CSE 6242: Data & Visual Analytics Final Report

Team 046: Gabriel Gros, Camille Migozzi, Aubin Rey, Tien Lu, Ting-Yang Kao

1 INTRODUCTION

With the increasing volume of car accident data available, there is an opportunity to harness this information for a deeper understanding of safety on roadways. Our project aims to provide a comprehensive analysis of car accident data to develop insights that can lead to safer routes and smarter decisions on the road. This progress report serves as an overview of the key concepts, methodologies, and innovations that our team has been working on. We will provide a glimpse into our approach and the potential impact of our research. To sum up, our innovations are as follows:

- Real-time Safety Index: Develop a real-time safety index that combines accident data, weather conditions, time of day, traffic density, and other relevant factors to provide up-to-the-minute safety assessments for different routes. This innovation could allow users to make informed decisions based on current road conditions, reducing the risk of accidents.
- Predictive Route Planning: Implement algorithms that predict the probability of accidents on specific routes based on historical data. By providing predictive route planning, users can choose safer paths that minimize the probability of encountering accidents, ultimately enhancing road safety.
- Dynamic Data Visualization: Create interactive and dynamic data visualizations that enable users to explore accident patterns, factors, and safety indices for different regions. Such visualizations can help authorities and individuals gain a deeper understanding of road safety, leading to more informed decisions.
- Community Engagement: Introduce a community engagement feature that allows users to report and share real-time information about road conditions, accidents, or hazards. By crowdsourcing data, this innovation can enhance the accuracy and timeliness of safety information, further improving road safety.

2 LITERATURE SURVEY

Recent research in road safety, particularly in car accident data analysis, has garnered significant attention. This literature review categorizes key studies in this field and outlines their unique approaches, contributions, and limitations.

- **Interactive Visualization of Car Accident Data:** A significant body of research has concentrated on creating interactive visualizations of car accident data. [1, 4, 12, 14, 17]. These visualizations often incorporate interactive elements, filters, and user-friendly features, making it easier for stakeholders to explore accident patterns. While these visualizations enhance data accessibility, they primarily serve as descriptive tools and may not provide the predictive capabilities necessary for proactively planning safer routes.
- **Analyzing Causative Factors of Accidents:** Another group of studies emphasizes in-depth analyses of the factors contributing to accidents. [2, 16]. These factors encompass a wide range of elements, including environmental conditions [9] and human behavior.[11]. While these studies provide valuable insights into the multifaceted causes of accidents, they may not generalize well to different regions for complex interactions between factors. This category is important in understanding the root causes of accidents, but may not directly facilitate route-specific safety predictions.
- **Comparative Analysis:** One research focuses on the comparative analysis of various statistical models to determine their effectiveness in predicting accident patterns likelihood. [8]. The other compares different geospatial regions [6] to highlight diverse risk factors contributing to accidents. These comparative analyses provide insights into the strengths and limitations of different modeling techniques. However, the specific emphasis may differ between the broader modeling approach and region-specific risk assessment.
- **Machine Learning-Based Safety Indices:** Machine learning models have been applied based on car accident data [10, 15]. These models serve a dual purpose: they analyze accident patterns [7] to identify similarities among accidents or predict and suggest route choices[3]. However, these approaches face challenges such as incomplete consideration of relevant factors in the training data due to geographic constraints. While machine learning-based models

are great in providing real-time safety assessments for route planning, addressing data biases and enhancing generalizability remain critical challenges.

There is also an existing navigation system that can allow users to travel more safely. The PASSAGE system [5] considered the historical crime events in Atlanta and provided route predictions by leveraging distance and safety from one point to another. However, in our project, we focus on the accidents instead of the crimes.

3 PROPOSED METHOD

There will be four parts to our processes: Data Preprocessing, Data Interactive Visualization, Safety Index, and Path Planning.

