

Article

Speed Behaviour Approaching Pedestrian Crossing in Urban Area

Monica Meocci * , Camilla Mazzi , Andrea Paliotto , Francesca La Torre  and Alessandro Marradi 

Department of Civil and Environmental Engineering, University of Florence, Via di Santa Marta 3, 50139 Firenze, Italy; camilla.mazzi@unifi.it (C.M.); andrea.paliotto@unifi.it (A.P.); francesca.latorre@unifi.it (F.L.T.); alessandro.marradi@unifi.it (A.M.)

* Correspondence: monica.meocci@unifi.it; Tel.: +39-0552758872

Abstract

Pedestrian safety at urban crosswalks remains a major public concern, as both vehicle speeds and roadway characteristics strongly influence drivers' behaviour when approaching these locations. This study investigates driver behaviour patterns when approaching pedestrian crossings by integrating operating speed with key road-layout features derived from a naturalistic driving experiment conducted in Florence. A dataset of 401 observations was analysed using an unsupervised clustering framework specifically designed to handle mixed numerical and categorical variables. After preprocessing, the optimal number of clusters was identified using an elbow-based model selection applied to the K-Prototypes algorithm. The analysis produced four distinct clusters, primarily differentiated by operating speed and secondarily by contextual variables such as lane number, lane width, and acceleration behaviour. Lower-speed clusters were associated with single narrow-lane configurations, whereas higher-speed clusters were characterised by wider or multilane segments and more frequent acceleration near crossings. Information Gain analysis confirmed the dominant role of lane-related attributes, while the presence of crosswalks alone did not systematically reduce speeds. Complementary clustering excluding speed resulted in fewer clusters, indicating that speed adds essential granularity to behavioural segmentation. These findings highlight the interplay between road design and driver behaviour and provide evidence-based insights to support crosswalk configurations that mitigate high-speed conflicts in urban settings.

Keywords: pedestrian safety; speed; pedestrian crossing; urban area layout; cluster analysis



Academic Editor: Ján Dižo

Received: 6 February 2026

Revised: 13 March 2026

Accepted: 18 March 2026

Published: 26 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction and Background

Road crashes in urban areas remain one of the major public health concerns worldwide, with a particularly severe impact on vulnerable road users. According to international road safety reports, a substantial share of traffic-related fatalities occurs in urban areas (more than 70%), where complex interactions between motorised vehicles and vulnerable road users (VRUs) are concentrated [1]. Research reports indicate that pedestrian deaths account for approximately 22–23% of total fatalities [2]. The total number of pedestrians involved in crashes is not explicitly reported, particularly for medium- or low-severity injuries; therefore, the percentage of pedestrians involved in crashes is reasonably higher than the rate observed for fatal crashes.

Within this context, pedestrian crossings represent one of the most critical locations in the urban environment because they are conflict points between users with markedly

different levels of protection and exposure to risk. In Italy, approximately 36% of pedestrian crashes are estimated to occur at pedestrian crossings [3]. Despite being explicitly designed to facilitate pedestrian mobility and safety, crosswalks are frequently involved in severe crashes, especially in dense urban environments characterized by high traffic demand and mixed land use.

A large body of literature has consistently highlighted the central role of vehicle speed in pedestrian safety. Both the likelihood of a collision and, more importantly, the severity of its consequences increase sharply with speed, following a strongly non-linear relationship. Even relatively small increases in operating speed are associated with a disproportionate rise in the probability of fatal or severe injury outcomes for pedestrians [4]. As a result, speed management is widely recognized as a cornerstone of urban road safety strategies. However, empirical evidence also suggests that the mean operating speed alone is insufficient to fully explain crash occurrence and severity at pedestrian crossings, pointing to the need for a more comprehensive investigation of driver behaviour and contextual factors [5–11].

In this regard, inappropriate driver behaviours when approaching pedestrian crossings, such as failure to decelerate, maintaining or increasing speed, and limited yielding compliance, have been identified as critical contributors to crash occurrence. These behaviours are often shaped by drivers' expectations and perceptions of the road environment, which are, in turn, influenced by geometric and functional characteristics of the road. Features such as the number of lanes, lane width, crosswalk design, and the presence or absence of traffic-calming elements can implicitly encourage higher operating speeds and reduce drivers' propensity to adopt precautionary behaviours near pedestrian crossings [6,11–13]. Driver distraction may also contribute to crashes, because diverting one's gaze away from pedestrians can significantly increase perception and reaction times.

Consequently, there is a growing consensus on the necessity of investigating pedestrian crash risk through integrated approaches that jointly consider driver behaviour, operating speed, and road characteristics. Understanding how these factors interact is essential for identifying distinct behavioural regimes and for supporting the design of effective countermeasures aimed at reducing pedestrian exposure to high-speed conflicts.

Data-driven and experimental studies conducted in real traffic conditions play a crucial role in this process, as they allow for a direct observation of driver responses to different crosswalk configurations and road environments [11,14,15]. Such evidence is fundamental for informing urban road design and policy decisions oriented toward safer infrastructure and sustainable mobility. Insights derived from empirical observations allow planners and policymakers to move beyond purely theoretical or model-based assumptions. They also provide a clearer understanding of how drivers behave in situ under real operational constraints. Although empirical studies on this topic are numerous [13], they also exhibit several important limitations. In many cases, the analytical methods employed require extensive datasets and detailed contextual information that are not always available or feasible to collect in real-world settings [16]. In other instances, studies are able to characterise observed patterns in driver–pedestrian interactions but fall short of fully explaining the underlying variability of the phenomenon, especially when complex behavioural, environmental, and infrastructural factors are involved [17,18]. These constraints highlight the need for more robust, multimethod research capable of capturing the multifaceted nature of pedestrian safety dynamics. This knowledge is essential for identifying design elements that may unintentionally promote risky behaviours that are directly related to human factors, which are among the most critical determinants of road safety [19].

Evidence from real-world studies can therefore guide the implementation of targeted countermeasures, including traffic-calming interventions [20], optimized crosswalk geome-

try [21], improved sightlines, and multimodal street layouts that prioritize non-motorised users [22,23]. Ultimately, data-driven insights strengthen the evidence base for policies aiming to rebalance urban mobility systems toward safer, more walkable, and more sustainable environments [15].

Building on this evidence, the recent literature has employed a wide range of methodological approaches to investigate pedestrian–vehicle interactions, particularly at crosswalks. Traditional observational studies remain one of the most widely used techniques, relying on manual or video-based recording to capture yielding behaviour, approach speeds, gap acceptance, and conflict dynamics under naturalistic traffic conditions. These studies often provide fine-grained insight into driver responses but are typically constrained by limited sample sizes and substantial effort required for data collection and annotation. To overcome these limitations, surrogate safety measures, such as time-to-collision (TTC), post-encroachment time (PET), deceleration profiles, and conflict severity indices, have been extensively adopted as reliable proxies for crash likelihood in real-world environments. Complementary to these approaches, controlled field experiments and instrumented crosswalk installations have been used to isolate the effect of specific geometric or regulatory treatments, including curb extensions, raised crossings, or advanced yield markings [24].

In parallel, the increasing availability of high-resolution trajectory data, multimodal traffic sensors, and naturalistic driving datasets has facilitated the growth of data-driven analytical approaches [13,21,25,26]. Statistical models, ranging from generalised linear models to hierarchical and mixed-effect formulations, have been widely applied to disentangle the contribution of driver behaviour, traffic conditions, and built environment variables [17]. More recently, machine learning techniques such as random forests, gradient boosting, and neural networks have emerged as powerful tools for identifying latent patterns and nonlinear interactions that traditional models struggle to capture, especially in complex urban settings [18,27]. At the same time, microsimulation tools and agent-based models have been used to reproduce pedestrian–vehicle interactions under controlled scenarios, enabling sensitivity analyses that are not feasible in live traffic environments [28]. Taken together, these different methodological approaches highlight the increasing sophistication of research in this area, while also underscoring the need for integrative frameworks capable of bridging empirical observation, advanced analytics, and mechanistic modelling to fully capture the dynamics of pedestrian safety at urban crosswalks.

Recent studies have increasingly employed cluster analysis to investigate pedestrian–vehicle interactions, especially at crosswalks by exploiting the possibility of directly capturing the patterns described by the data. Some contributions adopt vision-based approaches that extract kinematic variables from video observations and use data-mining and cluster techniques to identify recurring patterns of interaction and potential risk scenarios [29]. Other works apply cluster techniques to classify driver and pedestrian behavioural profiles, for example, by examining the effects of enforcement cameras on yielding pattern and risk perception at unsignalized crosswalks [30].

In the field of crash analysis, cluster-based methods have been used to generate pre-crash pedestrian scenarios, particularly under complex conditions such as nighttime environments [31]. Additional studies combine multiple correspondence analysis and cluster analysis to derive typologies of pedestrian injury patterns and relate them to contextual and demographic variables [32]. From a spatial perspective, cluster techniques have also been applied to identify high-risk pedestrian environments or statistically significant hot spots in the urban network [33].

More broadly, cluster analysis has been used in data-driven safety research to group crash patterns and provide a basis for targeted countermeasures and behavioural interpretation [34]. Similar approaches have also been adopted to classify urban areas with

comparable pedestrian risk factors, supporting the identification of structural characteristics associated with elevated crash likelihood [35]. Together, these applications demonstrate the value of clustering techniques in reducing the complexity of high-dimensional data, uncovering latent structures in pedestrian–vehicle interactions, and supporting the development of more focused and context-sensitive urban safety interventions.

Despite the wide range of methodological approaches adopted in the literature, several important gaps remain. Most existing studies analyse driver behaviour, roadway design, or conflict dynamics separately, often relying on regression-based models or predictive machine learning techniques to estimate crash likelihood or severity [17,18,27]. While these approaches are effective for prediction and hypothesis testing, they are generally less suited to the exploratory identification of latent behavioural patterns emerging from the interaction between heterogeneous variables such as speed, infrastructure characteristics, perceptual responses, and contextual factors.

