



Quantitative Risk Assessment of Pedestrians – Train Accidents in the Colombo Division Railway Network

R.M.B.I. Rathnayake ^{a*}, S.G.J. Senarathne ^a

*^aDepartment of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya,
20400, Sri Lanka*

^{a*}Email: binushirathnayake@gmail.com

^aEmail: jagath.senarathne@sci.pdn.ac.lk

Abstract

Rail transportation stands as a critical mode for both passenger and freight movement in Sri Lanka, acting as a major catalyst for economic activity while linking different regions across the country. Train accidents cause tragic loss of life, injuries, and lead to noteworthy financial and social disturbances. Unfortunately, there has been a lack of comprehensive studies on railway accidents in Sri Lanka, likely due to restricted data accessibility. Therefore, advancing research in this area and gaining a deeper understanding of the underlying causes and patterns of train accidents is essential. The main objective of this study is to quantify pedestrian–train accident risk across the Colombo Division railway network and to determine the risk status of each line segment of the network. The Colombo division comprises 5 railway lines and includes 194 stations. Through segmentation of each railway line based on stations, a numerical assessment of risk is conducted for each segment. Statistical techniques, including bootstrapping and K-means clustering, are employed to classify the calculated risk scores into three categories: high, medium, and low. The reliability of these risk categories is validated using Kruskal-Wallis rank sum test and Dunn’s test with Bonferroni correction. The results indicate that, out of the total 194 segments, 165 are classified as low risk, 10 as medium risk, and 19 as high risk segments. Among these, the Coastal line has the highest count of high risk segments. Although, the Northern line and Kelani Valley line exhibit no high risk segments, Puttalam line has 1, while the Main line shows 7 high risk segments. These findings offer valuable insights to relevant authorities for identifying railway line segments that are highly vulnerable to accidents and for guiding government efforts in implementing safety measures for railway operations in Sri Lanka. Future research could enhance this study by including analyses of train accidents involving vehicles, animals, and other properties.

Keywords: Bonferroni Correction; Bootstrapping; Dunn's Test; K-means Clustering; Kruskal-Wallis Rank Sum Test

1. Introduction

Rail transportation plays a vital role in Sri Lanka's connectivity and economic development, serving as a key mode for both passenger travel and freight movement. However, railway safety remains a critical concern, as train related accidents can lead to severe human, social, and economic consequences. Ensuring the safety of passengers, railway personnel, and the public requires proactive measures, among which identifying high risk railway lines is a fundamental step.

Despite the essential nature of this issue, existing literature on railway accidents in Sri Lanka is relatively limited. Notably, [1] examined various contributory factors, such as noise levels, visibility, and track geometry, affecting accidents involving individuals, along with demographic correlations like age and gender of victims. Similarly, [2] investigated fatalities on railways from a medical perspective, analysing autopsy data over a decade to understand injury types and causes of death. While these studies provide valuable insights into accident characteristics and victim profiles, they do not offer a comprehensive risk assessment framework for railway lines.

In contrast, studies conducted in other countries have adopted more structured approaches to railway risk analysis. For instance, [3] developed a railway crossing safety assessment model using Stochastic Petri Nets, Time Nets, and GIS. [4] applied fuzzy reasoning and fault tree analysis to create a comprehensive model, while [5] used regression based models to quantify risk at station platforms. These methods emphasise model based assessment using large datasets and sophisticated techniques, often requiring extensive data that may not always be available in developing contexts.

To address data limitations, simpler yet effective approaches like [6] equation, calculating risk as the product of likelihood and severity, have proven valuable. Additionally, clustering techniques such as K-means [7], latent class clustering [8], and text clustering [9] have been applied in international studies to detect patterns, classify accident types, and even predict future incidents.

Given the lack of similar comprehensive or predictive risk analysis in Sri Lanka, this study aims to fill that gap by focusing on risk assessment across five key railway lines in the Colombo division: the Kelani Valley line, Northern line, Puttalam line, Main line, and Coastal line. This study employs both a quantitative risk scoring method (based on likelihood and severity) and clustering techniques to analyse accident data and identify high risk locations. It also determines appropriate intervals for risk categories and evaluates the presence of significant differences in accidents, injuries, and deaths across these categories. Unlike previous studies, this research not only analyses accident characteristics but also produces actionable insights for prioritising safety interventions.

By contextualising global methodologies within the specific data environment and operational challenges of Sri Lanka, this study contributes a novel and practical approach to railway risk management. The novelty of this research lies in proposing a risk assessment approach using bootstrapped K-means clustering to evaluate risk in railway lines. Such a methodological approach has not previously been applied in this context. We also highlighted that this is the first study in Sri Lanka to assess railway line risk using a rigorous statistical approach. The findings

have the potential to inform national transportation policy, enhance passenger and crew safety, reduce economic disruptions, and serve as a benchmark for continuous safety improvements in the country's railway system.

