstrated that analysing seasonal crash patterns supports the identification of high-risk locations and periods, allowing safety interventions to be adapted to changing environmental conditions throughout the year.

2.1.4. Circumstances of Crashes

The circumstances surrounding a traffic crash, including the types of vehicles involved and the behavior of pedestrians and cyclists, play an important role in determining both crash occurrence and injury severity. Understanding these factors provides valuable insights for developing targeted road safety measures.

KIM ET AL. (2010), using mixed logit models to analyse pedestrian injury severity, and SIDDIQUI ET AL. (2012), through a macroscopic spatial analysis of pedestrian and bicycle crashes, demonstrated that vehicle type is an important determinant of crash severity, particularly for vulnerable road users. Their findings indicate that heavy vehicles, such as trucks and buses, are associated with higher fatality rates because of their greater mass, longer stopping distances, and limited maneuverability, whereas passenger cars generally result in less severe injuries but remain the vehicle type most frequently involved in pedestrian crashes because of their predominance in urban traffic. MA ET AL. (2022) further highlighted that motorcycles present additional risks because their relatively high operating speeds and maneuverability reduce the time available for pedestrians to react.

ZIAKOPOULOS & YANNIS (2020), through a review of road safety studies, highlighted that bicycle crashes generally result in less severe injuries than motor vehicle crashes but can still cause significant harm, particularly in shared spaces and at locations where bicycle lanes intersect pedestrian paths. MA ET AL. (2022) further reported that public transport vehicles, especially buses, are frequently involved in crashes near bus stops and pedestrian crossings, where boarding, alighting, and high pedestrian activity increase the potential for conflicts.

BASILE ET AL. (2010), through the application of the Analytic Hierarchy Process (AHP) to pedestrian crossings, demonstrated that pedestrian behaviour is a key factor influencing crash occurrence. Their findings indicate that unsafe pedestrian behaviours, including jaywalking, ignoring traffic signals, and crossing outside designated locations, substantially increase crash risk, particularly at non-signalized crossings. THEOFILATOS & YANNIS (2014), through a review of traffic flow and crash occurrence, further highlighted that cyclists are more likely to be involved in crashes when riding in mixed traffic without complying with traffic regulations or when interactions with pedestrians are poorly managed.

SCHWEBEL ET AL. (2012), through a review of behavioural risk factors affecting pedestrian safety, identified distraction caused by mobile device use as an important contributor to pedestrian crashes because it reduces situational awareness and delays reaction times in complex urban environments. These findings emphasize that improving road safety requires not only appropriate infrastructure but also education, enforcement, and greater awareness to encourage safer interactions between pedestrians, cyclists, and motor vehicles (CLIFTON ET AL., 2009; PUCHER & DIJKSTRA, 2003).

2.2. Predictive Models and Network-Based Approaches

2.2.1. Probability Models for Crashes

Probability models have been widely applied in traffic safety research to estimate crash occurrence, identify contributing factors, and support the development of effective safety interventions. These models relate crash frequency or severity to explanatory variables such as traffic volume, roadway characteristics, environmental conditions, and human behavior.

Classical statistical models, particularly Poisson regression, have been extensively used to analyze crash frequency because traffic crashes are discrete and relatively infrequent events. KRAIDI & EVDORIDES (2020) applied Poisson regression to estimate pedestrian crash risk in urban environments, identifying traffic volume, speed variation, and pedestrian violations as significant predictors. A different approach was proposed by BASILE, PERSIA, & USAMI (2010), who applied the Analytic Hierarchy Process (AHP) to evaluate pedestrian crossing safety by integrating infrastructure characteristics such as visibility, road width, and pedestrian refuges into a composite safety index. Unlike traditional crash-frequency models, this approach enables safety assessment even where detailed traffic data are unavailable.

To overcome the limitations of conventional statistical models, Bayesian and spatial approaches have been developed to account for spatial dependence and unobserved heterogeneity. HUANG & ABDEL-ATY (2010) introduced a Bayesian hierarchical framework capable of incorporating multilevel spatial and temporal information, thereby improving crash prediction in complex urban environments. AGUERO-VALVERDE & JOVANIS (2006), through spatial analysis of fatal and injury crashes, demonstrated the importance of incorporating spatial dependence into crash models. ZIAKOPOULOS & YANNIS (2020), through a review of spatial road safety studies, further emphasized that spatial models provide a more detailed understanding of crash risk by capturing variations associated with road density, network configuration, and proximity to high-risk locations.

Recent studies have expanded probability modeling by integrating behavioral analysis, simulation techniques, and machine learning methods. MA ET AL. (2022), through a scientometric review of pedestrian safety research, reported an increasing use of simulation models to reproduce pedestrian–vehicle interactions and evaluate safety measures before implementation. THEOFILATOS & YANNIS (2014), through a review of traffic flow and road safety studies, demonstrated that the relationship between traffic flow and crash occurrence is often non-linear, supporting the need for more flexible modeling approaches beyond traditional linear assumptions. Similarly, SHRINIVAS ET AL. (2023), through a systematic review and meta-analysis of pedestrian injury severity studies, highlighted the application of logistic regression, Bayesian hierarchical models, and other statistical approaches to evaluate pedestrian injury severity using a comprehensive range of behavioral, environmental, and traffic-related variables. MUSSONE & ALIZADEH MEINAGH (2023) compared Artificial Neural Networks (ANNs), Generalized Linear Mixed Effects (GLME), and Multinomial Regression (MNR), concluding that while ANNs achieve superior predictive performance for complex non-linear relationships, statistical models remain valuable because of their greater interpretability. Overall, recent developments indicate that combining statistical, spatial, and machine learning approaches provides a more comprehensive framework for crash prediction and urban road safety analysis.

2.2.2. Network-Based Crash Weighting

Network-based crash modelling has become an increasingly important approach in transportation safety research because it explicitly considers the structural characteristics of road networks. PORTA, CRUCITTI, & LATORA (2006), through graph-theoretic analysis of urban street networks, demonstrated that representing road systems as graphs composed of nodes and edges enables crash occurrence to be analysed in relation to network connectivity, accessibility, and topology rather than solely through geographic location. MARSHALL & GARRICK (2011), through a statistical analysis of street network design and crash rates, demonstrated that road network configuration plays an important role in crash occurrence.

Crash records are generally available as geographic coordinates and therefore must first be associated with network elements before graph-based analyses can be performed. OKABE & SUGIHARA (2012), through the development of network-based spatial analysis methods, described several crash-to-network assignment procedures, including nearest-node allocation, nearest-edge allocation, and orthogonal projection. XIE & YAN (2013) demonstrated that representing crashes directly in network space improves the analysis of traffic accident patterns compared with conventional planar methods, highlighting the advantages of network-based representations for transportation safety analysis.

PORTA, CRUCITTI, & LATORA (2006) introduced a graph-theoretic representation of urban street networks in which roads are represented as edges and intersections as nodes. Building upon this framework, crash information can be incorporated into the network through edge-based or node-based weighting schemes. Edge-based weighting assigns crashes to road segments, providing a direct representation of crash concentration along traffic corridors, whereas node-based weighting associates crashes with intersections and junctions where traffic conflicts are more likely to occur. Because many graph-theoretic measures are calculated at the node level, node-based weighting facilitates the integration of crash information with graph-theoretic analyses and supports the investigation of relationships between network structure and road safety. Overall, the reviewed literature demonstrates that integrating crash information with graph-based representations of road networks enables analyses that simultaneously consider crash occurrence and network structure. Consequently, network-based crash weighting provides the foundation for subsequent applications involving centrality measures, clustering techniques, and predictive modelling.

2.2.3. Centrality Indices

PORTA, CRUCITTI, & LATORA (2006), through graph-theoretic analysis of urban street networks, demonstrated that centrality indices provide quantitative measures of the structural importance of nodes and edges within transportation networks. Their work showed that these measures are valuable for analysing network connectivity and accessibility. ZIAKOPOULOS & YANNIS (2020), through a review of spatial road safety studies, further highlighted that centrality measures are useful for identifying locations that are potentially associated with elevated crash risk.

Betweenness Centrality. FREEMAN (1977) introduced betweenness centrality as a measure of the extent to which a node lies on the shortest paths connecting other nodes within a network. Nodes with high betweenness act as important connectors because a large proportion of network movements passes through them. ZIAKOPOULOS & YANNIS (2020), through a review of spatial road safety studies, further identified betweenness centrality as a useful indicator for recognising potentially high-risk locations within urban road networks.

Closeness Centrality. PORTA, CRUCITTI, & LATORA (2006), through graph-theoretic analysis of urban street networks, described closeness centrality as a measure of how easily a node can reach all other nodes in a network through the shortest paths. Their graph-theoretic analysis showed that nodes with high closeness are generally more accessible and contribute to efficient movement across the network. ZIAKOPOULOS & YANNIS (2020) further highlighted that highly accessible intersections often attract greater traffic demand, making closeness centrality a useful indicator for assessing network accessibility and identifying potentially critical locations.

Recent developments have extended traditional centrality concepts through the introduction of new graph-based indices. MUSSONE ET AL. (2021) introduced PathRank, which evaluates node importance by considering the contribution of multiple weighted paths throughout the network rather than relying solely on shortest paths. The same study also introduced ICENTR, a centrality measure that incorporates both node and edge weights to provide a more comprehensive representation of network importance. By integrating structural connectivity with weighted network information, these indices enhance the analysis of complex transportation systems and support advanced applications in crash analysis, clustering, and transportation safety assessment.

2.2.4. Clustering Techniques

ZIAKOPOULOS & YANNIS (2020), through a review of spatial road safety studies, demonstrated that clustering techniques are valuable for identifying groups of road segments, intersections, or crash events that share similar characteristics. By grouping locations with comparable roadway, traffic, or crash patterns, these techniques facilitate the identification of high-risk areas and support the development of targeted safety interventions.

KRAIDI & EVDORIDES (2020) divided urban roads into short segments to evaluate pedestrian crash risk at a local scale, enabling more detailed analyses of roadside activities and traffic conditions. BASILE, PERSIA, & USAMI (2010) evaluated pedestrian crossings using the Analytic Hierarchy Process (AHP) based on infrastructure characteristics such as visibility, road width, and pedestrian refuges. Although they employed different methodologies, both studies demonstrated the value of partitioning transportation systems into comparable spatial units to support localized safety assessment and decision-making.

MUSSONE ET AL. (2017) employed Principal Component Analysis (PCA) and Self-Organizing Maps (SOM) to analyse crash characteristics and classify variables associated with crash severity, including roadway features, vehicle types, and driver characteristics. Their study demonstrated that these techniques reduce the complexity of large datasets while highlighting the factors that contribute most to crash occurrence.

MUSSONE & ALIZADEH MEINAGH (2023) combined machine learning models with oversampling methods to address class imbalance in crash datasets, improving the prediction of severe crashes. Their findings indicated that advanced data-driven approaches enhance the identification of crash-prone scenarios and support more reliable safety analyses. ZIAKOPOULOS & YANNIS (2020) further emphasized that grouping locations with similar crash characteristics enables more efficient allocation of road safety interventions.

As demonstrated by ZIAKOPOULOS & YANNIS (2020) and MUSSONE ET AL. (2017), clustering techniques provide an effective framework for organizing crash data according to shared roadway, traffic, and safety characteristics. By reducing the complexity of large crash datasets and revealing underlying patterns, these methods support subsequent predictive analyses and contribute to more effective transportation safety planning.

2.2.5. Model Validation and Performance Assessment

KRAIDI & EVDORIDES (2020), through the application of Poisson regression to pedestrian crash prediction, demonstrated that the evaluation of crash prediction models is essential for assessing their reliability, predictive performance, and applicability in transportation safety research. Their study assessed model performance using the Likelihood Ratio Test, Akaike Information Criterion (AIC), and Mean Squared Prediction Error (MSPE), highlighting the importance of balancing model fit with predictive capability. BASILE, PERSIA, & USAMI (2010) evaluated the consistency of their Analytic Hierarchy Process (AHP) using the Consistency Ratio (CR) and sensitivity analysis, ensuring the reliability of the weighting process and the resulting safety rankings.

LORD & MANNERING (2010), through a comprehensive review of statistical crash frequency models, emphasized the importance of selecting appropriate goodness-of-fit statistics and model comparison criteria when evaluating crash prediction models, particularly for over-dispersed crash data. Their review highlighted the need to assess both model fit and predictive capability while accounting for unobserved heterogeneity.

KIM ET AL. (2010), using mixed logit models to analyse pedestrian injury severity, demonstrated how flexible statistical modelling frameworks can better capture the influence of multiple interacting factors on crash outcomes, while also highlighting the importance of appropriate model selection criteria when analysing crash severity.

SHRINIVAS ET AL. (2023), through a systematic review and meta-analysis of pedestrian injury severity studies, synthesized evidence from numerous pedestrian safety investigations to identify the variables that most consistently influence pedestrian injury severity. Their findings demonstrated the importance of comprehensive validation and comparative evaluation when interpreting crash prediction results across different modelling approaches.

Overall, the findings reported by the reviewed studies indicate that no single evaluation metric is sufficient to assess predictive performance comprehensively. Instead, combining goodness-of-fit measures, prediction accuracy indicators, and comparative evaluation criteria provides a more reliable basis for selecting appropriate crash prediction models and supporting transportation safety analysis (LORD & MANNERING, 2010).

3 Network-Based Crash Prediction Methodology

This chapter presents the research methodology adopted to analyze the spatial distribution of urban road crashes and develop predictive models for identifying high-risk locations within the NoLo district of Milan. The proposed methodology was based on a network-oriented approach that integrates crash data with the structural properties of the road network, enabling the computation of indicators that capture both crash distribution and network topology. The methodology consists of two preliminary steps applied once to the full dataset:

1. Road Network Preparation
2. Data Partitioning and Scenarios Definition

Following these, a series of analytical steps are applied for each defined scenario, including crash-to-network assignment, computation of crash-based and centrality indicators, hierarchical clustering, probabilistic crash prediction, and model evaluation.

The network-based methodology is particularly well suited to the NoLo district for three main reasons:

- The area is characterized by a high density of intersections and short road segments, which increases the importance of network topology in influencing crash occurrence.
- There are strong interactions between different road users, including vehicles, pedestrians, and cyclists, especially at intersections and crosswalks.
- Detailed and continuous traffic volume data, such as Average Annual Daily Traffic (AADT), are unavailable at the neighborhood level, limiting the applicability of traditional exposure-based crash analysis methods.

By relying on network structure and crash-based indicators rather than traffic flow data, the proposed methodology provides a robust framework for crash analysis and prediction in a dense urban environment where conventional data sources are incomplete.

3.1. Road Network Preparation

The road network of the NoLo district was obtained from OpenStreetMap (OSM) (Figure 3.1), an open and widely used spatial database for urban and transportation studies. OSM represents road networks using nodes, which store geographic coordinates, and ways, which are ordered sequences of nodes that describe road segments.

Figure 3. 1 Road network of NoLo extracted from OpenStreetMap

The network data were extracted using an automated query database through the Overpass API and subsequently imported into MATLAB for post processing (figure 3.2). In addition to road centerlines and intersections, point features representing pedestrian crosswalks were also retrieved. The extracted data were then spatially filtered using the NoLo boundary polygon to ensure that only roads and crosswalks located within the study area were retained. Figure 3.2 presents the network after extraction and import into MATLAB, prior to any topological simplification.

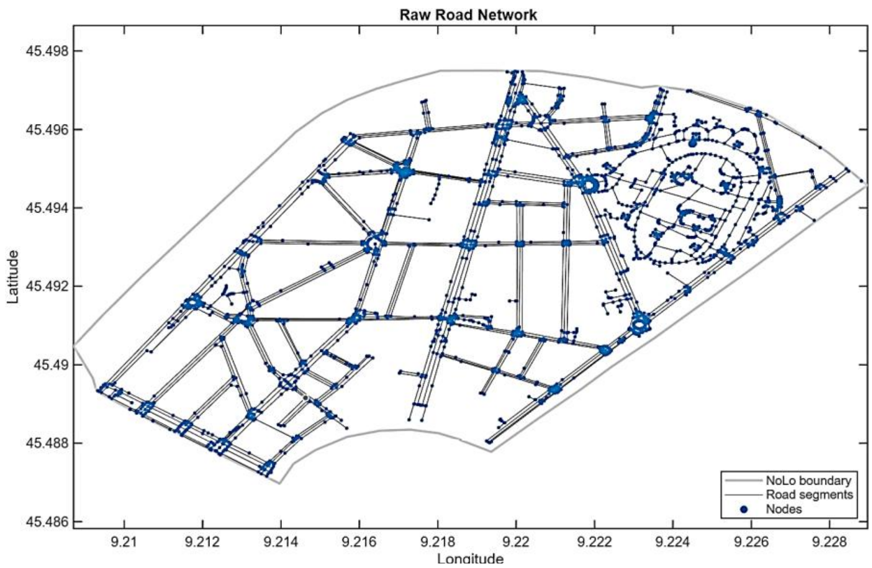

Figure 3. 2 Road network graph of NoLo district, prior to network refinement

Raw OSM data often contain mapping artefacts that can negatively affect network-based analyses if not properly addressed. Therefore, before analysis, the road network extracted from OpenStreetMap was refined and simplified to ensure geometric accuracy and topological consistency. The refinement procedure consisted of four steps:

1. duplicated nodes and edges were removed to avoid double counting of connections and distortions in connectivity measures.
2. Disconnected segments and isolated components not connected to the road network were also eliminated, as they do not contribute to through-movement or crash propagation within the study area.
3. OpenStreetMap frequently represents streets as sequences of short segments with intermediate vertices that do not correspond to actual intersections. These intermediate points, identified as degree-2 nodes, were removed by merging adjacent edges along continuous road segments. The degree of a node, defined as the number of nodes directly connected through an edge, provides fundamental structural information: nodes of degree 1 typically represent dead-ends, while nodes of degree 3 or 4 correspond to standard T-junctions or four-way intersections commonly found in dense urban street grids such as NoLo.
4. Very short or redundant edges were also removed unless they represented meaningful connectors.

After refinement and simplification, the final network consists of 118 nodes and 155 edges (Figure 3.3) The resulting graph was topologically coherent and ready for the subsequent steps.

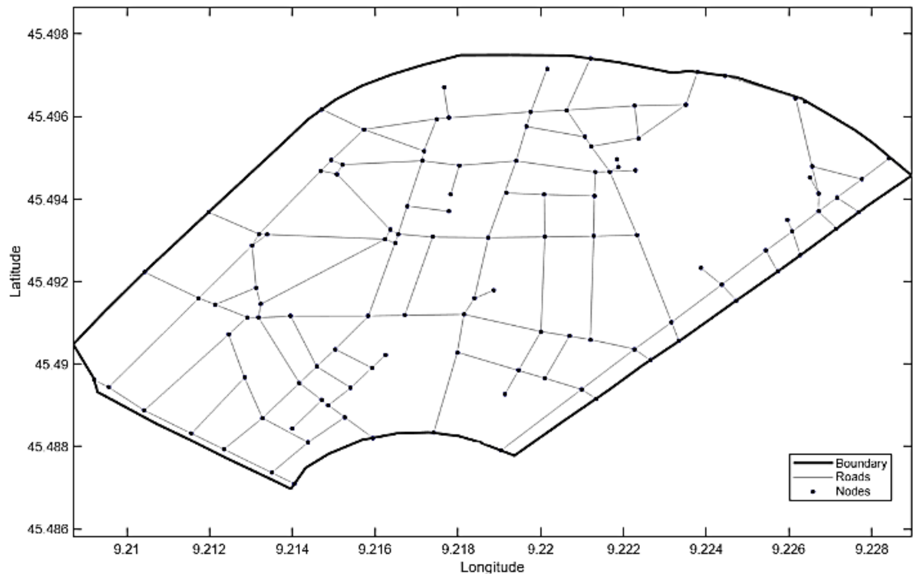

Figure 3. 3 Road network graph of NoLo , after network cleaning

Following network simplification, pedestrian crosswalks were integrated into the road network as independent spatial elements. In OpenStreetMap, crosswalks are represented as point features and were preserved in the dataset, as they constitute critical locations of interaction between pedestrians and vehicles. Figure 3.4 presents the road network with integrated crosswalks.

With the road network fully constructed and simplified, the next step involves the preparation and integration of the crash dataset. The crash dataset used in this thesis originates from the Archivio Crash NoLo (AIN), a municipal geodatabase containing police-reported road crashes for the Municipality of Milan. Each record includes geographic coordinates expressed in longitude and latitude (WGS84), the year of occurrence, and descriptive attributes such as crash type and severity.

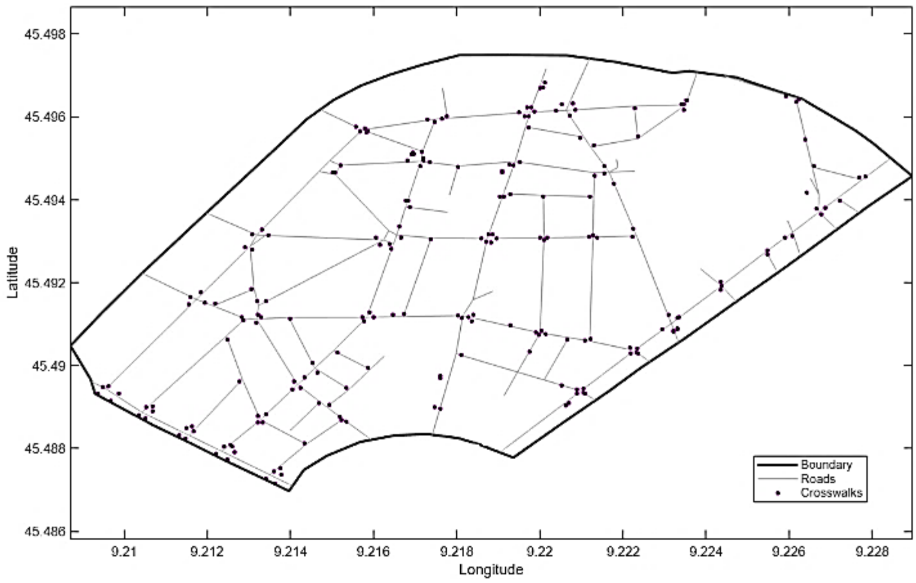

Figure 3. 4 Road Network Graph of NoLo with Crosswalks

Crash records from 2017 to 2021 were extracted to provide a multi-year representation of crash behavior within the NoLo district. This temporal window is sufficiently long to capture stable spatial patterns and recurring hotspots, while remaining consistent with the current configuration of the road network.

To ensure spatial consistency with the study area, all crash records were filtered using the official NoLo administrative boundary (boundary.mat). Records located outside the district were excluded.

After pre-processing, each crash record was defined by a spatial location, a temporal attribute, and a unique identifier. At this stage, crashes are not yet associated with the road network.

Figure 3.5 illustrates the crash dataset overlaid on the simplified road network of the NoLo district, including pedestrian crosswalks.

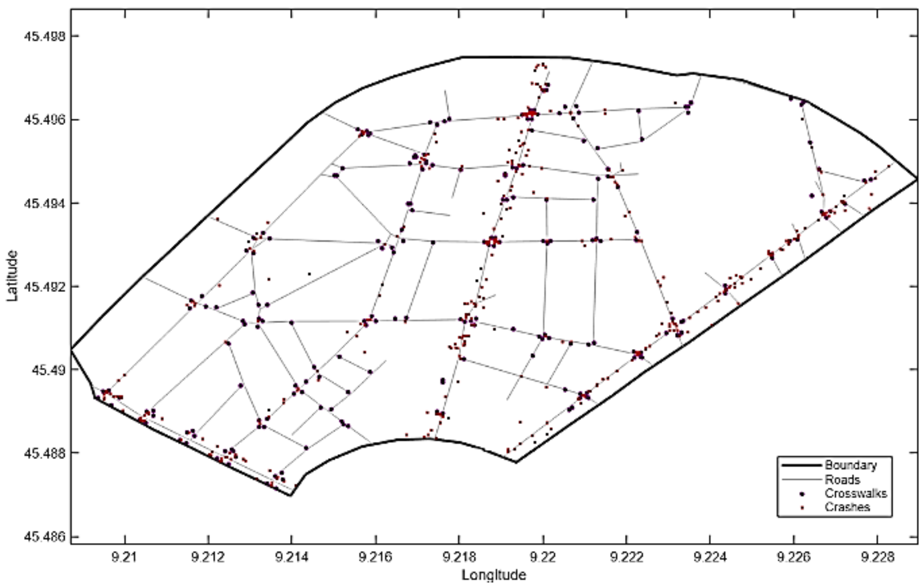

Figure 3. 5 Road Network Graph of NoLo with Crosswalks and Crashes

The processed crash dataset produced in this section serves as the direct input for the crash-to-network assignment procedure described in Section 3.3, where crashes were mapped onto the road network to support the computation of network-based indicators.

Having established a clean and topologically consistent network, the road system can be formally represented as a mathematical graph:

$$G = (V, E)$$

Where V is the set of nodes, and E is the set of edges.

In this representation, each node in V represents a road intersection and is identified by an integer, while each edge is a road segment connecting two adjacent intersections.

Pedestrian crosswalks are not included as graph nodes. Instead, their midpoint locations are stored separately and projected onto the nearest road segment during the crash-to-network assignment procedure described in Section 3.3.

All edges are treated as undirected, as the objective of the analysis is to examine network connectivity and topological structure rather than direction-dependent traffic flows.

The use of a cleaned and simplified network ensures that all subsequent analyses are based on a structure that accurately reflects real-world connectivity and interaction points within the NoLo district.

3.2. Data Partitioning

The crash dataset covering the full five-year period (2017–2021) was first merged into a single dataset for each of the three crash categories: pedestrian, bicycle, and other vehicle crashes. These were treated as distinct cases throughout the analysis, as it cannot be assumed that they follow the same underlying process or respond identically to network structure. The merged dataset for each crash type was then partitioned into six scenarios: five temporal scenarios and one random scenario.

In each temporal scenario, one year was designated as the testing period while the remaining four years were merged to form the training set. A rolling validation framework was adopted as follows:

- **Scenario 1 (S1):** Testing 2017 — Training 2018–2021
- **Scenario 2 (S2):** Testing 2018 — Training 2017 and 2019–2021
- **Scenario 3 (S3):** Testing 2019 — Training 2017–2018 and 2020–2021
- **Scenario 4 (S4):** Testing 2020 — Training 2017–2019 and 2021
- **Scenario 5 (S5):** Testing 2021 — Training 2017–2020
- **Random Scenario:** A randomly selected partition in which 20% of crashes are allocated to the testing set and the remaining 80% are used for training.

Across all scenarios, the topology of the NoLo road network remains fixed. Only the temporal allocation of crash events varies between training and testing sets.

For each scenario, the following steps are applied:

1. Crash Assignment to the Network: each crash was assigned to its spatially closest road segment. This procedure identifies the road segment on which the crash is located and its relative position along that segment.

2. Indices Computation: crash weights were computed for both the training and testing sets, and subsequently used to derive the following six indicators:

- Raw Edge Crash Weight (RE)
- Raw Node Crash Weight (RN)
- Icentr, Edge (IE)
- Icentr, Node (IN)
- Betweenness Centrality (BET)
- Closeness Centrality (CLS)

3. Hierarchical Clustering: clustering was applied to group network nodes into clusters. The number of clusters was explored over a range of 2 to 50, chosen to cover both fine-grained and coarse spatial segmentations of the network. The optimal number of clusters was determined based on prediction performance, as described in the following sections.

4. Probability Density Function: Probability density functions were derived from the crash distribution within each cluster, based exclusively on the training data. These functions were then used to predict crash occurrences within each cluster.

5. Model Evaluation: The predicted values were compared against observed crash counts in the testing set using Root Squared Error (RSE), Root Mean Square Error (RMSE), and the Modified Root Mean Square Error (RSMEM) as evaluation metrics.

6. Selection of Optimal Candidates: The optimal number of clusters, indicator, and scenario were selected through a systematic evaluation across all six scenarios, with particular focus on Scenario 5 (S5) and the Random Scenario (SR) as the most representative configurations. Scenario 5 reflects a realistic prediction framework by using the most recent crash data available for training and predicting the final year of the study period, while the Random Scenario provides a robust validation by randomly sampling crashes from all years without relying on a specific temporal ordering. Cluster numbers between 5 and 25 were evaluated for each indicator, and candidate solutions were identified based on the lowest RSMEmod value.

7. Optimal Configuration Selection: Two additional criteria were incorporated: the Crash Prevention Rate (CPR) and the Homogeneity Ratio (HR), described in the following sections. So, the final selection was based on the combined assessment of RSMEM, CPR, and HR. The configuration achieving the best overall performance across all criteria was selected as the optimal solution.

3.3. Crash & Crosswalk Assignment

3.3.1. Crash-to-Network Assignment

Crash locations recorded in the Archivio Crash NoLo (AIN) dataset were represented as geographic point coordinates (longitude, latitude). However, these points do not generally lie exactly on the road centerlines extracted from OpenStreetMap. To integrate crash data into the graph-based representation of the NoLo road network, each crash must be associated with the road segment to which it is spatially closest.

This association is performed using an orthogonal projection approach, which is widely adopted in network-based crash analysis. The objective of this procedure was to convert raw crash point data into a network-referenced representation by identifying:

- the road segment (edge) on which the crash is located
- the relative position of the crash along that segment

These outputs provide the foundation for all subsequent edge-based and node-based crash exposure measures.

Orthogonal Projection Method

A road segment e=(a,b) in the network is defined by two nodes:

- start node: a=(ax,ay)
- end node: b=(bx,by)

and a crash point with crash location: p=(px,py)

Step 1: Computation of the Projection Parameter

The relative position of the crash along the segment was determined by computing the normalized projection parameter *t*:

$$t = \frac{(p-a)\cdot(b-a)}{||b-a||^2}$$

This gives the relative position of the crash along the edge:

- t = 0 indicates that the crash is closest to node *a*
- t = 1 indicates proximity to node b
- 0 < t < 1 indicates that the crash projects to an interior point of the segment.
- t < 0 or t > 1, the orthogonal projection lies outside the road segment, indicating that the crash does not belong to that edge.

Step 2: Computation of the Projected Point

Once the value of *t* is known, the exact coordinates of the projected location on the road segment were obtained using:

$$p_{\text{proj}} = a + t(b-a)$$

This expression represents a linear interpolation between the two endpoints of the edge.

Step 3: Computation of the Squared Perpendicular Distance

The squared distance between the crash point and its projected point is:

$$d^2 = ||p - p_{\text{proj}}||^2$$

Using the squared distance avoids unnecessary square-root computations while preserving distance ranking.

Step 4: Edge Selection

For each crash, the projection procedure was applied to all edges in the network. The edge that minimizes the perpendicular distance was selected as the crash-bearing segment.

This process was implemented in MATLAB by iterating over all edges, computing the projection parameter t and the corresponding distance d^2, and storing the identifier of the selected edge.

Figure 3.6 presents the outcome of the crash-to-network assignment procedure, displaying only the projected crash locations aligned with the road network. This representation constitutes the network-consistent crash dataset used in all subsequent analyses.

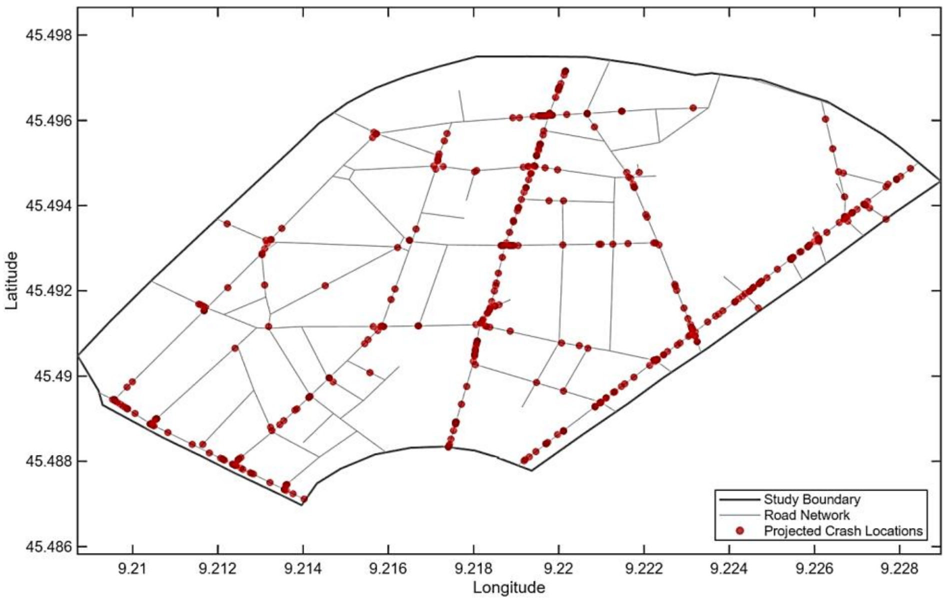

Figure 3. 6 Final crash location onto NoLo road network after orthogonal projection

3.3.2. Crosswalk-to-network assignment

Pedestrian crosswalks were represented as point features corresponding to their geometric centerpoints. As with crash locations, these points do not necessarily lie exactly on the road centerlines extracted from OpenStreetMap due to digitization offsets and geometric simplifications.