Furthermore, although clustering techniques have been applied in pedestrian safety research to classify crash typologies, behavioural patterns, or high-risk environments [31–35], many of these studies rely on methods primarily designed for homogeneous numerical datasets or require substantial data transformation when dealing with mixed numerical and categorical information. However, real-world driver–pedestrian interactions are inherently multidimensional and characterised by heterogeneous data structures combining operational, behavioural, infrastructural, and perceptual variables derived from naturalistic observations [25,26].

As a consequence, there is still a lack of analytical frameworks capable of jointly handling mixed-type datasets while preserving interpretability and enabling the identification of coherent behavioural patterns directly applicable to urban safety design and policy evaluation. Addressing this gap is essential for moving beyond isolated variable-based analyses and toward a more integrated understanding of driver behaviour at pedestrian crossings.

These considerations highlight the need for analytical frameworks capable of jointly capturing behavioural and contextual dimensions. This gap directly motivates the present study.

To address these limitations, this study adopts a clustering approach specifically designed for heterogeneous datasets. In contrast to traditional clustering techniques such as K-Means or hierarchical clustering, which are primarily suited for numerical variables [36], the K-Prototypes algorithm allows the simultaneous analysis of numerical and categorical variables within a unified framework [37]. This feature is particularly relevant in pedestrian safety research, where behavioural responses, roadway characteristics and contextual conditions are intrinsically heterogeneous.

Compared with alternative modelling approaches such as mixed-effects regression, classification trees, or predictive machine learning models [17,27], the selected method is not primarily aimed at prediction or causal inference. Instead, it focuses on the exploratory identification of latent behavioural regimes emerging from complex interactions among variables. This makes it particularly suitable for uncovering interpretable patterns in naturalistic driving data and for supporting the design of targeted infrastructural and policy interventions based on observed behavioural typologies rather than solely on aggregated statistical relationships.

The main objective of this study is to explore and characterise drivers' behavioural patterns when approaching pedestrian crossings by jointly considering operating speed, roadway-related characteristics and psychophysiological factors involved in vehicle–pedestrian interaction derived from a naturalistic driving experiment conducted in Florence within the SWALK (Safe crossWALKs) project, funded by the University of Florence. Evidence from previous research, together with the first SWALK findings [18], shows that road layout variables alone are insufficient to fully explain pedestrian–vehicle interactions. In

this study an unsupervised cluster approach is adopted to identify homogeneous groups of observations based on both numerical and categorical information. By applying clustering techniques specifically designed for mixed-type data, the study aims to uncover latent structures in the data, quantify the relative contribution of speed and contextual variables, and provide an interpretable segmentation of driver behaviour under different crosswalk configurations. A complementary analysis excluding speed is further conducted to assess the robustness of the identified patterns and to understand the specific contribution of speed and road features.

The contribution of this study is threefold. From a methodological perspective, it demonstrates the applicability of clustering techniques for mixed-type data to the analysis of driver behaviour at pedestrian crossings, providing a structured analytical framework that jointly integrates speed profiles, infrastructural characteristics, and behavioural responses. From an empirical perspective, the study presents evidence derived from a naturalistic driving experiment conducted in a real urban environment, allowing direct observation of driver behaviour under realistic operational conditions. From a policy-oriented perspective, the identification of interpretable behavioural clusters offers actionable insights for the design of traffic-calming strategies, crosswalk configurations, and urban safety interventions aimed at mitigating high-risk vehicle–pedestrian interactions.

By explicitly linking behavioural patterns to roadway and contextual characteristics, the proposed approach contributes to bridging the gap between data-driven behavioural analysis and practical urban safety design, supporting evidence-based decision-making for safer and more inclusive urban mobility systems.

2. Methodology

This section presents the methodological framework adopted in the study. First, the design of the on-field experiment and the characteristics of the study site are described. Subsequently, the structure of the database and the variables used in the analysis are introduced and briefly described. Finally, the clustering methodology adopted to identify drivers' behavioural patterns when approaching pedestrian crossings is presented.

2.1. On-Field Experiment

This subsection describes the experimental campaign conducted to collect the empirical data used in the analysis. The on-field experiment was designed to observe drivers' behaviour under naturalistic conditions while approaching pedestrian crossings along an urban corridor. The following subsections present the characteristics of the study site, the design of the experiment, and the ethical considerations related to participant involvement.

2.1.1. Street Descriptions

The on-site experiment was conducted along Via Pistoiese, an urban collector road in the north-western suburban area of Florence. This street was selected because it is considered one of the most critical corridors in the suburban area of Florence. Its criticality derives from its dual role: providing fast access to the city centre and accommodating frequent crossing movements between the two sides of the road, driven by the high concentration of commercial activities. In this context, the high level of interaction between pedestrians and motorised users makes this road one of the locations with the highest number of pedestrian crashes in the Florence urban area [38].

The road is located in a flat area and follows a predominantly straight alignment, with a total length of approximately 4 km. The street can be divided into two main segments characterised by different cross-sectional layouts. The eastern segment is approximately 2 km long and features a cross-section consisting of two wide lanes (5.5 m each), one for

each direction of travel. The western segment, also approximately 2 km long, features a cross-section consisting of three standard-width lanes (3.0 m each), arranged alternately in a 1 + 2 configuration depending on traffic direction. In this configuration, the two traffic directions are separated by a central curb.

Signalized intersections are frequent along the corridor. Pedestrian crossings are also numerous along the corridor, including traditional zebra crossings and raised coloured crossings (sinusoidal speed tables). The latter have a sinusoidal profile, complying with the requirements of Ministerial Decree 6792 of 05.11.2001 [39] governing vertical alignment design (Figures 1 and 2). The raised crossings were designed as vertical alignment variations rather than traditional raised platforms, in accordance with Italian technical standards [40,41].

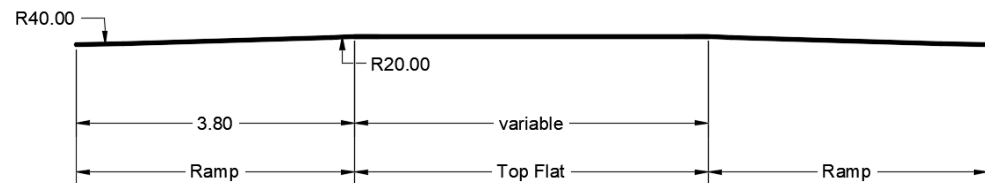


Figure 1. Sinusoidal pedestrian crossing cross section.



Figure 2. Sinusoidal pedestrian crossing—street view.

Adjacent to the carriageway, two 2 m-wide lateral areas are generally used for parking, including both authorised and unauthorised parking, while two 1.5 m-wide sidewalks complete the overall road cross-section.

2.1.2. Design of Experiment

The on-site test was carried out by recruiting a group of participants who drove along the street in both directions using their own vehicles and were instructed to drive as naturally as possible.

The recruitment campaign resulted in a sample of 36 volunteers who conducted the experiment during non-peak hours, when traffic levels were minimal (traffic is typically heavy on this street). The characteristics of the driver sample are summarized in Table 1. Thirteen of the 36 drivers can be classified as expert drivers, as they were either professional taxi drivers or had an annual mileage exceeding 20,000 km. Sixteen of the 36 participants can be classified as young drivers, as they had held a driving licence for less than 10 years and were under 25 years of age.

Table 1. Participant information.

Information		Value
age	Min	21
	Max	78
	Average	40.6
Expert drivers	Tot	13
	Men	11
	Women	2
Young drivers	Tot.	16
	Men	11
	Women	5

Almost all volunteers reported regularly driving on all types of roads (rural, urban, and highways). Eighty-six percent drive at least 5000 km per year, while 36% drive more than 20,000 km annually.

Each driver drove for approximately half an hour after receiving the necessary instructions to perform the test safely and after verifying that the devices installed in the vehicle (black box) and worn by the driver (eye-tracking glasses) were functioning properly. The volunteers were not informed about the purpose of the study.

During the experiment, each participant drove their own vehicle equipped with two synchronized devices: a black box capable of recording real-time driving data (e.g., speed, acceleration, deceleration, position, etc.) and an eye tracker device to monitor the driver's gaze. The collected information allowed us to describe both the dynamic and perceptive parameters of the drivers. Although eye-tracking data were collected during the experiment to monitor drivers' gaze behaviour, the present study focuses specifically on the analysis of operating speed and roadway contextual variables. The analysis of visual attention patterns and other psychophysiological indicators falls outside the scope of this paper and will be addressed in future research based on the same experimental dataset.

Each driver encountered a total of 19 pedestrian crossings in each travel direction along the study corridor. For the purposes of the present analysis, only crossings characterised by homogeneous geometric and environmental conditions were considered to minimise the influence of confounding factors such as nearby intersections or substantial changes in roadway configuration. This selection resulted in a subset of 12 comparable crosswalk encounters per driver (six in each direction), representing the observations used for the clustering analysis.

Because multiple crosswalk encounters were recorded for each participant, the dataset includes repeated observations for the same drivers. Consequently, the observations cannot be considered fully statistically independent, and a certain degree of intra-driver correlation may be present. However, the aim of the present study is exploratory rather than inferential, focusing on the identification of recurrent behavioural patterns emerging from the interaction between operating speed, infrastructural characteristics, and perceptual responses under comparable roadway conditions.

Within this framework, each crosswalk encounter is treated as an individual behavioural instance reflecting the combined influence of driver-specific tendencies and road context. While stable driver-related traits may contribute to behavioural variability, the clustering procedure is intended to capture behaviour–environment interaction patterns observable across the driver sample rather than to isolate individual driving styles. The potential influence of repeated measures is therefore acknowledged when interpreting the results, and the identified clusters should be considered as descriptive behavioural typologies representative of the experimental context.

2.1.3. Ethical Issues

The authors note that, at the time the experiment was conducted, the University of Florence recommended, but did not require, approval from the ethics committee for studies of this type. Consequently, no ethics committee approval was obtained for this experiment, as such approval is mandatory only for medical studies involving human subjects.