3.1 Data Preprocessing

To calculate the safety index of the roads in Atlanta, we need to match each accident to the nearest road where the accident took place. Since there are a great number of accidents and roads in Atlanta, we created a joint JSON file to store the matching results instead of a super big table. In the JSON file, the 'id's of roads are keys and the corresponding values are dictionaries. A list of matched accidents is stored in each of the dictionaries.

3.2 Data Interactive Visualization

The framework comprises two main components. Initially, Figure 1 and Figure 2 are static graphs that analyze patterns in the dataset, focusing on factors that impact accident frequency. The interactive Figure 3 offers users dynamic exploration by utilizing zoom functionalities. Geographical representation allows focusing on specific map regions for detailed examination.

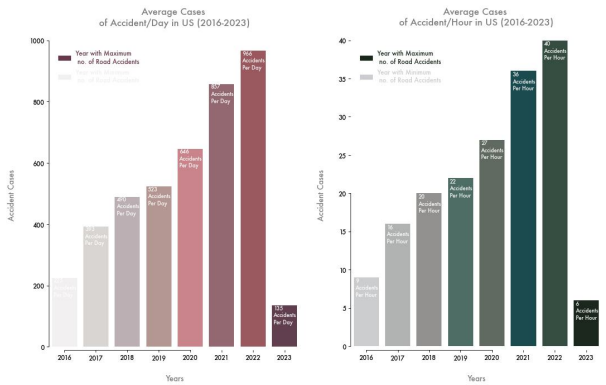

Figure 1: Accident count barchart.

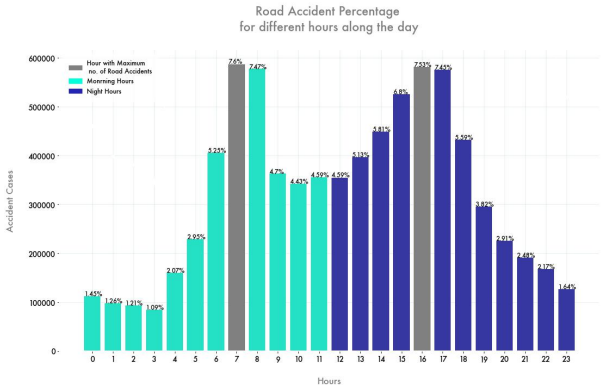

Figure 2: Accidents percentage for different hours in a day.

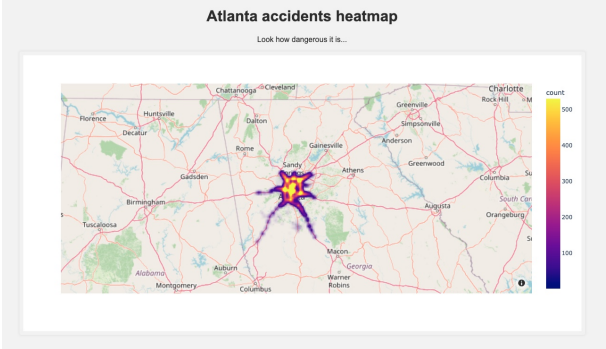

Figure 3: Interactive zoomable heatmap of accidents near Atlanta.

In Figure 1, we can see that the accident count increases over the year, even during the Covid era. In Figure 2, it's obvious to observe that accidents happen most frequently during rush hours. Figure 3 is a screenshot of an interactive graph that allows users to zoom in/out on the region to see the location density of the car accidents. There are more graphs and interactive plots on the web page, here we only list some representatives.

3.3 Safety Index

We proposed two different safety indexes in our project. One is based on the statistical data of the accident numbers in Atlanta in the past, while the other is based on a time series machine learning model that can predict the number of accidents in the future. The advantage of the first approach is that it can offer users different results according to the time that the users make the query.