2. Risk assessment of railway lines

Risk assessment in rail transport systems involves a comprehensive approach that focuses on thoroughly assessing potential risks at critical intersections and corridors. From a statistical perspective, risk refers to the likelihood or the probability of an undesirable event occurring and the negative consequences that can arise in a given situation. [6] provides the following insightful formula for a more precise definition of risk.

$$\text{Risk Score} = \text{Likelihood} \times \text{Severity} \quad (1)$$

In this equation, risk is quantified as the product of the likelihood of an accident occurring and the potential severity of its consequences. Following the study of [6], we defined the likelihood of an accident in a particular line segment of a railway network as the number of accidents occurred within the given line segment divided by the total number of accidents occurred in the entire network over a given time period. Similarly, the severity is calculated as the number of deaths occurred within the line segment divided by the total number of accidents occurred in that segment. This definition was selected to ensure comparability with prior studies, although it is acknowledged that alternative severity measures may provide different perspectives on risk. These segments are defined based on railway stations in a railway line. In particular, if there are three consecutive stations A, B, and C on a railway line, incidents related to the segment around station B (referred to as line segment B) are aggregated from the sections between the midpoints of stations A and B, and stations B and C. Subsequently, the number of accidents, injuries, and deaths is computed for each line segment and used to compute the risk (referred to as the risk score) for each segment separately.

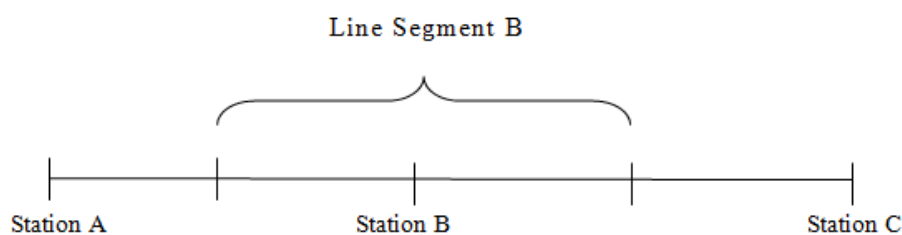


Figure 1: Segmentation of railway lines

For comparing railway line segments based on the proposed risk scores, accident data from 2021 and 2022 were used, primarily sourced from the main record book maintained by the Colombo (Maradana) operating unit. The record book includes information such as the date of the accident, nearest stations, a description of the incident, train number, and details regarding the number of injuries and fatalities. In addition to the operating unit records, supplementary data vital for the study were collected from the Sri Lanka railway headquarters. The supplementary data encompasses the length of nine significant railway lines, the number of stations along those lines, and train schedules.

After Segmentation, 194 railway line segments were identified in the Colombo division. Out of these line

segments, accidents did not occur in 62 segments over the entire two years. Therefore, these segments have been excluded from the analysis, leaving a total of 132 segments under consideration. According to Equations (1), certain segments obtained a zero risk score due to a severity level of zero (indicating no deaths). To ensure meaningful analysis, these zero values have been adjusted to a small value, specifically 0.0001. This substitution was applied solely to obtain meaningful risk values, as there were a few reported accidents in those line segments with zero fatalities.

Following the computation of risk scores, the aim was to classify them into three meaningful categories: high, medium, and low risk, such that policymakers and other respective authorities can use them to make necessary decisions to mitigate the risk of railway lines in the Colombo division. To determine the upper and lower bounds of these categories, a clustering technique combined with bootstrap sampling was applied. K-means clustering was selected as the clustering algorithm because the number of clusters was known beforehand. Also, it provides quick convergence and can be easily adapted or re-run if the dataset is updated. Unlike hierarchical methods, K-means allows reassigning data points during iterations, leading to more flexible and potentially more accurate clustering results.

For the clustering, a new dataset with four variables: accidents, injuries, and deaths occurred during the given time period and the risk score of each line segment is considered. This dataset consists of records for 132 line segments as mentioned above. A bootstrapping technique is used to generate 100 samples from the dataset, each with a sample size of 132. Subsequently, each sample is clustered using the K-means clustering algorithm into three clusters, based on the variables of accidents, injuries, and deaths, excluding the variable risk score. After the clustering process, the mean risk score for each cluster is determined, and the clusters are then labeled according to their mean risk scores as low risk, medium risk and high risk categories. This procedure is repeated for all 100 samples. Subsequently, considering the mean risk scores of 100 samples, 95% confidence intervals for three clusters are calculated. These confidence intervals are considered as the risk intervals for three risk categories.

3. Results and Discussion

As shown in Figure 2, the variable risk score spans a range from 0.0001 to 0.0161, encompassing very small values. The mean risk score is 0.0029, with a standard deviation of 0.0034, indicating small values with moderate variability. The distribution of risk score appears to be right skewed, indicating that most line segments have lower risk values, while relatively fewer segments exhibit higher risk values.