Participant recruitment was conducted exclusively on a voluntary basis. Each participant was informed about the testing procedures and their rights and provided written informed consent. The consent form consisted of two sections:

1. **Personal data:** This section collected basic personal information used solely for scheduling the test sessions. These data (name and surname) were stored only until the end of the experiment. Afterwards, all personal information was anonymised using an alphanumeric code for data analysis and processing. Participants were also informed that no photos or videos identifying them would be recorded.
2. **Test methodology:** This section described the procedure and clarified that participants were free to withdraw or stop the test at any time without explanation.

To ensure maximum protection, additional insurance coverage was taken out for all vehicles used during the test, supplementing the mandatory insurance already covering the participants' vehicles. No incidents occurred during the experiment.

2.2. Database Description

The database contains a large amount of information, only part of which was used for the present study. In particular, this research focuses on data related to the cross-sectional layout at pedestrian crossings. It includes variables describing both drivers' behaviour and road geometry, specifically:

1. Speed at the centre of the crosswalk (behavioural);
2. Acceleration/deceleration when approaching the pedestrian crossing (behavioural);
3. Presence of a traditional pedestrian crossing (geometric);
4. Presence of a sinusoidal crossing (geometric);
5. Number of lanes (geometric);
6. Lane width (geometric).

The database consists of a total of 540 records, almost evenly distributed among the described categories, as illustrated in Figure 3.

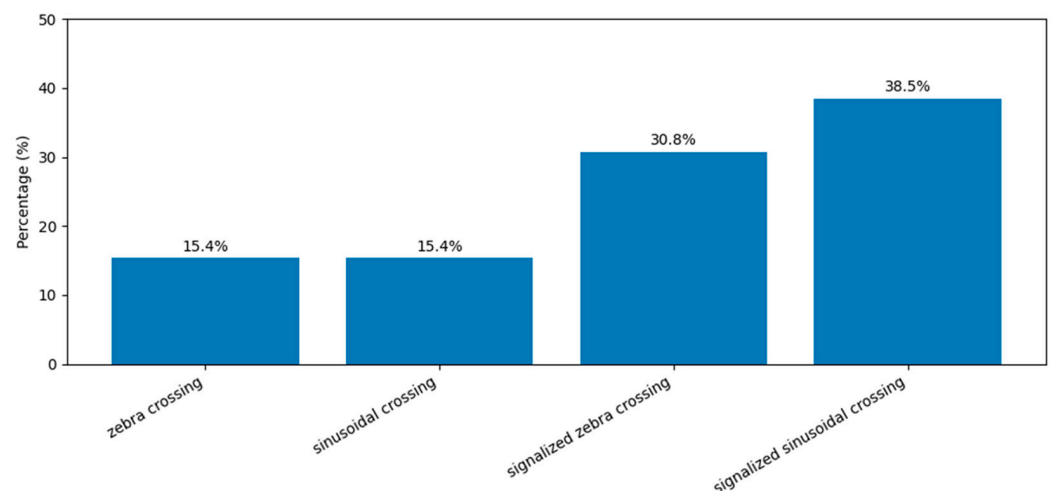


Figure 3. Pedestrian crossing distribution.

Speed was considered a continuous numerical variable, whereas the other variables were treated as binary categorical variables.

When drivers decelerate while approaching the pedestrian crossing, the variable takes the value 0, whereas if the driver maintains the same speed or accelerates when approaching the pedestrian crossing, the variable takes the value 1.

The presence of a traditional zebra crossing is coded as 1. The presence of a sinusoidal hump is also coded as 1.

The number of lanes can be equal to 1 or 2, and lane width was coded as 0 or 1 if the width was 3.50 m or 5.50 m, respectively. No other different lane widths are present in the cross sections analysed.

The high level of traffic frequently observed on the tested road occasionally affected the experiment; therefore, a preliminary analysis of the outliers was performed considering the operating speed at the centre of the crosswalk.

Although the experiment was conducted during non-peak hours, occasional traffic disturbances (e.g., short queues, temporary congestion, or signalized interruptions) were still observed along the study corridor. These conditions occasionally forced drivers to approach the pedestrian crossing at very low speeds that did not reflect voluntary behavioural adaptation to the crosswalk environment but rather external traffic constraints.

A preliminary screening of the dataset was therefore performed to identify observations clearly influenced by such traffic conditions. The excluded observations were all characterised by very low speeds (below approximately 20 km/h), with an average value of 10.46 km/h and a standard deviation of 6.38 km/h. These values are substantially lower than the typical operating speeds observed under unconstrained conditions and are representative of stop-and-go or queue-related situations rather than intentional speed modulation in response to crosswalk geometry. In a limited number of cases, very low approach speeds could also be associated with the actual presence or crossing of a pedestrian. However, such events can still be classified as traffic-related disturbances, as they impose an external constraint on drivers' natural approach behaviour. Although the behavioural response to an actively crossing pedestrian would represent an interesting phenomenon to investigate, these occurrences were too infrequent within the experimental dataset to support a statistically meaningful analysis. Consequently, these observations were conservatively classified as traffic-influenced conditions and excluded from the clustering dataset to preserve the comparability of quasi-free-flow behavioural responses across crosswalk approaches.

Therefore, these observations were not treated as statistical outliers in a strict sense but as traffic-constrained conditions not comparable with the quasi-free-flow behavioural responses that constitute the focus of the present analysis. Including such observations would have introduced systematic bias by masking drivers' natural approach behaviour and artificially inflating the variance associated with external traffic factors.

After removal of these traffic-affected observations, the final dataset used for the clustering analysis consisted of 400 records, representing comparable crosswalk approaches under homogeneous operational conditions.

The complete composition of the database information is described in Figure 4.

Figure 5 shows the speed distribution obtained from the analysed dataset. Given a speed limit of 50 km/h, the recorded speeds range from a minimum of 22 km/h to a maximum of 67 km/h, indicating a relatively wide variability in driving behaviour within the observed sample. The average speed is 42 km/h, which represents the central tendency of the data, while the standard deviation is 7.53 km/h.

The speed distribution appears roughly unimodal and symmetric, with most observations concentrated between 35 and 50 km/h. The histogram shows that the majority of values cluster around 40–45 km/h, indicating relatively consistent driving behaviour

among participants. Lower speeds (below 30 km/h) and higher speeds (above 55 km/h) occur much less frequently and represent only a small proportion of the dataset.

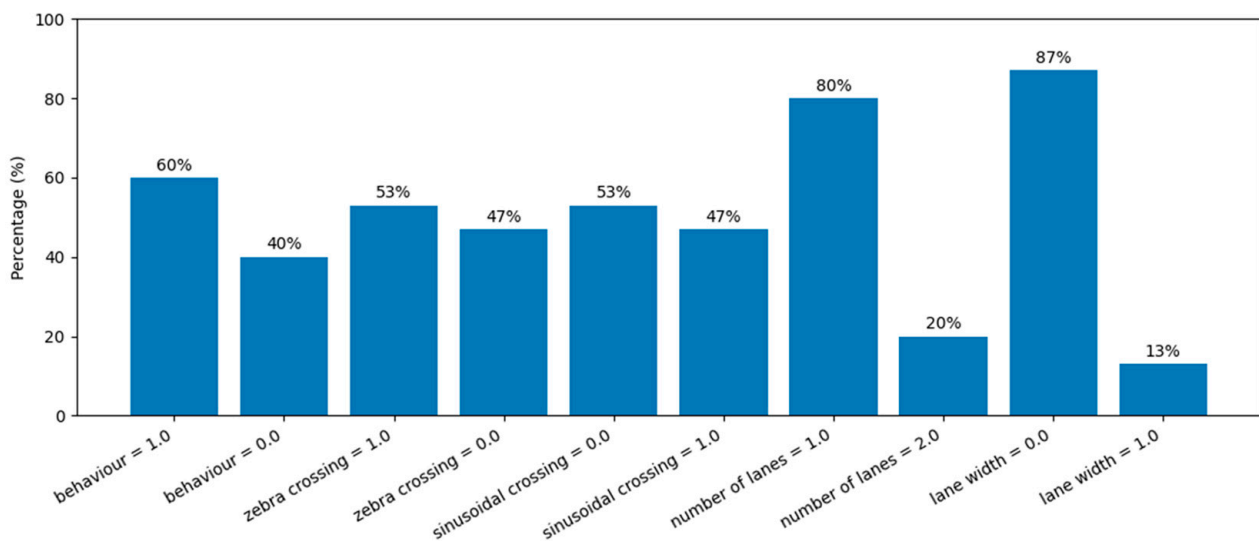


Figure 4. Graphical description of the database.

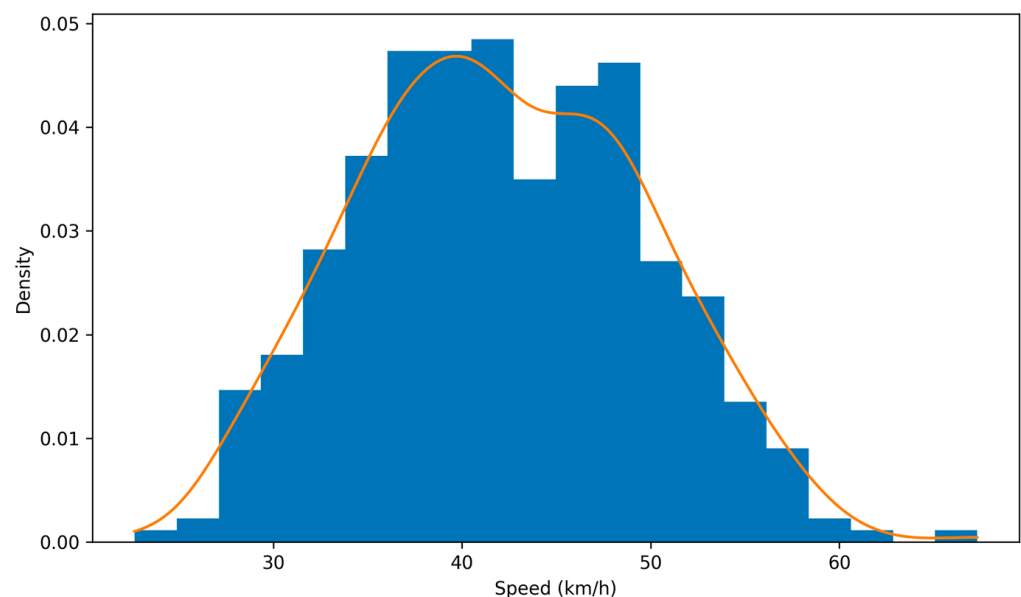


Figure 5. Speed distribution.