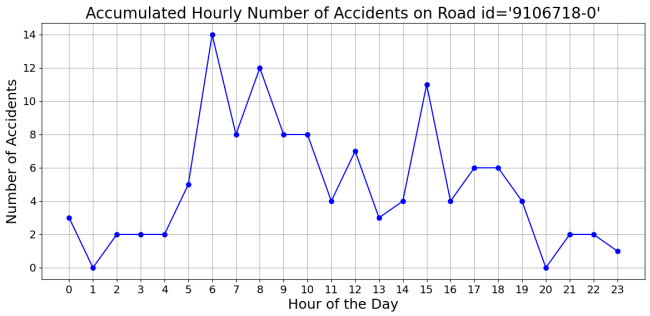

Figure 4: An example of accumulated hourly numbers of accidents

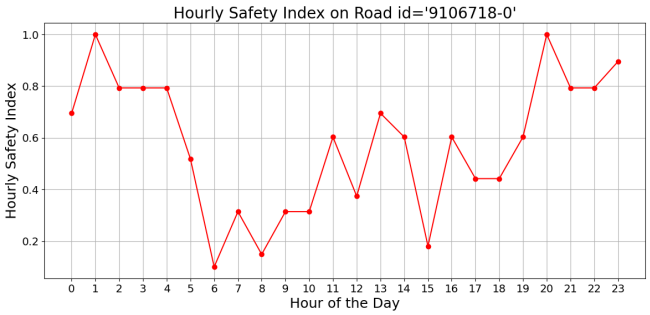

Figure 5: An example of the hourly safety indexes

The advantage of the second approach is that it can better reflect the trend of increasing accident numbers in recent years.

3.3.1 Statistical Method.

The statistical safety index in our project is referred to as the PASSAGE system [5]. They assign 48 distinct risk values to each edge in the graph of Atlanta, where each value corresponds to a 30-minute interval of the day. Similarly, in our case, we assign 24 different safety indexes to each edge in the graph of Atlanta. Each safety index represents a 1-hour interval of the day. This is because the safety of roads can differ from time to time. For example, the risk of accidents may increase during the rush hours and decrease in the mid-nights. Figure 4 shows an example of the accident counts on a road in Atlanta, and Figure 5 demonstrates the corresponding hourly safety indexes on the same road.

In order to transfer the hourly number of accidents into a safety index that ranges from 0 to 1, we applied the sigmoid function, depicted in Equation 1, with its coefficients α, β, and γ set to 2, 0.21, and -0.5, respectively. The variable x is the accumulated hourly number of accidents that happened on the road.

$$SI_{history} = f(x) = \alpha \left(\frac{1}{1 + e^{-\beta x}} + \gamma \right) \tag{1}$$

Despite that our statistical safety indexes can provide the safety of each hour on each road in Atlanta, we ignored the trend of how the number of accidents may evolve from 2016 to 2023 since we were just adding up the accident numbers in the past. Therefore, we will introduce our second safety index which aims to deal with the problem in the next section.

3.3.2 Machine Learning Method.

Our project aims to provide the safest route from point A to point B the safest route without being absurd. Since the majority of path finding algorithms are based on the minimization of a distance, we need to apply a function to the distance between each intersection (in other words, to the length of each road segment). For a directed graph $G = (V, E)$, this means applying a transformation $f : \mathbb{R}_+ \to \mathbb{R}_+$ to each edge $e \in E$. Taking into account the distance coupled with the dangerousness of the road, we obtain a new graph on which we can directly apply a path finding algorithm.

Since the literature did not propose a real safety index in an analytical way, we defined our own index as follows. For each edge of the graph:

$$f(\omega_e) = (1 + acc_e)^p \, \omega_e \tag{2}$$

with acc_e the number of accidents in the past few months and p a positive constant. One can observe that, as the number of accidents rises, the distance rises and as the number of accidents decreases, the distance changes less. But all in all, it penalizes a road if it has too many accidents. Moreover, if there is no accident, we have $f(\omega_e) = \omega_e$ which is logic. This approach is very efficient because it doesn't change the running time of path finding algorithms as computing the new graph $G' = (V, f(E))$ is a preprocessing step. In our investigation, we conducted a comprehensive study focusing on varying values of the p coefficient, as illustrated in Figure 8. The selection of this coefficient was aimed at a value substantial enough to discern notable changes during the update step. However, we deliberately prevent ourselves from opting for an excessively large value, to maintain a balance that ensures realistic predictions in our analysis.