To ensure topological consistency between crashes, crosswalks, and the road network, each crosswalk was therefore assigned to its nearest road segment using the same orthogonal projection approach described in Section 3.3.1. For each crosswalk's centerpoint, the perpendicular distance to all road segments in the network was computed, and the segment yielding the minimum distance was selected as the associated road edge. Figure 3.7 illustrates the crosswalk-to-network assignment for the NoLo district.

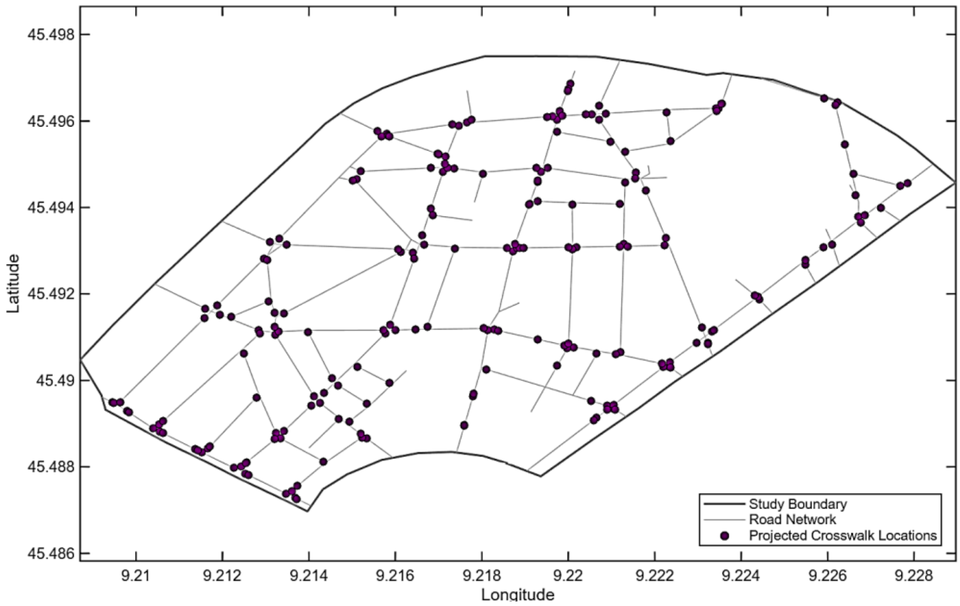

Figure 3. 7 Crosswalk-to-network assignment via orthogonal projection

3.3.3. Crash–Crosswalk Distance Computation

After both crash events and pedestrian crosswalks have been assigned to road segments through orthogonal projection, the spatial distance between crashes and crosswalks is computed under a strict same-edge constraint. Specifically, a crash was considered eligible for association with a crosswalk only if both elements are assigned to the same road segment in the network.

This constraint ensures topological consistency between crashes, crosswalks, and the road network, and prevents unrealistic associations across adjacent, parallel, or intersecting streets that were geometrically close in Euclidean space but disconnected in network terms.

For each crash event, the distance computation proceeded as follows:

1. The road segment (edge) to which the crash is assigned was identified.
2. All crosswalks assigned to the same road segment were selected as eligible candidates.
3. The nearest crosswalk on the same road segment was retained as the associated crossing facility.
4. If no crosswalk was present on the same road segment, the crash was classified as having no associated crosswalk.

The distance between each crash location and the associated pedestrian crosswalk midpoint was then computed using the Euclidean distance between their original spatial coordinates.

Figure 3.8 illustrates the crash–crosswalk distance computation for the NoLo district, with connecting lines representing the distance between each crash and its associated crosswalk. Crashes without an associated crosswalk on the same road segment were excluded from the distance computation and are shown separately.

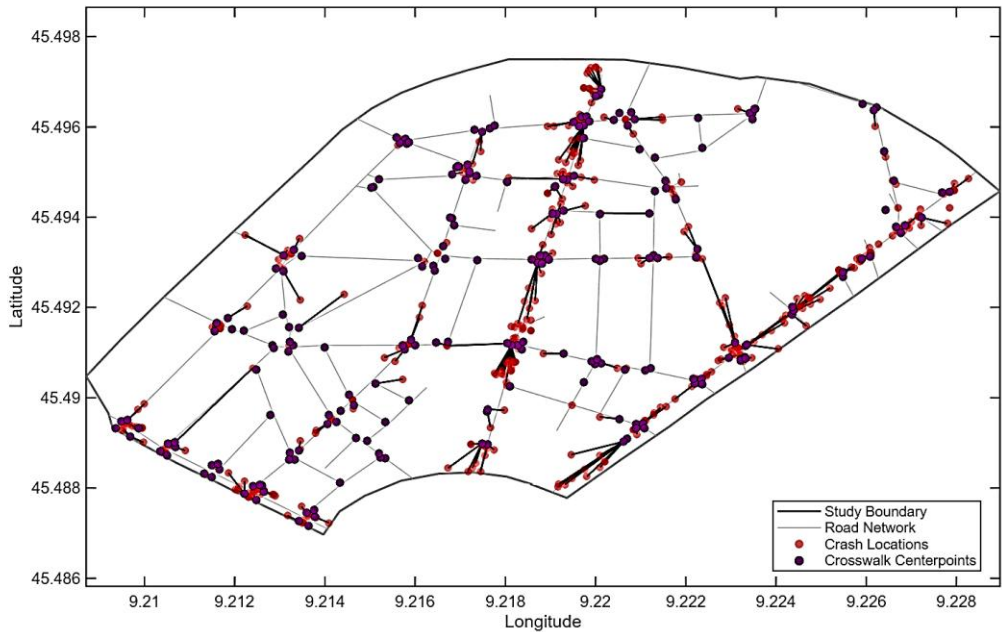

Figure 3.8 Crash–crosswalk distance computation

3.4. Crash Exposure Weighting Calculation

After the crash-to-network assignment described in Section 3.3, we transform these crash–edge relationships into quantitative crash exposure measures at both the edge and node levels.

Crash exposure weighting is required to:

- quantify crash accumulation along road segments
- translate segment-level crash information into intersection-level indicators
- provide weighted inputs for centrality indices calculations

Two levels of crash exposure are defined:

- Edge-based crash weights (we)
- Node-based crash indices (RE and RN)

1. Edge-Based Crash Weights (we)

Once each crash has been assigned to its nearest road segment through orthogonal projection, crashes were aggregated at the edge level to obtain an edge-based crash weight.

For an edge e in the network, the crash weight is defined as the number of crashes assigned to the edge:

$$w_e = \sum_{i=1}^{n} C_i$$

where Ci denotes a crash event assigned to edge e, and n is the total number of crashes associated with that edge.

Thus, we represented the total number of crashes projected onto edge e. This computation was implemented in MATLAB using an accumulation operation based on the vector crash_to_edge_map.

2. **Node-Based Crash Indices (RE and RN)**

While edge-based weights capture crash accumulation along road segments, many network-based analyses, particularly some centrality measures, require node-level crash exposure.

To translate crash information from edges to nodes, two complementary node-based indices were defined:

- RE (Raw Edge Assignment)
- RN (Raw Node Assignment)

Both indices rely on the projection parameter t computed during crash-to-edge assignment.

RE (Raw Edge Assignment): this method assigns each crash entirely to a single node, based on its position along the road segment. It reflects the intuitive idea that a crash close to a particular intersection should contribute fully to that intersection's risk profile.

For a crash projected onto the edge e=(a,b) with parameter $t \in [0,1]$:

- If $t < 0.5$, the crash belongs to start node a
- If $t \geq 0.5$, the crash belongs to end node b

This method produces a discrete and localized representation of crash exposure and is sensitive to small variations in crash position.

RN (Raw Node Assignment): the RN method distributes crash influence between the two endpoints of the edge according to the relative distance between the crash and each node. In this study, crash severity was also incorporated into the RN formulation to account for differences in crash impact.

For a crash k projected onto the edge e=(a,b) with projection parameter $t_k \in [0,1]$ and severity weight s_k, the contributions to the two nodes are defined as:

$$RN(a) \leftarrow RN(a) + s_k(1 - t_k)$$

$$RN(b) \leftarrow RN(b) + s_k t_k$$

A crash exactly between the two endpoints (t=0.5) contributes equally to both nodes. Crashes closer to one node contribute more strongly to that node.

A crash very close to node a (e.g., t=0.1) gives:

$$RN(a) = 0.9s_k, \quad RN(b) = 0.1s_k$$

This formulation produces a smoother and more stable representation of crash exposure compared to RE, as it distributes crash influence rather than assigning it entirely to a single node.

Normalization of Crash Exposure Measures

To ensure comparability across nodes and to stabilize subsequent centrality computations, the crash exposure measures are normalized.

The RE index, after aggregation at the node level, is normalized as:

$$RE_v^{norm} = \frac{RE_v}{\max(RE)}$$

Similarly, the RN index, derived from node-based crash weights is normalized as:

$$RN_v^{norm} = \frac{RN_v}{\max(RN)}$$

This distinction ensures that RE represents pure crash frequency localized at nodes, while RN represents severity-weighted and spatially distributed crash exposure. The normalization step expresses both indices on a common scale, enabling their use as inputs for centrality computation and clustering.

Both indices are retained and used as node weight vectors in the computation of crash-informed centrality measures, with all calculations implemented in MATLAB.

3.5. Centrality Indices Calculation

A. ICENTR

The ICENTR index, proposed by Mussone, Viseh, and Notari (2021), is a weighted graph centrality measure that evaluates network importance by simultaneously incorporating node and edge weights. In the present study, crash occurrences are first transformed into node and edge weights (Section 3.4), after which the ICENTR algorithm is applied.

According to Mussone et al. (2021), given a root node i_0, every node in the graph was assigned to a hierarchical level based on its topological distance from i_0. Nodes directly connected to i_0 form the first level, while nodes at level h are connected to nodes at level h−1 but not to any node at a lower level. Since the graph was fully connected, each node belongs to exactly one level.

Edges were assigned levels based on the levels of their two endpoint nodes. An edge connecting nodes from two different levels is assigned the higher of the two levels. An edge whose both endpoints belong to the same level was assigned to the next level above.

According to Mussone et al. (2021), the ICENTR value of the root node i0 is defined as, the ICENTR value for node i0 is defined as:

$$ICENTR(i_0) = \sum_{e \in E} IC(e)$$

where IC(e) represents the contribution of each edge e to the centrality of i_0. The contribution IC(e) of each edge was calculated based on the weights of its endpoint nodes and the edge itself, as well as their respective hierarchical levels:

$$\mathrm{IC}(e = \{u, v\}) = \begin{cases} \dfrac{w(u)}{c^{\mathrm{lev}(e)-1}} w(e) & \text{if } \mathrm{lev}(u) > \mathrm{lev}(v) \\ \dfrac{w(u) + w(v)}{c^{\mathrm{lev}(e)}} w(e) & \text{if } \mathrm{lev}(u) = \mathrm{lev}(v) \end{cases}$$

In this expression, w(u) and w(v) denote the weights of nodes u and v connected by edge e, and w(e) denotes the weight of edge e itself. The terms lev(u), lev(v), and lev(e) represent their assigned hierarchical levels relative to i_0. The parameter c is a positive constant that controls the sensitivity of the index to level distance and has been experimentally determined to be c = 2, which provides a balanced trade-off between sensitivity and interpretability.

A notable property of Icentr is its additivity with respect to edge weights: if the edge weights are decomposed into two separate sets w_1 and w_2, the Icentr value computed using the combined weights equals the sum of the values computed independently for each set. This additive property holds equivalently for node weights.

Formally, if $w1$ *and* $w2$ represents two different edge weightings, then for any node x, the following equation holds:

$$\mathrm{ICENTR}_{w_1+w_2}(x) = \mathrm{ICENTR}_{w_1}(x) + \mathrm{ICENTR}_{w_2}(x)$$

This additive property simplifies the analysis and interpretation of centrality indices when dealing with multiple weighting schemes.

B. Betweenness Centrality (BET)

Betweenness centrality quantifies how frequently a node lies on the shortest paths connecting all other pairs of nodes in the network (Freeman, 1977).

For a node x, betweenness is defined as:

$$BC(x) = \sum_{u \neq v \neq x} \frac{\sigma_{uv}(x)}{\sigma_{uv}}$$

where σ(u,v) is the total number of shortest paths between nodes u and v, and σ(u,v)(x) is the number of those paths that pass through node x.

Nodes with high BET values are expected to experience greater movement exposure due to their role in connecting multiple routes across the network. Conversely, nodes with low BET typically correspond to peripheral locations, dead-end streets, or local access roads with limited through-movement, and are therefore associated with lower interaction intensity and reduced crash risk.

C. Closeness Centrality (CLS)

Closeness centrality measures how close a node is to all other nodes in the network, based on the inverse of the total shortest-path distance from that node to all others (Porta, Crucitti, and Latora, 2006). Nodes with high closeness values are topologically central and can reach other locations within the network through relatively short paths, reflecting a high degree of spatial accessibility within the urban fabric.

The closeness centrality CC(x) for a node x is defined as:

$$\mathrm{CC(x)} = \frac{1}{\sum_{\mathrm{u \neq x}} \mathrm{d(u, x)}}$$

To obtain normalized values comparable across networks of different sizes, the formula is scaled by (n−1):

$$CC(x) = \frac{n-1}{\sum_{u \neq x} d(u,x)}$$

where:

- $d(u,x)$: shortest path distance between nodes u and x
- n is the total number of nodes in the network.

Unlike betweenness centrality, which captures through-movement, closeness centrality reflects spatial accessibility rather than flow concentration. In dense urban grids, nodes with high CLS values may experience elevated crash exposure due to the frequent circulation of vehicles, pedestrians, and cyclists attracted by their central and accessible position within the network.

Together, the four centrality measures: Icentr Edge, Icentr Node, Betweenness, and Closeness, provide a multidimensional representation of crash behavior, combining structural, spatial, and crash-informed perspectives. These indices form the feature matrix used in the subsequent clustering and predictive analyses.

3.6. Cluster Method Selection

Clustering constitutes a fundamental component of the proposed predictive framework. Given the relatively large number of network nodes (118) compared to the limited number of crash observations in each testing phase, direct node-level prediction would lead to unstable and highly sparse probability estimates. To address this issue, nodes are aggregated into clusters based on spatial proximity and structural similarity, allowing crash intensities to be estimated at a more stable group level.

The clustering process aims to identify groups of nodes that are similar in terms of both geographic location and network-based crash influence. Clustering was performed separately for each index, ensuring that the spatial grouping of nodes reflects both their position within the network and their structural importance according to the selected index.

Previous analysis demonstrate that hierarchical clustering is more suitable for this application due to:

- Its flexibility in handling irregular spatial distributions
- Its ability to reveal nested structures through dendrogram analysis
- Its robustness in heterogeneous urban networks
- Its independence from predefined centroid initialization

Average linkage with Euclidean distance was selected as the reference clustering configuration. This method provided the best balance between spatial compactness and differentiation of centrality-based risk levels. It avoided excessive merging while preventing fragmentation, resulting in spatially coherent and structurally meaningful clusters. Furthermore, among all tested clustering configurations, the Average–Euclidean method exhibited the greatest stability during simulations, generated the most homogeneous clusters with respect to the selected nodes, and produced the lowest prediction errors during the testing phase.

The Average–Euclidean method evaluates the similarity between two clusters using the average Euclidean distance between all pairs of observations belonging to the clusters. The Euclidean distance between two points A(x1,y1,z1) and B(x2,y2,z2) in three-dimensional space is calculated as:

$$D_{Euclidean}(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Once the distances between all points in two clusters are computed, their average is used to quantify cluster similarity. Clusters characterized by smaller average distances are merged first, ensuring that observations grouped within the same cluster remain relatively homogeneous.

Compared with complete linkage, which relies on the maximum distance between observations, and single linkage, which relies on the minimum distance, the Average–Euclidean approach provides a balanced compromise between cluster compactness and robustness to outliers. As a result, it preserves overall spatial coherence while reducing the influence of extreme observations that may distort cluster formation.

The method is particularly suitable for the multidimensional network representation adopted in this study, where clustering is performed using spatial coordinates together with network-based indicators. By considering average distances, the approach prevents clusters from becoming excessively scattered or overly influenced by a small number of closely located nodes, resulting in clusters that are both spatially cohesive and representative of the underlying network structure.

Clustering was performed independently for each index (IE, RE, RN, BET, CLS, and IN) in each scenario, generating distinct cluster configurations corresponding to each structural representation of crash influence. These cluster assignments are subsequently used for cluster-level crash aggregation and predictive modelling.

The optimal number of clusters is not fixed a priori but is determined during the prediction stage based on the combination that yields the lowest prediction error, the highest CPR value, and the most homogeneous cluster structure.

Through this methodological selection process, hierarchical clustering serves as a stable and interpretable bridge between network centrality indicators and the probabilistic crash prediction framework.

3.7. Probability Density Function

In the next step of the analysis, probability density functions (PDFs) were constructed to assign probabilistic weights to the clusters identified through the selected clustering method. The purpose of the PDF is to model the likelihood of future crash occurrence within each cluster based on the spatial distribution of the selected network indicator.

For each selected indicator and clustering configuration, the indicator values associated with all nodes belonging to the same cluster were aggregated. The cumulative indicator value of cluster (k) was computed as:

$$I_k = \sum_{i \in C_k} I_i$$

where:

- I_i is the value of the selected network indicator at node i;
- C_k is the set of nodes belonging to cluster k;
- I_k is the cumulative indicator value of cluster k.

The probability assigned to cluster k is then obtained by normalizing the cumulative indicator values:

$$P_k = \frac{I_k}{\sum_{k=1}^{K} I_k}$$

This formulation assigns higher probabilities to clusters characterized by larger cumulative indicator values. Consequently, clusters containing nodes with greater importance according to the selected network indicator are assigned a higher probability of receiving future crashes. Since each network indicator produces a different spatial distribution of node values, a distinct probability density function is generated for each indicator. Furthermore, the probability distribution also depends on the selected number of clusters, as different clustering configurations produce different spatial aggregations of the indicator values.

The resulting probability vector satisfies the normalization condition:

$$\sum_{k=1}^{K} P_k = 1$$

where K is the total number of clusters.

The probability density functions generated in this step are subsequently used to simulate the distribution of future crashes across the clusters. To reduce excessive dependence on the indicator-based probability distribution and improve the robustness of the forecasting process, a mixed probability formulation is introduced in the following subsection.

Mixed Probability Formulation:

To reduce excessive dependence on the indicator-based probability distribution and improve the robustness of the forecasting process, a mixed probability formulation was introduced. The final probability assigned to each cluster is expressed as a weighted combination of the indicator-based PDF and a uniform probability distribution:

$$\tilde{P}_k = \alpha P_k + (1-\alpha)\frac{1}{K}$$

where:

- P~k is the final probability assigned to cluster k
- α ∈ [0,1] is a calibration parameter controlling the balance between the indicator-based and uniform components
- P_k is the probability assigned to cluster k based on the cumulative value of the selected network indicator;
- 1 / K represents a uniform distribution across all clusters
- K is the total number of clusters.

When α=1, the distribution is entirely driven by historical crash frequencies. When α=0, crashes are distributed uniformly across clusters. For intermediate values of α\alpha$\alpha$, the final probability distribution represents a balance between the indicator-based allocation and a uniform distribution, allowing every cluster to retain a non-zero probability while preserving the structural information provided by the selected network indicator.

3.8. Cluster-Level Crash Prediction Estimation

The final step of the predictive framework consists of estimating the number of crashes expected within each cluster during the testing period. This step uses the probability density functions (PDFs) defined in Section 3.7 to generate a probabilistic distribution of crashes across the clustered network.

For each scenario, let $\tilde{P}_k$ denote the probability assigned to cluster k, and let N^{test} be the total number of crashes observed in the testing year. The expected number of crashes in each cluster is estimated as:

$$\hat{N}_k = \tilde{P}_k \cdot N^{test}$$

where $\hat{N}_k$ represents the predicted number of crashes in cluster k.

Stochastic Crash Allocation

To better simulate the variability of real-world crash occurrences, a stochastic approach was adopted using random sampling. Instead of assigning crashes deterministically, the total number of crashes was distributed across clusters according to the computed probability vector.

This process is implemented in MATLAB using the multinomial random sampling function (mnrnd), which generates random crash allocations based on the probability distribution $\tilde{P}_k$. The stochastic formulation allows the model to account for inherent randomness in crash occurrence while preserving the overall probability structure.

Simulation Parameters

The prediction process depends on several adjustable parameters:

- **Total crashes (Tot_c)**

Represents the total number of crashes to be distributed. It was derived from the testing dataset and may be scaled using a calibration factor.

- **Repetitions (Rep_extr)**

Defines the number of repeated stochastic simulations. Increasing this value improves the stability of the estimated results by averaging multiple random realizations.

- **Mixing parameter (α)**

Controls the balance between empirical and uniform probability distributions.

- **Scaling factor (Error_pred)**

Acts as a multiplicative adjustment applied to the total number of crashes, allowing calibration of the predicted totals relative to observed values.

Averaging and Final Estimation

To obtain stable predictions, the stochastic allocation process is repeated multiple times. The final predicted number of crashes for each cluster is computed as the average over all repetitions:

$$\hat{N}_k^{final} = \frac{1}{R}\sum_{r=1}^{R}\hat{N}_k^{(r)}$$

where:

- R is the number of repetitions
- $\hat{N}_k^{(r)}$ is the predicted number of crashes in cluster k at repetition r

This averaging process reduces the effect of random fluctuations and ensures convergence toward the expected probability distribution.

This cluster-level prediction approach provides a robust framework for estimating crash distribution across the network. By combining probabilistic modelling with stochastic simulation, the method captures both the structural patterns derived from historical data and the inherent variability of crash occurrence.

The predicted cluster-level crash values obtained in this step are subsequently compared with observed crashes in the testing dataset using performance metrics such as RMSE and modified RMSE, as described in the following section.

3.9. Predictive model evaluation

To evaluate the performance of the predictive model, it was essential to use error metrics that quantify the difference between predicted and observed crash occurrences across clusters. These metrics provide insight into the accuracy of the model and its ability to capture spatial variations in crash distribution.

In this study, three error measures were considered:

- Root Squared Error (RSE)
- Root Mean Squared Error (RMSE)
- Modified Root Mean Squared Error (RMSEmod)

These metrics are used to assess the effectiveness of the clustering-based prediction framework and to identify the optimal number of clusters.

A. Root Squared Error (RSE)

$$RSE = \sqrt{\sum_{k=1}^{K} \left(y_k^{\text{actual}} - y_k^{\text{predicted}}\right)^2}$$

- K: the total number of clusters (NOC)
- $y_k^{predicted}$: predicted number of crashes in cluster k
- y_k^{actual}: actual number of crashes in cluster k

RSE provides a global measure of the total prediction error across all clusters. It reflects the cumulative deviation between predicted and observed crash values.

However, RSE was not normalized by the number of clusters, meaning that:

- its value increases with the number of clusters
- it is not directly comparable across different NOC values

RSE is useful for evaluating the overall magnitude of prediction error. However, because it does not account for cluster size or normalization, it may not be suitable for selecting the optimal number of clusters. In particular, larger values of NOC tend to produce higher RSE values due to increased fragmentation of the data.

Therefore, RSE is mainly used as a supporting metric to understand global error trends, rather than as a primary criterion for model selection.

B. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(y_k^{\text{actual}} - y_k^{\text{predicted}}\right)^2}$$

- K: the total number of clusters (NOC)
- $y_k^{predicted}$: predicted number of crashes in cluster k
- y_k^{actual}: actual number of crashes in cluster k

RMSE measures the average prediction error across clusters by normalizing the squared differences by the number of clusters. This normalization allows for a more balanced evaluation compared to RSE, making RMSE more suitable for comparing model performance across different clustering configurations.

RMSE is widely used to evaluate predictive accuracy, as it provides a normalized measure of error that accounts for the number of clusters. It enables comparison between different values of NOC and helps identify configurations that yield lower prediction errors.

However, RMSE presents an important limitation in this context. As the number of clusters increases, the average number of crashes per cluster decreases. This can artificially reduce the RMSE value, even when the prediction quality does not improve. In extreme cases, when clusters contain very few crashes, the prediction error per cluster becomes small, leading to deceptively low RMSE values.

Therefore, RMSE may favor solutions with a large number of clusters, even if these configurations are not meaningful from a spatial or interpretative perspective.

For this reason, RMSE is used in conjunction with additional metrics, such as RMSEmod, to ensure a more reliable selection of the optimal number of clusters.

C. Modified Root Mean Squared Error (RMSEmod)

RMSEmod was introduced to address the limitations of RMSE in cluster-based prediction. Instead of normalizing the error by the total number of clusters, RMSEmod normalizes by the number of meaningful clusters, defined as those with actual crash values greater than 1.

This modification reduces the influence of clusters with zero or negligible actual crashes, which tend to increase artificially as the number of clusters grows.

$$RMSE_{mod} = \sqrt{\frac{1}{N_m}\sum_{k=1}^{K}\left(y_k^{\text{actual}} - y_k^{\text{predicted\_mod}}\right)^2}$$

- K: the total number of clusters (NOC)
- $\boldsymbol{y_k^{predicted}}$: predicted number of crashes in cluster k
- $\boldsymbol{y_k^{actual}}$: actual number of crashes in cluster k
- N_m : number of clusters with actual crash values greater than 1

As the number of clusters increases, many clusters may contain very few or zero actual crashes. In such cases, RMSE may produce misleadingly low values since errors in these clusters are minimal. RMSEmod corrects this effect by focusing only on clusters that carry significant predicted crash values.

This makes RMSEmod more suitable for:

- identifying the optimal number of clusters
- ensuring that selected clustering configurations remain meaningful and interpretable
- avoiding over-fragmentation of the network

By filtering out low-impact clusters, RMSEmod provides a more realistic measure of prediction performance and supports a balanced selection of NOC across different scenarios.

3.10. Optimal Configuration Selection

The optimal configuration was determined through a systematic evaluation of all six predictive scenarios and all tested indicators. Particular attention is given to Scenario 5 (S5) and the Random Scenario (SR), as these represent the most realistic temporal configurations for crash prediction. For each indicator, a range of cluster numbers between 2 and 50 was evaluated, and the corresponding prediction performance was assessed.

The selection process was based on three complementary criteria: prediction accuracy, crash prevention efficiency, and cluster homogeneity. This multi-criteria approach ensures that the chosen configuration provides not only accurate crash prediction but also meaningful and interpretable spatial segmentation of the NoLo road network.

For each indicator, configurations yielding the lowest RMSEmod values within the range of 5 to 25 clusters were identified as candidate solutions in S5 and SR and compiled into a comparison table for side-by-side evaluation. RMSEmod was adopted as the primary prediction accuracy metric because it focuses the error computation on clusters with significant crash concentrations, thereby reducing the influence of clusters with negligible crash occurrence.

However, prediction accuracy alone is not sufficient to determine the optimal configuration, as low prediction errors do not necessarily guarantee effective identification of high-risk areas. Therefore, two additional criteria, namely Crash Prevention Efficiency (CPE) and cluster homogeneity, were incorporated into the evaluation process. These criteria assess how effectively a model identifies crash concentrations while minimizing the proportion of the network requiring intervention and ensuring that the resulting clusters remain spatially coherent and structurally meaningful.

The Crash Prevention Rate (CPR) measures the percentage of actual crashes captured within the top predicted high-risk clusters and is defined as:

$$CPR = \frac{\text{Actual crashes in top clusters}}{\text{Total actual crashes}} \times 100$$

The Node Coverage Ratio (NCR) quantifies the proportion of network nodes included within the top predicted clusters relative to the total number of nodes in the network and is defined as:

$$NCR = \frac{\text{Nodes in top clusters}}{\text{Total network nodes}} \times 100$$

These two metrics are combined into the Crash Prevention Efficiency (CPE), which evaluates the ability of the model to capture crash concentrations while minimizing the portion of the network requiring intervention:

$$CPE = CPR + (100 - NCR)$$

Higher CPE values indicate configurations that capture a larger proportion of crashes while requiring interventions on a smaller portion of the network.

In addition to prediction accuracy and crash prevention efficiency, the quality of the generated clusters was also assessed through a homogeneity analysis. Cluster homogeneity reflects the internal connectivity and spatial coherence of the resulting groups. A cluster is considered homogeneous when its nodes maintain an adequate level of connectivity with other nodes belonging to the same cluster. To evaluate this condition, minimum intra-cluster connectivity requirements were defined according to node degree:

- Degree-1 nodes were required to be connected to at least one node within the same cluster.
- Degree-2 nodes were required to be connected to at least one node within the same cluster.
- Degree-3 and degree-4 nodes were required to maintain at least two intra-cluster connections.
- Degree-5 nodes were required to be connected to at least three nodes within the same cluster.

Configurations satisfying these connectivity requirements were considered homogeneous, whereas fragmented or poorly connected configurations were considered less suitable, even when prediction metrics were favorable. In ambiguous cases, spatial context and engineering judgment were used to assess cluster suitability.

The final selection of the optimal scenario, centrality indicator, and number of clusters was based on the combined assessment of RMSEmod, CPE, and cluster homogeneity. A visual inspection of the resulting cluster maps was also performed to verify the spatial coherence of the identified high-risk zones.

4 Crash Data Analysis

This chapter presents the descriptive analysis of road crashes in the NoLo (Nord di Loreto) district over the period 2017–2021, with the objective of identifying the main temporal, spatial, and modal crash patterns before the application of the network-based methodology presented in the following chapter.

The crash records were obtained from the ISTAT/municipal road crash dataset, which includes information on crash location, road-user category, year and hour of occurrence, injury severity, road type, road geometry, pavement and road surface conditions, weather conditions, intersection type, signage presence, vehicle types involved, and the reported circumstances of drivers, vehicles, and pedestrians.

Where appropriate, the results for NoLo are compared with Milan-wide statistics to highlight similarities and local differences. The analysis is restricted to crashes recorded between 2017 and 2021 and is organized according to the three principal road-user categories: pedestrians, cyclists, and motor vehicles.

During the five-year study period, 495 road crashes were recorded in NoLo, comprising 93 pedestrian crashes, 84 bicycle crashes, and 319 motor-vehicle crashes. The distribution of crashes by year and road-user category is presented in the following section.

4.1. Pedestrians Crashes Descriptive Analysis

This section presents a detailed descriptive analysis of pedestrian crashes in NoLo, with the objective of identifying their principal temporal, spatial, environmental, and behavioral characteristics. Pedestrian crashes are examined separately from bicycle and motor-vehicle crashes due to the high exposure of pedestrians in NoLo's dense urban environment and their heightened vulnerability to injury.

4.1.1. Spatial Classification

Classification by Spatial Distribution

The spatial distribution of pedestrian crashes in NoLo is strongly influenced by the structure and hierarchy of the district's street network. As illustrated in Figure 4.1, crashes are concentrated along the principal urban corridors, particularly Via Padova and Viale Monza.

Via Padova exhibits the highest concentration of pedestrian crashes, extending through the central and eastern portions of the district. Viale Monza also records a substantial number of crashes, reflecting its role as one of NoLo's primary arterial roads. A smaller concentration is observed along Viale Brianza near the western boundary of the district.

Overall, the spatial distribution indicates that pedestrian crashes are concentrated along the district's main traffic corridors rather than being uniformly distributed across the network. This pattern highlights the importance of these corridors in the observed pedestrian crash distribution and provides the basis for the network-based analysis presented in the following chapters.

