Proceedings of the 2024 ASME Joint Rail Conference
JRC2024
May 13–15, 2024, Columbia, SC, USA

JRC2024-130028

SPECTRAL CLUSTERING IN RAILWAY CROSSING ACCIDENTS ANALYSIS

Ethan Villalobos[1,†], Hector Lugo III[1,†], Biqian Cheng[2], Miguel Gutierrez[2], Constantine Tarawneh[1], Ping Xu[1], Jia Chen[2], Evangelos E. Papalexakis[2,*]

[1]University of Texas Rio Grande Valley, Edinburg, TX
[2]University of California Riverside, Riverside, CA

ABSTRACT

This study employs graph mining and spectral clustering to analyze patterns in railway crossing accidents, utilizing a comprehensive dataset from the US Department of Transportation. By constructing a graph of implicit relationships between railway companies based on shared accident localities, we apply spectral clustering to identify distinct clusters of companies with similar accident patterns. This offers nuanced insight into the underlying structure of these incidents. Our results indicate that "Highway User Position" and "Equipment Involved" play pivotal roles in accident clustering, while temporal elements like "Date" and "Time" exert a diminished impact. This research not only sheds light on potential accident causation factors but also sets the stage for subsequent predictive safety analyses. It aims to serve as a cornerstone for future studies that aspire to leverage advanced data-driven techniques for improving railway crossing safety protocols.

Keywords: Graph mining, spectral clustering, similarity matrix.

1. INTRODUCTION

Railway transportation remains a cornerstone of modern infrastructure, facilitating the movement of goods and ensuring connectivity across distant locations. As such, its influence extends far beyond conveyance; railways are a testament to the evolution of engineering, economic progression, and a reflection of societal growth. However, as with any large-scale system that interfaces directly with the public, safety concerns arise, especially at intersections where railways cross paths with other modes of transportation. These intersections, commonly known as railway crossings, have historically been sites of concern due to the potential for financially disastrous and injury-inducing accidents.

Accidents at railway crossings often result in severe consequences, affecting not only the immediate stakeholders—train operators, vehicle drivers, and passengers, but also the wider community and the reputation of the railway industry itself. The significance of these incidents transcends mere statistical tallies; it extends to an in-depth comprehension of the complex interplay of contributory elements that precipitate an accident. In this context, public accident reports emerge as invaluable reservoir of data. They skillfully capture the multifaceted nature of railway accidents, encompassing environmental conditions, human error, and the technical specifics of the involved vehicles. Notably, reports accessible from reputable sources like the US Department of Transportation (USDOT) and Federal Railroad Administration (FRA) provide essential and comprehensive data crucial for our analytical endeavors [1].

When handling complex and voluminous data, traditional analytical methods may fall short in capturing nuanced relationships and patterns. In this paper, our primary focus thus is on unearthing these patterns and presenting a structured methodology that combines vast detailed data from a public dataset with the powers of more advanced techniques such as graph mining [2], spectral clustering [3] and factor-driven element categorization. The objective is to not only elucidate the current state of railway crossing accidents but also to uncover broader patterns and clusters to chart their evolution over time and identify potential areas for intervention. By the end, we aim to provide those involved with actionable insights that can inform safety protocols, policy decisions, and future research directions in the realm of railway safety.

To achieve this objective, we develop a robust framework for understanding complex relationships behind the railway accident data that combines the strengths of network theory and data science, enabling a deeper exploration into the intricate structures and interconnections that standard analyses might overlook. Specifically, we propose to first construct a graph that represents the implicit relationships between various railway companies, based on shared accident localities and other parameters. With the constructed graph, we gain a unique perspective on the intricacies of these accidents. We then utilize three techniques to

[†]Joint first authors
[*]Corresponding author
Documentation for `asmeconf.cls`: Version 1.37, February 13, 2025.

perform dimensionality reduction before clustering the data in fewer dimensions, i.e., 1) apply the Radial Basis Function (RBF) kernel to the constructed graph to study its intrinsic structure [3]; 2) apply t-Distributed Stochastic Neighbor Embedding (t-SNE) to the RBF-transformed graph to maintain the local structure of the data in a lower dimension [4]; 3) apply Principal Component Analysis (PCA) to project the data onto a lower-dimensional space that captures the most significant structure for clustering [5]. To determine the optimal number of clusters, we perform grid search and evaluate the clustering quality using the silhouette score [6]. The final clustering is done by KMeans method [7]. In parallel with our clustering steps, we further incorporate additional factors into our dataset by quantitatively assessing the relationship between clusters based on these factors using a customized Jaccard similarity function [8]. The concatenated dataset that merges the results of the cluster analysis with the original railway accident data, provides a multi-dimensional perspective of the problem. This comprehensive approach, combining sophisticated graph-based analysis with detailed attribute exploration, enables us to uncover more profound insights into the patterns and factors influencing railway crossing accidents.