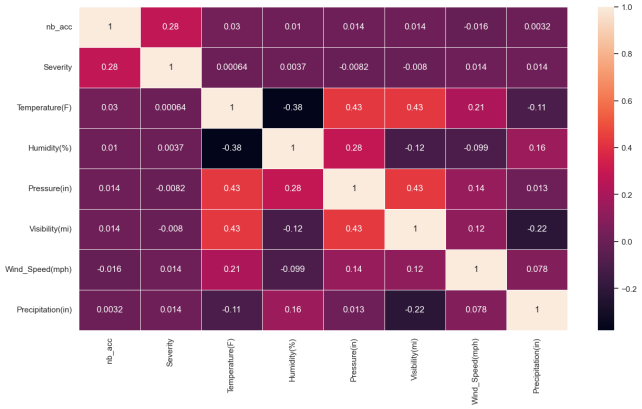

Figure 6: Correlation Matrix of the different features of our dataset

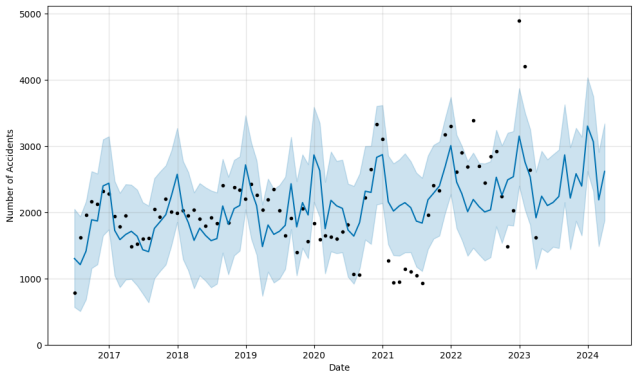

Figure 7: Accident Prediction for the year 2024 over all the United States

In Figure 6, we can see that the frequency of accidents exhibits a limited correlation with other variables in our dataset, such as Temperature, Humidity, or Pressure. Consequently, a deliberate decision has been made to prioritize the accident count as the primary feature for both the safety index and predictive modeling. This choice aims to prevent unnecessary complexity in the analytical framework.

However, as a period will be defined, it is static and it only depends on the sum of accidents over the past few months. Therefore, as the number of accidents on a given road can be modeled as a time series, we have thought about using Prophet, which is a very popular time series forecasting algorithm developed by Facebook's data science team in 2017 [13]. We predicted over the year 2024 for all our data concerning the USA in Figure 7. The core idea behind Prophet is to model time series as an additive combination of trend, seasonality, noise components, and the holiday effect, which is one of the main contributions of this state-of-the-art model. The trend represents the overall direction of the time series, the seasonality captures periodic patterns, the noise accounts for random fluctuations and the holiday effect simulates one-off phenomena. Moreover, Prophet utilizes a Bayesian framework to model time series data, estimating the posterior distribution of model parameters rather than just point estimates. This approach allows the algorithm to generate probabilistic forecasts that provide a measure of uncertainty. Therefore, Prophet assumes a time series to be modeled as follows:

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t) \tag{3}$$

with g the trend, s the seasonality, h the holiday effect and ϵ the noise.

More precisely, by grouping observations by their location and sorting them by date, we obtain a dataset based on relevant features from our *Kaggle* dataset with a correlation matrix analysis between the number of accidents and the average of other numerical features. This will conserve relevant features to train our model while minimizing its training cost. Therefore, the final edge weights can be rewritten as follows for each edge of the graph:

$$f'(\omega_e) = (1 + \mathbb{E}_e[\hat{y}|X])^p \, \omega_e \tag{4}$$

with $\mathbb{E}_e[\hat{y}|X]$ the expected number of accidents on road e and $p = 2$. To conclude on the safety index, we made a prediction on all the roads of Atlanta (an example can be seen in Figure 9) in order to use it for our path planning model.