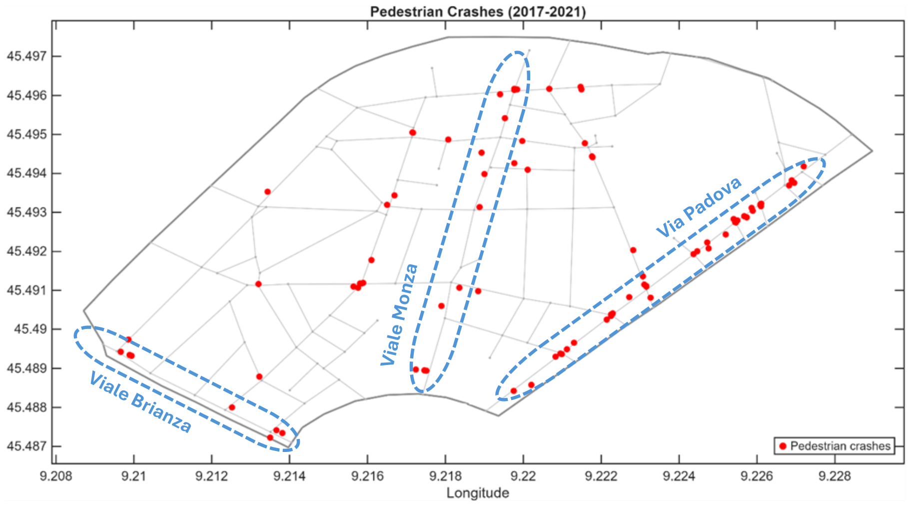

Figure 4.1 Spatial distribution of pedestrian crashes in NoLo (2017–2021)

Classification by Pavement Condition:

Table 4.1 presents the distribution of pedestrian crashes by pavement condition in NoLo between 2017 and 2021, compared with Milan-wide figures. All pedestrian crashes in NoLo occurred on paved roads (100%), consistent with the city-wide pattern, where almost all crashes also occurred on paved surfaces. This reflects the highly urbanized nature of NoLo and indicates that pavement condition is not a significant factor in explaining pedestrian crash occurrence, with traffic conditions, infrastructure, and pedestrian activity likely having a greater influence.

	Milan		**Nolo**	
Pavement Condition	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Paved road	5,537	99.7	93	100
Bumpy paved	4	0.1	0	0
Unpaved road	12	0.2	0	0

Table 4.1 Pedestrian crashes classification by pavement condition

Classification by Road Type:

Most pedestrian crashes in NoLo occurred on two-way carriageways, accounting for 50 crashes (54%), a higher proportion than the Milan-wide value of 46%. Roads with two separate carriageways accounted for 25 crashes (27%), while one-way carriageways recorded 18 crashes (19%). No pedestrian crashes were recorded on roads with more than two carriageways. Overall, the results indicate that pedestrian crashes in NoLo are primarily concentrated on conventional urban streets, particularly two-way carriageways and roads with two separate carriageways. This pattern reflects the structure of the district's road network and the concentration of pedestrian activity along its main urban corridors (table 4.2).

	Milan		Nolo	
Road Type	Crashes number	Percentage (%)	Crashes number	Percentage (%)
One-way carriageway	1,080	19	18	19
Two-way carriageway	2,566	46	50	54
Two carriageways	1,051	19	25	27
More than two carriageways	856	15	0	0

Table 4.2 Pedestrian crashes classification by road type

Classification by Road Surface Condition:

Most pedestrian crashes in NoLo occurred on dry road surfaces (73 crashes, 78.5%), consistent with the Milan-wide distribution (79%). Wet road surfaces accounted for 19 crashes (20.4%), closely matching the corresponding city-wide proportion (20.8%). Only one crash (1.1%) occurred on a snowy surface, while no crashes were recorded on icy or slippery surfaces. Overall, pedestrian crashes in NoLo predominantly occurred on dry road surfaces, consistent with the Milan-wide pattern (table 4.3).

	Milan		Nolo	
Road surface	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Dry	4385	79	73	78.5
Wet	1154	20.8	19	20.4
Slippery	2	0	0	0
Icy	3	0.1	0	0
Snowy	9	0.2	1	1.1

Table 4. 3 Pedestrian crashes classification by road surface

Classification by Intersection Type:

Most pedestrian crashes in NoLo occurred on straight road segments outside intersections (44 crashes, 47%), exceeding the Milan-wide proportion (39%). Signalized intersections accounted for 25 crashes (27%), while intersections controlled by traffic lights or traffic police accounted for 7 crashes (8%). Crashes at conventional intersections and roundabouts were relatively few, and no crashes were recorded at unsignalized intersections, curves, or slopes. Overall, pedestrian crashes in NoLo were concentrated on straight road segments and signalized locations (table 4.4).

	Milan		**Nolo**	
Intersection type	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Intersection	312	5.6	6	6.5
Roundabout	65	1.2	2	2
Signalized intersection	1091	19.6	25	27
Intersection with traffic light or traffic policeman	1064	19.2	7	8
Unsignalized intersection	65	1.2	0	0
Non - Intersection type	Milan		Nolo	
Straight road	2176	39.2	44	47
Curve	106	1.9	0	0
Narrow passage, off the road	5	0.1	0	0
Slope	1	0	0	0
Illuminated tunnel	2	0	0	0
Not illuminated tunnel	2	0	0	0
Undefined cases	664	12	9	10

Table 4.4 Pedestrian crashes classification by intersection type

Classification by Signage Presence:

Most pedestrian crashes in NoLo occurred on roads equipped with both vertical and horizontal signage (88 crashes, 94.6%), matching the Milan-wide proportion (95%). Roads with only vertical signage, only horizontal markings, or no signage accounted for the remaining 5.4% of crashes. Overall, pedestrian crashes in NoLo were predominantly recorded on roads with both vertical and horizontal signage, consistent with the Milan-wide distribution (table 4.5).

	Milan		Nolo	
Signage presence	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Absence of signs	161	2.9	2	2.2
Vertical signs only	44	0.8	1	1.1
Horizontal signs only	59	1.1	2	2.2
Both directions	5274	95	88	94.6
Undefined	15	0.3	0	0

Table 4.5 Pedestrian crashes classification by signage presence

4.1.2. Temporal Classification

Classification by Year of Occurrence

A total of 93 pedestrian crashes were recorded in NoLo during the study period, representing 1.7% of all pedestrian crashes in Milan. Crashes increased from 14 in 2017 to a peak of 26 in 2019, declined to 9 in 2020, and partially recovered to 19 in 2021. A similar temporal pattern was observed across Milan, with a marked decrease in 2020 followed by a recovery in 2021, consistent with the reduction and subsequent resumption of urban mobility during and after the COVID-19 pandemic (table 4.6).

	Milan		NoLo		NoLo / Milan
Year	Crashes number	Percentage (%)	Crashes number	Percentage (%)	Percentage (%)
2017	1,315	23.7	14	15.1	1.1
2018	1,364	24.6	25	26.9	1.8
2019	1,297	23.4	26	28	2
2020	665	12	9	9.7	1.4
2021	912	16.4	19	20.4	2.1
Total	5,553	100	93	100	1.68

Table 4.6 Pedestrian Crash Classification by Year of Occurrence

Classification by Hour of the Day

Pedestrian crashes in NoLo exhibit a clear daily pattern closely associated with urban mobility activity. Crash frequencies remain low during the night and early morning hours (00:00–06:00), when both pedestrian and vehicular movements are limited, and begin to increase from 07:00 onwards as commuting activity intensifies.

Two main peaks can be identified. The first occurs during the morning period, particularly at 08:00 (9%) and 10:00 (13%), reflecting increased travel associated with work, education, shopping, and access to public transport. The second and highest peak is recorded at 18:00 (15%), corresponding to the evening rush hour, when pedestrian and vehicular flows are simultaneously high. Elevated crash frequencies are also observed between 15:00 and 17:00, reflecting sustained afternoon mobility.

Compared with Milan as a whole, NoLo exhibits a similar temporal pattern characterized by low night-time crash frequencies and pronounced daytime peaks. However, the evening peak at 18:00 is proportionally higher in NoLo (15%) than in Milan (10%), suggesting a stronger concentration of pedestrian activity during this period within the district.

Overall, the results indicate that pedestrian crashes in NoLo are concentrated during periods of high urban activity, highlighting the close relationship between crash occurrence and pedestrian exposure (table 4.7).

	Milan		Nolo	
Hour	Crashes number	Percentage (%)	Crashes number	Percentage (%)
00:00	59	1.1	1	1.1
01:00	43	0.8	2	2.2
02:00	19	0.3	1	1.1
03:00	22	0.4	0	0
04:00	21	0.4	0	0
05:00	42	0.8	1	1.1
06:00	67	1.2	2	2.2
07:00	248	4.5	5	5.4
08:00	407	7.3	8	8.6
09:00	404	7.3	4	4.3
10:00	364	6.6	12	12.9
11:00	339	6.1	2	2.2
12:00	333	6	3	3.2
13:00	263	4.7	2	2.2
14:00	276	5	4	4.3
15:00	307	5.5	5	5.4
16:00	374	6.7	8	8.6
17:00	449	8.1	6	6.5
18:00	551	9.9	14	15.1
19:00	430	7.7	4	4.3
20:00	250	4.5	5	5.4
21:00	130	2.3	2	2.2
22:00	76	1.4	0	0
23:00	79	1.4	2	2.2

Table 4.7 Pedestrian crash classification by hour of the day

4.1.3. Weather Conditions Classification

Most pedestrian crashes in NoLo occurred under clear weather conditions (73 crashes, 78.5%), similar to the Milan-wide distribution (80.8%). Rain accounted for 14 crashes (15.1%), closely matching the corresponding city-wide proportion (14.2%). Snow conditions accounted for 2 crashes (2.2%), while no pedestrian crashes were recorded during fog, hail, or strong wind conditions. The remaining 4.3% of crashes occurred under other weather conditions. Overall, pedestrian crashes in NoLo predominantly occurred under clear weather conditions, consistent with the Milan-wide distribution. These results describe the observed distribution of crashes and should not be interpreted as indicating the effect of weather conditions on crash risk (table 4.8).

	Milan		**Nolo**	
Weather condition	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Clear	4488	80.8	73	78.5
Fog	51	0.9	0	0.0
Rain	786	14.2	14	15.1
Hail	1	0.0	0	0.0
Snow	23	0.4	2	2.2
Strong wind	2	0.0	0	0.0
Other	202	3.6	4	4.3

Table 4.8 Pedestrian crashes classification by weather condition

4.1.4. Vehicle Type Involved Classification

Pedestrian Crashes by First Vehicle Type Involved:

Private passenger cars were the most frequently involved vehicle type, accounting for 46 crashes (49%). Motorcycles represented the second largest category, accounting for 16 crashes (17%), slightly higher than the Milan-wide proportion of 14%.

Trucks accounted for 8 crashes (9%), while public transport vehicles, bicycles, mopeds, and micro-mobility vehicles each represented only a small share of cases.

A notable feature of the NoLo dataset is the proportion of hit-and-run crashes, which accounted for 9 cases (10%), exceeding the Milan-wide value of 6%. Undefined or other vehicle categories represented 5% of cases (table 4.9).

	Milan		Nolo	
Vehicle type (first vehicle)	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Private car	2905	52.3	46	49.5
Public service car	182	3.3	3	3.2
Rescue or police vehicle	14	0.3	0	0.0
Bus or trolleybus (urban)	51	0.9	2	2.2
Extra-urban bus	30	0.5	0	0.0
Tram	46	0.8	0	0.0
Truck	313	5.6	8	8.6
Articulated bus or articulated truck	3	0.1	0	0.0
Special vehicle	80	1.4	1	1.1
Bicycle (velocipede)	185	3.3	1	1.1
Moped	61	1.1	1	1.1
Motorcycle	769	13.8	16	17.2
Motorcycle with passenger	20	0.4	0	0.0
Motor truck or motor van	1	0.0	0	0.0
Animal or hand-drawn vehicle	1	0.0	0	0.0
Vehicle that fled (hit-and-run)	360	6.5	9	9.7
Quadricycle	11	0.2	0	0.0
Electric scooter	31	0.6	1	1.1
Electric bicycle	5	0.1	0	0.0
Undefined / other	485	8.7	5	5.4

Table 4.9 Pedestrian crashes classification by type of first vehicle involved

Pedestrian Crashes by Second Vehicle Type Involved:

All pedestrian crashes recorded in NoLo involved a single vehicle, with no second vehicle reported in any case. This finding is consistent with the Milan-wide dataset, where virtually all pedestrian crashes involved only one vehicle.

The results indicate that pedestrian crashes in NoLo are predominantly direct interactions between pedestrians and individual vehicles rather than multi-vehicle events (table 4.10).

	Milan		**Nolo**	
Vehicle type (second vehicle)	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Bicycle	1	0	0	0
Moped	1	0	0	0
No second vehicle involved	5551	100	93	100

Table 4.10 Pedestrian crashes classification by type of second vehicle involved

4.1.5. Vehicle Circumstances Classification

Circumstances for Vehicle Involvement Classification:

The most frequent circumstance in NoLo is the failure to give priority to pedestrians at crossings, accounting for 50 crashes (54%), which exceeds the Milan-wide proportion of 42%. Vehicles proceeding regularly accounted for 20 crashes (22%), while improper maneuvers and excessive speed represented 9% and 8% of cases, respectively. Signal violations and wrong-way driving each accounted for only a small proportion of crashes. Abnormal driver conditions were rare during the study period, with only one alcohol-related crash recorded and no cases associated with illness, narcotics, or vehicle defects (table 4.11).

	Milan		Nolo	
Circumstances for vehicle involved	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Did not give priority to pedestrians at crossings	2316	41.7	50	53.8
Proceeded regularly	1661	29.9	20	21.5
Maneuvered	491	8.8	8	8.6
Proceeded regularly with excessive speed	441	7.9	7	7.5
Maneuvered without respecting traffic signals	118	2.1	2	2.2
Proceeded regularly against traffic	65	1.2	1	1.1
Went off the road and hit the pedestrian	33	0.6	0	0.0
Hit the pedestrian with the load	29	0.5	0	0.0
Left the driveway without precaution	20	0.4	0	0.0
Overtook a vehicle that stopped to allow crossing	19	0.3	0	0.0
Overtaking a moving vehicle	14	0.3	0	0.0
Irregularly overtook a tram at the stop	10	0.2	0	0.0
Proceeded regularly without respecting speed limits	18	0.3	1	1.1
Skidding due to distracted driving	1	0.0	0	0.0
Unspecified circumstance	271	4.9	4	4.3
Undefined cases	46	0.8	0	0.0
Abnormal condition due to alcohol intoxication	35	0.6	1	1.1
Abnormal condition due to ongoing morbid conditions	1	0.0	0	0.0
Abnormal condition due to sudden illness	8	0.1	0	0.0
Abnormal condition due to narcotics or psychotropic substances	1	0.0	0	0.0

Table 4.11 Pedestrian crashes classified by vehicle circumstance

4.1.6. Pedestrians Circumstances Classification

The largest share of pedestrian crashes in NoLo occurred while pedestrians were crossing at unprotected marked crosswalks, accounting for 38 crashes (41%), substantially higher than the Milan-wide proportion of 31%. A further 11 crashes (12%) occurred at signalized crossings while pedestrians were respecting the traffic signals, whereas only 4% occurred while crossing against the signal.

Irregular crossings outside designated crossing locations accounted for 8 crashes (9%), below the corresponding Milan-wide value of 17%. In addition, 11 crashes (12%) occurred while pedestrians were walking or standing on sidewalks or platforms, and 10 crashes (11%) involved pedestrians walking in the middle of the carriageway.

The results indicate that a large proportion of pedestrian crashes in NoLo occur while pedestrians are using designated crossing facilities. This suggests that pedestrian safety is influenced not only by pedestrian behavior but also by driver compliance, crossing design, and the quality of interactions between pedestrians and vehicles at crossing locations (table 4.12).

	Milan		**Nolo**	
Circumstances of the pedestrian	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Crossed at pedestrian crossing respecting the signals	635	11.4	11	11.8
Crossed at pedestrian crossing without respecting the signals	444	8.0	4	4.3
Crossed at unprotected pedestrian crossing	1725	31.1	38	40.9
Crossed the road irregularly (outside crossing)	963	17.3	8	8.6
Crossed regularly but not at a pedestrian crossing	207	3.7	1	1.1
Suddenly emerged from behind a parked or stopped vehicle	145	2.6	1	1.1
Walked or stood on the sidewalk or platform	431	7.8	11	11.8
Walked in the middle of the road	339	6.1	10	10.8
Stopped, lingered, or played on the road	112	2.0	3	3.2
Walked on the road edge	150	2.7	1	1.1

Was walking against traffic	14	0.3	0	0.0
Got out of the vehicle with caution	18	0.3	1	1.1
Got out of the vehicle carelessly	20	0.4	0	0.0
Was working on the roadway protected by a special sign	6	0.1	0	0.0
Was working on the roadway not protected by a special sign	18	0.3	0	0.0
Was getting into a moving vehicle	14	0.3	0	0.0
Proceeded regularly	2	0.0	0	0.0
Unspecified / undefined	309	5.6	4	4.3

Table 4.12 Pedestrian crashes classified by the circumstances of pedestrians

4.1.7. Injured Pedestrians Classification

Classification by Physical State After the Crash:

Nearly all pedestrian crashes in NoLo resulted in at least one injured pedestrian (98.9%), while only one fatal crash (1.1%) was recorded during the study period, resulting in a death within the first 24 hours. No fatalities occurring after the first 24 hours were reported. This severity profile closely mirrors the Milan-wide distribution and indicates that pedestrian crashes in NoLo predominantly resulted in non-fatal injuries (table 4.15).

	Milan		**Nolo**	
Physical state after crash	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Injured	5483	98.7	92	98.9
Dead during the first 24 hours	51	0.9	1	1.1
Dead after the first 24 hours	19	0.3	0	0.0

Table 4.13 Pedestrian crashes classified by the physical state of pedestrians after crash

Pedestrian Crashes by Number of Injured Pedestrians per Crash

Most pedestrian crashes in NoLo involved a single injured pedestrian (79 crashes, 84.9%), followed by two injured pedestrians (11 crashes, 11.8%) and three injured pedestrians (2 crashes, 2.2%). One crash (1.1%) resulted in no injuries, and no crashes involved four or more injured pedestrians. Overall, pedestrian crashes in NoLo predominantly involved a single injured pedestrian, consistent with the Milan-wide distribution (table 4.14).

	Milan		Nolo	
Number of injured pedestrians	Crashes number	Percentage (%)	Crashes number	Percentage (%)
0	47	0.8	1	1.1
1	4719	85.0	79	84.9
2	705	12.7	11	11.8
3	70	1.3	2	2.2
4 or more	12	0.2	0	0.0

Table 4.14 Pedestrian crashes classified by the number of injured pedestrians per crash

4.1.8. Conclusion

The descriptive analysis of pedestrian crashes in NoLo between 2017 and 2021 reveals a clear risk profile shaped by the district's dense urban environment and high levels of pedestrian activity. A total of 93 pedestrian crashes were recorded, representing approximately 1.68% of all pedestrian crashes in Milan during the study period.

Crashes exhibit distinct temporal peaks during the morning and evening activity periods, particularly at 10:00 and 18:00, and are spatially concentrated along major corridors and straight road segments. Most crashes occurred on paved roads, under dry road surface conditions, and during clear weather.

Driver behavior emerged as a key contributing factor, with failure to yield to pedestrians at crossings accounting for 54% of crashes. On the pedestrian side, the largest share of crashes occurred while pedestrians were using designated crossing facilities, particularly unprotected pedestrian crossings (41%), whereas irregular crossing behavior accounted for only 9% of cases.

In terms of severity, 92 of the 93 crashes resulted in injuries, while one fatality was recorded during the study period. Overall, these findings provide a descriptive overview of pedestrian crash characteristics in NoLo and establish the basis for the network-based analysis and crash prediction framework developed in the following chapters.

4.2. Bicycle Crashes Descriptive Analysis

This section presents a descriptive analysis of bicycle crashes in the NoLo district between 2017 and 2021, with the objective of identifying their main temporal, spatial, environmental, and behavioral characteristics. Bicycle crashes are examined separately from pedestrian and motor-vehicle crashes due to the increasing role of cycling within NoLo's urban mobility system and the higher vulnerability of cyclists in mixed-traffic environments. The analysis aims to provide a comprehensive overview of the circumstances under which bicycle crashes occur and to identify the key factors influencing cyclist safety within the district.

4.2.1. Spatial Classification

Classification by Spatial Distribution of Crashes

As shown in Figure 4.2, bicycle crashes in NoLo are primarily concentrated along the district's main urban corridors, particularly Via Padova and Viale Monza, which serve as the principal movement axes within the study area. These corridors accommodate high levels of traffic activity and frequent interactions between cyclists and motor vehicles, contributing to elevated crash occurrence.

Moderate crash concentrations are also observed along connecting streets and near Viale Brianza. In addition, a localized concentration of crashes is visible in the area surrounding Parco Trotter, likely reflecting increased bicycle activity associated with local and recreational trips.

Overall, the spatial distribution indicates that bicycle crashes in NoLo are not randomly distributed but are concentrated along major corridors and areas characterized by high levels of mobility and interaction between different road users.

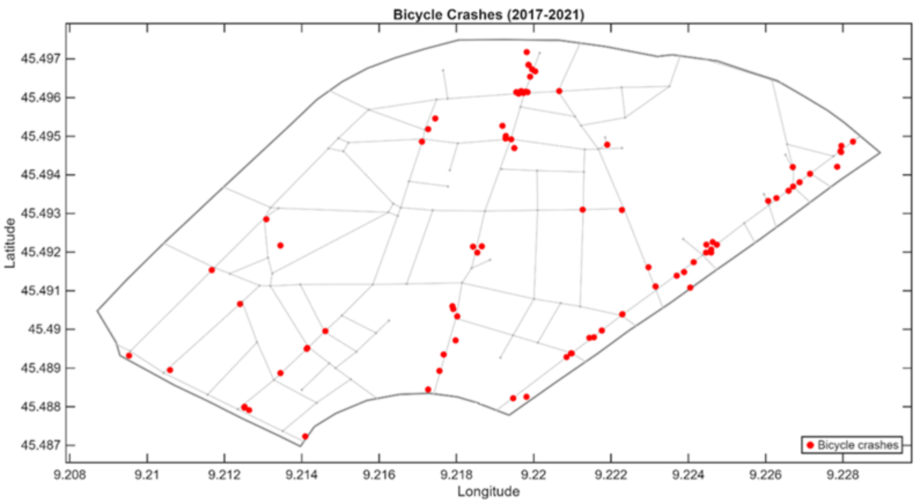

Figure 4. 2 Spatial distribution of Bicycle crashes in NoLo

Classification by Pavement Condition:

All bicycle crashes in NoLo occurred on paved roads (100%), consistent with the Milan-wide distribution, where virtually all bicycle crashes also occurred on paved surfaces. No crashes were recorded on bumpy paved or unpaved roads (table 4.1).

	Milan		Nolo	
Pavement Condition	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Paved road	5046	99.6	84	100.0
Bumpy paved road (ruined)	9	0.2	0	0.0
Unpaved or damaged road	12	0.2	0	0.0

Table 4.15 Bicycle crashes classification by road pavement condition

Classification by Road Type:

Most bicycle crashes in NoLo occurred on two-way carriageways (38 crashes, 45.2%), followed by roads with two separated carriageways (30 crashes, 35.7%). One-way carriageways accounted for 11 crashes (13.1%), while roads with more than two carriageways accounted for 5 crashes (6.0%). Compared with Milan, NoLo exhibited a higher proportion of crashes on roads with two separated carriageways and a lower proportion on one-way carriageways. Overall, bicycle crashes in NoLo were concentrated on major two-way roads and roads with separated carriageways (table 4.16).

	Milan		Nolo	
Road Type	Crashes number	Percentage (%)	Crashes number	Percentage (%)
One-way carriageway	1091	21.5	11	13.1
Two-way carriageway	2219	43.8	38	45.2
Two carriageways	1042	20.6	30	35.7
More than two carriageways	715	14.1	5	6.0

Table 4.16 Bicycle crashes classification by road type

Classification by Road Surface Condition:

Most bicycle crashes in NoLo occurred on dry road surfaces (81 crashes, 96.4%), compared with 89% across Milan. Wet road surfaces accounted for 3 crashes (3.6%), while no bicycle crashes were recorded on slippery, icy, or snowy surfaces. Overall, bicycle crashes in NoLo predominantly occurred on dry road surfaces, consistent with the observed distribution of crashes (table 4.17).

	Milan		Nolo	
Road Surface	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Dry	4501	88.8	81	96.4
Wet	544	10.7	3	3.6
Slippery	12	0.2	0	0.0
Icy	4	0.1	0	0.0
Snowy	6	0.1	0	0.0

Table 4.17 Bicycle crashes classification by road surface

Classification by Intersection Type:

Bicycle crashes in NoLo occurred predominantly on straight road segments (40 crashes, 47.6%). Signalized intersections accounted for 11 crashes (13.1%), while intersections controlled by traffic lights or traffic police accounted for 9 crashes (10.7%). Crashes at conventional intersections and roundabouts were relatively few, and no crashes were recorded at unsignalized intersections, curves, slopes, level or crossings. Undefined cases accounted for 20 crashes (23.8%). Compared with Milan, NoLo exhibited a higher proportion of bicycle crashes on straight road segments (47.6% versus 42.3%). Overall, bicycle crashes in NoLo were concentrated on straight road segments and regulated intersections (table 4.18).

	Milan		Nolo	
Intersection Type	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Intersection	281	5.5	3	3.6
Roundabout	113	2.2	1	1.2
Signalized intersection	732	14.4	11	13.1
Intersection with traffic light or traffic policeman	819	16.2	9	10.7
Unsignalized intersection	36	0.7	0	0.0
Level crossing	2	0.0	0	0.0
Straight road	2143	42.3	40	47.6
Curve	101	2.0	0	0.0
Narrow passage / off the road	14	0.3	0	0.0
Slope	3	0.1	0	0.0
Illuminated tunnel	10	0.2	0	0.0
Non-illuminated tunnel	3	0.1	0	0.0
Undefined cases	810	16.0	20	23.8

Table 4.18 Bicycle crashes classification by intersection type

Classification by Signage Presence:

Almost all bicycle crashes in NoLo occurred on roads with both vertical and horizontal signage (83 crashes, 98.8%), slightly higher than the Milan-wide proportion (95.4%). Only one crash (1.2%) occurred on a road with horizontal signage only, while no crashes were recorded on roads with no signage or only vertical signage. Overall, bicycle crashes in NoLo predominantly occurred on roads with both vertical and horizontal signage, consistent with the Milan-wide distribution (table 4.19).

	Milan		Nolo	
Signage Presence	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Absence of signs	132	2.6	0	0.0
Vertical signs only	33	0.7	0	0.0
Horizontal signs only	55	1.1	1	1.2
Both vertical and horizontal	4835	95.4	83	98.8
Undefined cases	12	0.2	0	0.0

Table 4.19 Bicycle crashes classification by signage presence

4.2.2. Temporal Classification

Classification by Year of Occurrence

A total of 84 bicycle crashes were recorded in NoLo during the study period, representing approximately 1.7% of all bicycle crashes in Milan. Bicycle crashes decreased from 18 in 2017 to 12 in 2019, then increased to 20 in 2020 and 21 in 2021. A similar trend was observed across Milan, although NoLo accounted for a higher share of city-wide bicycle crashes in 2020 (2.5%). The increase in bicycle crashes during 2020 and 2021 may reflect changes in travel behavior associated with the COVID-19 pandemic (table 4.20).

	Milan		NoLo		NoLo / Milan
Year	Crashes number	Percentage (%)	Crashes number	Percentage (%)	Percentage (%)
2017	1036	20.4	18	21.4	1.7
2018	990	19.5	13	15.5	1.3
2019	980	19.3	12	14.3	1.2
2020	812	16.0	20	23.8	2.5
2021	1249	24.7	21	25.0	1.7

Table 4.20 Bicycle crashes classification by year of occurrence

Classification by Hour of the Day

Bicycle crashes in NoLo were distributed throughout the day, with the highest concentration occurring during the afternoon. Crash occurrence was relatively low during mid night-time hour to early morning hours (00:00–07:00). Crash frequency increased during the morning period, with about 5% of crashes occurring at 08:00, 09:00, and 10:00. A further increase was observed during the late morning and early afternoon, with 9.5% of crashes recorded at both 11:00 and 13:00. The highest concentration occurred at 15:00 (10.7%), while crash frequencies remained relatively elevated between 14:00 and 18:00, with a secondary peak at 18:00 (8.3%). After 19:00, crash occurrence gradually declined. Compared with Milan, NoLo exhibited a broadly similar temporal pattern, characterized by low night-time and higher daytime crash frequencies, with the highest concentration occurring at 15:00 (table 4.21).

	Milan		**Nolo**	
Hour	Crashes number	Percentage (%)	Crashes number	Percentage (%)
00:00	72	1.4	1	1.2
01:00	47	0.9	1	1.2
02:00	28	0.6	0	0.0
03:00	17	0.3	0	0.0
04:00	16	0.3	0	0.0
05:00	25	0.5	2	2.4
06:00	46	0.9	1	1.2
07:00	172	3.4	0	0.0
08:00	350	6.9	4	4.8
09:00	381	7.5	4	4.8
10:00	264	5.2	4	4.8
11:00	308	6.1	8	9.5
12:00	281	5.5	4	4.8
13:00	297	5.9	8	9.5
14:00	301	5.9	6	7.1
15:00	344	6.8	9	10.7
16:00	310	6.1	5	6.0
17:00	396	7.8	5	6.0
18:00	394	7.8	7	8.3

19:00	319	6.3	2	2.4
20:00	283	5.6	4	4.8
21:00	206	4.1	4	4.8
22:00	122	2.4	3	3.6
23:00	88	1.7	2	2.4

Table 4.21 Bicycle crashes classification by hour of the day

4.2.3. Weather Conditions Classification

Most bicycle crashes in NoLo occurred under clear weather conditions (78 crashes, 92.9%), slightly higher than the corresponding Milan-wide proportion (88.9%). Rain accounted for 2 crashes (2.4%), while no bicycle crashes were recorded during fog, hail, snow, or strong wind conditions. The remaining 4 crashes (4.8%) occurred under other weather conditions. Overall, bicycle crashes in NoLo predominantly occurred under clear weather conditions, consistent with the Milan-wide distribution (table 4.22).

	Milan		**Nolo**	
Weather Condition	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Clear	4505	88.9	78	92.9
Fog	38	0.8	0	0.0
Rain	334	6.6	2	2.4
Hail	4	0.1	0	0.0
Snow	5	0.1	0	0.0
Strong wind	2	0.0	0	0.0
Other	179	3.5	4	4.8

Table 4.22 Bicycle crashes classification by weather conditions

4.2.4. Bicycle Crashes by Nature of Crash

Front-side collisions were the most frequent bicycle crash type in NoLo (28 crashes, 33.3%), consistent with the Milan-wide pattern (39.7%). Side collisions accounted for 18 crashes (21.4%), followed by collisions with temporarily stopped vehicles (13 crashes, 15.5%) and parked vehicles (8 crashes, 9.5%). Rear collisions and bicycle mechanical problems each accounted for 5 crashes (6.0%), while collisions with accidental obstacles represented 4 crashes (4.8%). Front collisions and pedestrian impacts were relatively uncommon, and no crashes involved trains, sudden braking, or falls from the bicycle. Overall, bicycle crashes in NoLo were predominantly associated with interactions between cyclists and vehicles, particularly front-side and side collisions (table 4.23).

	Milan		Nolo	
Nature of crash	**Crashes number**	**Percentage (%)**	**Crashes number**	**Percentage (%)**
Front collision	139	2.7	2	2.4
Front-side collision	2013	39.7	28	33.3
Side collision	775	15.3	18	21.4
Rear collision	292	5.8	5	6.0
Pedestrian hit	190	3.7	1	1.2
Collision with temporarily stopped vehicle	701	13.8	13	15.5
Collision with parked vehicle	194	3.8	8	9.5
Collision with accidental obstacles	233	4.6	4	4.8
Collision with train	1	0.0	0	0.0
Leakage / problems in the bicycle	509	10.0	5	6.0
Sudden braking	7	0.1	0	0.0
Fall from vehicle	13	0.3	0	0.0

Table 4. 23 Bicycle crashes classification by nature of crash

4.2.5. Vehicle Type Involved Classification

Bicycle Crashes by First Vehicle Type Involved:

In 50.0% of bicycle crashes in NoLo, the bicycle itself was recorded as the first vehicle, slightly higher than the Milan-wide proportion (47.4%). Among motor vehicles, private cars were the most frequently involved (21 crashes, 25.0%), followed by trucks and motorcycles (5 crashes each, 6.0%). Mopeds and public cars each accounted for 2 crashes (2.4%), while vehicles that fled and electric scooters were involved in one crash each (1.2%). Overall, bicycle crashes in NoLo were primarily associated with bicycles and private cars, consistent with the Milan-wide distribution (table 4.24).