2.3. Data Analysis

The analysis was carried out within an unsupervised clustering framework specifically designed for mixed-type datasets (i.e., datasets comprising both numerical and categorical variables). Such data structures are common in applied research and pose non-trivial methodological challenges, as standard cluster techniques typically rely on distance measures that are not directly applicable to heterogeneous feature spaces. In particular, transforming categorical variables into numerical values may introduce an artificial ordering or misrepresent the true relationship between categories. To address these issues, a clustering approach capable of jointly handling numerical and categorical variables without imposing arbitrary transformations was adopted, as formalized in the K-Prototypes clustering framework [37].

Before the clustering procedure, an extensive data preprocessing phase was implemented to enhance data quality, internal consistency, and robustness of the subsequent

analysis. Missing values in numerical variables were imputed using the median value, a robust statistic that mitigates the influence of skewed distributions and extreme values. For categorical variables, missing observations were replaced with the mode, ensuring that the most representative category for each variable was preserved and that the overall distributional structure of the data was minimally altered. Numerical features were retained on their original continuous scale, while categorical variables were treated as purely nominal, explicitly avoiding any implicit ordinal interpretation.

The optimal number of clusters was determined through a formal model selection procedure based on the evaluation of the clustering objective function over a predefined range of candidate values of K . Specifically, the K-Prototypes algorithm was applied iteratively for increasing values of K , and the corresponding values of the within-cluster dissimilarity (i.e., the objective function) were recorded. The optimal cluster configuration was identified using the elbow (or knee) criterion, which aims to detect the point beyond which further increases in the number of clusters lead to marginal improvements in model fit following established clustering validation practices [42]. Operationally, this point was defined as the maximum perpendicular distance from the curve representing the objective function values to the straight line connecting the values observed at the minimum and maximum considered numbers of clusters. This criterion provides an objective and reproducible way to balance model parsimony against within-cluster homogeneity, thereby avoiding reliance on purely subjective visual inspection of the elbow plot.

Once the optimal number of clusters had been determined, the final clustering solution was obtained by applying the K-Prototypes algorithm to the full dataset. This algorithm represents an extension of the classical K-Means approach, specifically designed to handle datasets composed of both numerical and categorical variables by integrating the corresponding dissimilarity measures into a single optimization framework.

For numerical variables, dissimilarity between observations and cluster centroids is computed using the squared Euclidean distance, which allows the algorithm to effectively capture variations in continuous attributes. For categorical variables, a simple matching dissimilarity measure is used, whereby discrepancies between the observed categories and the cluster-specific modes are penalised.

By combining these two types of dissimilarities, the K-Prototypes algorithm minimizes an overall objective function that balances the contribution of numerical and categorical variables, enabling a coherent and consistent clustering of mixed-type data. The mathematical formulation of this objective function is reported in Equation (1).

$$J = \sum_{i=1}^n \sum_{k=1}^K \delta_{ik} \left(\sum_{j \in \mathcal{N}} (x_{ij} - \mu_{kj})^2 + \gamma \sum_{j \in \mathcal{C}} \mathbb{I}(x_{ij} \neq m_{kj}) \right) \quad (1)$$

where:

- n is the number of observations;
- $\mathcal{N}$ and $\mathcal{C}$ represent the sets of numerical and categorical variables, respectively;
- μ_{kj} and m_{kj} correspond to the numerical centroids and categorical modes of cluster k ;
- δ_{ik} is an indicator variable denoting cluster membership;
- γ is a weighting parameter that regulates the relative influence of numerical and categorical components.

The weighting parameter γ regulates the relative contribution of numerical and categorical components in the overall dissimilarity function and therefore plays a key role in the clustering structure. In practical implementations of the K-Prototypes algorithm, γ is not arbitrarily imposed but is commonly determined through an automatic scaling heuristic that balances the influence of heterogeneous variables.

In the present study, γ was calibrated using the standard data-driven scaling rule embedded in the K-Prototypes implementation, which relates the dispersion of numerical variables to the number of categorical features included in the clustering procedure. In particular, the parameter was determined according to Equation (2).

$$\gamma = \frac{\sigma_{num}^2}{p_{cat}} \quad (2)$$

where:

- σ_{num}^2 represents the average variance of the numerical variables;
- p_{cat} denotes the number of categorical variables considered in the analysis.

This formulation allows the categorical mismatch penalty to be scaled proportionally to the variability of the numerical attributes, ensuring that neither numerical nor categorical components dominate the overall distance metric.

This automatic calibration provides a reproducible and data-driven approach for harmonising heterogeneous feature spaces and is widely adopted in clustering applications involving mixed-type datasets. The resulting γ value was therefore retained for all clustering iterations to guarantee internal consistency and robustness of the final partition.

Through an iterative process of cluster assignment updates and prototype re-estimation, the algorithm converges to a locally optimal partition of the data, in which observations within the same cluster exhibit high internal homogeneity with respect to both numerical attributes and categorical profiles. This joint optimization framework allows the clustering solution to simultaneously capture similarities in numerical magnitudes and categorical patterns, thereby ensuring that the intrinsic statistical properties of each variable type are preserved throughout the analysis. As a result, the adopted methodology yields an effective and interpretable segmentation of mixed-type data, avoiding the loss of information that may arise from ad hoc variable transformations or oversimplified distance metrics, as originally formalized in the K-Prototypes framework proposed by Huang [37].

To assess the robustness and reproducibility of the clustering solution, an explicit stability analysis was performed. As partition-based clustering algorithms such as K-Prototypes may converge to local optima depending on initialization, the clustering procedure was repeated across multiple random seeds with several internal initializations for each run. This approach allows verification that the resulting partition is not an artefact of a specific initial configuration but reflects intrinsic patterns in the data.

Cluster similarity across repeated runs was quantified using the Adjusted Rand Index (ARI), a widely adopted metric for comparing clustering partitions. The ARI measures the agreement between different clustering solutions while accounting for chance, with values close to 1 indicating strong similarity. By evaluating clustering consistency across multiple runs, this procedure provides a quantitative assessment of the stability and reproducibility of the identified cluster structure.

Following the identification of the final cluster structure, an additional post-clustering interpretability analysis was performed to evaluate the relative contribution of road-related variables in characterizing each cluster, and consequently in describing drivers' behavioural patterns when approaching different types of pedestrian crossings. To this end, an entropy-based Information Gain (IG) measure was employed to quantify the explanatory power of categorical variables within each cluster. For each categorical feature, Shannon entropy was first computed over the entire dataset, providing a measure of the overall uncertainty associated with that variable [43]. Subsequently, conditional entropy was calculated within each cluster, reflecting the degree of homogeneity induced by the clustering process. The Information Gain was then defined as the reduction in entropy achieved by conditioning on cluster membership, computed as the difference between the global entropy and the

cluster-specific one, as formalized in Equation (3). Higher IG values indicate variables that are particularly informative in distinguishing cluster-specific behavioural profiles and thus play a key role in characterising drivers' responses across different pedestrian crossing configurations.

$$IG_c(X) = H(X)_{\text{global}} - H(X | c) \quad (3)$$

where the Entropy for the categorical variable can be estimated as in Equation (4).

$$H(X) = - \sum_i p_i \log_2(p_i) \quad (4)$$

To facilitate interpretation and improve the comparability of the relative contribution of different variables, Information Gain values were also normalized with respect to the baseline entropy of each categorical feature. The normalized Information Gain (NIG) expresses the proportion of uncertainty reduction achieved by conditioning on cluster membership relative to the initial entropy of the variable, as reported in Equation (5).

$$NIG_c(X) = \frac{IG_c(X)}{H(X)_{\text{global}}} \quad (5)$$

This metric captures the extent to which a given categorical variable becomes more homogeneous within a cluster relative to its distribution in the overall population. Specifically, a higher IG (and correspondingly higher NIG) indicates that the variable exhibits a more concentrated and less uncertain distribution when conditioned on cluster membership, thereby playing a stronger role in characterizing the internal structure of the cluster. In this sense, elevated values of the IG index identify variables that contribute most substantially to the semantic differentiation of clusters by highlighting dominant categorical patterns and reducing entropy within each group [43].

Importantly, this interpretability approach is fully consistent with the unsupervised nature of the K-Prototypes framework, as it relies exclusively on the empirical distributional properties of categorical variables and does not introduce any form of supervision, labelling, or outcome-driven bias. The IG is computed solely as a post-clustering diagnostic measure and does not affect the clustering optimisation process itself, thereby preserving the integrity of the original partitioning derived from the joint numerical–categorical dissimilarity measure [37].

For each cluster, the two categorical variables exhibiting the highest IG values were identified as the most explanatory features. This selection criterion provides a concise yet informative summary of the dominant structural attributes defining each cluster, facilitating interpretation while avoiding over-parameterisation. By focusing on the most informative categorical variables, this approach enhances the readability and communicability of the clustering results, particularly in applied contexts where the goal is to link cluster structures to meaningful behavioural or contextual patterns.

3. Results

The clustering analysis was conducted on an updated dataset consisting of 400 observations and six variables, including one numerical variable (speed) and five categorical attributes (behaviour, zebra crossing, sinusoidal hump, number of lanes and lane width). The use of the K-Prototypes algorithm is appropriate in this context, as it enables the joint analysis of numerical and categorical variables within a unified clustering framework.