3.3.3 *Discussion.*

Actually, we can combine the two methods of safety index calculation to have both of their advantages. We can divide a day into 24 intervals, and apply the second method to predict the "hourly" number of accidents. However, we faced an issue of insufficient data. Since there are millions of roads in Atlanta, the number of accidents within each hour interval of a road will mostly be close to 0. The small numbers may lead to biases of the prediction. Therefore, we kept the two safety indexes in our project for now, and if we obtain more data in the future, we can further recalculate a more precise and representative safety index by combining the two.

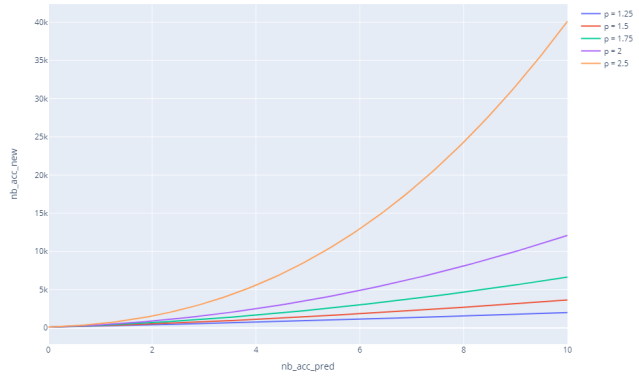

Figure 8: Plot of the different values of the p coefficient

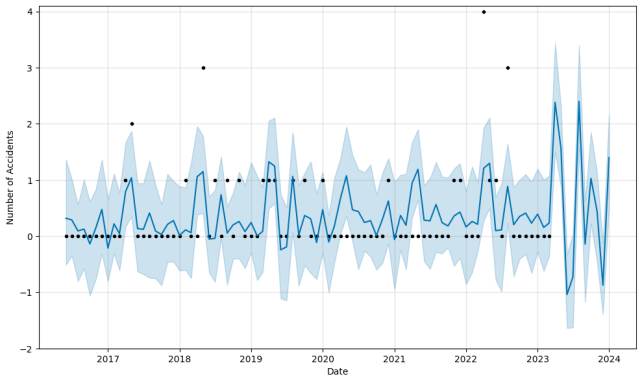

Figure 9: Accident Prediction for Next Month on one road

3.4 Path Planning

With a graph updated with safety-based weights, we can find the safest itinerary using various shortest path algorithms:

- Dijkstra: Efficient for sparse graphs with a priority queue, ideal for minimizing travel time or cost. Complexity: O((V + E) * log(V)).
- A* : Combines Dijkstra's and greedy best-first search advantages, offering the best itinerary from a source to a destination using a heuristic function. Time complexity depends on the heuristic but can be similar to Dijkstra in the worst case.
- Bellman-Ford: Less efficient than Dijkstra (O(V * E) complexity) but handles negative edge weights, which is not applicable in our case since the safety index is positive.
- Floyd-Warshall: Computes pairs of shortest paths for all nodes, but it's not relevant for our use case and has a high complexity of O(V^3).

After evaluating the pros and cons of each algorithm, we chose Dijkstra's algorithm since it is the one that it is the simplest one that fits best our problem. This algorithm needs a cost function to evaluate the cost of moving from one node to a neighbor. We formulated three distinct cost functions to compare three different types of itineraries:

- Fastest Itinerary: This itinerary, calculated from point A to point B, is determined by considering the travel time of each edge. The travel time is derived by multiplying the length of the edge with the speed limit corresponding to the road type of that edge. It's important to note that this computation assumes an ideal travel scenario without accounting for potential traffic slowdowns. Consequently, the Dijkstra minimization aims to reduce the overall travel time, aligning with the common objective of most computed itineraries on maps and navigation apps like Waze.

$$\text{Cost}(node, neighbor) = \text{length}(node, neighbor) * \text{speed limit}(node, neighbor) \quad (5)$$