	Milan		Nolo	
Vehicle Type (First Vehicle)	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Private car	1455	28.7	21	25.0
Public car	78	1.5	2	2.4
Rescue or police car	4	0.1	0	0.0
Bus or trolleybus	13	0.3	0	0.0
Regular or non-scheduled extra-urban buses	10	0.2	0	0.0
Tram	5	0.1	0	0.0
Truck	168	3.3	5	6.0
Articulated bus / truck	5	0.1	0	0.0
Special vehicle	17	0.3	1	1.2
Road tractor	2	0.0	0	0.0
Velocipede / bicycle	2400	47.4	42	50.0
Moped	38	0.8	2	2.4
Motorcycle only	310	6.1	5	6.0
Motorcycle with passenger	21	0.4	0	0.0
Animal or hand-drawn vehicle	1	0.0	0	0.0
Vehicle that fled	183	3.6	1	1.2
Quadricycle	10	0.2	0	0.0
Electric scooter	16	0.3	1	1.2
Electric bicycle	19	0.4	0	0.0
Undefined cases	312	6.2	5	6.0

Table 4.24 Bicycle crashes classified by type of first vehicle involved

Bicycle Crashes by Second Vehicle Type Involved:

The bicycle itself was the most frequently recorded second vehicle (42 crashes, 50.0%), followed by private cars (16 crashes, 19.0%). Trucks, motorcycles, and hit-and-run vehicles each accounted for one crash (1.2%), while no cases involved buses, articulated vehicles, mopeds, or electric bicycles. Crashes without a second recorded vehicle accounted for 23 cases (27.4%). Overall, the distribution was broadly consistent with the Milan-wide pattern, with bicycles and private cars representing the most common second vehicles (table 4.25).

	Milan		Nolo	
Vehicle Type (Second Vehicle)	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Private car	742	14.6	16	19.0
Public car	63	1.2	0	0.0
Rescue or police car	10	0.2	0	0.0
Bus or trolleybus in urban service	11	0.2	0	0.0
Regular or non-scheduled extra-urban bus / Electric scooter / Tram	5	0.1	0	0.0
Truck	74	1.5	1	1.2
Articulated bus or articulated truck	2	0.0	0	0.0
Special vehicle	9	0.2	0	0.0
Velocipede / bicycle	2678	52.9	42	50.0
Moped	10	0.2	0	0.0
Motorcycle only	67	1.3	1	1.2
Motorcycle with passenger	2	0.0	0	0.0
Motor truck / van / Quadricycle	1	0.0	0	0.0
Vehicle that fled / hit-and-run	61	1.2	1	1.2
Electric bicycle	19	0.4	0	0.0
Undefined / No second vehicle	1302	25.7	23	27.4

Table 4.25 Bicycle crashes classified by type of second vehicle involved

Bicycle Crashes by Third Vehicle Type Involved

No third vehicle was recorded in 82 bicycle crashes (97.6%), consistent with the Milan-wide distribution (97.9%). Only 2 crashes (2.4%) involved a hit-and-run vehicle as the third party, while no other third vehicle types were recorded. Overall, bicycle crashes in NoLo predominantly involved no third vehicle, indicating that most crashes involved one or two parties (table 4.26).

	Milan		Nolo	
Vehicle Type (Third Vehicle)	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Private car	51	1.0	0	0.0
Public car	4	0.1	0	0.0
Bus or trolleybus in urban service	4	0.1	0	0.0
Tram	2	0.0	0	0.0
Truck	1	0.0	0	0.0
Special vehicle	2	0.0	0	0.0
Velocipede / bicycle	37	0.7	0	0.0
Motorcycle only	4	0.1	0	0.0
Vehicle that fled / hit-and-run	2	0.0	2	2.4
Electric scooter	1	0.0	0	0.0
Undefined / No third vehicle involved	4959	97.9	82	97.6

Table 4.26 Bicycle crashes classified by type of third vehicle involved

4.2.6. Vehicle Driver Circumstances Classification

Driver Circumstances (First Vehicle Involvement) Classification:

Unspecified circumstances accounted for 6 crashes (7.1%) in NoLo, compared with 5.1% across Milan. At intersections, the most common circumstance was proceeding regularly without turning (11 crashes, 13.1%), while failure to maintain a safe distance, failure to give priority, failure to respect give-way signs, and excessive speed each accounted for between 2% and 4% of crashes. Outside intersections, proceeding regularly and driving distractedly or indecisively were the dominant circumstances, each representing 20 crashes (23.8%), considerably higher than the corresponding Milan-wide proportions. Skidding without collision accounted for 5 crashes (6.0%), while undefined cases represented 4 crashes (4.8%). No bicycle crashes in NoLo were associated with mechanical failures, alcohol intoxication, medical conditions, or driver dazzling (table 4.27).

	Milan		Nolo	
Circumstance for First Vehicle involved	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Unspecified circumstance	256	5.1	6	7.1
INTERSECTION (between 2 vehicles moving)				
Proceeded regularly without turning	525	10.4	11	13.1
Was driving distractedly or indecisively	271	5.3	0	0.0
Without maintaining safe distance	62	1.2	3	3.6
Without giving priority to vehicle from right	116	2.3	2	2.4
Without respecting stop sign	81	1.6	1	1.2
Without respecting give-way signal	181	3.6	2	2.4
Against traffic	32	0.6	1	1.2
Without respecting traffic/access prohibition signs	34	0.7	1	1.2
Excessively speeding	65	1.3	2	2.4
Without respecting speed limits	2	0.0	0	0.0
Turned right regularly	21	0.4	1	1.2
Turned right irregularly	122	2.4	1	1.2
Turned left regularly	8	0.2	0	0.0
Turned left irregularly	50	1.0	0	0.0
Overtaking at intersection	7	0.1	0	0.0
WITHOUT INTERSECTIONS (between 2 vehicles moving)				
Proceeded regularly	403	8.0	20	23.8
Was driving distractedly or indecisively	269	5.3	20	23.8
Without maintaining safe distance	108	2.1	3	3.6
Excessively speeding	82	1.6	2	2.4
Without respecting speed limits	2	0.0	0	0.0
Not close to right edge of road	32	0.6	1	1.2

Against traffic	18	0.4	1	1.2
Without respecting traffic/access prohibition signs	15	0.3	1	1.2
Overtook regularly	5	0.1	0	0.0
Overtook irregularly on the right	4	0.1	0	0.0
Overtaking irregularly when cornering or on a hill	1	0.0	0	0.0
Overtook irregularly a vehicle that was overtaking another	5	0.1	0	0.0
Overtook illegally without observing prohibition sign	4	0.1	0	0.0
Maneuvered in conversion	24	0.5	1	1.2
Maneuvered to enter traffic flow	56	1.1	0	0.0
Maneuvered to turn left	38	0.7	1	1.2
Maneuvered backwards to stop or park	10	0.2	0	0.0
Joined other two-wheeled vehicles irregularly	11	0.2	0	0.0
VEHICLE INVOLVEMENT IN PRESENCE OF PEDESTRIAN				
Proceeded regularly	67	1.3	0	0.0
Proceeded regularly with excessive speed	31	0.6	0	0.0
Without respecting speed limits	2	0.0	0	0.0
Proceeded regularly against traffic	25	0.5	0	0.0
Was maneuvering	16	0.3	0	0.0
Was maneuvering without respecting traffic light	5	0.1	0	0.0
Left driveway without precaution	3	0.1	0	0.0
Went off road and hit the pedestrian	2	0.0	0	0.0
Did not give priority to pedestrians at crossings	21	0.4	0	0.0

Hit the pedestrian with the load	2	0.0	0	0.0
Irregularly overtook a tram at the stop	1	0.0	0	0.0
VEHICLE INVOLVEMENT CRASHING NON-MOVING CAR				
Proceeded regularly	537	10.6	0	0.0
Driving distractedly or indecisively	314	6.2	0	0.0
Without maintaining safe distance	37	0.7	0	0.0
Against traffic	18	0.4	0	0.0
Excessively speeding	73	1.4	0	0.0
Without respecting traffic/access prohibition signs	7	0.1	0	0.0
Overtaking another moving vehicle	2	0.0	0	0.0
Carelessly crossed level crossing	1	0.0	0	0.0
VEHICLE INVOLVEMENT WITHOUT COLLISION (1st involver)				
Skidding to avoid collision	203	4.0	2	2.4
Skidding due to distracted driving	253	5.0	3	3.6
Skidding due to excess speed	24	0.5	0	0.0
Suddenly brakes resulting in injuries	5	0.1	0	0.0
Person falls from vehicle due to door opening	2	0.0	0	0.0
Person falls from vehicle due to getting out while moving	1	0.0	0	0.0
Person falling due to inadequate positioning	6	0.1	0	0.0
Undefined cases	491	9.7	4	4.8
Additional Driver and Vehicle Conditions				

Breakage of brakes / lack of visual devices	2	0.0	0	0.0
Steering breakage / burst or excessive tire wear	2	0.0	0	0.0
Abnormal for alcohol intoxication	4	0.1	0	0.0
Abnormal due to ongoing morbid conditions	2	0.0	0	0.0
Dazzled	31	0.6	0	0.0

Table 4.27 Bicycle crashes classified by first-vehicle circumstances and driver conditions

Circumstances for Second Vehicle

Among intersection-related circumstances, the most common category was proceeding regularly without turning (8 crashes, 9.5%), followed by driving distractedly or indecisively (7 crashes, 8.3%). Failure to respect stop signs and give-way signs each accounted for 3 crashes (3.6%), while the remaining categories were infrequent. Outside intersections, proceeding regularly was the most common circumstance (10 crashes, 11.9%), followed by driving distractedly or indecisively (7 crashes, 8.3%). Maneuvering to enter the traffic flow accounted for 3 crashes (3.6%), while excessive speed accounted for 2 crashes (2.4%). Regarding the condition of the second vehicle, vehicles stopped in a regular position represented the largest category (11 crashes, 13.1%), followed by vehicles stopped in an irregular position (7 crashes, 8.3%) and accidental obstacles (6 crashes, 7.1%). Pedestrian-related circumstances were uncommon, and no crashes were associated with bicycle equipment deficiencies, alcohol intoxication, or medical conditions. Overall, bicycle crashes involving a second vehicle were mainly associated with normal traffic movements, stopped vehicles, and roadside obstacles (table 4.28).

	Milan		**Nolo**	
Circumstance	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Unspecified circumstance	256	5.1	6	7.1
INTERSECTION (between 2 vehicles moving)				
Proceeded regularly without turning	790	15.6	8	9.5
Was driving distractedly or indecisively	297	5.9	7	8.3
Without maintaining safe distance	9	0.2	0	0.0

Without giving priority to vehicle from right	63	1.2	1	1.2
Without respecting stop sign	66	1.3	3	3.6
Without respecting give-way signal	78	1.5	3	3.6
Against traffic	90	1.8	1	1.2
Without respecting traffic lights or officer signals	36	0.7	0	0.0
Without respecting traffic/access prohibition signs	36	0.7	1	1.2
Excessively speeding	14	0.3	0	0.0
Turned right regularly	12	0.2	0	0.0
Turned right irregularly	29	0.6	0	0.0
Turned left regularly	10	0.2	1	1.2
Turned left irregularly	46	0.9	2	2.4
Overtaking at intersection	4	0.1	0	0.0
WITHOUT INTERSECTIONS (between 2 vehicles moving)				
Proceeded regularly	595	11.7	10	11.9
Was driving distractedly or indecisively	230	4.5	7	8.3
Without maintaining safe distance	12	0.2	0	0.0
Excessively speeding	22	0.4	2	2.4
Without respecting speed limits	2	0.0	0	0.0
Not close to right edge of road	55	1.1	1	1.2
Against traffic	34	0.7	1	1.2
Without respecting traffic/access prohibition signs	19	0.4	1	1.2
Overtook regularly	2	0.0	0	0.0

Overtook irregularly on the right	1	0.0	0	0.0
Maneuvered in conversion	20	0.4	1	1.2
Maneuvered to enter traffic flow	82	1.6	3	3.6
Maneuvered to turn left	23	0.5	0	0.0
Maneuvered backwards to stop or park	11	0.2	0	0.0
Joined other two-wheeled vehicles irregularly	2	0.0	0	0.0
Pedestrian state				
Walked or stood on sidewalk or platform	42	0.8	0	0.0
Walked regularly on edge of road	12	0.2	1	1.2
Was walking against traffic	4	0.1	0	0.0
Was walking in middle of road	18	0.4	0	0.0
Stopped, lingered or played on road	4	0.1	0	0.0
Was getting into a moving vehicle	1	0.0	0	0.0
Got out of vehicle with caution	2	0.0	0	0.0
Got out of vehicle carelessly	2	0.0	0	0.0
Suddenly came out from behind parked vehicle	6	0.1	0	0.0
Crossed at pedestrian crossing respecting signals	13	0.3	2	2.4
Crossed at pedestrian crossing without respecting signals	11	0.2	2	2.4
Crossed at unprotected pedestrian crossing	20	0.4	0	0.0
Crossed road regularly not at pedestrian crossing	4	0.1	0	0.0
Crossed road irregularly	36	0.7	2	2.4

Vehicle B state				
Accidental obstacle	163	3.2	6	7.1
Vehicle stopping in regular position	379	7.5	11	13.1
Vehicle stopping in irregular position	415	8.2	7	8.3
Vehicle stopping without required sign	45	0.9	3	3.6
Vehicle stopped regularly reported	15	0.3	0	0.0
Fixed obstacle in roadway	61	1.2	0	0.0
Train crossing at level	1	0.0	0	0.0
Hole	1	0.0	0	0.0
VEHICLE INVOLVEMENT WITHOUT COLLISION (2nd involver)				
Accidental obstacle	95	1.9	6	7.1
Pedestrian	48	0.9	1	1.2
Animal avoided	8	0.2	1	1.2
Vehicle	39	0.8	2	2.4
Hole avoided	46	0.9	1	1.2
Without obstacles, pedestrians, or other vehicles	217	4.3	3	3.6
Fixed obstacle	40	0.8	1	1.2
Undefined cases	373	7.4	3	3.6
ADDITIONAL DRIVER AND VEHICLE CONDITIONS				
Lack of visual devices of bicycles	1	0.0	0	0.0
Abnormal for alcohol intoxication	10	0.2	0	0.0
Abnormal due to ongoing morbid conditions	3	0.1	0	0.0

Table 4. 28 Bicycle crashes classified by vehicle circumstance and involvement

4.2.7. Injured Cyclists Classification

Classification by Physical State After the Crash

All bicycle crashes recorded in NoLo resulted in injuries (100%), with no fatalities reported during or after the first 24 hours. This pattern is consistent with the Milan-wide distribution, where fatalities accounted for only a very small proportion of bicycle crashes (table 4.29).

	Milan		Nolo	
Physical State After Crash	Crashes number	Percentage (%)	Crashes number	Percentage (%)
Injured	5053	99.7	84	100.0
Dead during the first 24 hours	12	0.2	0	0.0
Dead after the first 24 hours	2	0.0	0	0.0

Table 4.29 Bicycle crashes classified by the physical state of cyclists after crash occurrence

Classification by Number of Injured Cyclists per Crash:

Most bicycle crashes in NoLo involved a single injured cyclist (76 crashes, 90.5%), while 7 crashes (8.3%) involved two injured cyclists and 1 crash (1.2%) involved three injured cyclists. No crashes involved four or more injured cyclists. Overall, bicycle crashes in NoLo predominantly involved a single injured cyclist, consistent with the Milan-wide distribution (table 4.30).

	Milan		Nolo	
Number of Injured Cyclists	Crashes number	Percentage (%)	Crashes number	Percentage (%)
0	11	0.2	0	0.0
1	4652	91.8	76	90.5
2	369	7.3	7	8.3
3	29	0.6	1	1.2
4	6	0.1	0	0.0

Table 4.30 Bicycle crashes classified by the number of injured cyclists per crash

4.2.8. Conclusion

The descriptive analysis of bicycle crashes in NoLo between 2017 and 2021 identified 84 crashes, representing approximately 1.66% of all bicycle crashes in Milan. Crashes were concentrated along the district's principal corridors, particularly Via Padova and Viale Monza, and occurred mainly on two-way carriageways and roads with separated carriageways.

Bicycle crashes were distributed throughout the day, with the highest frequencies recorded during the late morning and afternoon. Unlike pedestrian crashes, bicycle crashes increased during 2020 and 2021. Most crashes occurred on paved roads, under dry road surface conditions and clear weather, consistent with the Milan-wide distribution.

Front-side and side collisions were the most common crash types, while private cars were the most frequently involved motor vehicles. Driver circumstances were dominated by normal traffic movements and distracted or indecisive driving, with stopped vehicles and roadside obstacles also representing common circumstances in crashes involving a second vehicle. No crashes were associated with alcohol intoxication, medical conditions, or vehicle defects.

All bicycle crashes resulted in injuries, with no fatalities recorded during the study period, and most involved a single injured cyclist. Overall, these findings provide the descriptive basis for the network-based analysis and crash prediction framework presented in the following chapter.

4.3. Crosswalk-Crashes Distance Analysis

Crosswalks represent critical interaction points between pedestrians, cyclists, and motorized traffic within urban road networks. Although they are designed to improve safety and facilitate road crossings, they also concentrate movements of different road users and may therefore become locations where conflicts and crashes are more likely to occur. Understanding the spatial relationship between crashes and crosswalk locations can provide valuable insight into the mechanisms contributing to road safety risks in urban environments.

To investigate this relationship in NoLo district, the distance from each pedestrian and bicycle crash to the nearest crosswalk was calculated. Crashes and crosswalks were first assigned to the road network, and distances were measured considering only crosswalks located on the same road segment as the crash location.

Figure 4.3 shows that pedestrian crashes are highly concentrated near crosswalk locations. Most crashes occur within a short distance of a crossing facility, while the number of crashes decreases rapidly as the distance increases. Only a limited number of pedestrian crashes occur farther from crosswalks, indicating a strong spatial association between pedestrian crash occurrence and crossing environments.

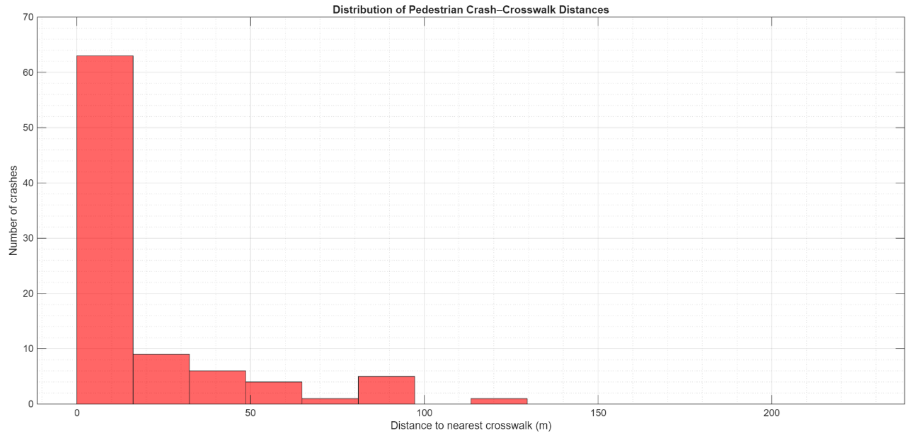

Figure 4.3 Distribution of pedestrian crash distances to the nearest crosswalk

Figure 4.4 presents the distribution of bicycle crash distances from the nearest crosswalk. Although bicycle crashes also exhibit a concentration near crosswalk locations, the distribution is more dispersed than that observed for pedestrians. Bicycle crashes occur across a wider range of distances, suggesting that bicycle safety is influenced not only by crossing locations but also by other roadway characteristics such as intersections, turning movements, traffic interactions, and roadway geometry.

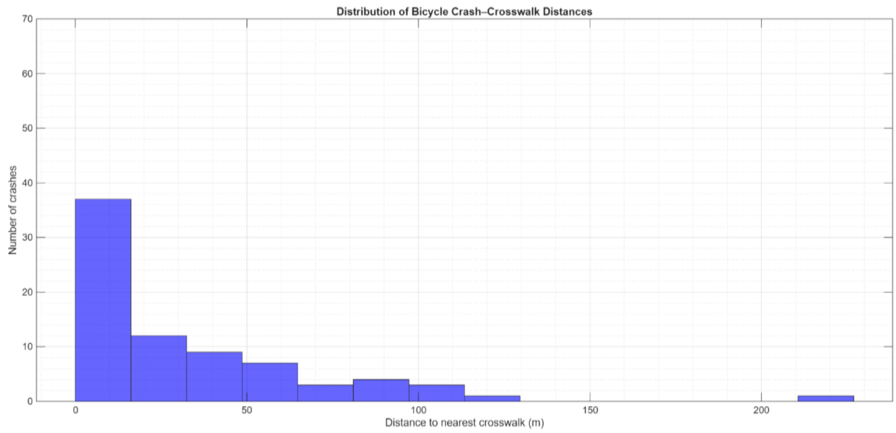

Figure 4.4 Distribution of bicycle crash distances to the nearest crosswalk

Overall, the results indicate that pedestrian crashes are more strongly associated with crosswalk locations than bicycle crashes. This finding suggests that crossing facilities represent critical safety locations for pedestrians in NoLo, whereas bicycle crash risk is distributed across a broader portion of the road network. These observations provide additional insight into the spatial mechanisms underlying crash occurrence and help explain the localized crash concentrations identified later in the network-based clustering and prediction analyses presented in Chapter 5.

5 Network-Based Analysis & Results

Chapter 3 presented the methodological framework adopted in this study, including the network-based crash prediction approach and the evaluation criteria used to assess model performance. Chapter 4 provided a descriptive analysis of crash patterns within the NoLo district, establishing the empirical context for the analyses that follow.

Building upon these foundations, this chapter presents the network-based analysis and prediction results for pedestrian, bicycle, and vehicle crashes. The performance of the different indicators and predictive scenarios is evaluated and compared in order to identify the most suitable configuration (best scenario, indicator and optimal number of clusters) for each crash category based on prediction accuracy, crash prevention efficiency, and cluster homogeneity. Particular attention is given to the spatial distribution of predicted crash concentrations with the identification of high-risk areas.

As a reference for the subsequent analyses, table 5.1 summarizes the annual distribution of pedestrian, bicycle, and vehicle crashes recorded in the NoLo district between 2017 and 2021.

Year	Pedestrian Crashes	Bicycle Crashes	Vehicle Crashes	All Crashes
2017	14	18	82	114
2018	25	13	67	105
2019	26	12	58	96
2020	9	20	48	77
2021	19	21	63	103
Total	93	84	319	495

Table 5. 1 Annual crash distribution by category in the NoLo district (2017–2021)

5.1. Pedestrian Crashes: Scenario Analysis and Results

The objective of this section is to apply the network-based clustering and probabilistic prediction framework to pedestrian crashes, evaluate its performance across different temporal configurations, and identify the optimal indicator and number of clusters for crash spatial prediction in the NoLo district.

5.1.1. Scenario 1

Scenario 1 evaluates the ability of the proposed framework to estimate pedestrian crash occurrence in 2017 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2017
- Training years: 2018–2021
- Observed pedestrian crashes in testing year: 14 crashes
- Pedestrian crashes in training years: 79 crashes

Figure 5.1 presents the spatial distribution of the 14 pedestrian crashes recorded in 2017, the crashes are distributed across different parts of the NoLo network, with no strong spatial concentration, reflecting the relatively low number of crashes recorded in this particular year.

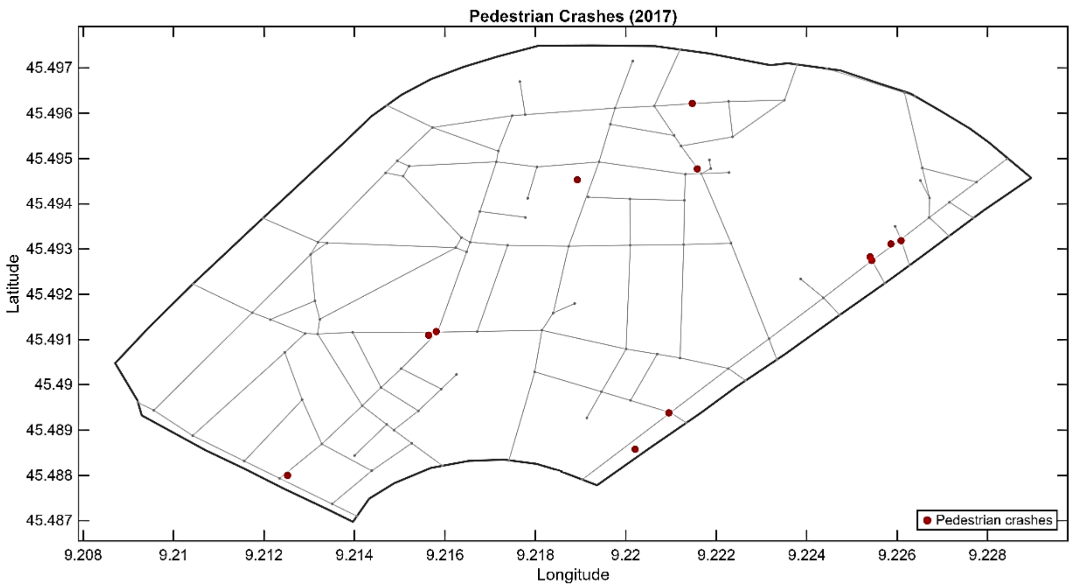

Figure 5.1 Spatial distribution of pedestrian crashes in 2017 (testing set, Scenario 1)

Figure 5.2 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 1. The results show that prediction performance varies considerably depending on both the selected indicator and the clustering configuration.

Betweenness Centrality (BTW) consistently produces the highest RMSEmod values across most clustering configurations, exhibiting highly unstable behavior particularly for small numbers of clusters. Given its inconsistent and unreliable behavior, BTW is considered unsuitable for pedestrian crash prediction and is therefore excluded from further analysis for this scenario.

In contrast, the indicators RE, RN, IE, and IN, together with CLS, demonstrate considerably lower and more stable RMSEmod values across the evaluated clustering range, among these, RN and RE exhibit the most consistent performance, maintaining low prediction errors across a wide range of cluster configurations. IE and IN also exhibit similar performance throughout the evaluated clustering range, maintaining relatively low and stable RMSEmod values. CLS shows less stable behavior up to 20 clusters, and its overall performance remains slightly less consistent than the other indicators.

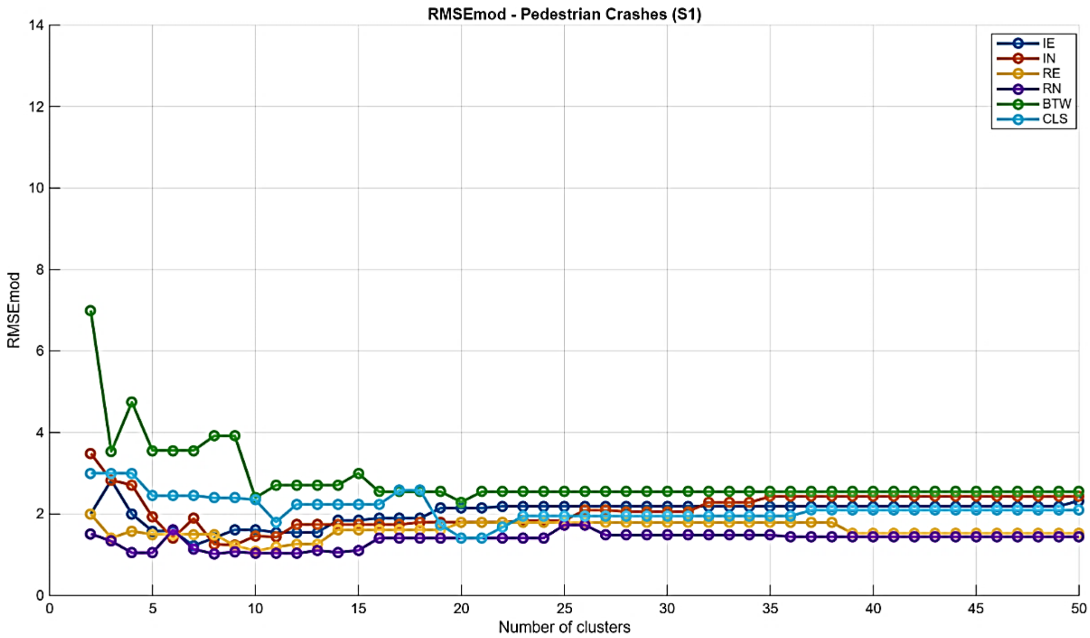

Figure 5.2 RMSEmod comparison across indicators for Scenario 1 (Pedestrian crashes)

5.1.2. Scenario 2

Scenario 2 evaluates the ability of the proposed framework to estimate pedestrian crash occurrence in 2018 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2018

- Training years: 2017, 2019–2021
- Observed pedestrian crashes in testing year: 25 crashes
- Pedestrian crashes in training years: 68 crashes

Figure 5.3 presents the spatial distribution of the 25 pedestrian crashes recorded in 2018, which constitute the testing dataset for this scenario. Compared to 2017, the crashes exhibit a wider spatial distribution across the NoLo network, with noticeable concentrations along the main corridors of Viale Monza and Via Padova.

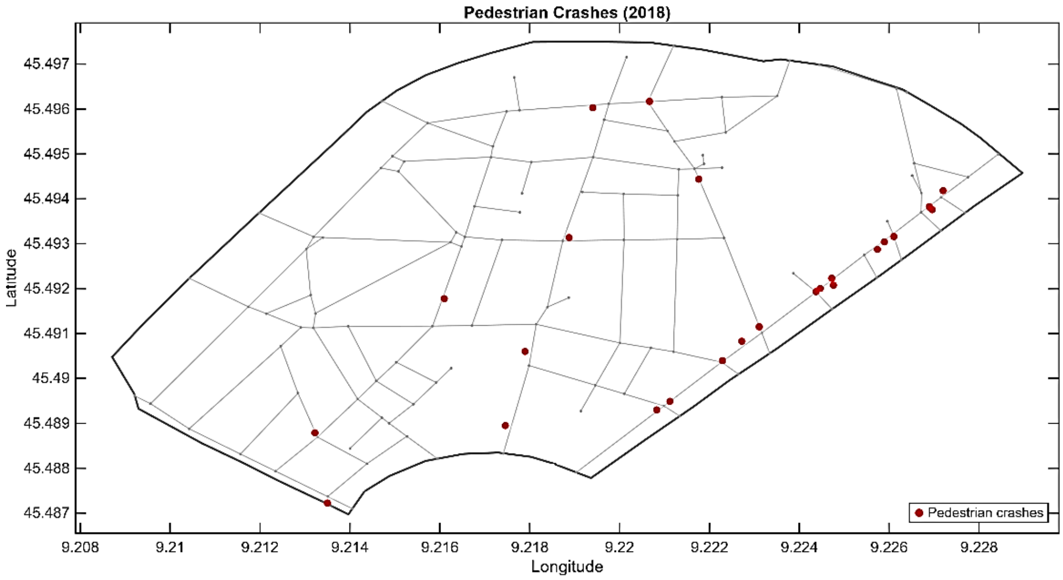

Figure 5.3 Spatial distribution of pedestrian crashes in 2018 (testing set, Scenario 2)

Figure 5.4 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 2. The results show a different behavior compared to Scenario 1, particularly for the topological measures.

Closeness Centrality (CLS) produces by far the highest RMSEmod values across all clustering configurations, exhibiting strong instability particularly for small numbers of clusters, with values reaching approximately 12 at NOC = 4. Given this highly unreliable behavior, CLS was excluded from further consideration in this scenario.

Among the remaining indicators, IN demonstrates the best overall performance, consistently producing the lowest RMSEmod values across the evaluated clustering range. RN follows closely, also showing stable and relatively low prediction errors throughout most configurations. IE performs good but with less stability and greater error, while RE shows acceptable but slightly higher errors and lowers stability than IE. BTW shows a moderate improvement in this scenario, and performs better than RE within certain clustering ranges, though it remains less competitive than IN and RN.