The optimal number of clusters was determined using the elbow (knee) analysis of the K-Prototypes cost function evaluated over a range of candidate values of K . The results show a substantial reduction in clustering cost when increasing K from 2 to 4, followed by progressively smaller improvements for higher values of K . Based on the point of

maximum curvature of the cost curve, the elbow was identified at $K = 4$, indicating an appropriate balance between model complexity and within-cluster homogeneity. The final solution with four clusters was further supported by the convergence behaviour of the algorithm, as multiple initialization runs consistently converged to very similar final cost values, suggesting a stable clustering structure (Figure 6).

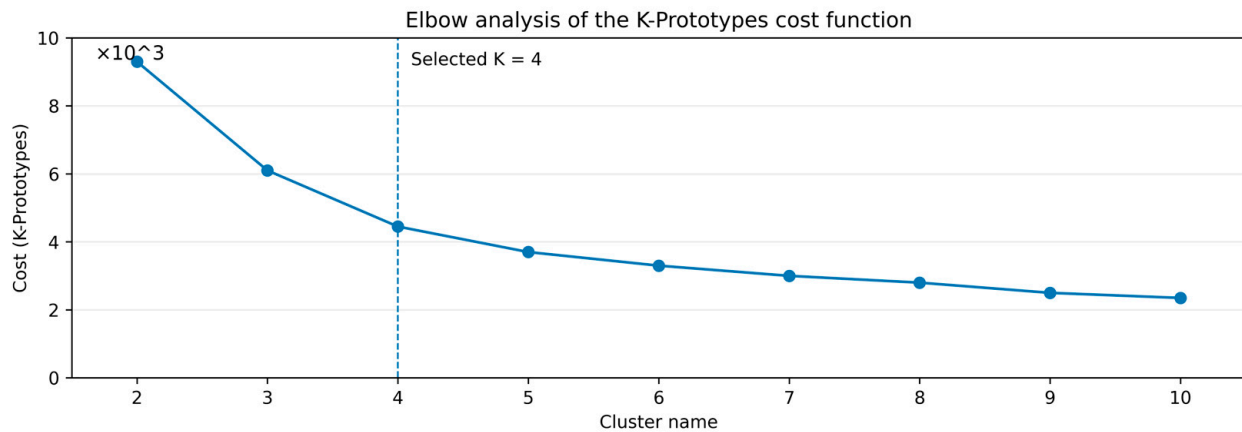


Figure 6. Selection of the optimal number of clusters by the elbow method (speed and categorical variables).

The stability assessment confirmed the robustness of the clustering solution with $K = 4$. Repeated executions of the K-Prototypes algorithm across 30 different random seeds consistently identified comparable cluster structures. The similarity between partitions obtained from different initializations was quantified using the Adjusted Rand Index (ARI), yielding a mean value of 0.980 (minimum 0.721; maximum 1.000; standard deviation 0.052).

The high mean ARI value indicates a very strong agreement between clustering solutions, demonstrating that the identified behavioural typologies are largely independent of the random initialization of the algorithm. The limited variability observed in a small number of runs does not alter the overall cluster interpretation, confirming the robustness and reproducibility of the four-cluster configuration.

The resulting clusters exhibit a reasonably balanced size distribution, with Cluster 1 being the largest group (138 observations), followed by Cluster 3 (121 observations), Cluster 0 (85 observations), and Cluster 2 (57 observations). Although Cluster 2 is comparatively smaller, none of the clusters appear degenerate or trivial, indicating that the partition captures meaningful subgroups within the data.

An examination of numerical profiles reveals a clear stratification of clusters according to operating speed. Cluster 0 is characterized by the lowest average speed (≈ 32 km/h), while Cluster 1 represents an intermediate speed regime (≈ 39 km/h). Clusters 3 and 2 correspond to higher speed levels, with mean values of approximately 46 km/h and 54 km/h, respectively. This monotonic ordering confirms that speed constitutes the primary continuous axis of separation among clusters and plays a dominant role in defining their structure (Table 2).

Table 2. Cluster information.

Cluster ID	Cluster Size	Average Speed
0	85	32 km/h
1	138	39 km/h
2	57	54 km/h
3	121	46 km/h

The categorical variables, summarized through their cluster-specific modal values, provide additional insight into the interpretation of the clustering outcome by highlighting the dominant contextual configurations associated with different operating speed regimes. In particular, Clusters 0 and Cluster 1, which are characterised by the lowest observed operating speeds (below 40 km/h), exhibit the highest prevalence of roadway segments defined by a single narrow traffic lane, with proportions ranging from approximately 85% to 92%. This finding suggests a strong association between constrained roadway geometry and lower speed preferences, consistent with well-established relationships between lane width, perceived risk, and speed choice [44,45].

Within these two low-speed clusters, the type of pedestrian crossings is approximately evenly distributed, with shares close to 50%. This balanced distribution indicates that, under geometrically constrained conditions, the type of pedestrian crossing does not emerge as a primary discriminating factor within the clustering structure. This should not be interpreted as evidence of behavioural irrelevance; rather, it suggests that within the multidimensional framework considered, other contextual variables, particularly roadway geometry, play a more dominant role in differentiating speed regimes. Instead, drivers appear to adjust their operating speed primarily in response to the physical characteristics of the roadway, while crosswalks alone do not introduce an additional differentiating effect in this context, at least not when considered in isolation.

Conversely, Cluster 2 and Cluster 3 display a pronounced reduction in the proportion of segments featuring a single narrow lane, reaching the lowest value observed across all identified clusters. These clusters are instead associated with roadway configurations characterised by wider lanes or multiple lanes, which are typically conducive to higher operating speeds. In parallel, both clusters exhibit an increased share of drivers showing acceleration behaviour when approaching pedestrian crossings, indicating a distinct behavioural response compared to the lower-speed clusters. This pattern suggests that in less constrained roadway environments, drivers may perceive greater manoeuvring freedom, which can translate into speed increases even in the presence of pedestrian infrastructure [46,47].

Similarly to the patterns observed in the lower-speed clusters, variability in the presence of pedestrian crossings does not delineate a clear or consistent separation between Clusters 2 and 3. The absence of a systematic differentiation across all clusters suggests that operating speed preferences are more strongly associated with overall roadway configuration, particularly lane number and lane width, than with the presence of crosswalks considered in isolation. However, this finding should be interpreted cautiously, as the clustering approach identifies segmentation patterns rather than testing causal relationships or statistical significance.

Overall, the updated clustering results reveal a coherent and interpretable structure characterised by a dominant speed-driven hierarchy, complemented by a secondary categorical differentiation that distinguishes two broader contextual regimes. The presence of clusters sharing identical categorical profiles but exhibiting distinct speed levels highlights the ability of the K-Prototypes algorithm to disentangle continuous effects even when categorical conditions remain constant. At the same time, the emergence of distinct categorical patterns among higher-speed clusters underscores the role of contextual roadway features in shaping driver behaviour. Collectively, these findings confirm the robustness of the clustering solution and provide a solid foundation for subsequent interpretability analyses and domain-specific interpretation of the results, as illustrated in Figures 7 and 8.

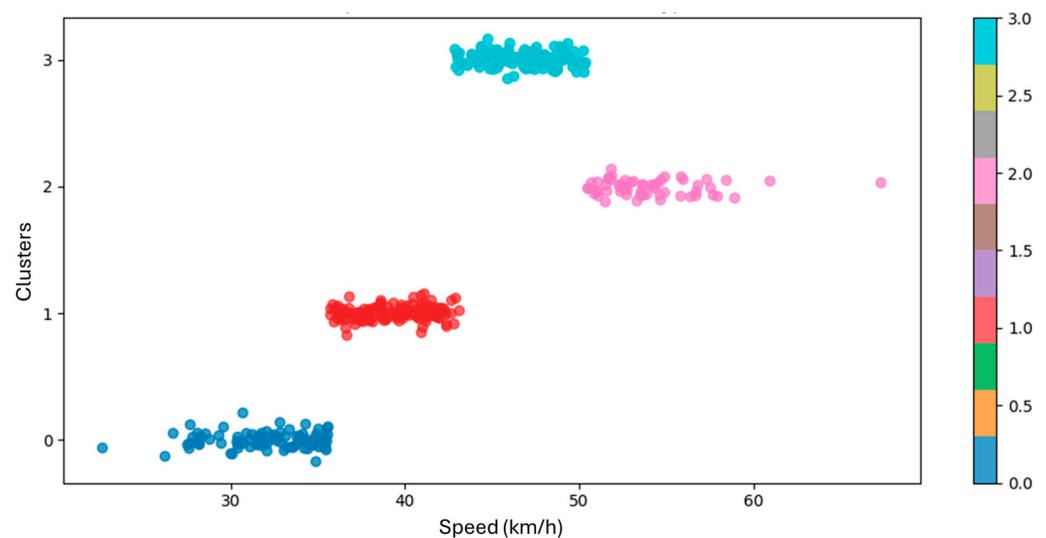


Figure 7. Cluster distribution.