- Safest Itinerary using Past Accident History: This particular route is determined by minimizing the occurrence of accidents within a specified one-hour time frame. The incident count has been normalized beforehand, ranging between 0 and 1 (see section 3.3.1): 1.0 indicates a safe road, while 0.0 denotes a high frequency of accidents. To prioritize safety, we aim to minimize the overall danger, and therefore, the cost function employed by Dijkstra's algorithm contains the term $1 - SI_{history}$. This factor is then multiplied by the travel time associated with a specific road segment. By integrating this approach, we can compute the fastest itinerary while accounting for the safety implications of each road segment. Moreover, it allows for the creation of a dynamic itinerary that adjusts throughout the day based on the user's chosen travel time.

$$\begin{aligned}\text{Cost}(node, neighbor) =&(1 - SI_{history}(node, neighbor, hour)) \\ &* \text{length}(node, neighbor) \\ &* \text{speed limit}(node, neighbor)\end{aligned} \tag{6}$$

- Safest Itinerary using Future Accident Predictions: This route is determined by employing a second safety index (see section 3.3.2), specifically designed to predict future accidents. The value of this index corresponds to an adjusted distance, resulting in a cost function identical to that of the first fastest itinerary. However, in this case, the length parameter is updated to reflect the most current information. By utilizing this approach, we aim to provide a route that prioritizes safety by considering predictive accident data, while maintaining consistency with the fastest itinerary paradigm.

$$\begin{aligned}\text{Cost}(node, neighbor) =&SI_{prediction}(node, neighbor) \\ &* \text{speed limit}(node, neighbor)\end{aligned} \tag{7}$$

4 EXPERIMENTAL RESULTS

In our experimentation, we implemented an HTML / JavaScript code that leverages the Leaflet library, integrating data retrieved from the OpenStreetMap (OSM) API. This code serves as the foundation for an interactive web map application, empowering users to designate their origin and destination points. Users have the flexibility to input these points either by clicking directly on the map or utilizing the search bar to find specific addresses.

Once the departure and destination points are specified, our system computes three distinct itineraries, as previously outlined. This computation is facilitated by the Graphology library, capable of executing Dijkstra's algorithm on a pre-loaded graph and accommodating custom cost functions. The graph itself, composed of nodes and edges, is initially loaded from a CSV file using the D3 library. This seamless integration of technologies enables a dynamic and user-friendly exploration of optimized travel routes.

The application's features are illustrated in the image depicted in Figure 10. Notably, the fastest itinerary

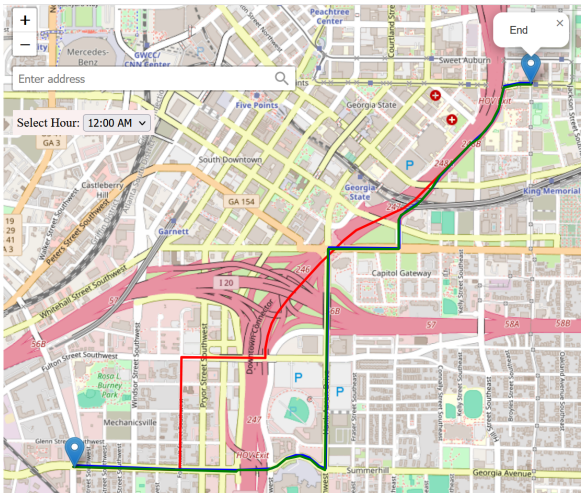

Figure 10: Screenshot of our web app. A departure point and an arrival point have been selected, and three itineraries have been computed: the fastest itinerary (in red), the statistically safest itinerary at 12 AM (in blue), and the safest itinerary using next month's accident predictions (in green).

prioritizes the use of motorways at the earliest opportunity, whereas the safer itineraries strategically opt for alternative major arteries, delaying entry onto the motorway. This decision aligns with a logical rationale: motorways tend to experience high traffic flow and speeds, contributing to a logically higher incidence of accidents. However, the safer itineraries, while emphasizing safety, still consider travel time, leading them to eventually incorporate a segment on the motorway. A more precise metric, rather than relying solely on the number of accidents per road, would involve considering the frequency of accidents per person-hour specific to that road.