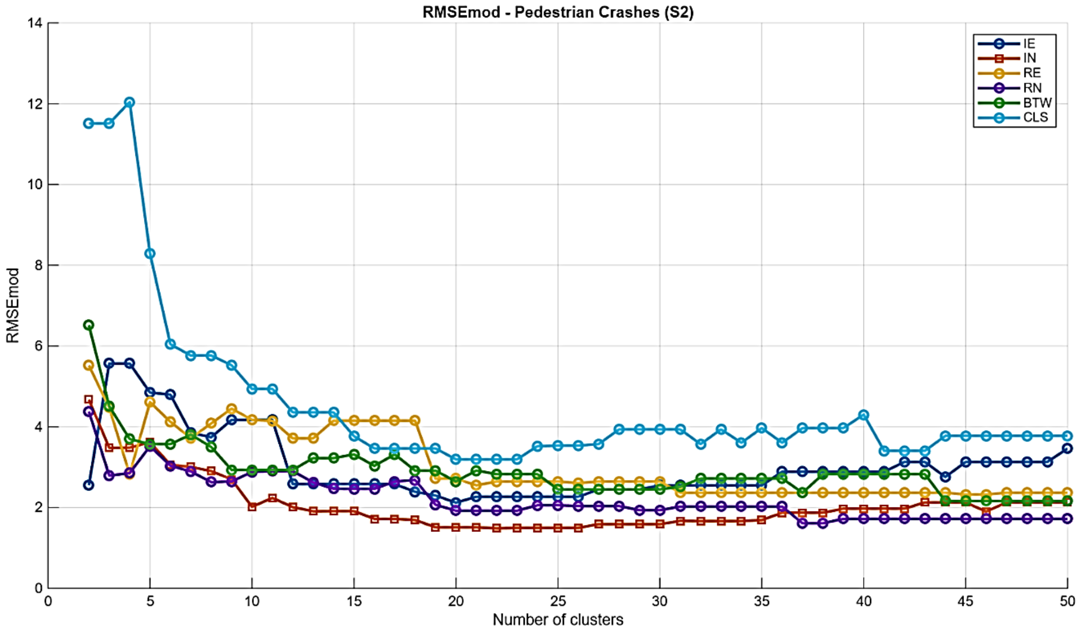

Figure 5.4 RMSEmod comparison across indicators for Scenario 2 (Pedestrian crashes)

5.1.3. Scenario 3

Scenario 3 evaluates the ability of the proposed framework to estimate pedestrian crash occurrence in 2019 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2019
- Training years: 2017–2018 and 2020–2021
- Observed pedestrian crashes in testing year: 26 crashes
- Pedestrian crashes in training years: 67 crashes

Figure 5.5 presents the spatial distribution of the 26 pedestrian crashes recorded in 2019, which constitute the testing dataset for this scenario. The crashes are mainly concentrated along Via Padova and in the western and southwestern sectors of the NoLo network, with fewer occurrences observed in the central area. This spatial pattern indicates the presence of localized pedestrian crash hotspots associated with major urban corridors and high pedestrian activity zones.

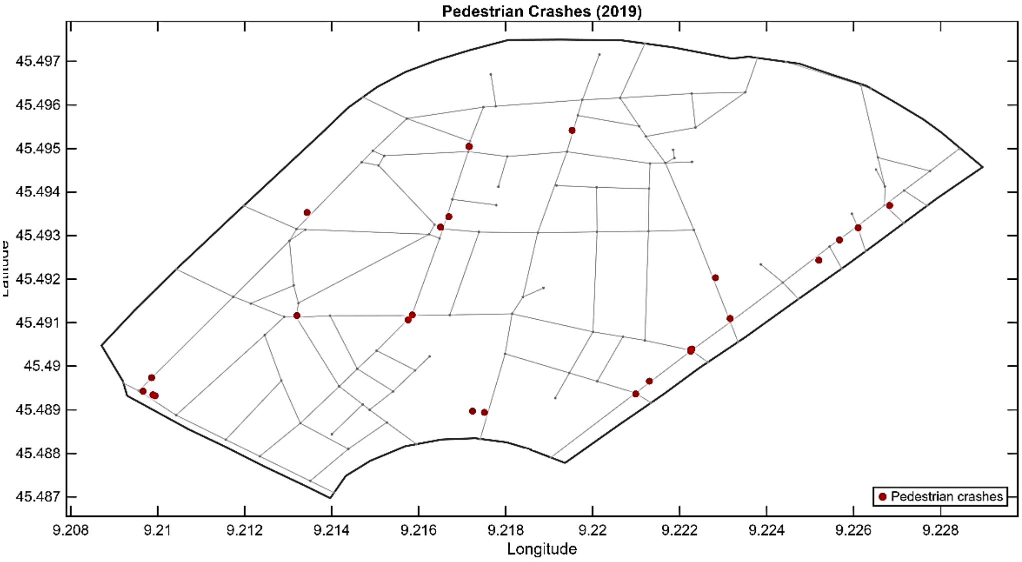

Figure 5.5 Spatial distribution of pedestrian crashes in 2019 (testing set, Scenario 3)

Figure 5.6 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 3.

Betweenness Centrality (BTW) consistently produces the highest RMSEmod values across most clustering configurations, showing strong instability for small and intermediate numbers of clusters. Given this unreliable behavior, BTW is excluded from further consideration in this scenario.

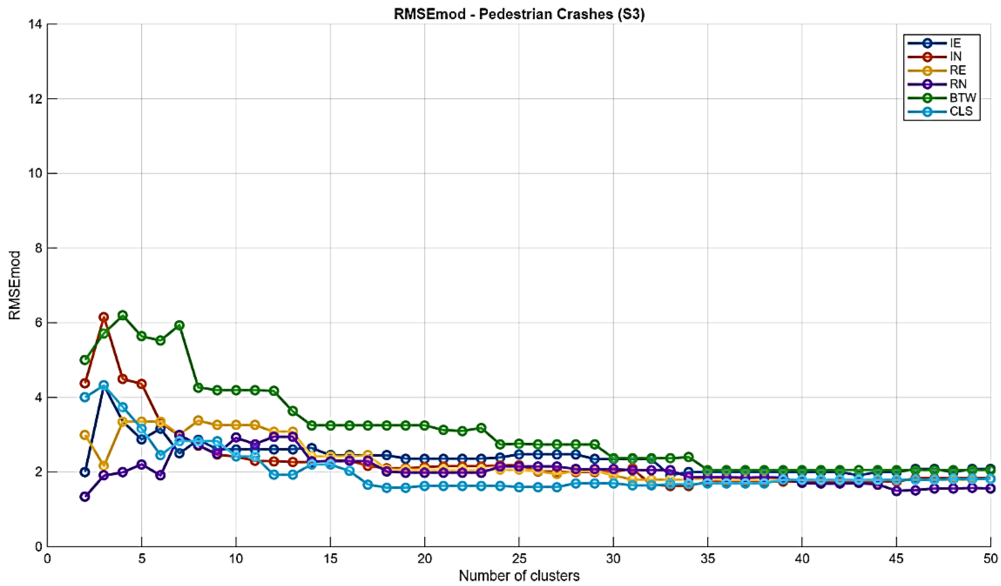

Figure 5.6 RMSEmod comparison across indicators for Scenario 3 (Pedestrian crashes)

Among the remaining indicators, RN and RE show some instability at small numbers of clusters, but stabilize and maintain low prediction errors beyond approximately 14 clusters. IE and IN show similar behavior, though with slightly more variability at smaller numbers of clusters. Notably, CLS shows a strong improvement compared to

Scenario 2, performing competitively with the remaining indicators throughout most of the evaluated clustering range.

5.1.4. Scenario 4

Scenario 4 evaluates the ability of the proposed framework to estimate pedestrian crash occurrence in 2020 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2020
- Training years: 2017–2019 and 2021
- Observed pedestrian crashes in testing year: 9 crashes
- Pedestrian crashes in training years: 84 crashes

Figure 5.7 presents the spatial distribution of the 9 pedestrian crashes recorded in 2020, which constitute the testing dataset for this scenario. The number of crashes is remarkably low compared to the other years of the study period, and the spatial distribution is highly concentrated in the central part of the NoLo district. This sharp reduction in crash occurrence is directly attributed to the COVID-19 pandemic, which led to significantly reduced mobility and traffic activity throughout the city during this period. Due to the exceptionally low crash count, the results of Scenario 4 are considered less representative of typical crash patterns and therefore carry limited weight in the final selection of the optimal configuration.

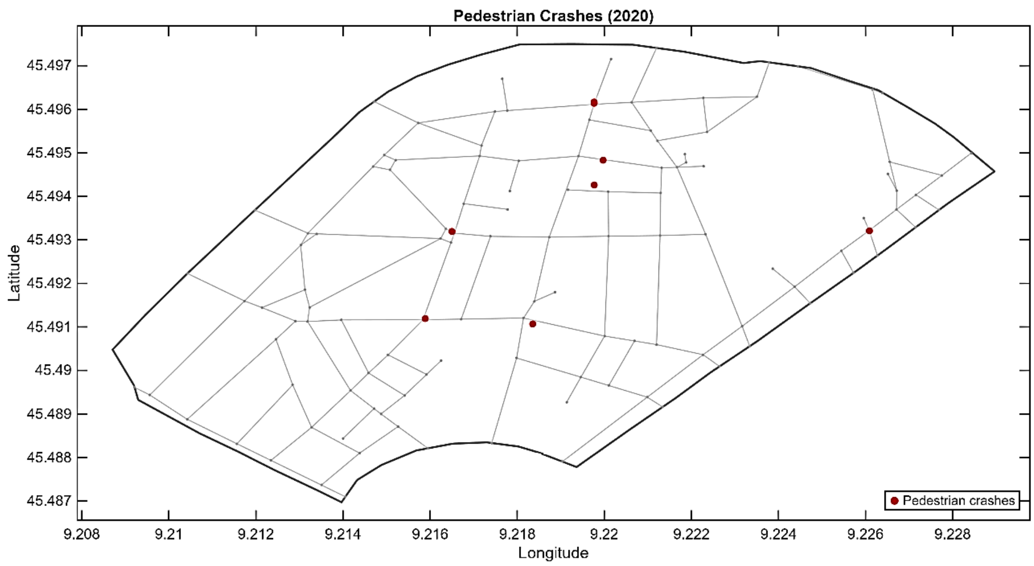

Figure 5.7 Spatial distribution of pedestrian crashes in 2020 (testing set, Scenario 4)

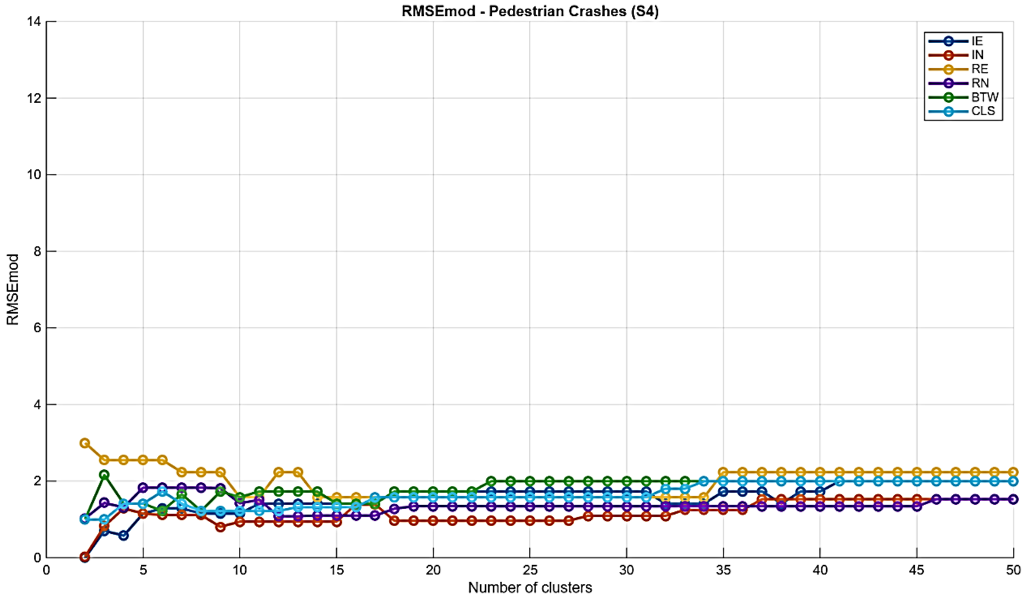

Figure 5.8 RMSEmod comparison across indicators for Scenario 4 (Pedestrian crashes)

Figure 5.8 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 4. As a direct consequence of the very low number of crashes in the testing year, all indicators produce generally low and comparable RMSEmod values across most clustering configurations. Unlike the previous scenarios, no indicator stands out as clearly superior or inferior, and all evaluated indices demonstrate acceptable predictive performance throughout the evaluated clustering range.

5.1.5. Scenario 5

Scenario 5 evaluates the ability of the proposed framework to estimate pedestrian crash occurrence in 2021 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2021
- Training years: 2017–2020
- Observed pedestrian crashes in testing year: 19 crashes
- Pedestrian crashes in the training years:74

Figure 5.9 presents the spatial distribution of the 19 pedestrian crashes recorded in 2021, which constitute the testing dataset for this scenario. The crashes show a relatively wide spatial spread across the NoLo network, with concentrations observed in the central and northeastern parts of the district, consistent with the overall spatial pattern observed across the full study period.

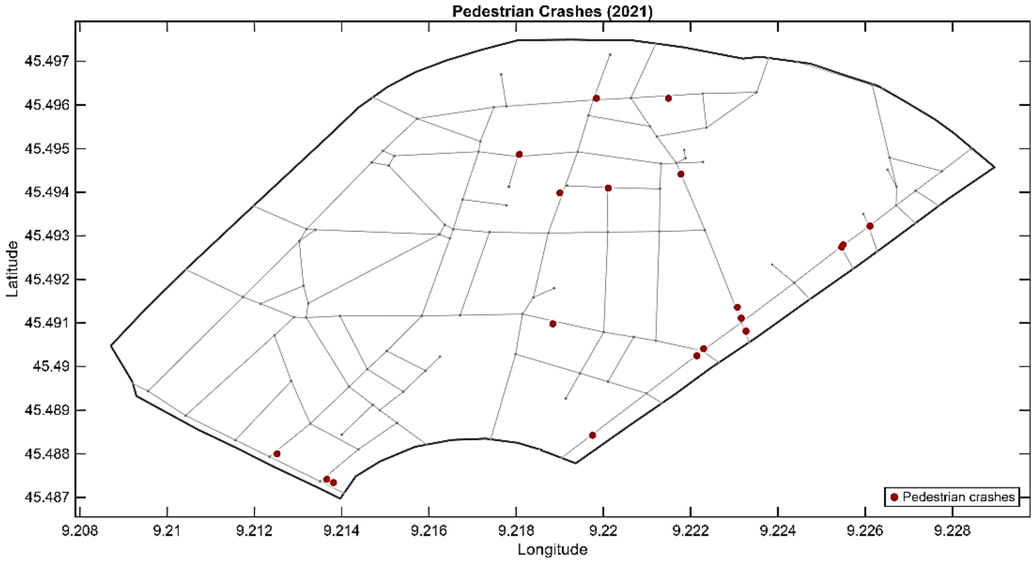

Figure 5.9 Spatial distribution of pedestrian crashes in 2021 (testing set, Scenario 5)

Figure 5.10 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 5. Betweenness Centrality (BTW) and Closeness Centrality (CLS) exhibit higher and less stable errors particularly for small numbers of clusters, with values reaching approximately 8 and 7 respectively at NOC = 2, before stabilizing progressively as the number of clusters increases. Given this unstable behavior at small cluster counts, these two indicators are less reliable for small clustering configurations.

Among the remaining indicators, all demonstrate acceptable and competitive performance throughout most of the evaluated clustering range. RN consistently produces the lowest RMSEmod values across most configurations, emerging as the best performing indicator in this scenario. RE follows closely as the second best performer, showing stable and low prediction errors throughout the range. IN performs well across most configurations, while IE shows slightly less stable behavior with some fluctuations particularly in the intermediate clustering range.

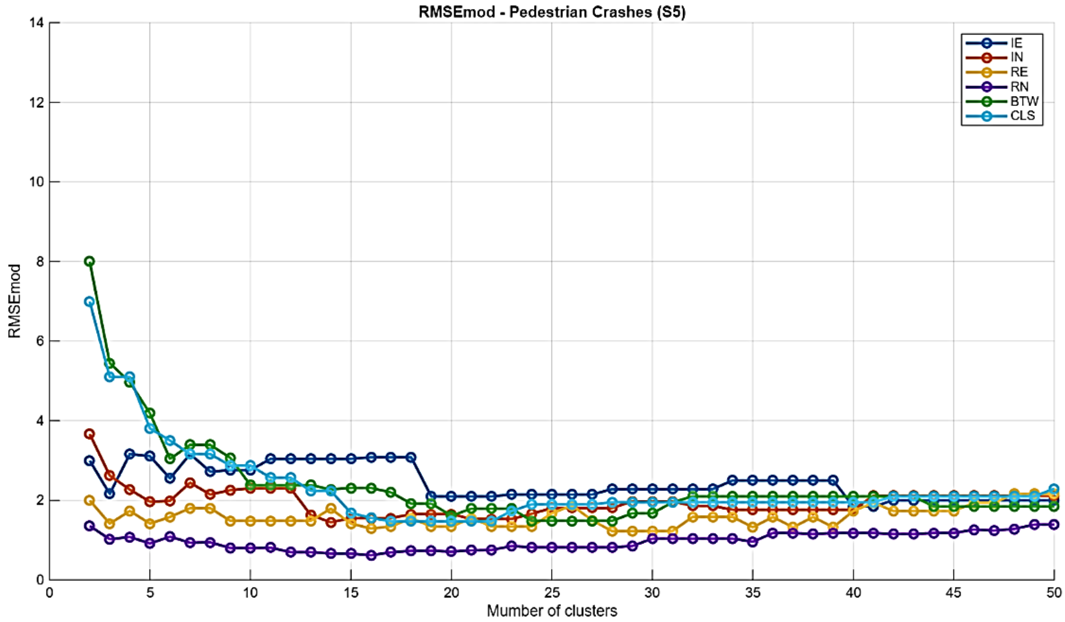

Figure 5.10 RMSEmod comparison across indicators for Scenario 5 (Pedestrian crashes)

After comparing the performance across all scenarios, the optimal indicator and number of clusters will be selected based on the overall behavior observed throughout the five temporal scenarios, with particular emphasis on Scenario 5 and the Random Scenario. Scenario 5 is considered the most realistic temporal configuration, as it uses all past years for training and evaluates the model on the most recent year, reflecting a genuine forecasting situation. The Random Scenario is equally representative, as the training data incorporates crashes from all years of the study period without being associated with any specific temporal ordering, providing a robust and comprehensive validation of the framework.

5.1.6. Random Scenario

The Random Scenario differs from the five temporal scenarios in its data splitting approach. Rather than using a year-based division, all 93 pedestrian crashes recorded across the full study period are merged and randomly divided into a training set comprising 80% of the data.

- Training set: 80% of the data (93 crashes), corresponding to approximately 74 crashes
- Testing set: 20% of the data, corresponding to 19 crashes

This configuration provides a supplementary validation that is independent of temporal ordering, and together with Scenario 5 represents the most reliable basis for the final selection of the optimal indicator and number of clusters.

Figure 5.12 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in the Random Scenario.

Betweenness Centrality (BTW) again produces the highest RMSEmod values, showing strong instability for small numbers of clusters with values reaching approximately 8 at NOC = 2, before gradually stabilizing at larger cluster counts. Given this unreliable behavior, BTW is excluded from further consideration.

In contrast, all remaining indicators, namely RN, RE, IN, IE, and CLS, demonstrate low and comparable RMSEmod values across most of the evaluated clustering range. RN consistently produces the lowest values throughout most configurations, while RE, IN, and IE all perform competitively and show stable behavior particularly within the intermediate clustering range. CLS shows less stable behavior and higher errors within the clustering range of approximately 5 to 15 clusters, before converging toward the other indicators at larger cluster counts.

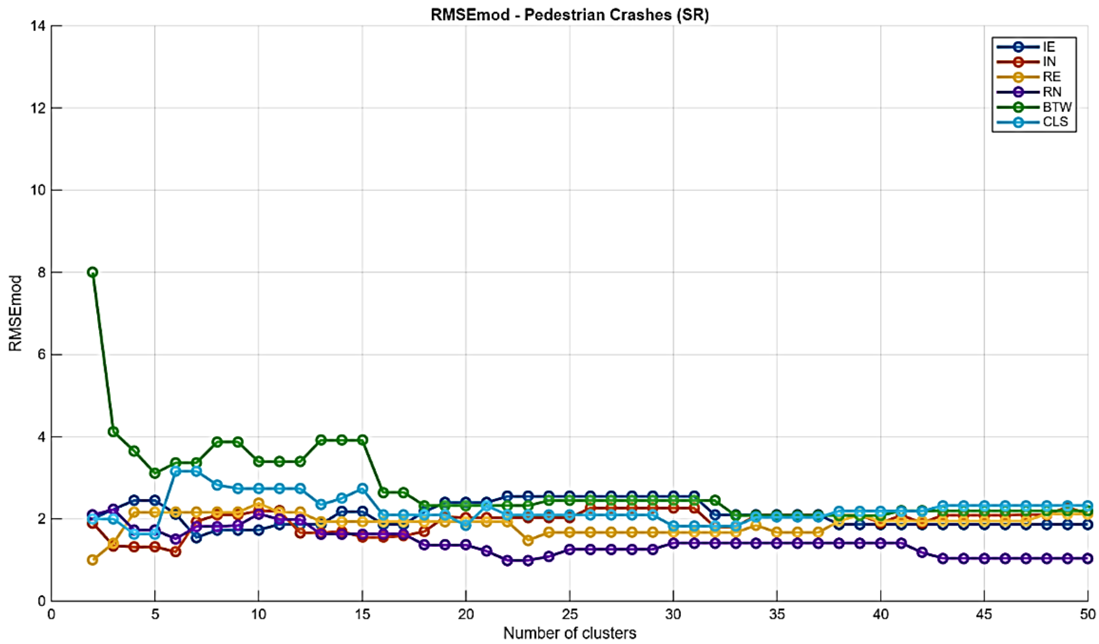

Figure 5.11 RMSEmod comparison across indicators for Random Scenario (Pedestrian crashes)

5.1.7. Results and Selected Configuration

Following the evaluation of all five temporal scenarios and the random scenario, this section presents the comparative analysis of the candidate configurations and identifies the optimal indicator and number of clusters for pedestrian crash prediction in the NoLo district.

The preliminary analysis of RMSEmod performance across all scenarios confirmed that BTW exhibits consistently unstable and poor predictive behavior and is therefore excluded from the final selection. Given the unstable behavior of CLS observed in some scenarios, the detailed evaluation focuses primarily on RN, RE, IE, and IN, considering only Scenario 5 and the Random Scenario

For each indicator, the number of clusters corresponding to the minimum RMSEmod value within the range of 8 to 25 clusters was identified separately for Scenario 5 and the Random Scenario. The selected configurations were then evaluated using three complementary criteria: the error metrics RMSEmod, Cluster Performance Efficiency (CPE), and the Homogeneity Ratio (HR). The objective was not only to minimize prediction error, but also to obtain spatially coherent clusters while maintaining a high crash capture capability for critical zones, reflected by high CPE and high HR values.

Table 5.2 presents the evaluation results for all candidate configurations across Scenario 5 and the Random Scenario. For each configuration, the top two clusters represent the most critical zones within the network, identified as those with the highest predicted crash concentration. These top clusters are therefore the primary output of the prediction framework, highlighting the areas where safety interventions should be prioritized.

Scenario	Index	NOC	RMSEmod	CPR (%)	NCR (%)	CPE	HR (%)
S5	IN	13	1.6306	49.47	17.80	131.67	72.03
S5	IN	14	1.4480	41.46	8.47	132.99	66.95
S5	IN	15	1.5502	41.58	8.47	133.11	64.41
S5	IN	17	1.5502	41.58	8.47	133.11	62.71
S5	IN	21	1.5310	27.89	8.47	119.42	56.78
SR	IN	6	1.1987	65.22	38.14	127.08	82.20
SR	IN	14	1.7016	42.11	25.42	116.69	69.49
S5	IE	7	3.1623	42.11	49.15	92.95	83.05
S5	IE	19	2.0976	15.79	11.86	103.93	61.02
SR	IE	7	1.5492	42.11	45.76	96.35	79.66
SR	IE	19	2.3979	21.05	8.47	112.58	65.25
S5	RN	9	0.801	40.26	5.08	135.18	77.12
S5	RN	11	0.810	40.26	5.08	135.18	70.4
S5	RN	12	0.698	40.26	5.08	135.18	69.5
S5	RN	16	0.6194	28.11	3.39	124.72	61.86
S5	RN	22	0.7521	24.79	3.39	121.40	53.39
SR	RN	16	1.6332	31.58	5.08	126.50	65.25
SR	RN	22	0.9951	13.37	5.93	107.44	51.69
S5	RE	16	1.2909	26.32	3.39	122.93	65.25
S5	RE	23	1.3416	26.32	3.39	122.93	56.78
SR	RE	16	1.9365	36.84	5.08	131.76	61.02
SR	RE	23	1.4832	26.32	4.24	122.08	54.24

Table 5. 2 Summary of candidate configurations for pedestrian crash prediction

The comparison of the candidate models demonstrates that the RN indicator provides the most suitable representation of pedestrian crash distribution within the NoLo road network. Among the evaluated configurations, the model based on the RN indicator with NOC = 11 under Scenario 5 achieves the best compromise between hotspot identification efficiency, network coverage, and spatial consistency.

Although some configurations exhibit lower prediction errors or slightly higher Homogeneity Ratios, these solutions are generally associated with lower Crash Prediction Efficiency (CPE) values or require a substantially larger portion of the network to capture a similar proportion of crashes. Conversely, some alternative configurations achieve comparable CPE values but at the expense of reduced spatial interpretability or less balanced overall performance.

The configuration RN–11 under Scenario 5 achieves a CPR of 40.26%, an NCR of 5.08%, a CPE of 135.18, and a Homogeneity Ratio of 70.4%. These results indicate that the model is capable of capturing a substantial proportion of pedestrian crashes while involving only a very limited fraction of the network nodes. The low NCR value demonstrates that pedestrian crash risk is concentrated within a small number of highly relevant locations, while the relatively high CPE confirms the efficiency of the selected clusters for identifying critical crash-prone areas. Furthermore, the satisfactory Homogeneity Ratio indicates that the clustering structure remains spatially coherent and interpretable.

Based on the combined evaluation of prediction accuracy, Crash Prediction Ratio (CPR), Network Coverage Ratio (NCR), Crash Prediction Efficiency (CPE), and Homogeneity Ratio (HR), the final selected configuration for pedestrian crash prediction is:

- Indicator: RN
- Scenario: S5
- Number of Clusters (NOC): 11
- Top Clusters: 9 and 10

Figure 5.12 presents the spatial distribution of the eleven clusters across the network. The clustering solution reveals a clear spatial organization of pedestrian crash risk. In particular, Clusters 9 and 10 emerge as the most critical pedestrian crash zones and represent the primary locations for targeted safety interventions.

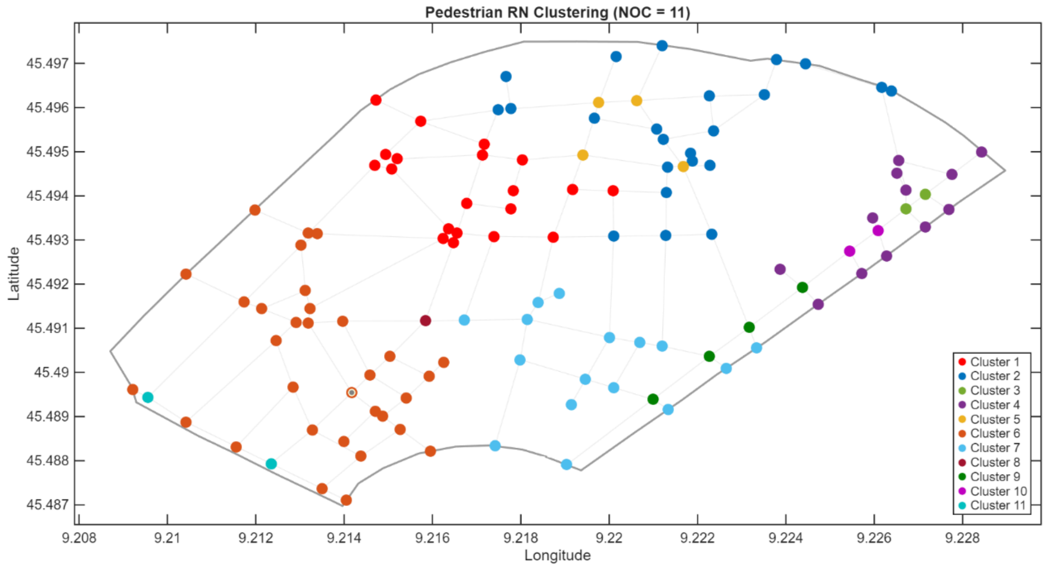

Figure 5. 12 Pedestrian crash clustering results for the selected RN–11 model under Scenario 5

Figure 5.14 compares the predicted and observed pedestrian crash frequencies for the eleven clusters obtained under Scenario 5 using the RN indicator. The results show a generally satisfactory agreement between predicted and actual crash distributions. The model successfully identifies Clusters 9 and 10 as the dominant pedestrian crash concentration zones, which correspond to the locations with the highest observed crash frequencies. While some discrepancies are present in several lower-risk clusters, the overall spatial hierarchy of pedestrian crash risk is preserved, confirming the ability of the RN-based approach to identify the most relevant pedestrian crash hotspots within the NoLo network.

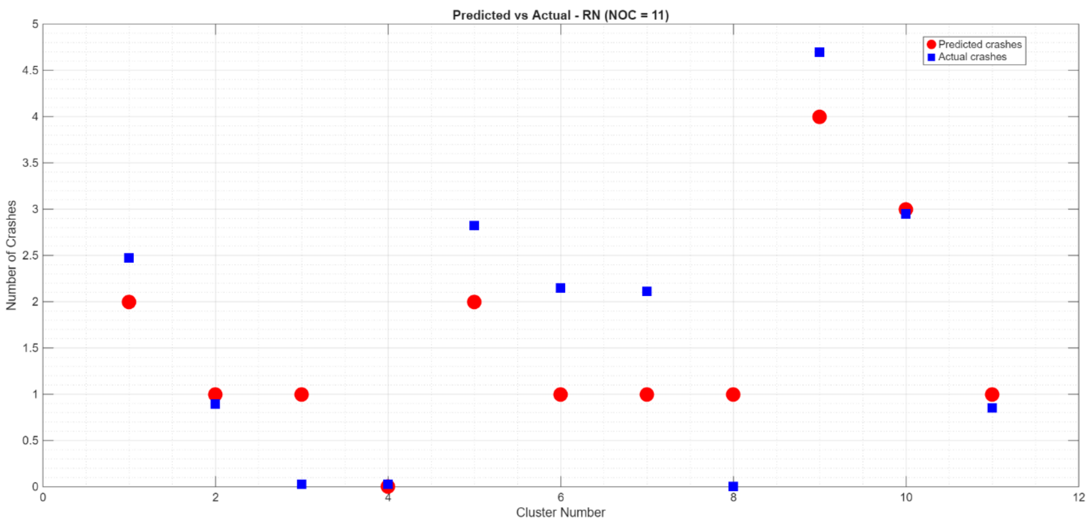

Figure 5. 13 Predicted vs actual pedestrian crash distribution across the 11 clusters of the selected RN model.

5.2. Bicycle Crashes: Scenario Analysis and Results

The scenario-based framework is subsequently applied to bicycle crashes, which represent an increasingly relevant component of urban road safety in the NoLo district, reflecting the growing use of cycling as a mobility mode in dense urban environments. The objective of this section is to apply the network-based clustering and probabilistic prediction framework to bicycle crashes, evaluate its performance across different temporal configurations, and identify the optimal indicator and number of clusters for crash spatial prediction in the NoLo district.

5.2.1. Scenario 1

Scenario 1 evaluates the ability of the proposed framework to estimate bicycle crash occurrence in 2017 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2017
- Training years: 2018–2021
- Observed bicycle crashes in testing year: 18 crashes
- Bicycle crashes in training years: 66 crashes

Figure 5.14 presents the spatial distribution of the 18 bicycle crashes recorded in 2017, which constitute the testing dataset for this scenario. The crashes are mainly distributed along Via Padova with some scattered crashes across other parts of the network.