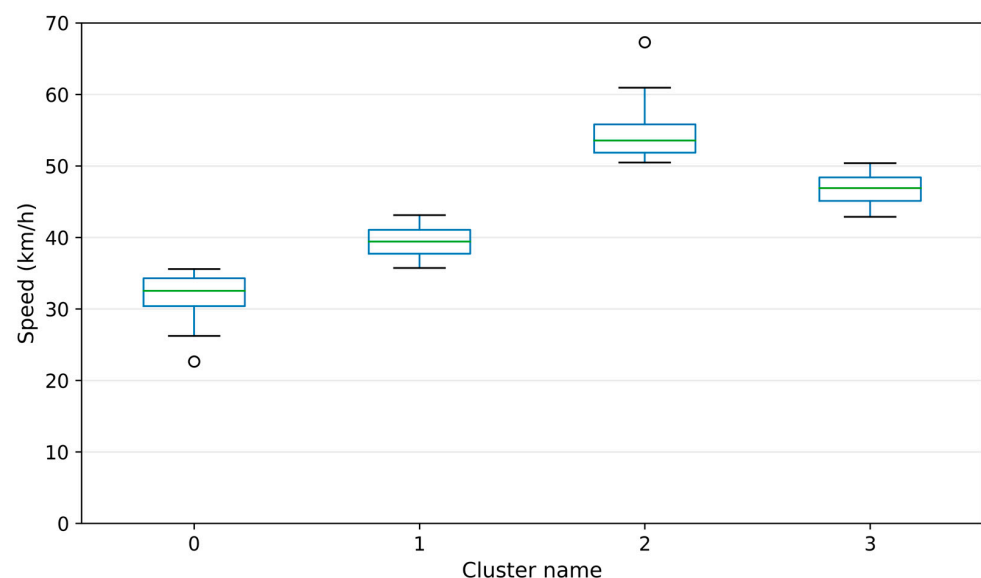


Figure 8. Speed distribution across the four identified clusters (boxplot representation).

Figure 8 illustrates the distribution of the numerical variable operating speed across the four identified clusters. The corresponding boxplots highlight a clear and systematic separation among clusters along the speed dimension, with median values progressively increasing from Cluster 0 to Cluster 2, while Cluster 3 represents an intermediate yet distinctly high-speed regime. This ordered pattern confirms that operating speed constitutes a primary discriminating dimension of the clustering structure. The limited overlap observed among the interquartile ranges further indicates a high degree of within-cluster homogeneity, coupled with statistically meaningful differences across clusters, thereby supporting the internal consistency and interpretability of the clustering solution.

In addition to differences in central tendency, Figure 8 reveals a notable increase in speed dispersion as cluster-level operating speed increases. Specifically, higher-speed clusters are characterized by wider interquartile ranges and longer whiskers, reflecting greater variability in observed speeds. This increase in variance suggests more heterogeneous driving behaviour under less constrained roadway conditions and carries relevant implications in terms of traffic safety, as higher speed variability is commonly associated with increased conflict risk and reduced predictability of driver responses.

To further enhance the interpretability of the clustering results, an IG analysis was conducted for the categorical road-related variables. The estimation of the IG index enables the identification of those variables that contribute most strongly to defining the internal structure of each cluster by quantifying the reduction in uncertainty achieved when conditioning on cluster membership. To facilitate comparison across variables with different baseline entropy levels, the corresponding normalized Information Gain (NIG) values were also computed. The results of this analysis are summarized in Table 3, where both IG and normalized IG values are reported for each cluster, providing a concise and informative characterisation of the dominant contextual features underlying the observed speed regimes.

Table 3. Variable importance per cluster (two most important variable).

Cluster ID	Baseline Entropy	IG	Normalized IG	Variable
0	0.730815	0.320399	0.438413	Number of lanes
0	0.997626	0.172660	0.173071	Zebra crossing
1	0.730815	0.172186	0.235608	Number of lanes
1	0.563358	0.089785	0.159375	Lane width
2	0.973237	0.023785	0.024439	Acceleration behaviour
2	0.997626	0.008532	0.008552	Zebra crossing
3	0.563358	0.151929	0.269684	Lane width
3	0.997626	0.066167	0.066325	Zebra crossing

The values obtained confirm the trend previously described, namely that the presence of a single narrow traffic lane is predominantly associated with clusters characterized by lower operating speeds. This finding is consistent with the notion that geometrically constrained roadway environments induce higher perceived risk and reduced manoeuvring space, thereby encouraging more cautious driving behaviour. Conversely, clusters associated with the highest operating speeds are primarily characterized by different road-related variables, notably greater lane width and a higher prevalence of acceleration behaviour when approaching pedestrian crossings. These results suggest that speed choice is more strongly associated with the combined effect of roadway geometry and driver perception of available space than with isolated infrastructural elements. Nevertheless, given the exploratory nature of the clustering framework, these associations should be interpreted as indicative patterns rather than causal relationships.

The observed patterns are in line with recent empirical evidence identifying lane width, number of lanes, and overall roadway layout as key determinants of higher operating speeds in urban environments. Several studies have shown that wider lanes and multi-lane configurations are systematically associated with increased speeds and more aggressive driving behaviour, even in the presence of traffic-calming elements or pedestrian infrastructure [44,45]. In this context, acceleration behaviour can be interpreted as a behavioural response to an environment perceived as permissive, reinforcing the link between road design and speed-related risk exposure. Similar findings have been reported in recent safety-oriented analyses, which highlight how urban road environments with generous cross-sections tend to encourage speed variability and aggressive manoeuvres, thereby increasing crash risk and injury severity [46,47].

To further assess the specific contribution of speed to the overall clustering structure, the analysis was replicated by excluding the numerical speed variable and performing the clustering procedure exclusively on categorical attributes. This alternative specification aims to isolate the extent to which structural and contextual roadway characteristics alone are capable of defining homogeneous groups, independently of speed-related information.

Given the exploratory and unsupervised nature of the analysis, this comparison is not intended to establish causal relationships but rather to evaluate the robustness, stability, and interpretability of the clustering solutions under alternative model specifications. By contrasting the full and reduced-variable clustering outcomes, it becomes possible to assess whether the identified cluster structure is intrinsically driven by roadway configuration or primarily shaped by speed dynamics, thereby strengthening the overall interpretive validity of the results.

When speed is excluded, the optimal number of clusters decreases from four to three, as indicated by the elbow criterion applied to the K-Modes cost function (Figure 9). This reduction suggests that, in the absence of the continuous speed dimension, the categorical variables alone support a coarser partition of the data. In other words, some of the distinctions observed in the full specification appear to be primarily driven by differences in speed rather than by differences in categorical configurations.

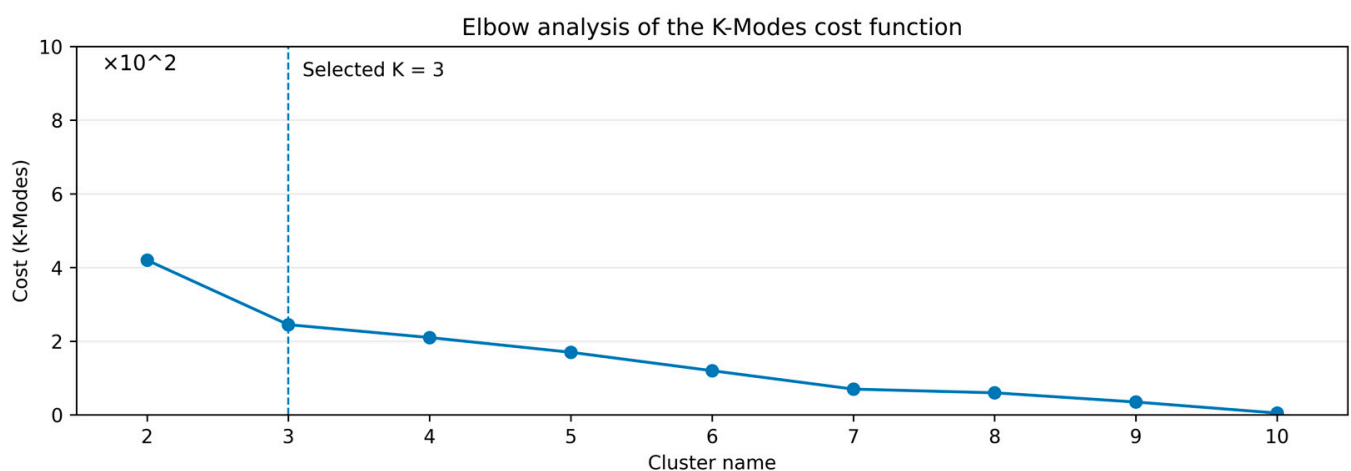


Figure 9. Selection of the optimal number of clusters by elbow methods (only categorical variables).

The three clusters obtained using categorical variables only are reasonably balanced in terms of size, comprising 212, 102, and 87 observations, respectively. This relatively even distribution suggests that the clustering solution is not driven by extreme class imbalance and that each cluster captures a substantive subset of the dataset. Moreover, the clusters exhibit distinct modal profiles across the categorical dimensions, with particularly pronounced differences observed in relation to the type of pedestrian crossing (e.g., zebra crossings versus sinusoidal speed humps) and the number of traffic lanes. These differences indicate that, even in the absence of speed-related information, the clustering procedure is able to identify meaningful and interpretable structural patterns grounded in roadway configuration and traffic control characteristics.

At the same time, the exclusion of the numerical speed variable leads to a reduction in the overall granularity of the clustering solution. Compared to the mixed-variable specification, the categorical-only clustering yields a smaller number of clusters with broader internal heterogeneity, reflecting the loss of a key continuous dimension that previously contributed to finer distinctions among observations. This outcome is consistent with the role of speed as a primary discriminating variable in the full model, while also confirming that categorical attributes alone retain sufficient information to delineate coherent contextual regimes.

Importantly, operating speed remains informative when examined ex post. Although not used during the clustering process, the three clusters derived solely from categorical variables exhibit clearly differentiated speed distributions when analysed a posteriori. In particular, mean operating speeds vary across clusters, ranging approximately between

40 km/h and 45 km/h, as illustrated in Figure 10. This result suggests that roadway and crosswalk characteristics implicitly encode speed-related information, reinforcing the interpretation that roadway geometry and traffic control features exert a systematic influence on speed choice. The emergence of distinct speed patterns *ex post* further supports the robustness of the clustering approach and highlights the complementary role of numerical and categorical variables in capturing the multifaceted nature of driver behaviour.