While our analytical approach provides valuable insights and enhances our ability to predict future incidents, it is essential to note that our current methodology does not engage in causal modeling per se. Instead, our analysis aids in the formulation of hypotheses regarding the factors influencing railway accidents. These hypotheses are not conclusive assertions of causality but starting points for further investigation and rigorous testing. Significant contributions of our study to railway safety include:

- **Identification of key factors:** Developing from graph mining and factor-driven element categorization, our analysis underscores the significant roles of "Highway User Position" and "Equipment Involved" in accident clustering. This finding steers the focus toward critical areas needing enhanced safety regulations and more targeted preventive measures.
- **Re-evaluating the role of temporal factors:** Challenging traditional beliefs, our study finds that "Date" and "Time" have a limited impact on accident causation. This revelation emphasizes the necessity to prioritize human and equipment-related factors over temporal aspects in accident analysis.
- **Impact of environmental conditions:** The study highlights the moderate yet notable influence of "Visibility" and "Weather Conditions" on accident occurrences. This aspect is particularly crucial in regions with extreme weather, suggesting a need for tailored safety strategies under varying environmental conditions.
- **Directions for future research:** The hypotheses and methodologies developed in this study pave the way for future exploration, particularly in employing machine learning techniques in the realm of predictive analysis through the implementation of a Kernel Ridge Regression model. These methods promise to enhance the predictive capabilities and facilitate the development of proactive safety protocols in railway crossings.

2. PROPOSED FRAMEWORK

This section first introduces the dataset studied in this paper and then provides details of the techniques utilized in our proposed framework.

2.1 Data acquisition and preprocessing

Our study leverages a comprehensive, publicly available data-set from Kaggle, titled "US Highway Rail Road Crossing Accident" [1]. This dataset, compiled by the US Department of Transportation, provides extensive details on railway crossing incidents throughout the United States from January 1, 1975, to February 28, 2021. It includes a wealth of information such as geographic locations, time frames, types of crossings, accident specifics, vehicle types, and highway user data, amongst others. Given the depth and breadth of this data-set, our preprocessing involves several crucial steps to ensure the data was primed for effective analysis. Our primary goal during preprocessing is to organize and consolidate the data, focusing on key attributes relevant to our study.

To do this, we utilize the Python library pandas [9] for data manipulation and aggregation, grouping by relevant categories such as "Railroad Code", "Incident Year", and "State Name", etc. This approach allows us to create a structured framework for our subsequent graph-based analysis with a focus on temporal and geographical distribution patterns. We also perform essential data cleaning tasks to ensure accuracy and consistency across the dataset. By organizing the data in this manner, we are able to identify and analyze trends and relationships within the context of both time and location.

2.2 Graph construction

The core of our methodology lies in the construction of a graph representing the implicit relationships between various railway companies, which allows us to analyze and visualize the complex network of relationships and interactions among the railway companies, providing insights into patterns and trends in railway crossing accidents. Specifically, we construct an undirected graph $G = (V, E)$, where each vertex $i \in V$ corresponds to a railway company, and each edge $e(i, j) \in E$ represents a shared accident locality between companies i and j. The weight of each edge, denoted as $e(i, j)$, is assigned based on the total number of accidents that occurred between the two companies, which is mathematically expressed as

$$e(i, j) = \text{Accident}_{\text{shared}}(i, j). \tag{1}$$

The granularity of the data per accident is carefully considered, particularly the geographic specifics such as the county or state of occurrence. This level of detail in our approach ensures a comprehensive and nuanced portrayal of the network of relationships. It enables us to not just delineate a complex web of inter-company connections but also to discern the variations and patterns that emerge in these relationships across diverse geographical landscapes and temporal scales.