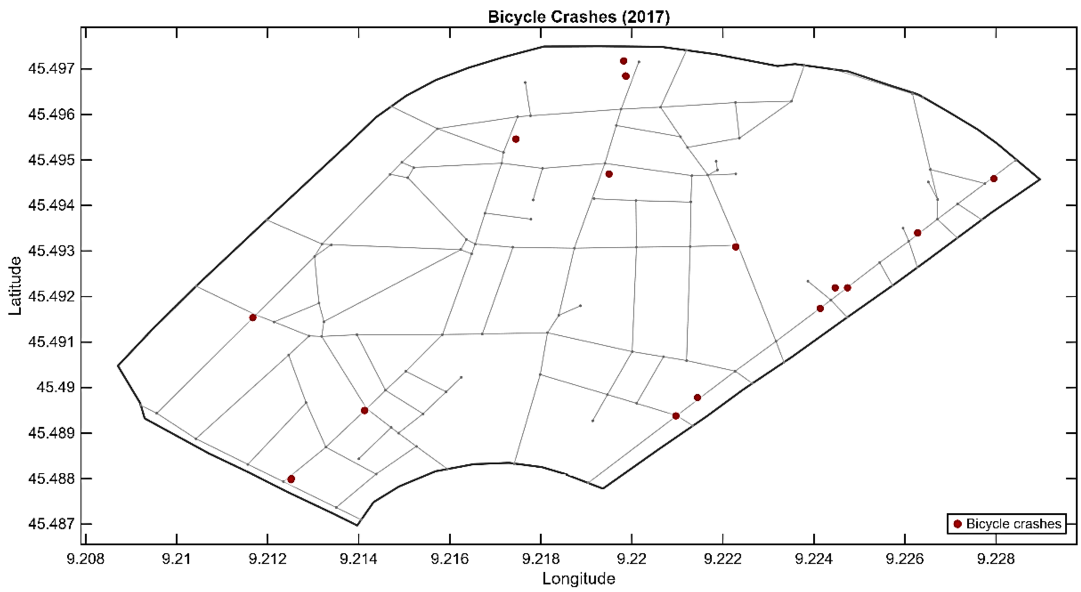

Figure 5.14 Spatial distribution of bicycle crashes in 2017 (testing set, Scenario 1)

Figure 5.15 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 1. Betweenness Centrality (BTW) and Closeness Centrality (CLS) consistently produce the highest RMSEmod values, exhibiting strong instability particularly for small numbers of clusters, with values reaching approximately 6 and 7 respectively at NOC = 2. Given their unreliable behavior, both BTW and CLS are excluded from further consideration in this scenario.

Among the remaining indicators, RE, RN, and IN demonstrate the best overall performance, consistently producing low and stable RMSEmod values across most of the evaluated clustering range. IE shows acceptable performance at small and large numbers of clusters; however it exhibits higher and less stable errors within the range of approximately 7 to 23 clusters, making it less reliable within this intermediate range compared to the crash-informed indicators.

After comparing the performance across all scenarios, the optimal indicator and number of clusters will be selected based on the overall behavior observed throughout the five temporal scenarios and the random scenario, with particular emphasis on Scenario 5 and the Random Scenario.

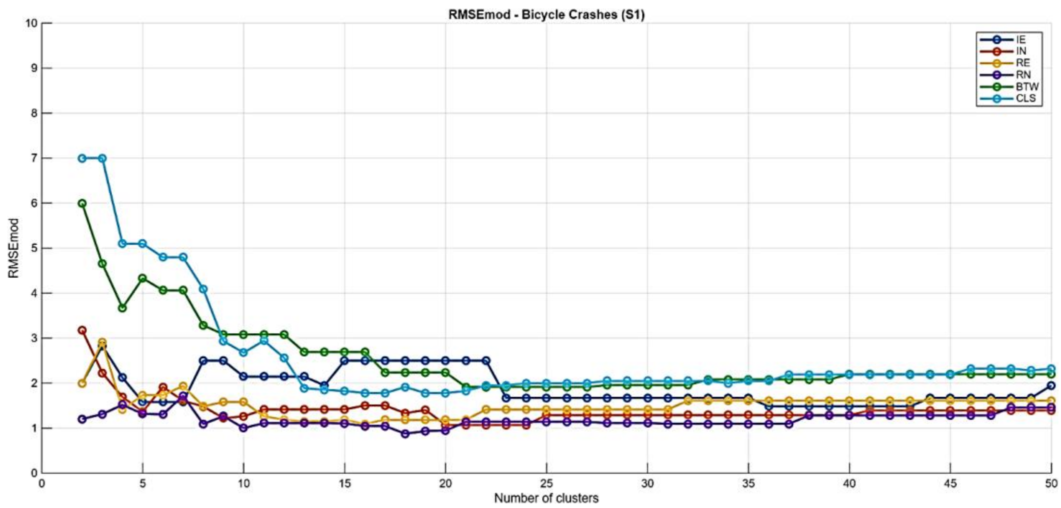

Figure 5.15 RMSEmod comparison across indicators for Scenario 1 (Bicycle crashes)

5.2.2. Scenario 2

Scenario 2 evaluates the ability of the proposed framework to estimate bicycle crash occurrence in 2018 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2018
- Training years: 2017, 2019, 2020, and 2021
- Observed bicycle crashes in testing year: 13 crashes
- Bicycle crashes in training years: 71 crashes

Figure 5.16 presents the spatial distribution of the 13 bicycle crashes recorded in 2018, which constitute the testing dataset for this scenario. The crashes are mainly concentrated along Via Padova and Via Monza, reflecting the corridor-oriented nature of bicycle crash distribution in the NoLo district.

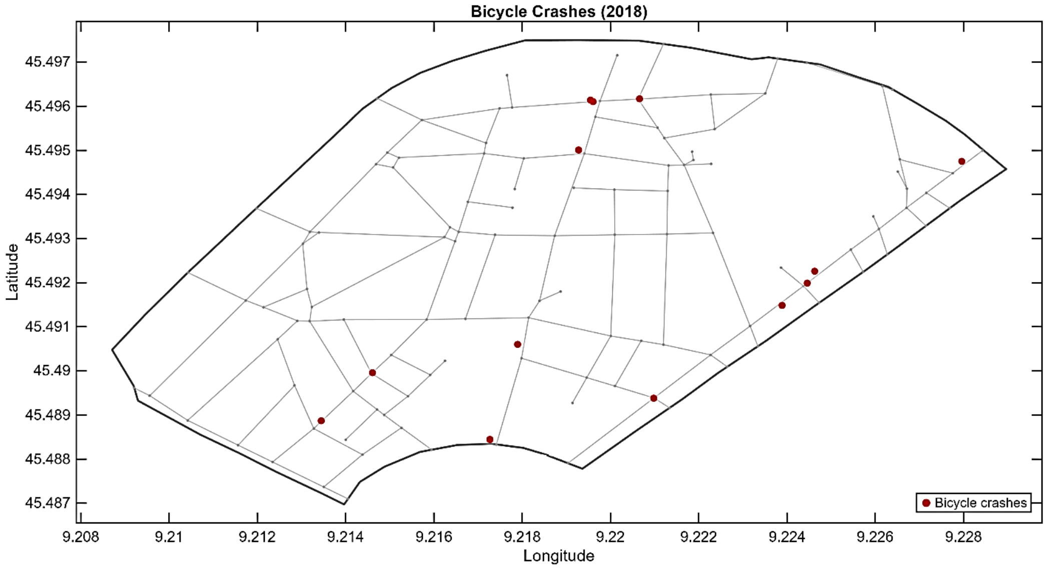

Figure 5.16 Spatial distribution of bicycle crashes in 2018 (testing set, Scenario 2)

Figure 5.17 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 2.

Betweenness Centrality (BTW), Closeness Centrality (CLS), and ICentr index (IE) consistently exhibit the most unstable behavior and the highest RMSEmod values across the evaluated clustering range, making them unreliable indicators for this scenario. ICentr index (IN) shows relatively high error values at low cluster counts but stabilizes and improves progressively as the number of clusters increases, demonstrating acceptable behavior in the mid-to-high range. Among all indicators, RN delivers the best overall performance, consistently producing the lowest RMSEmod values across most of the evaluated clustering range. RE follows closely, showing good and stable behavior throughout, making it the second most reliable indicator in this scenario.

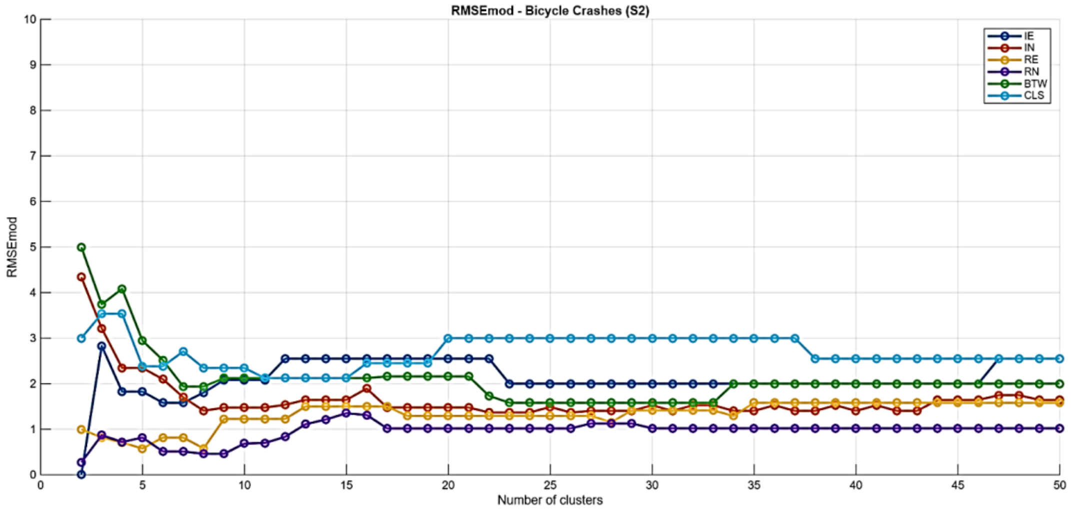

Figure 5.17 RMSEmod comparison across indicators for Scenario 2 (Bicycle crashes)

5.2.3. Scenario 3

Scenario 3 evaluates the ability of the proposed framework to estimate bicycle crash occurrence in 2019 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2019
- Training years: 2017, 2018, 2020, and 2021
- Observed bicycle crashes in testing year: 12 crashes
- Bicycle crashes in training years: 72 crashes

Figure 5.18 presents the spatial distribution of the 12 bicycle crashes recorded in 2019, which constitute the testing dataset for this scenario. The crashes are mainly concentrated at one intersection of Via Monza and along the Via Padova corridor.

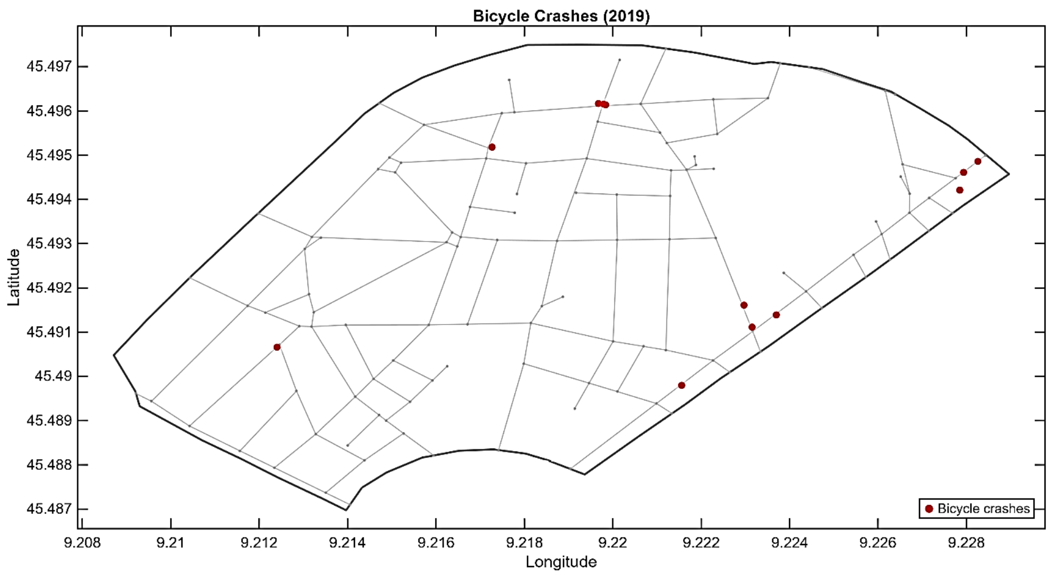

Figure 5.18 Spatial distribution of bicycle crashes in 2019 (testing set, Scenario 3)

Figure 5.19 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 3

From figure 5.19 we can say, betweenness Centrality (BTW) consistently produces the highest RMSEmod values, showing strong instability particularly for small numbers of clusters with values reaching approximately 5 at NOC = 4. Given this unreliable behavior, BTW is excluded from further consideration in this scenario.

Among the remaining indicators, RE and RN demonstrate the best overall performance, consistently producing the lowest and most stable RMSEmod values across most of the evaluated clustering range. IN follows closely, showing competitive performance particularly within the intermediate and larger clustering range. IE and CLS show comparatively higher error values throughout most configurations, making them less reliable than the crash-informed indicators in this scenario.

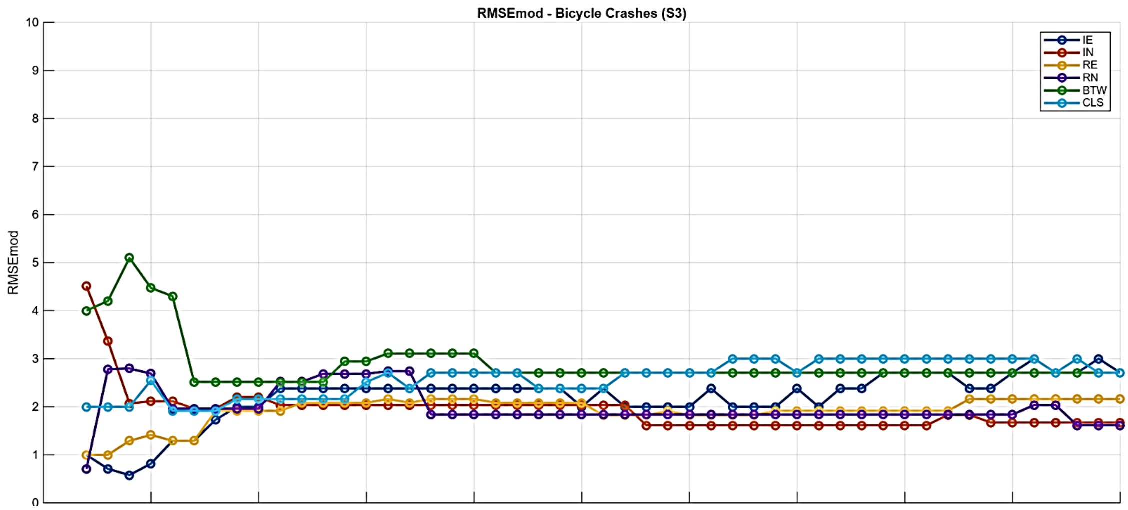

Figure 5. 19 RMSEmod comparison across indicators for Scenario 3 (Bicycle crashes)

5.2.4. Scenario 4

Scenario 4 evaluates the ability of the proposed framework to estimate bicycle crash occurrence in 2020 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2020
- Training years: 2017, 2018, 2019, and 2021
- Observed bicycle crashes in testing year: 20 crashes
- Bicycle crashes in training years: 64 crashes

Figure 5.20 presents the spatial distribution of the 20 bicycle crashes recorded in 2020, which constitute the testing dataset for this scenario. The crashes are mainly concentrated along Via Monza and Via Padova. It is worth noting that unlike pedestrian crashes, bicycle crashes in 2020 did not decrease during the COVID-19 pandemic period, but instead remained relatively high, likely reflecting the increased use of cycling as an alternative mobility mode during the restrictions.

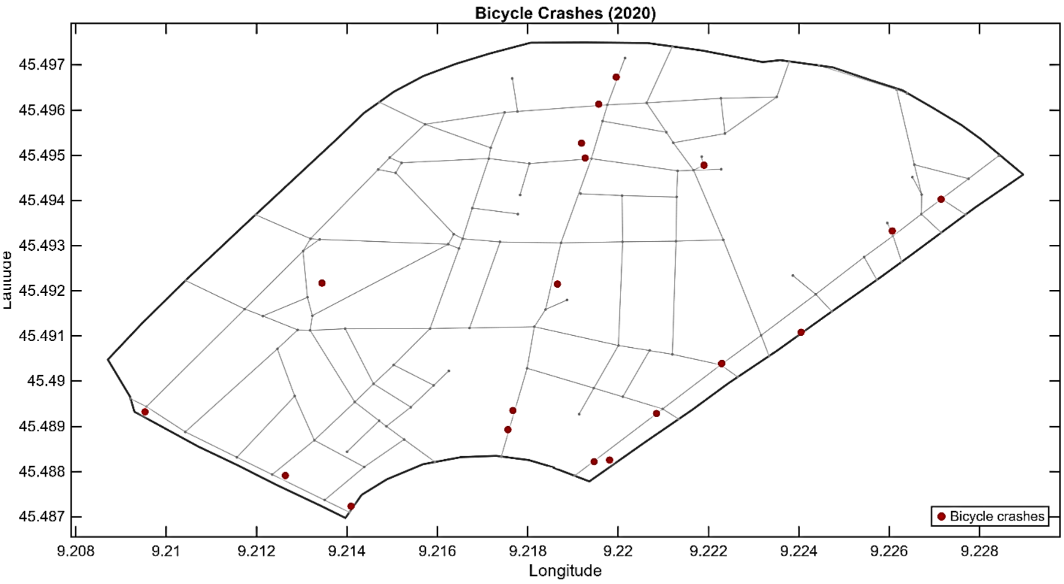

Figure 5. 20 Spatial distribution of bicycle crashes in 2020 (testing set, Scenario 4)

Figure 5.21 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 4. Betweenness Centrality (BTW) consistently produces the highest RMSEmod values, showing strong instability particularly for small numbers of clusters with values reaching approximately 8 at NOC = 2. Given this unreliable behavior, BTW is excluded from further consideration in this scenario.

Among the remaining indicators, RN and IN demonstrate the best overall performance, consistently producing the lowest and most stable RMSEmod values across most of the evaluated clustering range. RE shows competitive performance within the intermediate and larger clustering range, though its errors are slightly higher at small numbers of clusters. CLS shows comparatively higher error values throughout most configurations, making it less competitive than the other indicators.

After comparing the performance across all scenarios, the optimal indicator and number of clusters will be selected based on the overall behavior observed throughout the five temporal scenarios and the random scenario, with particular emphasis on Scenario 5 and the Random Scenario.

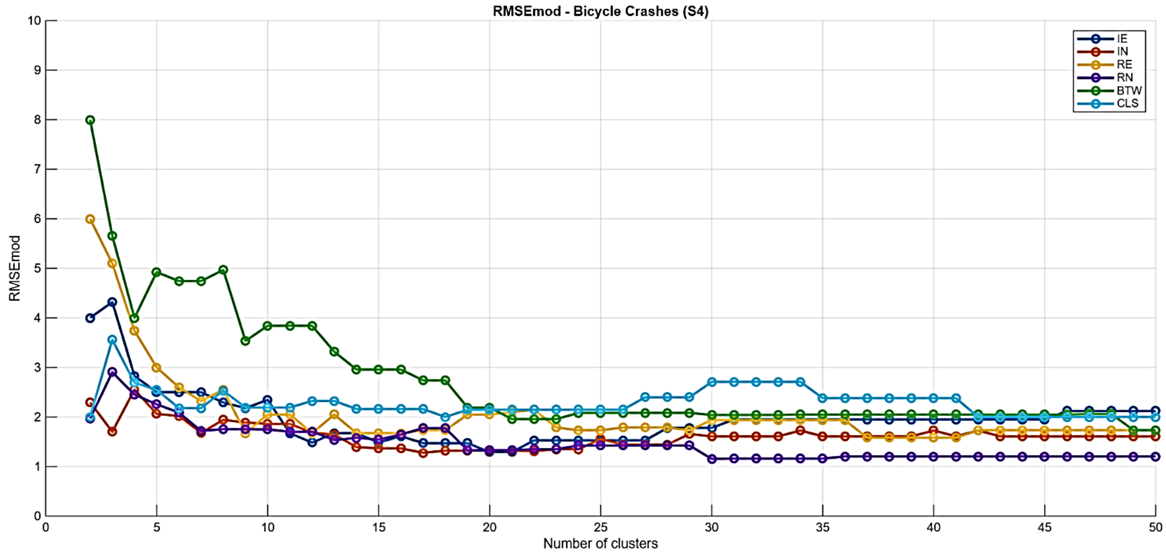

Figure 5. 21 RMSEmod comparison across indicators for Scenario 4 (Bicycle crashes)

5.2.5. Scenario 5

Scenario 5 evaluates the ability of the proposed framework to estimate bicycle crash occurrence in 2021 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2021
- Training years: 2017–2020
- Observed bicycle crashes in testing year: 21 crashes
- Bicycle crashes in training years: 63 crashes

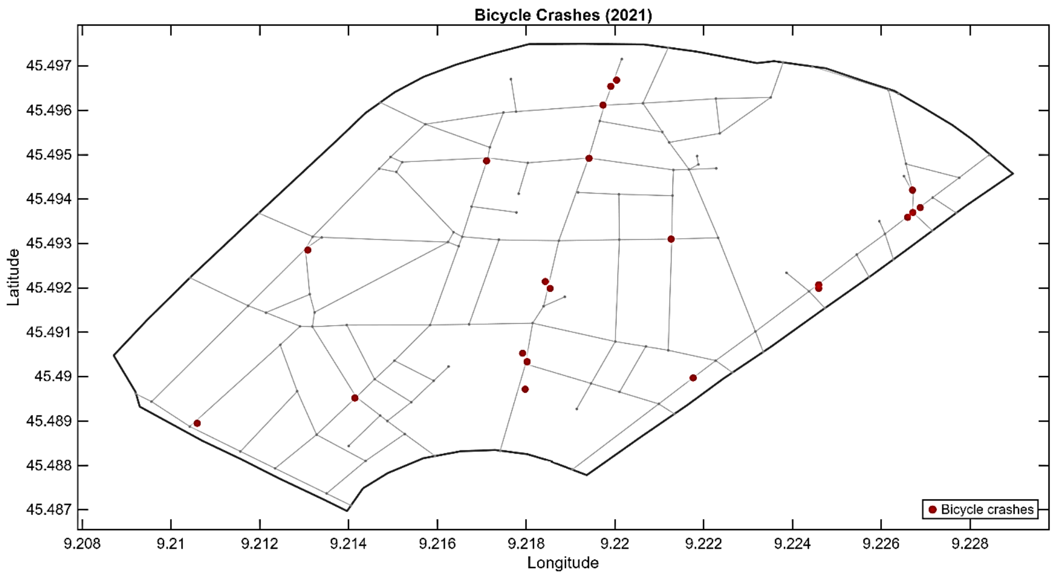

Figure 5. 22 Spatial distribution of bicycle crashes in 2021 (testing set, Scenario 5)

Figure 5.22 presents the spatial distribution of the 21 bicycle crashes recorded in 2021, which constitute the testing dataset for this scenario. The crashes are mainly concentrated along Via Padova and Via Monza, confirming the corridor-oriented nature of bicycle crash distribution consistently observed across the study period.

Figure 5.23 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 5. Betweenness Centrality (BTW) and Closeness Centrality (CLS) consistently produce the highest RMSEmod values, showing strong instability particularly for small numbers of clusters with values reaching approximately 5 and 6 respectively at NOC = 2. Given their unreliable behavior, both BTW and CLS are excluded from further consideration in this scenario.

RE shows comparatively higher error values throughout most configurations, making it less competitive than the other crash-informed indicators in this scenario.

Among the remaining indicators, IN, IE, and RN demonstrate the best overall performance, consistently producing the lowest and most stable RMSEmod values across most of the evaluated clustering range. After comparing the performance across all scenarios, the optimal indicator and number of clusters will be selected based on the overall behavior observed throughout the five temporal scenarios and the random scenario, with particular emphasis on scenario 5 and the random scenario.

The Random Scenario is equally representative, as the training data incorporates crashes from all years of the study period without being associated with any specific temporal ordering, providing a robust and comprehensive validation of the framework.

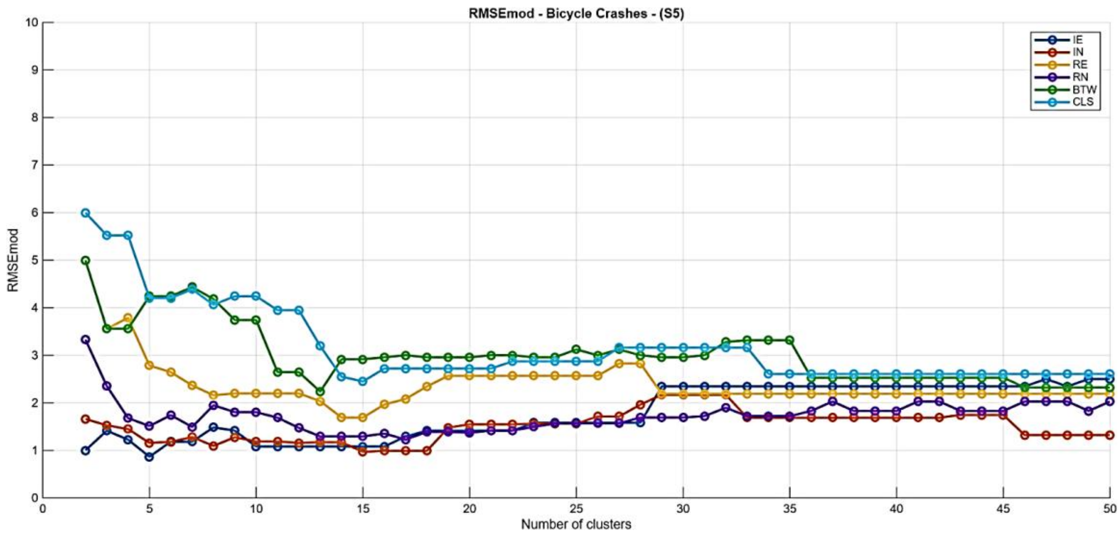

Figure 5.23 RMSEmod comparison across indicators for Scenario 5 (Bicycle crashes)

5.2.6. Results and Selected Configuration

Following the evaluation of all five temporal scenarios and the Random Scenario, this section presents the comparative analysis of the candidate configurations and identifies the optimal indicator and number of clusters for bicycle crash prediction in the NoLo district.

The preliminary analysis of RMSEmod performance across all scenarios confirmed that BTW and CLS exhibit consistently unstable and poor predictive behavior and are therefore excluded from the final selection, the detailed evaluation focuses primarily on RN, RE, IE, and IN, considering only Scenario 5 and the Random Scenario. Scenario 5 is considered the most realistic temporal configuration, as it uses all past years for training and evaluates the model on the most recent year, reflecting a genuine forecasting situation. The Random Scenario is equally representative, as the training data incorporates crashes from all years of the study period without being associated with any specific temporal ordering, providing a robust and comprehensive validation of the framework.

For each indicator, the number of clusters corresponding to the minimum RMSEmod value within the range of 5 to 25 clusters was identified separately for Scenario 5 and the Random Scenario. The selected configurations were then evaluated using three complementary criteria: the error metrics RMSEmod and RSE, the Cluster Prevention Efficiency (CPE), and the Homogeneity Ratio (HR). The objective was not only to minimize prediction error, but also to obtain spatially coherent clusters while maintaining a high crash capture capability, reflected by high CPE and high HR values.

Scenario	Index	NOC	RMSEmod	CPR (%)	NCR (%)	CPE (%)	HR (%)
S5	IN	8	1.0935	28.1	35.5	92.5	73.7
S5	IN	15	0.9703	32.1	13.5	118.6	69.5
SR	IN	8	2.3341	14.7	29.6	85.1	74.6
SR	IN	15	1.8278	11.7	15.2	96.5	66.1
S5	IE	5	0.866	61.9	61.5	100.4	82.2
S5	IE	6	1.18	28.5	36.4	92.1	78
S5	IE	15	1.0801	33.3	14.4	118.9	70.4
SR	IE	5	3.8471	23.5	64.4	59.1	87.3
SR	IE	8	2.6458	11.7	33.0	78.7	80.5
SR	IE	15	2.7929	11.7	16.9	94.8	69.5
SR	IE	23	1.8257	17.6	7.63	110	56.8
S5	RE	6	2.646	61.9	45.7	116.1	89.8
S5	RE	8	2.16	28.5	5.08	123.5	86.4
S5	RE	9	2.1984	28.5	5.08	123.5	85.6
S5	RE	14	1.69	9.52	3.85	105.7	76.3
SR	RE	5	1.936	58.8	11.8	147	81.3
SR	RE	15	1.472	29.4	3.39	126	68.6
SR	RE	20	1.29	23.5	2.54	121	62.7
S5	RN	5	1.5123	46.0	46.6	100	90
S5	RN	11	1.6875	22.2	4.24	118	82
S5	RN	14	1.2937	12.6	3.85	108.8	78.8
S5	RN	15	1.2937	17.4	1.69	115.8	77.9
S5	RN	17	1.227	9.76	1.69	108.1	74.5
SR	RN	7	1.5842	38.2	31.3	106.9	83.9
SR	RN	8	1.779	47.0	7.63	139.4	83.9
SR	RN	9	1.564	39.5	6.78	132.8	82.2
SR	RN	10	1.3723	39.7	3.39	136.4	80.5

Table 5. 3 Summary of candidate configurations for bicycle crash prediction

Table 5.3 presents the evaluation results for all candidate bicycle crash prediction configurations. For each configuration, the top clusters correspond to the network areas with the highest predicted crash concentration and therefore represent the primary output of the prediction framework for identifying priority intervention zones. Among all tested configurations, the RE indicator with NOC = 5 under the Random Scenario (SR) achieved the highest Crash Prevention Efficiency (CPE) value of 146.96%, indicating the strongest ability to concentrate crashes within a limited portion of the network. This configuration captures 58.82% of bicycle crashes while covering only 11.86% of network nodes, resulting in the highest efficiency among all evaluated models. In addition, the model maintains a high Homogeneity Ratio (HR) of 81.36%, confirming that the identified clusters remain spatially coherent and well connected.

Although some configurations produce lower prediction errors, their Crash Prediction Efficiency values remain substantially lower than that achieved by RE–5 (SR). Since the primary objective of the prediction framework is hotspot identification rather than exact crash count estimation, greater importance is assigned to the ability of a model to concentrate crashes within a small portion of the network while maintaining spatial consistency. From this perspective, the RE indicator provides the most effective representation of bicycle crash concentration patterns in NoLo. Based on the combined evaluation of prediction accuracy, Crash Prevention Ratio (CPR), Network Coverage Ratio (NCR), Crash Prevention Efficiency (CPE), and Homogeneity Ratio (HR), the final selected configuration for bicycle crash prediction is:

- Indicator: RE
- Scenario: SR (Random Scenario)
- Number of Clusters (NOC): 5
- Top Clusters: 2 and 4

Figure 5.25 presents the spatial distribution of the five clusters across the NoLo road network. The clustering solution highlights a limited number of spatially concentrated bicycle crash zones, supporting the interpretation that bicycle crashes are associated with specific conflict points rather than diffuse network-wide patterns.

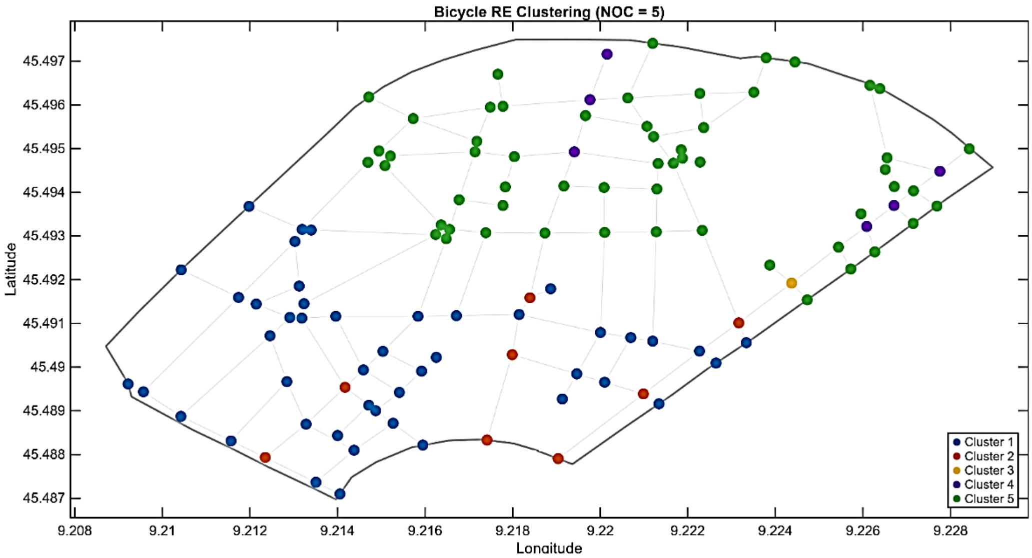

Figure 5. 24 Bicycle crash clustering results for the selected RE–5 model under the Random Scenario.