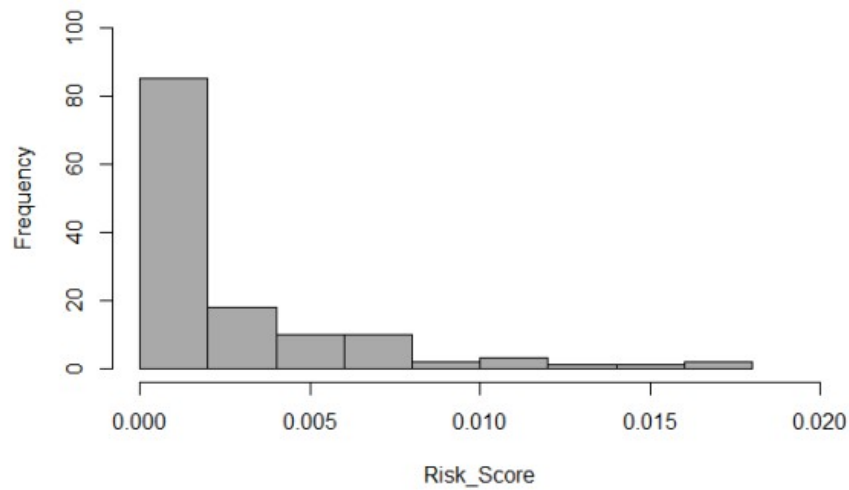


Figure 2: Distribution of risk score

The Table 1 outlines the lower and upper bounds of the three risk categories discussed in Section 2. As the expected range of risk values spans from zero to infinity, the lower bound for the low risk category was set to zero (with the actual value being 0.0027), while the upper bound for the high risk category was considered as infinity (with the actual value being 0.0086).

Table 1: Low and upper bounds of risk categories

Risk category	Lower bound	Upper bound
Low	0.0000	0.0041
Medium	0.0045	0.0056
High	0.0071	$+\infty$

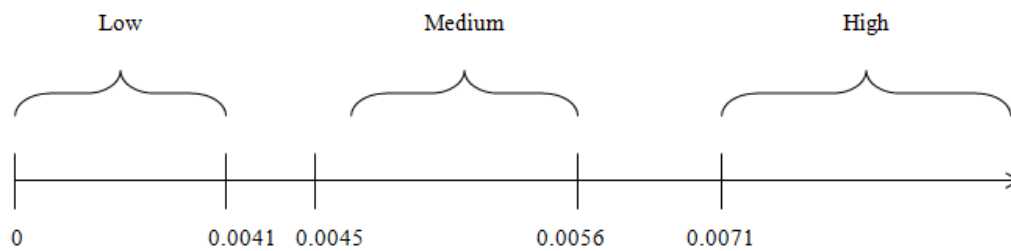


Figure 3: Risk category intervals

As shown in Figure 3, there is a noticeable gap between the upper limit of the low risk category and the lower limit of the medium risk category. As a result, it becomes challenging to determine whether a risk score that falls within this range should be classified as low or medium. Similar uncertainty exists between the upper limit of the medium risk category and the lower limit of the high risk category, making it difficult to definitively assign some line segments to a specific risk category.

According to the classification described above, risk categories were identified for 132 railway line segments. Initially, 62 segments were excluded from the analysis because no accidents were reported in those segments over the two year period. These segments were therefore directly classified as low risk. After classifying all 194 segments, 165 segments were identified as low risk, 10 as medium risk, and 19 as high risk areas.

Figure 4 visualises the geographic locations of stations corresponding to line segments categorised as low, medium, or high risk. As can be seen, the Coastal line (Colombo Fort-Beliatta) has the highest number of high risk segments, followed by the Main line (Colombo Fort-Polgahawela). This is due to the large number of accidents that occurred on those lines during 2021 and 2022. Over the two year period, the Coastal line recorded 259 accidents (45.84%), while the Main line reported 201 accidents (35.58%). In terms of deaths, the Coastal line recorded 83 deaths (41.50%), and the Main line reported 78 deaths (39.00%).



Figure 4: Geographical distribution of stations (representing line segments) categorised by risk

3.1. Assessing the impact of each factor on risk assessment

As a part of the study, the factors including accidents, deaths, and injuries were analysed to investigate the impact of each factor on the risk category. This assessment helped to determine how individual factors contribute to the differentiation between different risk categories. To examine this, hypothesis testing was conducted. Specifically, to assess the presence of significant differences in accidents, deaths, and injuries across different risk categories, Kruskal-Wallis rank sum test is employed. Using this test, the following hypothesis was tested.

- H_0 : There is no difference in the median accidents/injuries/deaths of all three risk categories.
- H_1 : There is a significant difference in the median accidents/injuries/deaths across at least two risk categories.