2.3 Spectral clustering

Our analysis progress with the application of spectral clustering, a method suitable for identifying inherent groupings within

complex network data [3]. The initial step in spectral clustering involves translating the complex network of graph G into a form that can be mathematically and computationally managed. This is accomplished by generating the adjacency matrix A that is associated with the graph, where each element a_{ij} indicates the presence (a_{ij} =1) or absence ($a_{ij} = 0$) of an edge between nodes i and $j, \forall i, j \in V$. The generation of the adjacency matrix is a crucial intermediary step as it encapsulates the presence or absence of edges between the nodes in the graph with a two-dimensional array. To incorporate the edge weight in the adjacency matrix, we further define a weighted adjacency matrix W, where each element w_{ij} quantifies the connection weight between nodes i and j. Mathematically, we have

$$w_{ij} = a_{ij} \times e(i, j). \tag{2}$$

Note that the weighted adjacency matrix is symmetric and the its elements not only reflect the presence of connections between different nodes but also the strength of the connections.

We then apply a Radial Basis Function (RBF) kernel to W to enhance the graph's intrinsic structure:

$$W_{\text{RBF}} = \exp(-\gamma \cdot W^2), \tag{3}$$

where γ is a scale parameter that determines the kernel's width.

Subsequently, we apply t-Distributed Stochastic Neighbor Embedding (t-SNE) [4], a nonlinear dimensionality reduction technique to W_{RBF} to maintain the local structure of the data while viewing it in a lower dimension. Principal Component Analysis (PCA) is then employed to project the data onto a lower-dimensional space that captures the most significant structure for clustering [5].

In seeking the optimal number of clusters in the reduced dimension, a grid search strategy was employed, utilizing silhouette scores as a measure of clustering quality [6]:

$$s = \frac{b - a}{\max(a, b)} \tag{4}$$

where a is the mean intra-cluster distance, and b is the mean nearest-cluster distance for each sample.

The final step of spectral clustering involves the normalized Laplacian matrix of the graph, which is given by

$$L = I - D^{-1/2} W_{\text{RBF}} D^{-1/2}, \tag{5}$$

where I is the identity matrix, and D is the degree matrix of the graph G. Upon determine the number of principle components M and the optimal number of clusters N, we then perform KMeans clustering to cluster form N clusters using the first M principal eigenvectors of L.

2.4 Integration of additional factors

In parallel with our clustering analysis, we incorporate additional factors into our data-set. To quantitatively assess the relationship between clusters based on these factors, we employ a customized Jaccard similarity function [8]. The standard Jaccard similarity index is defined as the size of the intersection divided by the size of the union of the sample sets. However, in our case, we adapted this measure to suit the unique characteristics of our data-set, particularly focusing on categorical data like "Highway User Position", "Equipment Involved", etc. The customized Jaccard matrix funtion is thus defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{6}$$

where A and B are sets of categorical attributes for two different clusters. This adaptation is necessary to capture the nuances of our specific data-set, where traditional numerical similarity measures might not be entirely applicable.

The concatenated dataset, which merges the results of the cluster analysis with the original railway accident data, provided a multi-dimensional perspective of the problem. This amalgamation is crucial as it allows us to examine the clusters not only in the context of their graph-based relationships but also in terms of real-world attributes and accident characteristics. By doing so, we could explore deeper into the specific attributes contributing to railway crossing accidents within each cluster, offering a richer and more informative analysis. Specifically, the final step of our proposed framework involves in-depth exploration of the resulting concatenated dataset using similarity matrices, which leads to the formulation of several hypotheses regarding the factors influencing accident clustering.

3. RESULTS

Facilitated by graph mining techniques and similarity matrices, we are able to cluster railway companies into four distinct groups, enabling comprehensive data analysis. These techniques are crucial for unraveling the complex relationships and patterns within our data. For instance, the similarity matrices for each factor, displayed as tables 1 to 8, provide in-depth insights into how various attributes influence the formation of clusters. This approach involves calculating a baseline similarity matrix that includes all factors and then assessing the impact of each factor when it is excluded. Such an analysis allows us to determine the unique contribution of each factor to the clustering of railway accidents.

The values within these matrices are interpreted as follows:

- **Positive Values:** Indicate that the absence of a factor leads to a decrease in similarity between clusters compared to the baseline, signifying a significant role in distinguishing between clusters.
- **Negative Values:** Suggest that removing a factor actually increases the similarity, implying that the factor might be adding noise or redundancy.
- **Zero Values:** Imply no unique contribution to the differentiation between clusters when all factors are considered.