Figure 5.26 compares the predicted and observed bicycle crash frequencies across the five clusters for the selected RE–5 configuration. The model correctly identifies Clusters 2 and 4 as the most critical crash concentration zones. Cluster 4 shows particularly good agreement between predicted and observed values, while Cluster 2 slightly overestimates crash occurrence. Some discrepancies remain for Clusters 1 and 5; however, the overall ranking of the most critical clusters is preserved, which is the principal objective of the hotspot identification framework.

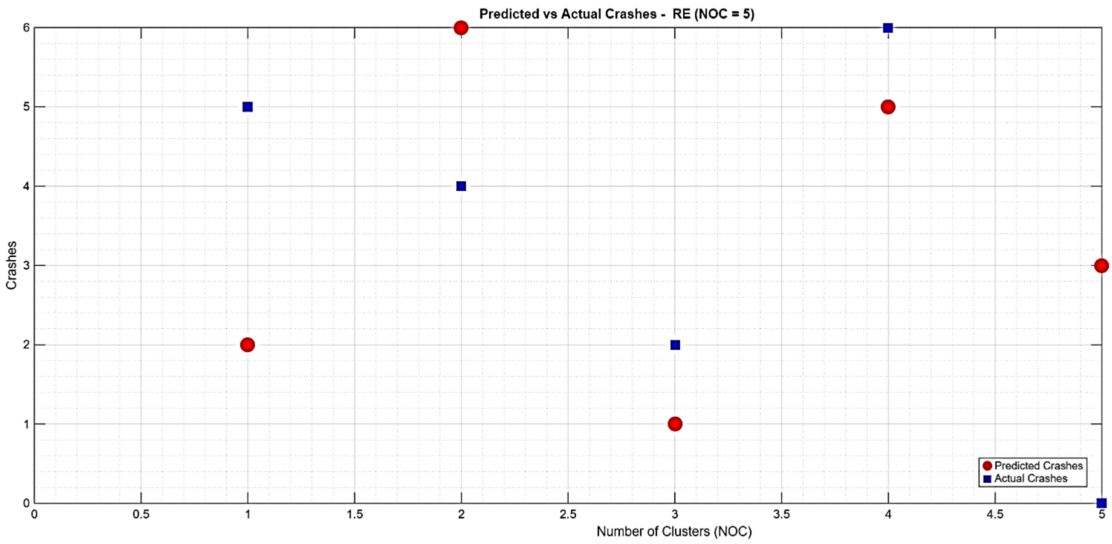

Figure 5. 25 Predicted vs actual bicycle crash distribution across the 5 clusters of the selected RE model.

5.3. Vehicle Crashes: Scenario Analysis and Results

The scenario-based framework is finally applied to vehicle crashes, which represent the largest crash category in the NoLo district with 319 recorded crashes over the study period 2017–2021. The objective of this section is to apply the network-based clustering and probabilistic prediction framework to vehicle crashes, evaluate its performance across different temporal configurations, and identify the optimal indicator and number of clusters for crash spatial prediction in the NoLo district.

For each scenario, the network is partitioned into clusters based on the selected indicators, and probability density functions are constructed at the cluster level using the training dataset. These distributions are then used to estimate the number of crashes within each cluster for the corresponding testing year. Model performance is evaluated using RSE and RMSEmod across varying numbers of clusters, and the best configurations are carried forward to a second level of evaluation using CPE, HR, and visual inspection of the cluster spatial distribution.

5.3.1. Scenario 1

- Testing year: 2017
- Training years: 2018–2021
- Observed vehicle crashes in testing year: 82 crashes
- Vehicle crashes in training years: 236 crashes

Figure 5.27 presents the spatial distribution of the 82 motor vehicle crashes recorded in 2017, which constitute the testing dataset for this scenario. The crashes are broadly distributed across the NoLo network, with concentrations visible along the main arterial corridors including Via Padova in the northeastern section, Viale Monza in the central portion of the district, and Viale Brianza in the southwestern section, as well as scattered occurrences across secondary streets throughout the network.

Figure 5.28 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 1. Betweenness Centrality (BTW) and Closeness Centrality (CLS) consistently produce the highest RMSEmod values, exhibiting extreme instability at small numbers of clusters, with values reaching approximately 30 and 14 respectively at NOC = 2. Both indicators converge only gradually at larger cluster counts but remain above the other indicators throughout most of the evaluated range. Given their unreliable behavior, both BTW and CLS are excluded from further consideration in this scenario.

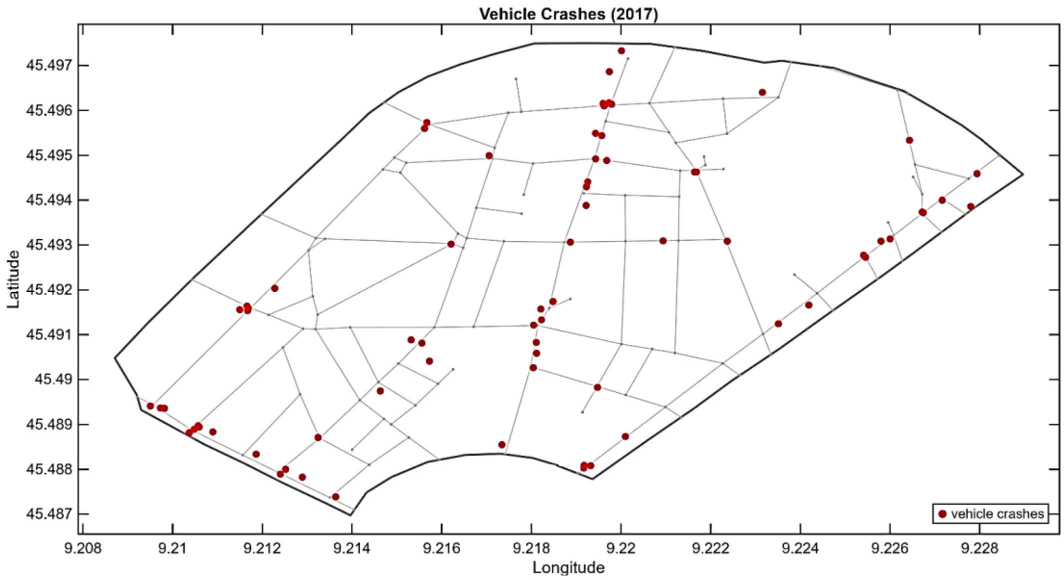

Figure 5. 26 Spatial distribution of vehicle crashes in 2017 (testing set, Scenario 1)

Among the remaining indicators, RN demonstrates the best overall performance, consistently producing the lowest RMSEmod values across most of the evaluated clustering range, converging toward approximately 1 at large cluster counts.

RE follows closely, showing stable and low error values throughout. IN and IE both show higher error values at small cluster counts, before stabilizing and converging toward the other indicators at intermediate and larger cluster counts.

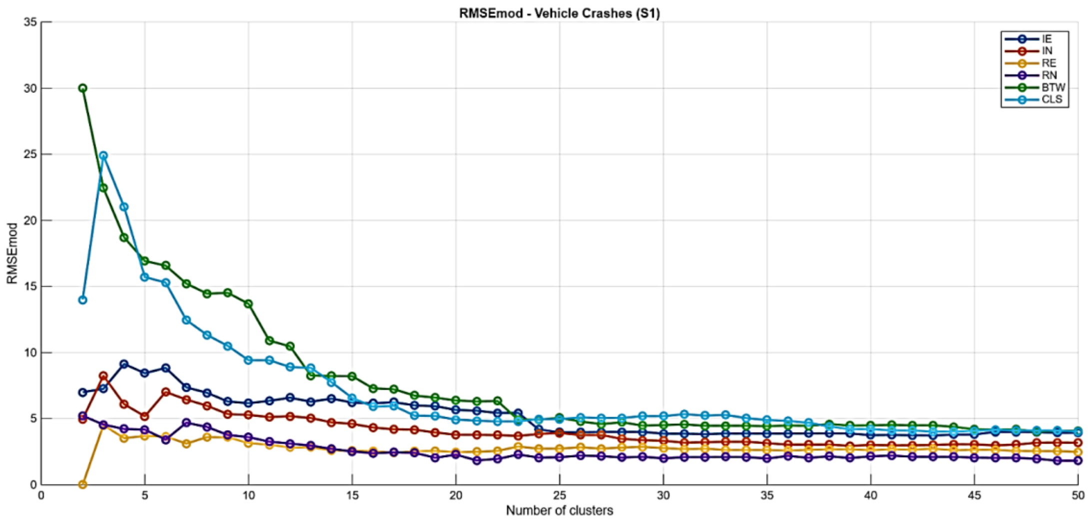

Figure 5. 27 RMSEmod comparison across indicators for Scenario 1 (Vehicle crashes)

5.3.2. Scenario 2

Scenario 2 evaluates the ability of the proposed framework to estimate motor vehicle crash occurrence in 2018 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2018
- Training years: 2017, 2019, 2020, and 2021
- Observed motor vehicle crashes in testing year: 67 crashes
- Motor vehicle crashes in training years: 252 crashes

Figure 5.29 presents the spatial distribution of the 67 motor vehicle crashes recorded in 2018, which constitute the testing dataset for this scenario. The crashes are distributed across the NoLo network with concentrations visible along Viale Monza in the central portion of the district, Via Padova in the northeastern section, and Viale Brianza in the southwestern section, alongside scattered occurrences across secondary streets.

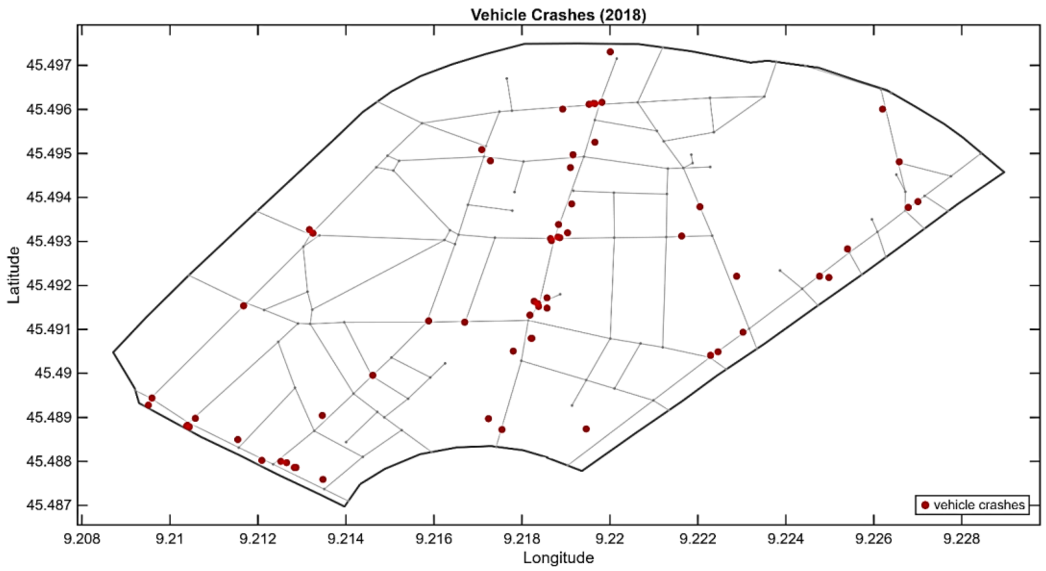

Figure 5. 28 Spatial distribution of vehicle crashes in 2018 (testing set, Scenario 2).

Figure 5.30 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 2. Betweenness Centrality (BTW) and Closeness Centrality (CLS) again produce the highest RMSEmod values, with CLS reaching approximately 24 at NOC = 3 and BTW reaching approximately 19 at NOC = 2, both exhibiting strong instability at small cluster counts before converging slowly at larger values. Given their unreliable behavior, both BTW and CLS are excluded from further consideration in this scenario. Among the remaining indicators, RN and RE demonstrate the best overall performance by a notable margin, producing consistently low and stable RMSEmod values close to 1 across virtually the entire evaluated clustering range. IN and IE both show higher error values at small cluster counts, reaching approximately 10 at NOC = 2.

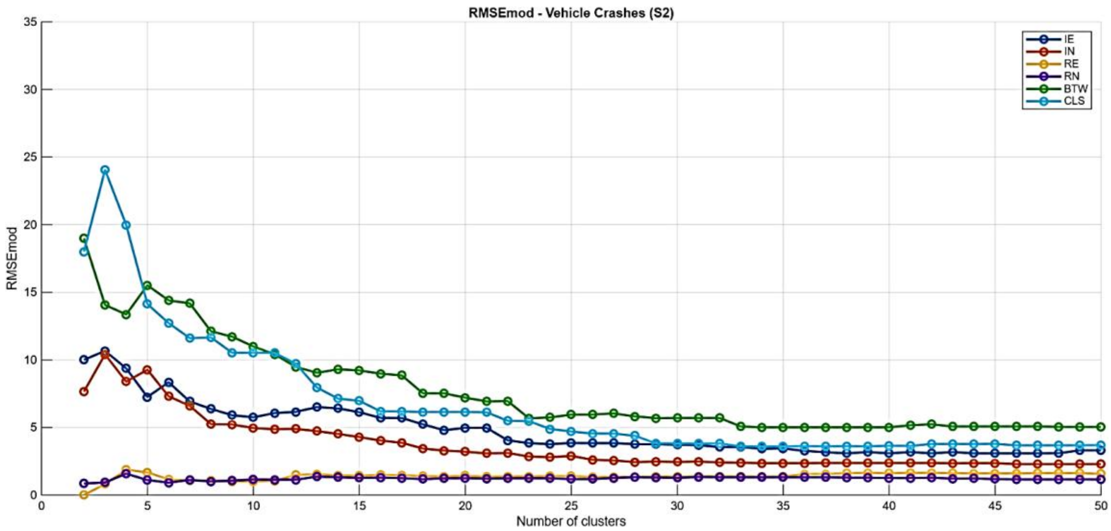

Figure 5. 29 RMSEmod comparison across indicators for Scenario 2 (Vehicle crashes).

5.3.3. Scenario 3

Scenario 3 evaluates the ability of the proposed framework to estimate motor vehicle crash occurrence in 2019 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2019
- Training years: 2017, 2018, 2020, and 2021
- Observed motor vehicle crashes in testing year: 58 crashes
- Motor vehicle crashes in training years: 261 crashes

Figure 5.31 presents the spatial distribution of the 58 motor vehicle crashes recorded in 2019, which constitute the testing dataset for this scenario. The crashes show a more concentrated distribution compared to previous years, with a notable cluster along Viale Monza in the central portion of the district and along Via Padova in the northeastern section. A secondary concentration is also visible in the southwestern area near Viale Brianza, while the overall crash density across secondary streets is comparatively lower than in 2017 and 2018.

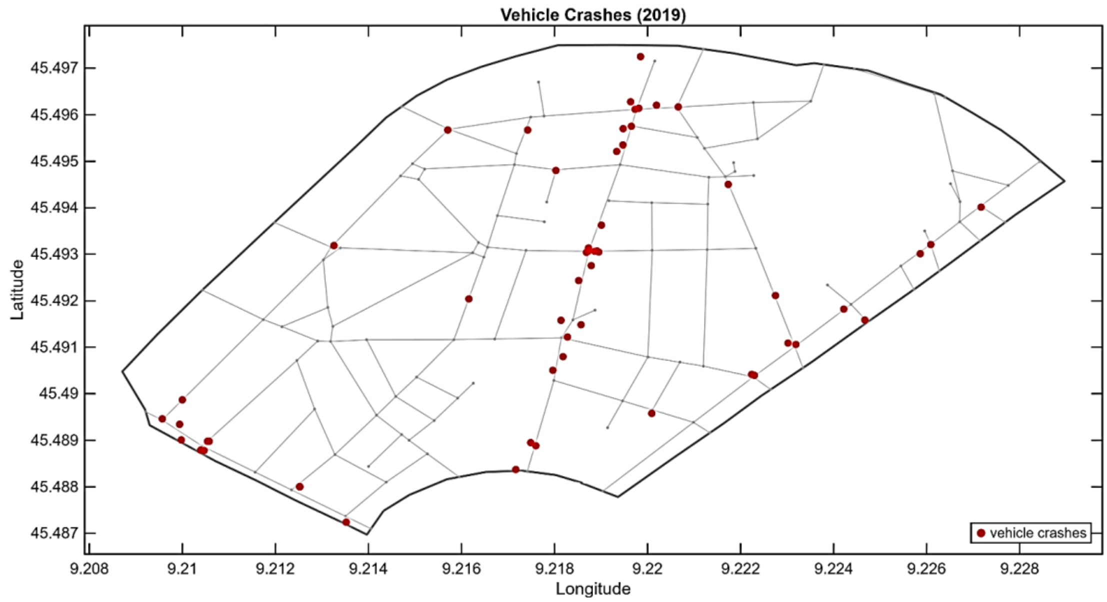

Figure 5. 30 Spatial distribution of vehicle crashes in 2019 (testing set, Scenario 3).

Figure 5.32 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 3.

Betweenness Centrality (BTW) and Closeness Centrality (CLS) again produce the highest RMSEmod values, with BTW reaching approximately 26 and CLS approximately 19, both exhibiting strong instability at small cluster counts. Given their unreliable behavior, both BTW and CLS are excluded from further consideration in this scenario.

Among the remaining indicators, RN and RE demonstrate the best overall performance, consistently producing the lowest RMSEmod values across the entire evaluated clustering range, converging toward approximately 1 at large cluster counts. IN and IE show higher error values at small cluster counts before stabilizing at intermediate and larger cluster counts, making them less reliable than RN and RE in this scenario.

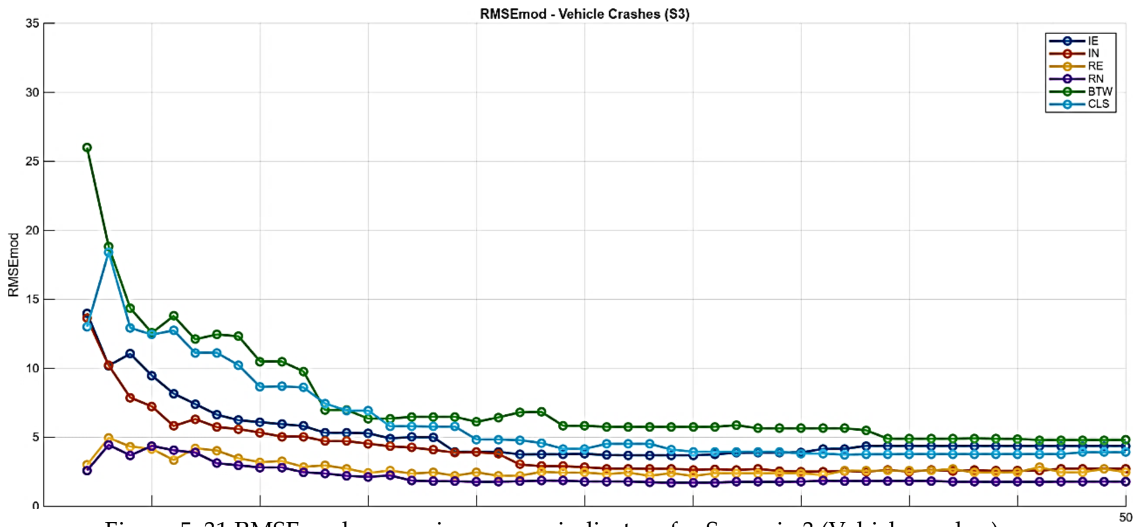

Figure 5. 31 RMSEmod comparison across indicators for Scenario 3 (Vehicle crashes).

5.3.4. Scenario 4

Scenario 4 evaluates the ability of the proposed framework to estimate motor vehicle crash occurrence in 2020 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2020
- Training years: 2017, 2018, 2019, and 2021
- Observed motor vehicle crashes in testing year: 48 crashes
- Motor vehicle crashes in training years: 271 crashes

Figure 5.33 presents the spatial distribution of the 48 motor vehicle crashes recorded in 2020, which constitute the testing dataset for this scenario. The crashes show a notably lower overall density compared to other years, consistent with the substantial reduction in vehicular traffic during the COVID-19 pandemic lockdown period. Despite this overall reduction, crash concentrations remain visible along Viale Monza in the central portion of the district and along Via Padova in the northeastern section, confirming that these corridors retain their role as primary crash locations even under reduced traffic conditions. A smaller number of crashes are distributed across secondary streets and the southwestern area near Viale Brianza.

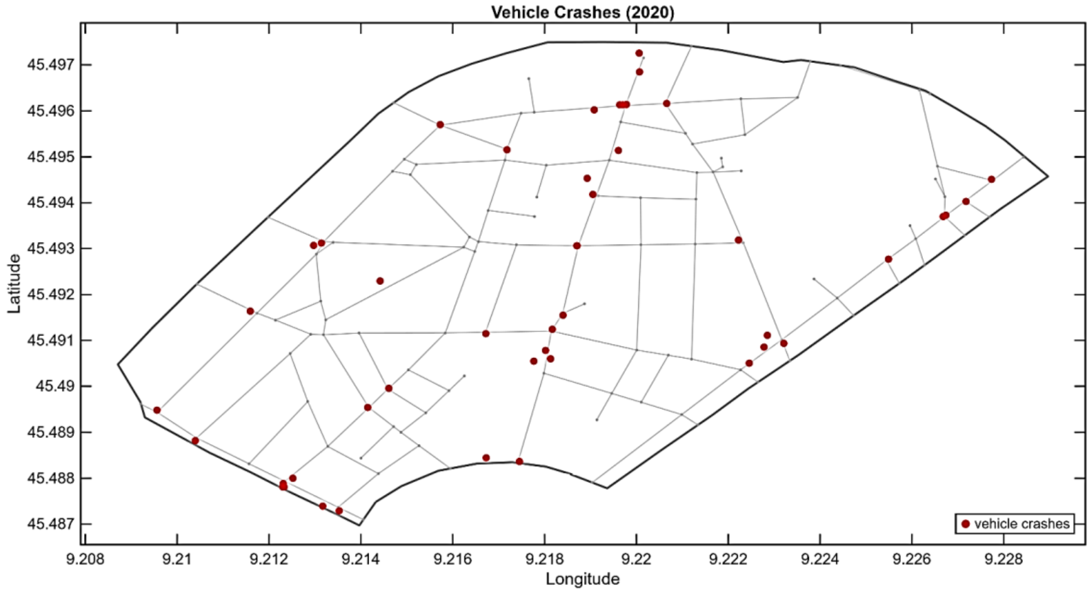

Figure 5. 32 Spatial distribution of vehicle crashes in 2020 (testing set, Scenario 4)

Figure 5.34 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 4.

Betweenness Centrality (BTW) again produces the highest RMSEmod values, reaching approximately 20 at NOC = 2 and declining only gradually across the evaluated range, remaining consistently above all other indicators. Given this unreliable behavior, BTW is excluded from further consideration in this scenario. Closeness Centrality (CLS) also shows elevated instability at small cluster counts, reaching approximately 11 at NOC = 3, before converging more rapidly than in previous scenarios toward the other indicators at intermediate cluster counts.

Among the remaining indicators, RN and RE demonstrate the best overall performance, producing the lowest and most stable RMSEmod values across virtually the entire evaluated clustering range, converging toward approximately 0 at larger cluster counts. IN and IE both show slightly higher error values at small cluster counts before stabilizing rapidly, demonstrating competitive performance within the intermediate and larger clustering range

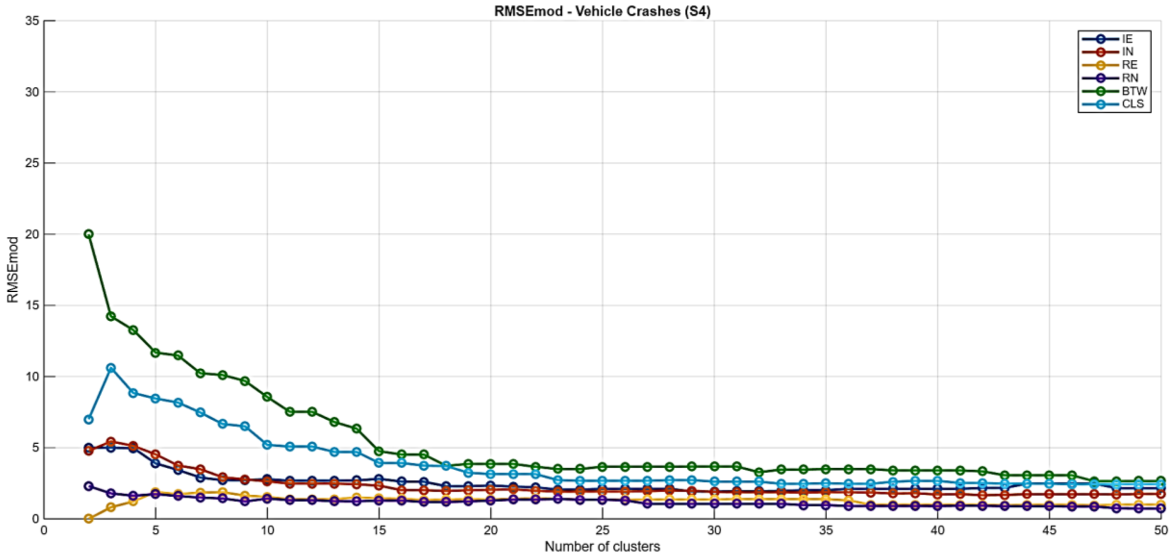

Figure 5. 33 RMSEmod comparison across indicators for Scenario 4 (Vehicle crashes).

5.3.5. Scenario 5

Scenario 5 evaluates the ability of the proposed framework to estimate motor vehicle crash occurrence in 2021 using information learned from the remaining years of the study period. The scenario configuration is summarized as follows:

- Testing year: 2021
- Training years: 2017–2020
- Observed motor vehicle crashes in testing year: 63 crashes
- Motor vehicle crashes in training years: 256 crashes

Figure 5.35 presents the spatial distribution of the 63 motor vehicle crashes recorded in 2021, which constitute the testing dataset for this scenario. The crashes are broadly distributed across the NoLo network, with concentrations visible along Viale Monza in the central portion of the district, along Viale Brianza in the southwestern section, and along Via Padova in the northeastern section, confirming the corridor-oriented nature of motor vehicle crash distribution consistently observed across the study period.

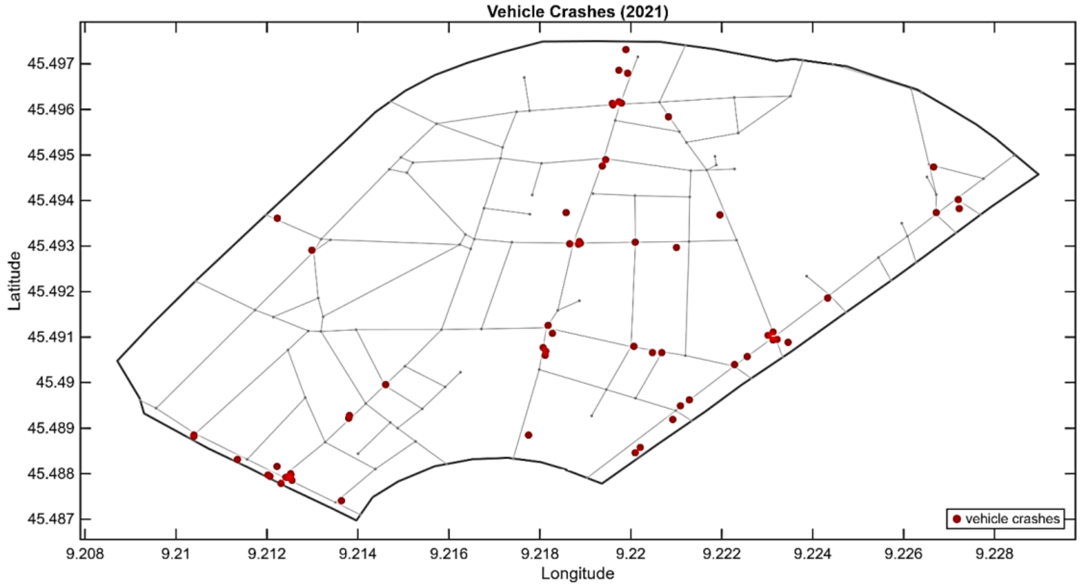

Figure 5. 34 Spatial distribution of vehicle crashes in 2021 (testing set, Scenario 5).

Figure 5.36 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in Scenario 5. Betweenness Centrality (BTW) and Closeness Centrality (CLS) again produce the highest RMSEmod values, with both reaching approximately 15 to 16 at NOC = 2 and exhibiting instability at small cluster counts before gradually converging at larger values. Given their unreliable behavior, both BTW and CLS are excluded from further consideration in this scenario.

Among the remaining indicators, RN and RE demonstrate the best overall performance, consistently producing the lowest and most stable RMSEmod values across most of the evaluated clustering range. IN and IE both exhibit higher error values at small cluster counts, with IE showing a notable peak around NOC = 3, before progressively stabilizing and converging toward RN and RE performance levels at larger cluster counts.

After comparing the performance across all scenarios, the optimal indicator and number of clusters will be selected based on the overall behavior observed throughout the five temporal scenarios and the Random Scenario, with particular emphasis on Scenario 5 and the Random Scenario. The Random Scenario is equally representative, as the training data incorporates crashes from all years of the study period without being associated with any specific temporal ordering, providing a robust and comprehensive validation of the framework.

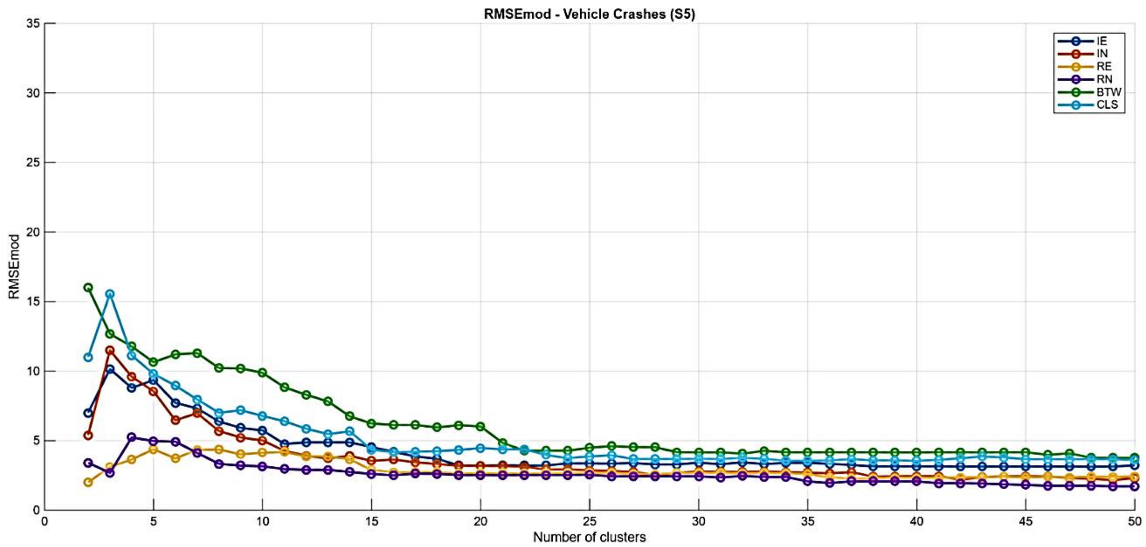

Figure 5. 35 RMSEmod comparison across indicators for Scenario 5 (Vehicle crashes).