5 CONCLUSIONS

In summary, our project is dedicated to enhancing road safety through data analytics and visualization. We developed two kinds of safety indexes that factor in accidents and time for real-time route planning. By providing users with safer route recommendations, we aim to reduce accidents, ultimately making road travel safer and more informed for all. **All team members have contributed a similar amount of effort.**

REFERENCES

[1] Halima Akhter. 2015. Information extraction and interactive visualization of road accident related news. *International Journal of Computer Applications* 975 (2015), 8887.

[2] Mohamed Aljaban. 2021. Analysis of car accidents causes in the usa. (2021).

[3] Hannah Bast, Daniel Delling, Andrew Goldberg, Matthias Müller-Hannemann, Thomas Pajor, Peter Sanders, Dorothea Wagner, and Renato F Werneck. 2016. Route planning in transportation networks. *Algorithm engineering: Selected results and surveys* (2016), 19–80.

[4] Wei Chen, Fangzhou Guo, and Fei-Yue Wang. 2015. A survey of traffic data visualization. *IEEE Transactions on intelligent transportation systems* 16, 6 (2015), 2970–2984.

[5] Matthew Garvey, Nilaksh Das, Jiaxing Su, Meghna Natraj, and Bhanu Verma. 2016. Passage: A travel safety assistant with safe path recommendations for pedestrians. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*. 84–87.

[6] Hiroshi Hayakawa, Paul S Fischbeck, and Baruch Fischhoff. 2000. Traffic accident statistics and risk perceptions in Japan and the United States. *Accident Analysis & Prevention* 32, 6 (2000), 827–835.

[7] Georgios N Kouziokas. 2017. The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment. *Transportation research procedia* 24 (2017), 467–473.

[8] Dominique Lord and Fred Mannering. 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation research part A: policy and practice* 44, 5 (2010), 291–305.

[9] Fanny Malin, Ilkka Norros, and Satu Innamaa. 2019. Accident risk of road and weather conditions on different road types. *Accident Analysis & Prevention* 122 (2019), 181–188.

[10] Philippe Nitsche, Pete Thomas, Rainer Stuetz, and Ruth Welsh. 2017. Pre-crash scenarios at road junctions: A clustering method for car crash data. *Accident Analysis & Prevention* 107 (2017), 137–151.

[11] Burcu Oralhan and Ziya Gökalp Göktolga. 2018. Determination of the risk factors that influence occurrence time of traffic accidents with survival analysis. *Iranian journal of public health* 47, 8 (2018), 1181.

[12] Aqsa Qalb, Hafiz Syed Hamid Arshad, Muhammad Shafaat Nawaz, and Asra Hafeez. 2023. Risk reduction via spatial and temporal visualization of road accidents: a way forward for emergency response optimization in developing countries. *International journal of injury control and safety promotion* 30, 2 (2023), 310–320.

[13] Benjamin Letham Sean J.Taylor. 2017. Forecasting at Scale. *Facebook* (2017).

[14] N Shaadan, MIK Azhar Suhaimi, MI Hazmir, and EN Hamzah. 2021. Road accidents analytics with data visualization: a case study in Shah Alam Malaysia. In *Journal of Physics: Conference Series*, Vol. 1988. IOP Publishing, 012043.

[15] DA Skorobogatchenko, VV Borovik, and AI Frolovichev. 2021. Assessment automation of road traffic safety with account for road conditions of an individual itinerary. In *Journal of Physics: Conference Series*, Vol. 2091. IOP Publishing, 012051.

[16] Dongye Sun, Yunfei Ai, Yunhua Sun, and Liping Zhao. 2020. A highway crash risk assessment method based on traffic safety state division. *PLoS one* 15, 1 (2020), e0227609.

[17] Jirapon Sunkpho and Warit Wipulanusat. 2020. The role of data visualization and analytics of highway accidents. *Walailak Journal of Science and Technology (WJST)* 17, 12 (2020), 1379–1389.