The Kruskal-Wallis test revealed statistically significant differences across the risk categories for all three variables: accidents (p-value = 0.0000), injuries (p-value = 0.0004), and deaths (p-value = 0.0000).

As the Kruskal-Wallis test was rejected, a Post-Hoc test (Dunn's test with Bonferroni correction) was used for the pairwise comparisons of risk categories. According to the Post-Hoc test results, when considering the accidents and deaths variables, the low risk category exhibited a statistically significant difference compared to the medium risk category with the p value of 0.0000. However, no significant difference was observed between the high risk category and the medium risk category. On the other hand, for the injuries variable, the low risk category differed significantly from the high risk category with the p value of 0.0002. Nevertheless, no significant differences were observed between the low-medium risk categories (p value = 0.1676) and the high-medium risk categories (p value = 0.4298).

4. Conclusion

This research evaluated the risk associated with railway lines in the Colombo division by analysing train accidents involving pedestrians. A quantitative approach was applied using a risk score defined as the product of likelihood and severity. A railway line segment which consists of a station was considered as unit area for assessing the risk of the entire network, and risk category (low, medium, high) of the line segment was determined using K-means clustering, bootstrapping, and confidence intervals.

According to the analysis 85.05% railway line segments were categorised as low risk, 5.15% as medium risk, and 9.79% as high risk. High risk segments were primarily located along the Coastal line, while the Northern and Kelani Valley lines contained no high risk sections, likely due to lower train frequency. These high-risk segments are often situated near densely populated urban centers, contain informal pedestrian crossings, or experience high pedestrian and train traffic density, conditions that collectively elevate accident likelihood. Also, statistical analysis revealed significant differences in accident and fatality counts across certain risk categories. However, no significant differences were observed in some categories, likely due to the small range of risk values produced by the method.

This research provides a practical and data driven approach to railway risk assessment, filling a gap in existing safety analyses in Sri Lanka. By combining quantitative risk scoring with clustering and bootstrapping techniques, it offers a reliable method for identifying high risk segments. The use of station-based segmentation allows for localised insights, and the findings support targeted safety improvements. For example, this study provides support for decision-makers and relevant authorities to prioritise safety measures in high-risk segments, including increased surveillance, installation of warning systems, and targeted public awareness campaigns. The approach is simple, reproducible, and adaptable, making it valuable for broader applications in rail safety planning. With knowledge of the likelihood and severity, one can easily determine the corresponding risk categories.

A notable limitation encountered in the calculation of risk scores was the generation of very small values, which posed challenges in reporting. To address this, future work could explore expressing these values as ratios, offering a more interpretable and practical representation. Additionally, there is potential bias due to the absence of

exposure normalisation in the current analysis; future studies should aim to integrate factors such as train frequency, pedestrian volume, and segment length when such data become available. A forward looking aspect of this research involves the creation of a risk assessment model. This model could leverage the identified risk categories and pertinent railway incident factors to interpret the aspects of the railway industry that may need modification. As a conclusion, this research extension has the potential to contribute significantly to the development of rail safety in countries, providing insights and recommendations for improving overall railway industry safety.

References

- [1] A. Asoka, "A Review of trespassing accident and suicide on railway track for pedestrian safety," 2018.
- [2] P. Paranitharan and W. Perera, "Deaths on rail roads: a study from Colombo North teaching hospital, Sri Lanka," *Sri Lanka Journal of Forensic Medicine, Science & Law*, vol. 9, no. 2, pp. 10-20, 2018.
- [3] S. Ishak, W. Yue and S. Somenahalli, "The methodology development of railway level crossing safety systems – South Australia case study," in *Computers in Railways XI*, 2008.
- [4] Y. Chen, "Improving railway safety risk assessment study," 2013.
- [5] S. Terabe, T. Kato , H. Yaginuma , N. Kang and K. Tanaka, "Risk assessment model for railway passengers on a crowded platform.," *Transportation Research Record*, vol. 2673, no. 1, pp. 524-531, 2019.
- [6] A. Alikhani, "Proposed risk control in strategic management," *International Journal of Energy and Environment*, vol. 3, no. 3, pp. 112-121, 2009.
- [7] L. Shi, X. Yang, X. Chang, J. Wu and H. Sun, "An improved density peaks clustering algorithm based on k nearest neighbors and turning point for evaluating the severity of railway accidents.," *Reliability Engineering & System Safety*, vol. 233, 2023.
- [8] Y. Kang and A. Khattak, "Cluster-based approach to analyzing crash injury severity at highway–rail grade crossings," *Transportation Research Record*, vol. 2608, no. 1, pp. 58-69, 2017.
- [9] T. Williams, J. Betak and B. Findley, "Text mining analysis of railroad accident investigation reports," in *ASME/IEEE Joint Rail Conference*, 2016.