This nuanced understanding of factors, achieved through graph mining and matrix analysis, enhances our interpretation of the most pertinent factors in railway crossing accidents. We examine the distribution of clusters in Figure 1 and analyze the temporal and geographical spread of accidents in Figures 2 and

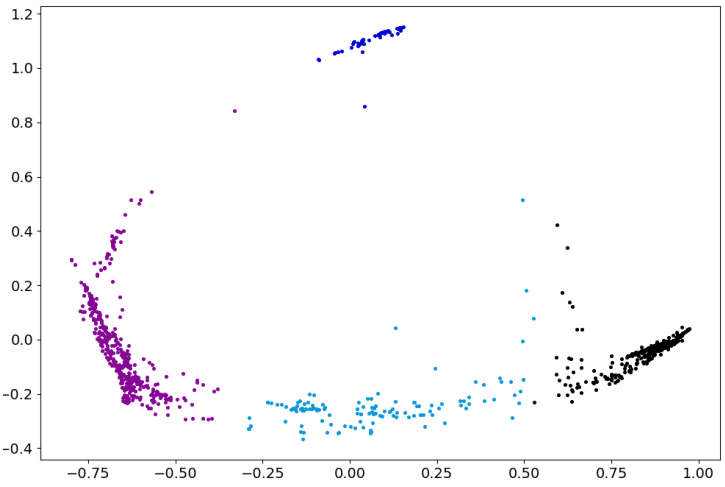

FIGURE 1: Spectral clustering visualization, demonstrating the grouping of railway companies based on shared accident characteristics.

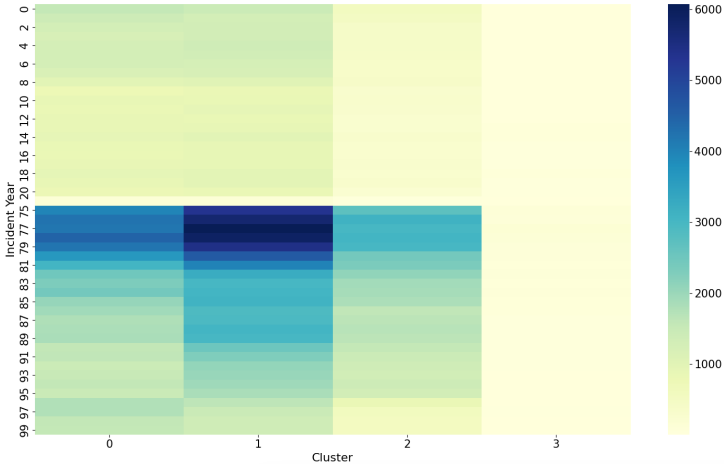

FIGURE 2: Temporal cluster distribution, illustrating the change in accident clustering over time, emphasizing periods of higher incidence and identifying potential cyclic trends within the data.

3. Additionally, tools like seaborn [10] are utilized to create heat maps, effectively illustrating the distribution and influence of various factors across the clusters.

Our analysis of the railway crossing accident data has led to the formulation of several key hypotheses. These hypotheses are centered around the impact of various factors on accident causation and clustering.

- **Hypothesis 1: "Highway User Position" and "Equipment Involved" are important Factors.** Our analysis reveals that both "Highway User Position" and "Equipment Involved" are significant in differentiating accident clusters, as shown in Tables 1 and 6. These factors consistently exhibit high relative importance values, indicating their pivotal roles in accident causation and clustering. The variations in highway user positions, including pedestrians, cyclists, and motorists, coupled with the diverse types of equipment involved (such as trucks, cars, and bicycles), are key elements in understanding the dynamics of railway crossing accidents. These findings suggest a strong correlation between these factors and the occurrence of accidents, highlighting the need for targeted investigations into their specific impact and patterns.