5.3.6. Random scenario

The Random Scenario differs from the five temporal scenarios in its data splitting approach. Rather than using a year-based division, all 319 motor vehicle crashes recorded across the full study period are merged together and randomly divided into:

- Training set: 80% of the data, corresponding to approximately 255 crashes
- Testing set: 20% of the data, corresponding to approximately 64 crashes

Figure 5.37 presents the RMSEmod variation as a function of the number of clusters for all evaluated indicators in the Random Scenario. Betweenness Centrality (BTW) again produces the highest RMSEmod values by a considerable margin, reaching approximately 30 at NOC = 2 and declining only gradually, remaining consistently above all other indicators throughout the evaluated range. Closeness Centrality (CLS) similarly exhibits strong instability at small cluster counts, reaching approximately 19 at NOC = 3, before converging more progressively at larger values. Given their unreliable behavior, both BTW and CLS are excluded from further consideration.

Among the remaining indicators, RN demonstrates the best overall performance, consistently producing the lowest and most stable RMSEmod values across virtually the entire evaluated clustering range, converging toward approximately 1 at large cluster counts. RE follows closely, showing stable and low error values throughout and representing the second most reliable indicator in this scenario. IN and IE both exhibit higher error values at small cluster counts, reaching approximately 10 to 12 at NOC = 2, before stabilizing progressively and converging toward RN and RE performance levels at larger cluster counts.

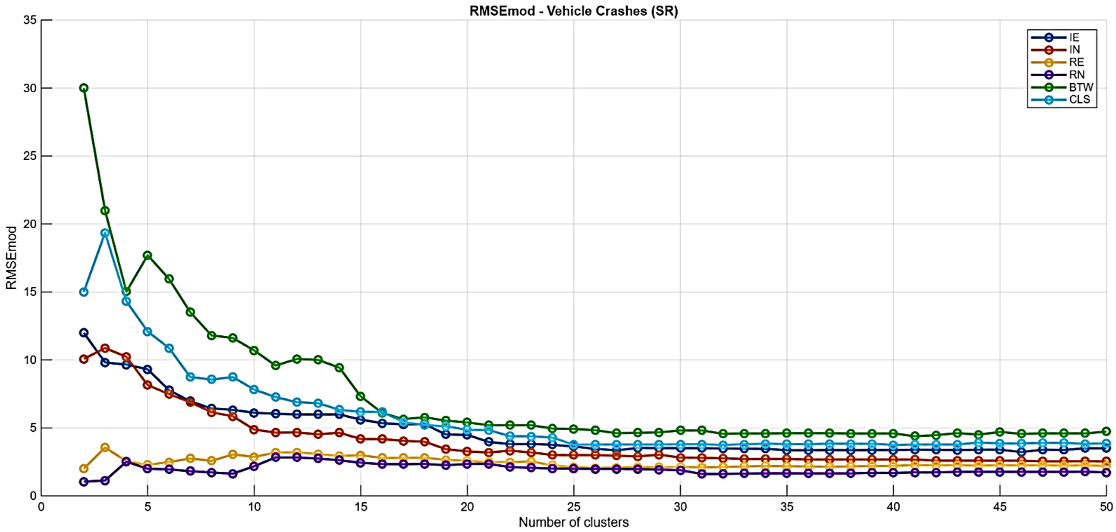

Figure 5. 36 RMSEmod comparison across indicators for the Random Scenario (Vehicle crashes).

5.3.7. Results and Selected Configuration

Following the evaluation of all five temporal scenarios and the Random Scenario, this section presents the comparative analysis of the candidate configurations and identifies the optimal indicator and number of clusters for motor vehicle crash prediction in the NoLo district.

The preliminary analysis of RMSEmod performance across all scenarios confirmed that BTW and CLS exhibit consistently unstable and poor predictive behavior and are therefore excluded from the final selection. The detailed evaluation focuses primarily on RN, RE, IE, and IN, considering only Scenario 5 and the Random Scenario. Scenario 5 is considered the most realistic temporal configuration, as it uses all past years for training and evaluates the model on the most recent year, reflecting a genuine forecasting situation. The Random Scenario is equally representative, as the training data incorporates crashes from all years of the study period without being associated with any specific temporal ordering, providing a robust and comprehensive validation of the framework.

For each indicator, the number of clusters corresponding to the minimum RMSEmod value within the range of 5 to 25 clusters was identified separately for Scenario 5 and the Random Scenario. The selected configurations were then evaluated using three complementary criteria: the error metrics RMSEmod and RSE, the Crash Prevention Efficiency (CPE), and the Homogeneity Ratio (HR). The objective was not only to minimize prediction error, but also to obtain spatially coherent clusters while maintaining a high crash capture capability, reflected by high CPE and high HR values.

Table 5.4 presents the evaluation results for all candidate vehicle crash prediction configurations. For each configuration, the top two clusters represent the most critical areas within the network, identified as those with the highest predicted crash concentration and therefore considered priority locations for road safety interventions.

Among the evaluated models, the RN indicator with NOC = 10 under Scenario 5 provides the most balanced combination of crash capture capability, prediction efficiency, and spatial consistency. This configuration achieves a Crash Prediction Ratio (CPR) of 36.51%, indicating that more than one-third of all vehicle crashes are concentrated within the identified critical clusters while covering only 6.78% of the network nodes. As a result, the model reaches a Crash Prediction Efficiency (CPE) of 129.7%, one of the highest values among all tested configurations.

Although some alternative models achieve lower prediction errors or slightly higher CPR values, they either require substantially larger network coverage or exhibit lower hotspot concentration efficiency. Conversely, some highly concentrated models achieve lower NCR values but fail to capture a sufficient proportion of crashes. The selected RN–10 configuration therefore provides a more appropriate compromise between hotspot concentration, crash capture performance, and network representation. The model also maintains a relatively high Homogeneity Ratio (HR) of 82.2%, indicating that the resulting clusters remain spatially coherent and well connected. This characteristic is particularly important for practical implementation, as it facilitates the identification of geographically meaningful crash concentration zones. The results suggest that vehicle crashes in the NoLo district tend to be concentrated around specific network nodes and intersections rather than being uniformly distributed throughout the network. This behavior is consistent with the RN methodology, which directly incorporates node-based crash frequencies and therefore effectively captures localized crash concentration patterns. Based on the combined evaluation of prediction performance, the final selected configuration for vehicle crash prediction is:

- Indicator: RN
- Scenario: S5
- Number of Clusters (NOC): 10
- Top Clusters: 10 and 6

Scenario	Index	NOC	RMSEmod	CPR (%)	NCR (%)	CPE (%)	HR (%)
SR	IN	10	4.8757	28.5	26.2	102.3	74.5
SR	IN	20	3.2715	12.7	29.66	83.04	60.1
S5	IN	15	3.5436	13.8	29.6	84.18	64.4
S5	IN	20	3.1818	5.4	20.3	85.06	61.8
S5	IE	11	4.7566	15.8	24.7	64.02	77.9
S5	IE	20	3.201	6.35	29.0	21.86	61.0
SR	IE	11	6.0415	23.44	21.1	102.2	74.5
SR	IE	20	3.2016	29.69	20.3	109.3	58.4
S5	RE	9	4.013	28.57	20.3	108.2	84.7
S5	RE	14	3.66	25.4	2.54	122.8	82.2
S5	RE	10	4.136	28.13	20.3	107.7	84.7
S5	RE	15	2.872	25.4	2.54	122.8	78.8
SR	RE	8	2.578	46.88	33.05	113.8	85.5
SR	RE	14	2.9297	26.56	4.24	122.3	79.6
SR	RE	16	2.787	34.38	2.54	131.8	77.9
SR	RE	25	2.136	23.44	2.54	120.9	64.4
S5	RN	7	4.1187	36.19	36.44	99.75	88.9
S5	RN	16	2.5088	22.22	5.93	116.2	77.9
S5	RN	10	3.1494	36.51	6.78	129.7	82.2
SR	RN	5	1.9876	73.81	96.61	77.2	91.5
SR	RN	9	1.6372	49.05	34.75	114.3	86.4
SR	RN	19	2.2577	13.49	16.95	96.54	72.8

Table 5. 4 Summary of candidate configurations for vehicle crash prediction

Figure 5.38 shows the spatial distribution of the ten clusters obtained using the RN indicator (NOC = 10) under Scenario 5 for vehicle crash analysis in the NoLo district. The clustering solution partitions the road network into ten spatially coherent groups, highlighting localized vehicle crash concentration patterns. Clusters 10 and 6 were identified as the most critical clusters according to the selected prediction model and therefore represent the primary priority areas for vehicle safety interventions.

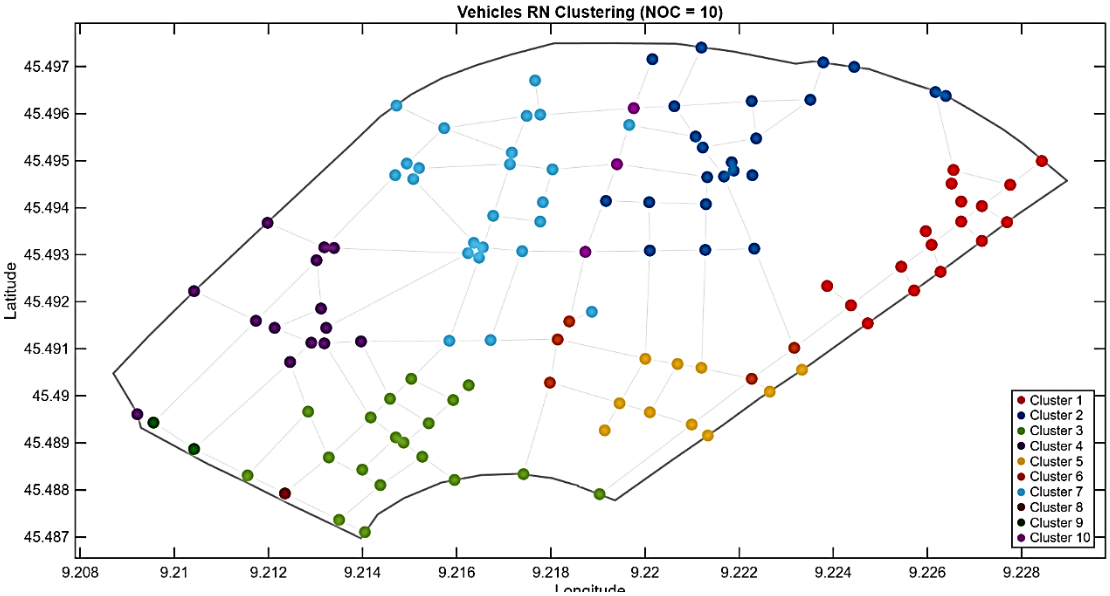

Figure 5. 37 Vehicle crash clustering results for the selected RN–10 model under Scenario 5

Figure 5.39 compares the predicted and observed vehicle crash frequencies for each cluster. The model successfully identifies Clusters 10 and 6 as the dominant vehicle crash hotspots, which correspond to the clusters with the highest observed crash frequencies. Cluster 10 records the highest predicted crash count and closely matches the observed crash distribution, while Cluster 6 is also correctly classified as a high-risk cluster. Although some discrepancies are observed for secondary clusters, the overall hierarchy of crash risk is preserved. This confirms the ability of the RN-based model to identify the principal vehicle crash concentration zones and provides further support for selecting the RN indicator with NOC = 10 under Scenario 5 as the final configuration for vehicle crash prediction.

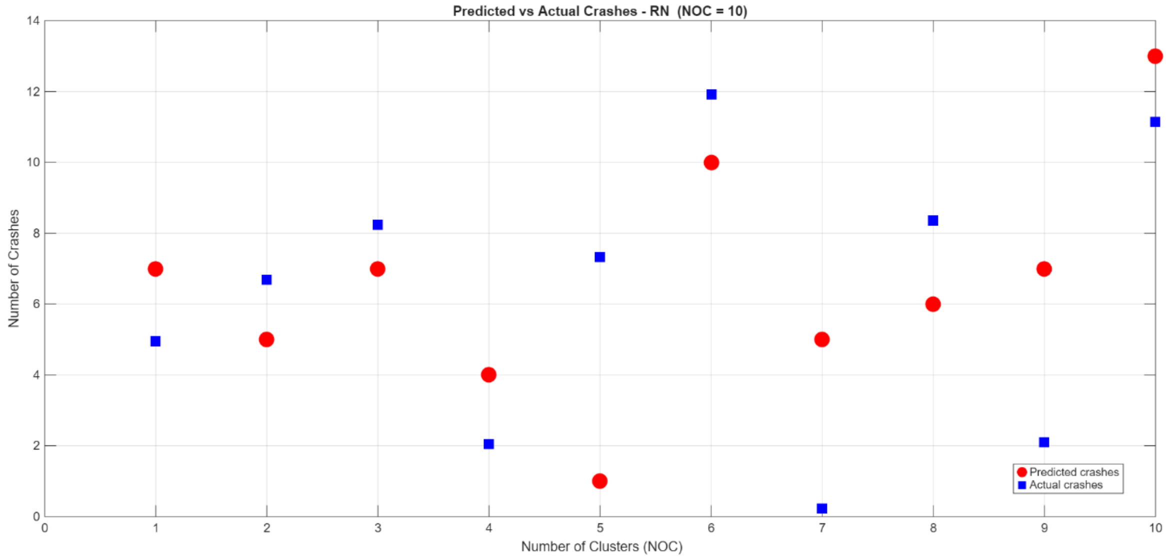

Figure 5. 38 Predicted vs actual vehicle crash distribution across the 10 clusters of the selected RN model.

5.4. Top Clusters

Finally, the selected clustering configurations identified through the model comparison are used to determine the most critical crash-prone zones within the NoLo district. These clusters correspond to the top-ranked clusters with the highest predicted crash frequencies and therefore represent the primary locations where future safety interventions should be prioritized.

Pedestrian crashes

Figure 5.40 illustrates the final pedestrian clustering configuration obtained using the RN indicator under Scenario 5 with NOC = 11. Based on the prediction results, Clusters 9 and 10 recorded the highest predicted pedestrian crash frequencies and are therefore identified as the most critical pedestrian crash zones within the NoLo network.

These clusters are mainly located along the eastern section of the district, particularly at Via Padova corridor, where pedestrian activity and vehicle-pedestrian interactions are concentrated.

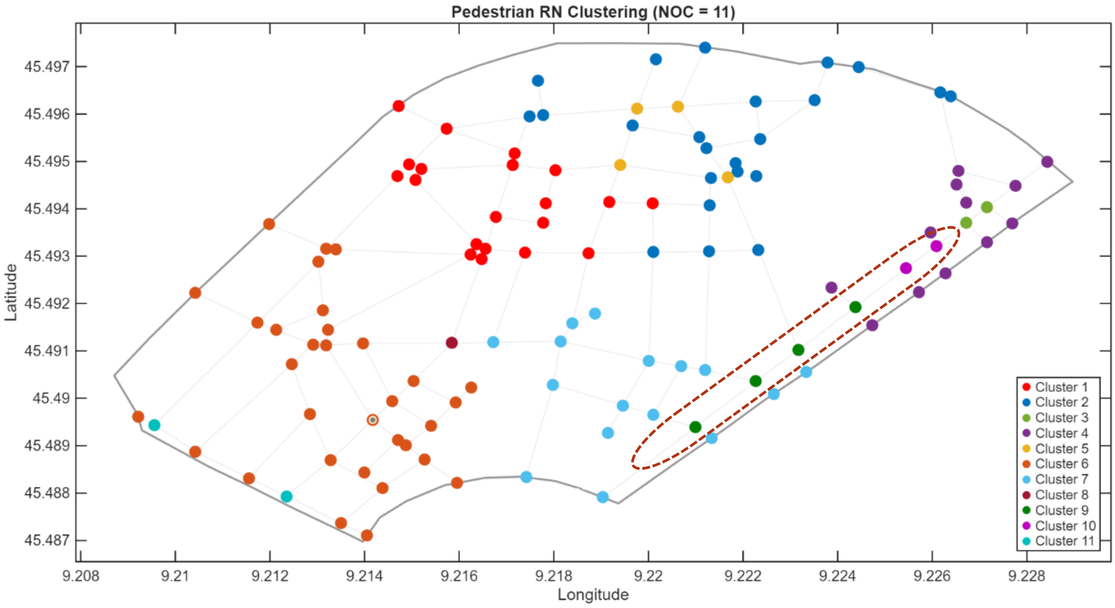

Figure 5. 39 Final pedestrian crash clustering configuration, highlighting the top clusters.

Figure 5.41 presents the final bicycle clustering configuration obtained using the RE indicator under the Random Scenario with NOC = 5. The prediction model identifies Clusters 2 and 4 as the clusters with the highest predicted bicycle crash frequencies. These clusters cover substantial portions of the western and northern sectors of NoLo and include several intersections and corridor segments where bicycle movements frequently interact with vehicular traffic.

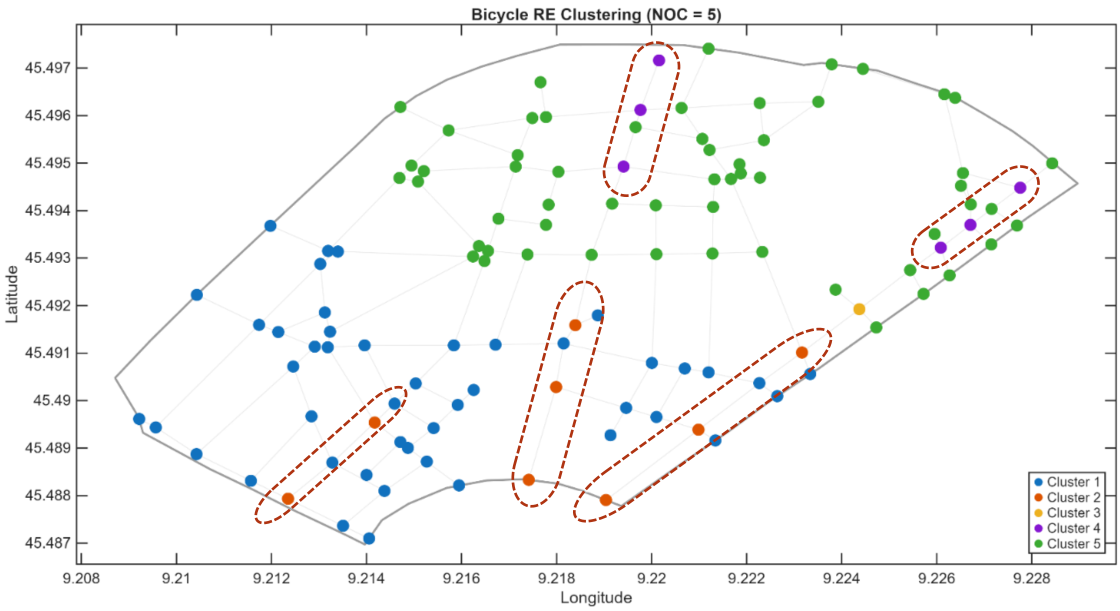

Figure 5. 40 Final bicycle crash clustering configuration, highlighting the top clusters.

Figure 5.42 shows the final vehicle clustering configuration obtained using the RN indicator under Scenario 5 with NOC = 10. The prediction model identifies Clusters 10 and 6 as the clusters with the highest predicted vehicle crash frequencies.

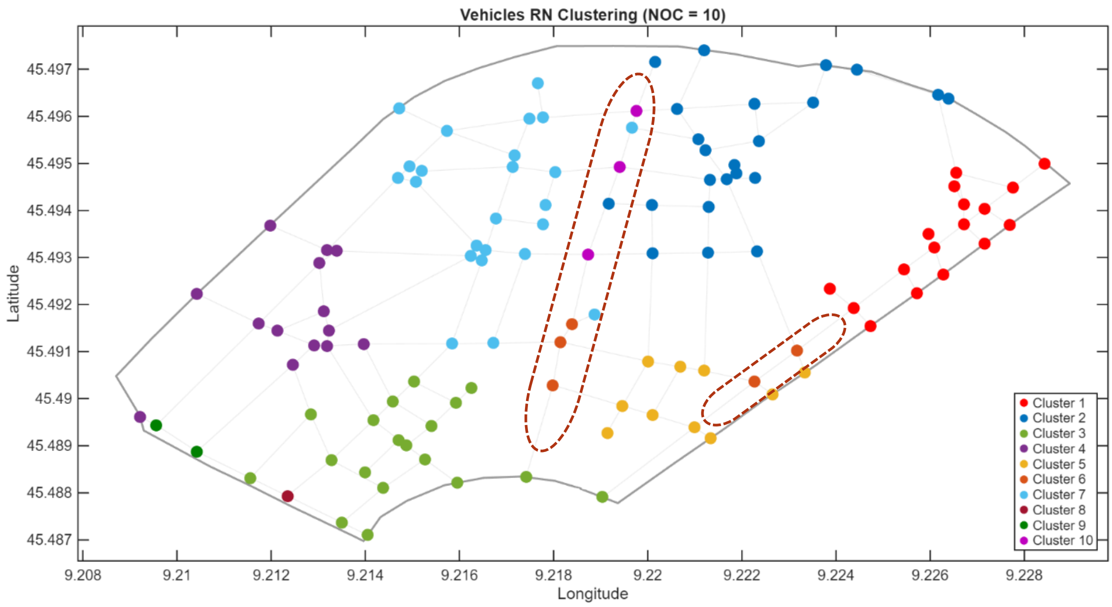

Figure 5. 41 Final vehicle crash clustering configuration, highlighting the top two clusters

5.5. Final Analysis

One of the main outcomes of the analysis is that all final selected models belong to the Raw Data family rather than the ICENTR family. This does not mean that ICENTR performed poorly. In fact, some ICENTR configurations, particularly for pedestrian crashes, achieved competitive results and CPE values that were close to those of the selected RN models. However, when prediction error, CPR, NCR, CPE, and HR were evaluated together, the Raw Data models consistently provided the most balanced and reliable performance for the NoLo district.

This outcome can be explained by the spatial structure of the district and the nature of crash occurrence within it. NoLo is a relatively compact urban area where crashes are strongly concentrated along a limited number of corridors, especially Via Padova, Viale Monza, and Viale Brianza. These corridors are not only important parts of the network but also areas with high pedestrian activity, cycling movements, traffic flow, crossings, commercial activity, and public transport access. Therefore, the actual crash locations already contain the combined effect of many local exposure factors.

ICENTR evaluates the structural importance of nodes and edges within the network. It is useful for identifying locations that are important from a connectivity and hierarchy point of view. However, a structurally important node is not always the most dangerous location. Crash occurrence also depends on local operational conditions, such as crossing demand, turning movements, traffic volume, visibility, parking activity, and interactions between different road users. These factors are not fully captured by network topology alone.

This explains why Raw Data indicators performed better in NoLo. They are based directly on where crashes actually occurred in previous years and therefore capture the real local conflict conditions of the district. In this sense, historical crash concentration acts as a practical proxy for exposure and risk factors that were not explicitly included in the model. This is particularly important in NoLo, where the main crash-prone locations are stable and repeatedly appear along the same corridors and intersections.

The crosswalk-distance analysis presented in Chapter 4 supports this interpretation. Pedestrian crashes were found to occur predominantly near crosswalks, while bicycle crashes also showed a clear concentration near crossing facilities and intersections. This confirms that many crashes are related to localized interaction points where pedestrians, cyclists, and vehicles converge. Such localized risk conditions are better represented by observed crash concentrations than by centrality measures alone.

Overall, the results show that ICENTR remains useful for understanding the structural organization of the NoLo network and identifying important areas from a connectivity perspective. However, for the purpose of crash prediction and hotspot identification, Raw Data indicators were more suitable. The dominance of RN and RE indicates that crash occurrence in NoLo is governed more by localized exposure conditions and recurring crash concentrations than by network hierarchy alone.

6 Conclusion

This thesis presented a network-based framework for the analysis and prediction of pedestrian, bicycle, and vehicle crashes within the NoLo district of Milan. By integrating five years of crash data (2017–2021) with a graph representation of the urban road network, the study investigated the spatial distribution of crashes, evaluated different node-weighting approaches, and developed predictive clustering models capable of identifying future high-risk locations.

The descriptive analysis revealed clear spatial patterns for all crash categories. Pedestrian crashes were concentrated primarily along Via Padova, Viale Monza, and Viale Brianza, reflecting the importance of major pedestrian corridors and crossing locations within the district. Bicycle crashes exhibited strong clustering along specific corridors and intersections where interactions with motorized traffic are frequent. Vehicle crashes were more numerous and spatially widespread but remained concentrated around the main urban corridors and principal intersections of the network.

To investigate these patterns, several network indicators were evaluated, including ICENTR-based measures and crash-based node-weighting approaches. The results demonstrated that ICENTR successfully captured important structural characteristics of the network and produced competitive results in several configurations, particularly for pedestrian crashes. This confirms that network hierarchy and connectivity contribute to the spatial organization of crash risk and help identify strategically important locations within the urban road system.

However, the final model selection consistently favored crash-based indicators. The selected configurations were RN–11 for pedestrian crashes, RE–5 for bicycle crashes, and RN–10 for vehicle crashes. These models achieved the best overall balance between prediction accuracy, Crash Prediction Ratio (CPR), Network Coverage Ratio (NCR), Crash Prediction Efficiency (CPE), and Homogeneity Ratio (HR).

The dominance of the Raw Data indicators suggests that crash occurrence in NoLo is influenced more strongly by localized exposure conditions than by network hierarchy alone. Historical crash distributions implicitly capture the combined effects of traffic demand, pedestrian activity, bicycle flows, crossing demand, land-use characteristics, public transport accessibility, parking activity, and roadway operations. As a result, the observed crash concentrations provide a direct representation of the actual conflict conditions experienced by road users.

The crosswalk-distance analysis further supports this interpretation. A large proportion of pedestrian crashes occurred very close to crosswalks, while bicycle crashes also showed a clear concentration near crossing facilities and intersections. These findings indicate that crash occurrence is strongly influenced by localized interaction points where pedestrians, cyclists, and vehicles converge. Such operational and behavioral factors are only partially represented by network topology but are naturally reflected in historical crash distributions.

The clustering methodology proved effective in transforming large crash datasets into interpretable spatial risk zones. The selected pedestrian model identified recurring conflict locations associated with crossings and high-activity urban corridors. The bicycle model highlighted highly localized conflict points where interactions with motorized traffic are concentrated. The vehicle model captured the major traffic corridors and intersections where crashes repeatedly occur. Together, these results provide a practical framework for identifying priority areas for safety interventions.

From an operational perspective, the identified clusters can support targeted improvements such as pedestrian crossing upgrades, traffic-calming measures, intersection redesign, cycling infrastructure enhancements, visibility improvements, and traffic management strategies. By focusing resources on the locations with the highest predicted crash risk, urban authorities can implement more efficient and evidence-based safety measures.

Overall, this thesis demonstrates that combining network analysis, crash-based node weighting, and clustering techniques provides an effective methodology for urban crash prediction. While network hierarchy remains an important explanatory factor, the results show that localized exposure conditions and recurring crash concentrations are the primary determinants of crash occurrence in NoLo. Consequently, models based on historical crash distributions provide the most reliable framework for identifying future hotspots and supporting road safety planning within the district.

Future research may extend this methodology to other districts of Milan or different urban environments to evaluate its transferability under varying network structures and land-use conditions. Additional variables such as traffic volumes, pedestrian counts, bicycle flows, public transport accessibility, and socioeconomic characteristics could also be incorporated to explicitly represent exposure. Furthermore, demographic variables such as age, gender, and user characteristics may improve the understanding of crash vulnerability and enhance prediction accuracy. Such developments would contribute to more comprehensive and transferable models capable of supporting data-driven road safety policies in complex urban environments.

Bibliography

Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention*, 38(3), 618–625.

AMAT – Agenzia Mobilità Ambiente Territorio (2023). Analisi degli incidenti stradali sulle strade di Milano, 2017–2021. Internal report. Milan, Italy: AMAT.

Andrey, J., Mills, B., Leahy, M., & Suggett, J. (2003). Weather as a chronic hazard for road transportation in Canadian cities. *Natural Hazards*, 28(2–3), 319–343.

Antoniou, C., Yannis, G., & Katsochis, D. (2013). Impact of meteorological factors on the number of injury accidents. *Journal of Traffic and Transportation Engineering*, 1(1), 12–20.

Basile, O., Persia, L., & Usami, D. (2010). A methodology to assess pedestrian crossing safety. *European Transport Research Review*, 2, 129–137.

Bergel-Hayat, R., Debbarh, M., Antoniou, C., & Yannis, G. (2013). Explaining the road accident risk: Weather effects. *Accident Analysis & Prevention*, 60, 456–465.

Clifton, K. J., Burnier, C. V., & Akar, G. (2009). Severity of injury resulting from pedestrian–vehicle crashes: What can we learn from examining the built environment? *Transportation Research Part D: Transport and Environment*, 14(6), 425–436.

Comune di Milano (2020). Piano Urbano della Mobilità Sostenibile (PUMS). Milan, Italy.

Elvik, R., Høye, A., Vaa, T., & Sørensen, M. (2009). *The Handbook of Road Safety Measures* (2nd ed.). Emerald Group Publishing Ltd., Bingley, UK.

European Commission (2022). Road Traffic Fatalities in the EU in 2022. Directorate-General for Mobility and Transport.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1), 35–41.

Huang, H., & Abdel-Aty, M. (2010). Multilevel data and Bayesian analysis in traffic safety. *Accident Analysis & Prevention*, 42(6), 1556–1565.

ISTAT (2022). Incidenti stradali in Italia 2022. Rome: Italian National Institute of Statistics.

ISTAT (2023). Popolazione residente per quartiere, Milano. Rome: Italian National Institute of Statistics.

Jung, S., Qin, X., & Noyce, D. A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis & Prevention*, 42(1), 213–224.

Kim, J. K., Ulfarsson, G. F., Shankar, V. N., & Mannering, F. L. (2010). A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. *Accident Analysis & Prevention*, 42(6), 1751–1758.

Kraidi, R., & Evdorides, H. (2020). Pedestrian safety models for urban environments with high roadside activities. *Safety Science*, 130, 104847.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), 291–305.

Ma, W., Alimo, P. K., Wang, L., & Abdel-Aty, M. (2022). Mapping pedestrian safety studies between 2010 and 2021: A scientometric analysis. *Accident Analysis & Prevention*, 174, 106744.

Marshall, W. E., & Garrick, N. W. (2011). Does street network design affect traffic safety? *Accident Analysis & Prevention*, 43(3), 769–781.

Mussone, L., Bassani, M., & Masci, P. (2017). Analysis of factors affecting the severity of crashes in urban road intersections. *Accident Analysis & Prevention*, 103, 112–122.

Mussone, L., & Alizadeh Meinagh, M. (2023). A crash data analysis through a comparative application of regression and neural network models. *Safety*, 9(2), 20.

Mussone, L., Viseh, H., & Notari, R. (2021). Novel centrality measures and applications to underground networks. *Physica A: Statistical Mechanics and its Applications*, 589, 126595.

Okabe, A., & Sugihara, K. (2012). *Spatial Analysis along Networks: Statistical and Computational Methods*. John Wiley & Sons, Chichester, UK.

Porta, S., Crucitti, P., & Latora, V. (2006). The network analysis of urban streets: A primal approach. *Environment and Planning B: Planning and Design*, 33(5), 705–725.

Pucher, J., & Dijkstra, L. (2003). Promoting safe walking and cycling to improve public health: Lessons from The Netherlands and Germany. *American Journal of Public Health*, 93(9), 1509–1516.

Schwebel, D. C., Davis, A. L., & O'Neal, E. E. (2012). Child pedestrian injury: A review of behavioral risks and preventive strategies. *American Journal of Lifestyle Medicine*, 6(4), 292–302.

Siddiqui, C., Abdel-Aty, M., & Choi, K. (2012). Macroscopic spatial analysis of pedestrian and bicycle crashes. *Accident Analysis & Prevention*, 45, 382–391.

Shrinivas, V., Bastien, C., Davies, H., Daneshkhah, A., & Hardwicke, J. (2023). Parameters influencing pedestrian injury and severity: A systematic review and meta-analysis. *Transportation Engineering*, 11, 100158.

Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*, 72, 244–256.

World Health Organization (2023). Global Status Report on Road Safety. Geneva: World Health Organization.

Xie, Y., & Yan, X. (2013). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 39, 33–43.

Yu, R., & Abdel-Aty, M. (2014). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety Science*, 63, 50–56.

Zhai, X., Huang, H., Sze, N. N., Song, Z., & Hon, K. K. (2019). Diagnostic analysis of the effects of weather conditions on pedestrian crash severity. *Accident Analysis & Prevention*, 122, 318–324.

Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. *Accident Analysis & Prevention*, 135, 105323.