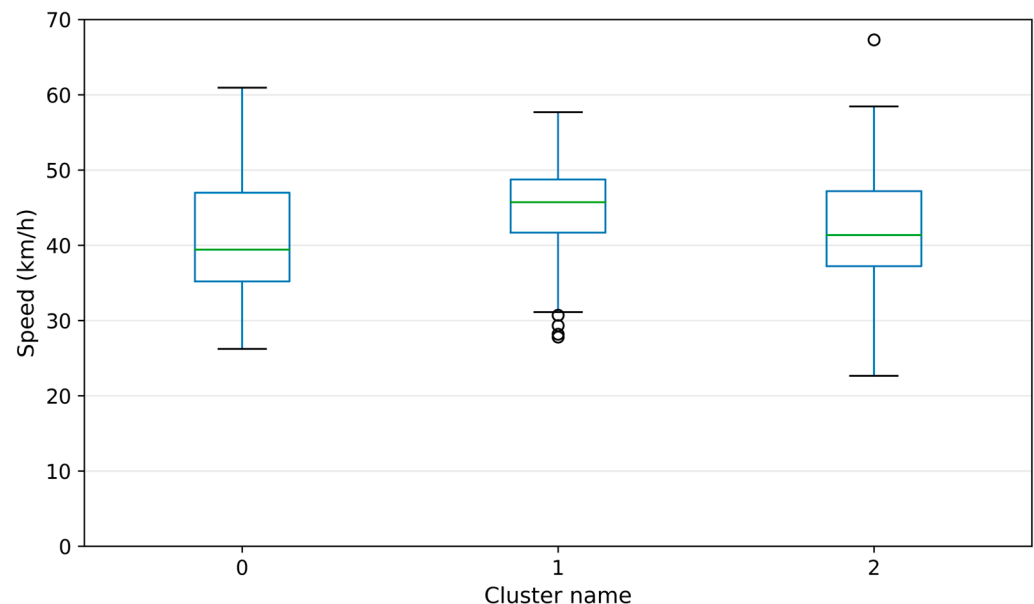


Figure 10. Speed distribution across the three identified clusters (boxplot representation).

This finding suggests that specific roadway configurations are systematically associated with distinct speed regimes, even when operating speed is not explicitly included as an input variable in the clustering algorithm. In other words, geometric and contextual road attributes appear to implicitly encode speed-related information, reinforcing the notion that drivers' speed choices are strongly shaped by the surrounding infrastructure and traffic control environment.

To further substantiate this interpretation with a quantitative assessment, an analysis of variance was conducted on operating speed across the categorical-only clusters.

To provide a quantitative assessment of the extent to which categorical roadway variables implicitly encode speed information, a one-way analysis of variance (ANOVA) was conducted on operating speed across the clusters obtained from the categorical-only specification. The results indicate statistically significant differences in mean speed among the three clusters ($F = 13.26$, $p < 0.001$).

The estimated effect size, expressed as eta squared ($\eta^2 = 0.063$), indicates a moderate proportion of variance in operating speed explained by the categorical clustering structure. These findings provide quantitative support for the interpretation that roadway geometry and contextual variables implicitly embed meaningful speed-related information, even when operating speed is not explicitly included as an input variable in the clustering procedure.

This result is consistent with previous empirical studies showing that lane width, cross-sectional design, and traffic calming elements act as powerful implicit speed regulators in urban contexts [44,48].

Despite the robustness and internal coherence of the analytical framework adopted, several limitations should be acknowledged. First, the use of the K-Prototypes algorithm entails reliance on a distance-based partitioning method that converges to locally optimal solutions and may exhibit sensitivity to initialization, potentially affecting cluster stability

even when multiple runs are performed. This limitation is common to centroid-based clustering techniques and has been widely discussed in the methodological literature on unsupervised learning [36]. Moreover, while the clustering approach effectively captures associative patterns among road characteristics, driver behaviour, and operating speed, it does not allow for the identification of causal relationships, as the analysis remains inherently observational and exploratory.

Additional constraints arise from the characteristics and spatial scope of the dataset itself. The observations analysed in this study were collected along a single urban corridor located in the city of Florence. Although the selected corridor presents a variety of roadway configurations and pedestrian crossing treatments, the results should be interpreted within the specific infrastructural, traffic, and behavioural context of the study area. Consequently, the identified clustering patterns should not be interpreted as universally applicable to all urban environments, as differences in urban layout, traffic conditions, cultural driving behaviour, and regulatory frameworks may influence driver responses in other contexts.

Although the adopted preprocessing strategies mitigate the impact of missing values, the relatively limited sample size and the presence of categorical variables with uneven or sparse distributions may influence both the granularity and the interpretability of the resulting clusters. Furthermore, the analysis incorporates a restricted set of behavioural and road-related variables. As a result, potentially relevant unobserved factors, such as traffic volume, pedestrian presence, visibility conditions, sightlines, or driver-specific attributes, may also play a role in shaping speed patterns but remain unaccounted for in the current specification. Finally, while the complementary clustering exercises excluding speed provide valuable insight into the robustness of the identified structures, they still rely on the same underlying dataset and therefore cannot fully disentangle the complex interactions between behavioural and contextual determinants.

Taken together, these limitations suggest that future research should aim to incorporate richer and more diverse data sources, including additional behavioural indicators and exposure variables, as well as alternative clustering and modelling techniques. Such extensions would help to further validate the observed patterns, enhance causal interpretability, and improve the external validity and transferability of the findings across different urban contexts.

4. Conclusions

This study applied a mixed-type clustering framework to characterise driver behaviour at pedestrian crossings using real-world data collected through a naturalistic driving experiment conducted in Florence. The results consistently indicate that operating speed represents the primary dimension differentiating driver behaviour, confirming its central role in shaping distinct behavioural regimes. In line with the cluster-level analyses, lane number and lane width emerged as the most influential contextual factors associated with these speed regimes, highlighting the dominant effect of roadway geometry on drivers' speed choices and acceleration behaviour.

Conversely, the presence of pedestrian crosswalks alone did not emerge as a primary discriminating variable within the clustering structure. This finding should not be interpreted as evidence that crosswalks have no influence on driver behaviour. Rather, it indicates that within the exploratory clustering framework adopted, crosswalk configuration did not represent the dominant factor separating behavioural groups when compared with broader roadway geometry and contextual characteristics. It is important to emphasize that the clustering analysis identifies latent segmentation patterns within the dataset but does not provide inferential testing of effect size or statistical significance. Therefore, the

results should be interpreted as evidence of relative prominence among variables within the adopted analytical framework rather than as proof of the absence of behavioural influence.

This finding, corroborated by both the mixed-variable and categorical-only clustering specifications, underscores that isolated traffic control elements may be insufficient to induce consistent speed reductions when not embedded within a geometrically constraining road environment. Instead, the results point to the importance of holistic roadway design, where cross-sectional characteristics implicitly regulate speed by shaping drivers' risk perception and available manoeuvring space.

The complementary clustering analysis excluding speed further confirmed the structural role of geometric and contextual features, demonstrating that meaningful and interpretable clusters can still be identified based only on categorical variables. At the same time, the ex-post analysis of speed distributions revealed systematic differences across these clusters, highlighting the additional explanatory power gained by explicitly incorporating speed into the clustering process. This comparison strengthens the robustness of the proposed framework and illustrates how numerical and categorical variables jointly contribute to capturing the multidimensional nature of driver behaviour.

Overall, these findings provide empirical support for the view that effective speed management at pedestrian crossings requires coordinated design strategies that integrate crosswalk treatments with appropriate lane configurations and roadway layouts. From a practical perspective, the results offer valuable evidence to inform the design of pedestrian crossings and surrounding road environments aimed at moderating vehicle speeds and enhancing pedestrian safety. However, the transferability of these findings should be considered with caution. The empirical analysis is based on observations collected along a single urban corridor in the city of Florence, and the identified behavioural patterns therefore reflect the specific infrastructural, traffic, and contextual conditions of the study area. While the observed relationships between roadway geometry and operating speed provide meaningful insights, their applicability to other urban environments may vary depending on differences in urban layout, traffic demand, regulatory frameworks, and local driving behaviour. Future research should extend this framework by integrating additional behavioural indicators, traffic and exposure conditions, and visual attention or eye-tracking data, as well as by exploring alternative clustering and machine learning techniques. Such extensions would allow for a more refined behavioural segmentation and a deeper understanding of pedestrian–vehicle interactions under different urban conditions.

Author Contributions: Conceptualization, M.M. and A.P.; methodology, M.M.; software, C.M.; validation, F.L.T. and A.P.; formal analysis, M.M. and C.M.; investigation, C.M.; resources, M.M.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, F.L.T. and A.M.; visualization, A.M.; supervision, M.M.; project administration, M.M.; funding acquisition, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by University of Florence—SWALK (Safe CrossWALK) project—Bando di ateneo progetti competitivi per Ricercatori a Tempo determinato (RTD) dell'Università di Firenze, anno 2020–2021.