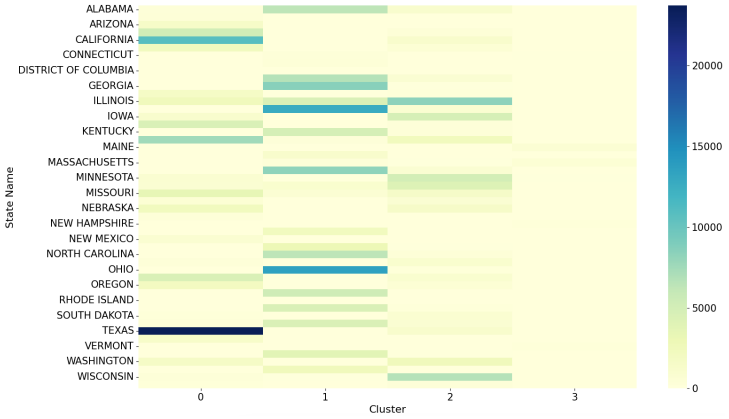

FIGURE 3: Geographic cluster distribution, visualizing the spatial distribution of railway accidents across different regions, revealing areas with higher frequencies and potential geographic risk factors, suggesting the influence of local traffic conditions, infrastructure quality, and regional safety policies on the incidence of accidents.

TABLE 1: Relative importance matrix for Highway User Position.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0.00	0.0029	0.0084	0.0413
Cluster 1	0.0029	0.00	0.0069	0.0401
Cluster 2	0.0084	0.0069	0.00	0.0375
Cluster 3	0.0413	0.0401	0.0375	0.00

- **Hypothesis 2: "Date" and "Time" have limited impact.** The relative importance matrices for "Date" and "Time," as presented in Tables 3 and 4, indicate a generally lower impact of these factors on clustering compared to others. This observation suggests that the temporal aspects of accidents, such as the specific date or time of occurrence, have limited influence on accident causation and clustering. This hypothesis implies that the likelihood and nature of accidents might be more closely related to factors directly associated with the individuals and equipment involved in the incidents, rather than being heavily dependent on the temporal context in which they occur.

- **Hypothesis 3: "Visibility" and "Weather Condition" may play a role.** The relative importance matrices for "Visibility" and "Weather Condition," shown in Tables 5 and 2, reveal a varied influence of these factors on clustering. These factors exhibit a moderate impact on accident causation, as indicated by their mixed positive and negative values in the matrices. Specific visibility conditions (clear, foggy) and diverse weather conditions (sunny, snowy) appear to con-

TABLE 2: Relative importance matrix for Equipment Involved.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0.00	0.0029	0.0084	0.0413
Cluster 1	0.0029	0.00	0.0069	0.0401
Cluster 2	0.0084	0.0069	0.00	0.0375
Cluster 3	0.0413	0.0401	0.0375	0.00

TABLE 3: Relative importance matrix for Date.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0.00	-0.0074	-0.0232	-0.0845
Cluster 1	-0.0074	0.00	-0.0242	-0.0856
Cluster 2	-0.0232	-0.0242	0.00	-0.0873
Cluster 3	-0.0845	-0.0856	-0.0873	0.00

TABLE 4: Relative importance matrix for Time.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0.00	0.0007	0.0042	-0.0349
Cluster 1	0.0007	0.00	0.0042	-0.0357
Cluster 2	0.0042	0.0042	0.00	-0.0385
Cluster 3	-0.0349	-0.0357	-0.0385	0.00

TABLE 5: Relative importance matrix for Visibility.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0.00	0.0029	0.0084	0.0127
Cluster 1	0.0029	0.00	0.0069	0.0115
Cluster 2	0.0084	0.0069	0.00	0.0090
Cluster 3	0.0127	0.0115	0.0090	0.00

tribute to the occurrence of accidents. This finding suggests that visibility and weather conditions might play a significant role in the dynamics of railway crossing accidents. It calls for more detailed exploration, especially in regions with extreme weather, to fully understand how these factors influence accident patterns and risks.

- **Hypothesis 4: "Equipment Type" and "Equipment Struck" are moderately important.** The analysis of the relative importance matrices for "Equipment Type" and "Equipment Struck," as presented in Tables 7 and 8, suggests a moderate influence of these factors on the clustering of accidents. These matrices show mixed positive and negative values, indicating that these factors have a varied impact on differentiating clusters. This implies that while the specific types of equipment involved and the manner in which they are struck in accidents (such as front or rear collisions) play a role in accident scenarios, their influence may not be as dominant as factors like "Highway User Position" or "Weather Condition." Further detailed analysis of these factors is necessary to understand their specific contributions to the dynamics of railway crossing accidents.