Institutional Review Board Statement: Ethical review and approval were waived for this study because, for this type of study, ethics committee approval was not mandatory, as the experimentation was not of a medical nature.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. ISTAT. *Incidenti Stradali in Italia—Anno 2024*; Istituto Nazionale di Statistica: Roma, Italy, 2025.
2. European Commission. *Road Safety: Facts and Figures—Roads Inside Urban Areas*; Directorate-General for Mobility and Transport: Brussels, Belgium, 2022. Available online: https://road-safety.transport.ec.europa.eu/system/files/2022-07/ff_roads_inside_urban_areas_20220707.pdf (accessed on 15 December 2025).
3. ISTAT. *Incidenti Stradali in Italia—Anno 2023*; Istituto Nazionale di Statistica: Roma, Italy, 2024.
4. Rosén, E.; Sander, U. Pedestrian fatality risk as a function of car impact speed. *Accid. Anal. Prev.* **2009**, *41*, 536–542. [\[CrossRef\]](#)
5. Tefft, B.C. Impact speed and a pedestrian's risk of severe injury or death. *Accid. Anal. Prev.* **2013**, *50*, 871–878. [\[CrossRef\]](#)
6. Abdel-Aty, M.; Al-Omari, M.E. Investigation of Driving Behavior at Alternative Intersection Designs and Safety Improvement: A Driver Simulator Study (No. UCF-2-Y3). 2021. Available online: http://safersim.nads-sc.uiowa.edu/final_reports/UCF-2-Y3_Final%20Report.pdf (accessed on 17 March 2026).
7. Sucha, M.; Dostal, D.; Risser, R. Pedestrian-driver communication and decision strategies at marked crossings. *Accid. Anal. Prev.* **2017**, *102*, 41–50. [\[CrossRef\]](#)
8. Sheykhfard, A.; Haghighi, F.; Papadimitriou, E.; Van Gelder, P. Analysis of the occurrence and severity of vehicle-pedestrian conflicts in marked and unmarked crosswalks through naturalistic driving study. *Transp. Res. Part F Traffic Psychol. Behav.* **2021**, *76*, 178–192. [\[CrossRef\]](#)
9. Branzi, V.; Meocci, M.; Domenichini, L.; La Torre, F. Drivers' performance in response to engineering treatments at pedestrian crossings. *Adv. Transp. Stud.* **2018**, *1*, 55.
10. La Torre, F.; Meocci, M.; Domenichini, L.; Branzi, V.; Tanzi, N.; Paliotto, A. Development of an accident prediction model for Italian freeways. *Accid. Anal. Prev.* **2019**, *124*, 1–11. [\[CrossRef\]](#)
11. Meocci, M.; Terrosi, A.; Paliotto, A.; Arrighi, R.; Petrizzo, I. Drivers' performance assessment approaching pedestrian crossings through the analysis of the speed and perceptive data recorded during on-field tests. *Heliyon* **2024**, *10*, e24249. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Vedagiri, P.; Kadali, B.R. Evaluation of Pedestrian–Vehicle Conflict Severity at Unprotected Midblock Crosswalks in India. *Transp. Res. Rec. J. Transp. Res. Board* **2016**, *2581*, 48–56. [\[CrossRef\]](#)
13. Papadimitriou, E.; Lassarre, S.; Yannis, G. Introducing human factors in pedestrian crossing behaviour models. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *36*, 69–82. [\[CrossRef\]](#)
14. Feldstein, I.T.; Dyszak, G.N. Road crossing decisions in real and virtual environments: A comparative study on simulator validity. *Accid. Anal. Prev.* **2020**, *137*, 105356. [\[CrossRef\]](#)
15. Guan, F.; Whitley, T.; Xu, H.; Wang, Z.; Chen, Z.; Hui, T.; Tian, Y. Vehicle-pedestrian interaction analysis for evaluating pedestrian crossing safety at uncontrolled crosswalks—A geospatial approach using multimodal all-traffic trajectories. *J. Saf. Res.* **2024**, *91*, 326–341. [\[CrossRef\]](#)
16. Poó, F.M.; Ledesma, R.D.; Trujillo, R. Pedestrian crossing behavior, an observational study in the city of Ushuaia, Argentina. *Traffic Inj. Prev.* **2018**, *19*, 305–310. [\[CrossRef\]](#)
17. Kashani, A.T.; Shariat-Mohaymany, A.; Ranjbari, A. Analysis of factors associated with traffic injury severity on rural roads in Iran. *J. Inj. Violence Res.* **2012**, *4*, 36. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Meocci, M.; Branzi, V.; Martini, G.; Arrighi, R.; Petrizzo, I. A Predictive Pedestrian Crash Model Based on Artificial Intelligence Techniques. *Appl. Sci.* **2021**, *11*, 11364. [\[CrossRef\]](#)
19. Mignot, D. Topical collection on human factors and safety. *Eur. Transp. Res. Rev.* **2017**, *9*, 41. [\[CrossRef\]](#)
20. Bunn, F.; Collier, T.; Frost, C.; Ker, K.; Roberts, I.; Wentz, R. Traffic calming for the prevention of road traffic injuries: Systematic review and meta-analysis. *Inj. Prev.* **2003**, *9*, 200–204. [\[CrossRef\]](#)
21. Fitzpatrick, K.; Turner, S.M.; Brewer, M.; Carlson, P.J.; Ullman, B.; Trout, N.D.; Park, E.S.; Whitacre, J.; Lalani, N.; Lord, D. *Improving Pedestrian Safety at Unsignalized Crossings*; Transit Cooperative Research Program (TCRP) Report 112 or NCHRP Report 562; Transportation Research Board: Washington, DC, USA, 2006.
22. Marshall, W.E.; Ferencak, N.N. Why cities with high bicycling rates are safer for all road users. *J. Transp. Health* **2019**, *13*, 100548. [\[CrossRef\]](#)
23. Kim, K.; Pant, P.; Yamashita, E. Measuring influence of accessibility on accident severity with structural equation modeling. *Transp. Res. Rec.* **2011**, *2236*, 1–10. [\[CrossRef\]](#)
24. Chen, P.; Zeng, W.; Yu, G.; Wang, Y. Surrogate Safety Analysis of Pedestrian-Vehicle Conflict at Intersections Using Unmanned Aerial Vehicle Videos. *J. Adv. Transp.* **2017**, *2017*, 5202150. [\[CrossRef\]](#)
25. Singh, H.; Kathuria, A. Analyzing driver behavior under naturalistic driving conditions: A review. *Accid. Anal. Prev.* **2021**, *150*, 105908. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Fu, T.; Miranda-Moreno, L.; Saunier, N. A novel framework to evaluate pedestrian safety at non-signalized locations. *Accid Anal Prev.* **2018**, *111*, 23–33. [\[CrossRef\]](#) [\[PubMed\]](#)

27. Santos, D.; Saias, J.; Quaresma, P.; Nogueira, V.B. Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. *Computers* **2021**, *10*, 157. [\[CrossRef\]](#)
28. Nassereddine, H. Modeling vehicle-pedestrian interactions at unsignalized intersections. *J. Transp. Saf. Secur.* **2025**, *17*, 664–682. [\[CrossRef\]](#)
29. Noh, B.; No, W.; Lee, J.; Lee, D. Vision-based potential pedestrian risk analysis on unsignalized crosswalk using data mining techniques. *Appl. Sci.* **2020**, *10*, 1057. [\[CrossRef\]](#)
30. Li, H.; Hu, H.; Zhang, Z.; Ren, G.; Liu, X. Impacts of enforcement cameras on pedestrians' risk perception and drivers' behaviors at non-signalized crosswalks. *J. Saf. Res.* **2022**, *81*, 313–325. [\[CrossRef\]](#)
31. Rangam, H.; Sivasankaran, S.K.; Balasubramanian, V. Generation of nighttime pedestrian fatal precrash scenarios at junctions in Tamil Nadu, India, using cluster correspondence analysis. *Traffic Inj. Prev.* **2024**, *25*, 870–878. [\[CrossRef\]](#)
32. Bajada, T.; Attard, M. A typological and spatial analysis of pedestrian fatalities and injuries in Malta. *Res. Transp. Econ.* **2021**, *86*, 101023. [\[CrossRef\]](#)
33. Bayiga Zziwa, E.; Mutto, M.; Guwatudde, D. Cluster analysis of the spatial distribution of pedestrian deaths and injuries by parishes in Kampala city, Uganda. *Int. J. Inj. Control. Saf. Promot.* **2023**, *30*, 419–427. [\[CrossRef\]](#)
34. Lee, C.; Abdel-Aty, M. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. *Accid. Anal. Prev.* **2005**, *37*, 775–786. [\[CrossRef\]](#)
35. Sun, M.; Sun, X.; Shan, D. Pedestrian crash analysis with latent class clustering method. *Accid. Anal. Prev.* **2019**, *124*, 50–57. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [\[CrossRef\]](#)
37. Huang, Z. Clustering large data sets with mixed numeric and categorical values. In Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference (PAKDD), Singapore, 23–24 February 1997; pp. 21–34.
38. Domenichini, L.; La Torre, F.; Tartaglia, V.; Branzi, V.; Fanfani, F. Safety improvements in Urban Areas based on Human Factors principles. A case study: Via Pistoiese. In Proceedings of the BAST–Ageing and Safe Mobility Conference, Gladbach, Germany, 27–28 November 2014; pp. 27–28.
39. Ministry of Infrastructures and Transports. *Norme Funzionali E Geometriche Per La Costruzione Delle Strade*; Decreto Ministeriale n.6792 del 05/11/2001; Ministry of Infrastructures and Transports: Roma, Italy, 2001.
40. Domenichini, L.; Branzi, V.; Meocci, M. Virtual testing of speed reduction schemes on urban collector roads. *Accid. Anal. Prev.* **2018**, *110*, 38–51. [\[CrossRef\]](#)
41. Ministry of Infrastructures and Transports. *Decreto Legislativo 30 Aprile 1992, n. 285—Nuovo Codice Della Strada*; Gazzetta Ufficiale Della Repubblica Italiana: Roma, Italy, 1992.
42. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2001**, *63*, 411–423. [\[CrossRef\]](#)
43. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [\[CrossRef\]](#)
44. Bella, F.; Silvestri, M. Effects of safety measures on driver's speed behavior at pedestrian crossings. *Accid. Anal. Prev.* **2015**, *83*, 111–124. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Marshall, W.E.; Garrick, N.W. Does street network design affect traffic safety? *Accid. Anal. Prev.* **2011**, *43*, 769–781. [\[CrossRef\]](#)
46. Gitelman, V.; Pesahov, F.; Carmel, R. Speed perception by drivers as dependent on urban street design; a case-study. *Trans. Transp. Sci.* **2020**, *11*, 5–18. [\[CrossRef\]](#)
47. Wang, X.; Zhou, Q.; Quddus, M.; Fan, T.; Fang, S.E. Speed, speed variation and crash relationships for urban arterials. *Accid. Anal. Prev.* **2018**, *113*, 236–243. [\[CrossRef\]](#)
48. Fitzpatrick, K.; Brewer, M.A.; Turner, S. Another Look at Pedestrian Walking Speed. *Transp. Res. Rec. J. Transp. Res. Board* **2006**, *1982*, 21–29. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.