These hypotheses, derived from our similarity matrix analysis, lay the groundwork for future research. To validate these hypotheses and draw more definitive conclusions, the application of machine learning techniques, such as classification or regression models, is proposed. These models will enable us to predict accident outcomes based on the identified factors, thereby enhancing our understanding of railway crossing accidents and contributing to more effective preventive strategies.

4. CONCLUSION

This research marks a substantial advancement in unraveling the complexities of railway crossing accidents. Utilizing graph mining techniques, we meticulously constructed a graph that captures the intricate relationships among various railway companies. Our application of spectral clustering effectively revealed distinct clusters within this network, each signifying a group of companies sharing similar accident profiles. This nuanced approach allowed us to delve deeper into the patterns and trends underlying these accidents. Our methodology and findings have significant implications for enhancing railway safety. They provide valuable insights for policymakers, railway companies, and public safety officials, advocating a shift toward data-driven analysis and predictive modeling. However, we recognize certain limitations, particularly the reliance on publicly available accident reports, which might not encompass the entirety of incidents or capture all relevant nuances. Additionally, our high-level approach of analyzing data through similarity matrices, while insightful, may not fully convey the complexity of individual accidents. Such an approach tends to generalize patterns and relationships, which could overlook unique, case-specific factors. Future research could address these gaps by incorporating more comprehensive data-sets, including unreported incidents and near-misses. This expansion would provide a more holistic view of railway crossing accidents. Additionally, integrating more granular, case-by-case analyses could complement the high-level insights from similarity matrices, offering a more nuanced understanding of each incident.

5. ACKNOWLEDGMENTS

This study was made possible by funding support provided by the NSF CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS) under Grant No. 2112650 and the University Transportation Center for Railway Safety (UTCRS) at UTRGV through the USDOT UTC Program under Grant No. 69A3552348340.

REFERENCES

[1] US Department of Transportation. "US Highway Rail Grade Crossing Accident Dataset." URL https://www.kaggle.com/datasets/yogidsba/us-highway-railgrade-crossing-accident?resource=download. Data retrieved from US DOT.

[2] Koutra, Danai and Faloutsos, Christos. *Individual and collective graph mining: principles, algorithms, and applications*. Springer Nature (2022).

[3] Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and computing* Vol. 17 (2007): pp. 395–416.

[4] van der Maaten, L. and Hinton, G. "Visualizing Data using t-SNE." *Journal of Machine Learning Research* Vol. 9 (2008): pp. 2579–2605.

TABLE 6: Relative importance matrix for Weather Condition.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0.00	-0.0080	-0.0136	0.0083
Cluster 1	-0.0080	0.00	-0.0050	0.0163
Cluster 2	-0.0136	-0.0050	0.00	0.0245
Cluster 3	0.0083	0.0163	0.0245	0.00

TABLE 7: Relative importance matrix for Equipment Type.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0.00	0.0029	-0.0011	0.0222
Cluster 1	0.0029	0.00	-0.0026	0.0210
Cluster 2	-0.0011	-0.0026	0.00	0.0273
Cluster 3	0.0222	0.0210	0.0273	0.00

[5] Jolliffe, I. T. *Principal Component Analysis*, 2nd ed. Springer Series in Statistics, Springer, New York (2002).

[6] Shahapure, Ketan Rajshekhar and Nicholas, Charles. "Cluster quality analysis using silhouette score." *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*: pp. 747–748. 2020. IEEE.

[7] MacQueen, J. B. "Some Methods for classification and Analysis of Multivariate Observations." *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: pp. 281–297. 1967.

[8] Ivchenko, GI and Honov, SA. "On the jaccard similarity test." *Journal of Mathematical Sciences* Vol. 88 (1998): pp. 789–794.

[9] McKinney, W. "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference*: pp. 56–61. 2010. URL https://pandas.pydata.org/.

[10] Waskom, M. et al. "Seaborn: Statistical Data Visualization." URL https://seaborn.pydata.org/.

TABLE 8: Relative importance matrix for Equipment Struck.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	0.00	0.0029	0.0084	-0.0064
Cluster 1	0.0029	0.00	0.0069	-0.0076
Cluster 2	0.0084	0.0069	0.00	-0.0101
Cluster 3	-0.0064	-0.0076	-0.0101	0.00