University of Twente

Managing Big data

Assessing the Impact of Geographical Features on Car Accidents in the Netherlands

Authors:
Tijs Zandt (s2121182)
Nynke Luijten (s2563363)
Pascal Vriend (s2739046)
Daan Schram (s2692759)

January 27, 2025

https://gitlab.utwente.nl/s2692759/mbd-project-group-5/-/tree/main

1 Introduction

As we begin 2025, we have reached the halfway point of the second decade of action for road safety. The UN General Assembly adopted resolution A/RES/74/299, titled "Improving global road safety" [1] in September 2020. The goal of this resolution is to reduce global road deaths and serious injuries by 50% by the end of the decade. However, according to the Directorate-General for Mobility and Transport [2], the European Union saw only a 1% decrease in road crashes in 2023 compared to the previous year, a result that falls short of the required 4.5% annual reduction needed to meet the targets set by the resolution. While the Netherlands reported a 7% decrease in road fatalities compared to 2022, this result is undermined by considering the long term 4% increase in fatalities compared to 2019.

To meet the targets, the European Commission has outlined a long-term vision to reduce road deaths to zero by 2050 [3], as a part of its "Vision Zero" initiative. This plan highlights several challenges, including infrastructure safety, vehicle safety, safe road use (such as speed control, alcohol and drug use, distractions and protective equipments), and the need for better road design. The Commission's proposal stresses the importance of creating 'forgiving' roads, those equipped with safety barriers to mitigate the impact of driving errors, improved road signs and markings, and advanced driver assistance systems. In March 2023, a more detailed proposal regarding driving license requirements and road traffic rules was introduced [4]. This includes measures such as accompanied driving for learners, zero tolerance for drink-driving among novice drivers and enhanced cooperation between member states to address traffic violations more effectively.

However, these plans overlook the influence of natural infrastructure and the surrounding environment on road safety. Research [5][6] has demonstrated that natural and built environmental features, such as trees, buildings and other roadside elements, can contribute to increased road accidents due to factors like reduced visibility and other safety hazards.

Understanding the cause of road accidents is essential for formulating an effective strategy to achieve the UN's road safety goals and reduce accidents overall. Environmental features, as one of the key influencing factors, have not been thoroughly considered in the current road safety plans.

This research seeks to analyze and find the features that most significantly impact road accidents and provide recommendations to the UN on which environmental factors should be prioritized for consideration in future road safety strategies.

This corresponds to the following research question: *Which features have the most significant impact on the number and outcome of road accidents?*
The URL found on the front page leads to the Git repository containing the code necessary for the analysis.

1.1 Related work

Research on road side accidents and general traffic safety are plentiful.

Several research has been performed into geospatial features and road accidents. For instance, Obelheiro et al. introduced a new type of Traffic Safety Analysis Zone System (TSAZ) to overcome limitations of the traditional Traffic Analysis Zones (TAZ) [6]. They extended the original system by accounting for environmental features and local variations. Their new models outperformed global models in predicting crashes by including several features. Similarly, Huang et al. applied Geographically Weighted Regression (GWR) to explore relationships between crashes and the built environment of Detroit. Several relationships were found between features such as road density, intersection density or local road mileage percentage and accident likelihood [7]. Last, Bijleveld et al. explored the influence of weather, particularly precipitation, on accidents in the Netherlands [8]. Their findings show how seasonal and weather-driven variations impact crash frequency.

In urban settings, several interactions between land-use design and accident frequency exist. For example, Klanjčić et al. explored how urban features relate to vulnerable road users across European cities [9]. They highlighted how promoting walking and cycling, joint with speed reduction policies, reduces both the severity and frequency of car accidents.

Rural areas on the other hand present their own unique challenges, particularly with roadside environmental features, vegetation and wildlife interactions. Reseach done by Kekena et al. showed how specific roadside vegetation attract specific kind of animals, contributing to a higher frequency in car accidents [10].

Several studies have also been performed in a Dutch context. Xiong's research at our university identified urban road segments with high accident frequencies, emphasizing the role of features as road density and the use of land diversity [11]. Next to this, a study performed by Asadi et al. highlighted the importance of spatial spillover effects in understanding car crash frequency in the Netherlands [5]. This refers to the phenomenon where safety conditions are influenced by neighboring regions, meaning that for instance improved infrastructure in one area can have cascading effects to adjacent areas.

The International Road Assessment Programme (IRAP) [12] highlights numerous case studies that provide practical examples from around the world of how deaths and serious injuries have been prevented. Most of these are focused on small urban areas with high traffic flow and highlight improvements made to infrastructure.

A recent review by Watson R. [13] used several techniques to perform geospatial tests to identify the geospatial factors that are relevant to road accidents in the city of Geelong in the state of Victoria, Australia. A strong spatial autocorrelation was found, and analysis suggested that there are more accidents in built-up areas. However, no comparison between specific features and their impact on road accidents was performed.

2 Materials and methods

2.1 Datasets

A large collection of datasets has been utilized and combined. Table 1 presents the datasets along with their respective uncompressed sizes and entry counts. These datasets are primarily sourced from the Nationaal Wegenbestand, a comprehensive dataset from the Netherlands containing detailed information about the country's road network. Maintained by the Rijkswaterstaat (the Dutch Ministry of Infrastructure and Water Management), it includes data on road types, sections, infrastructure elements, and road events, such as accidents.

The Nationaal Wegenbestand can be integrated with other related datasets, including road characteristics databases also documented by the Rijkswaterstaat. These datasets are linked through dynamic segmentation, which allows different datasets to be associated with specific road sections. A unique road section ID identifies the relevant characteristics for each section, enabling connections across databases. For even greater precision, the dataset includes methods for further differentiating road sections, such as when multiple characteristics apply to the same section, or by using a length characteristic that, combined with the starting point, allows for coverage of multiple road section.

In addition to the Nationaal Wegenbestand, two small external datasets have been incorporated. First, a population density dataset from the Dutch Centraal Bureau Statistiek (CBS) was used [14]. This dataset includes the amount of inhabitants per squared kilometer per municipality. The second is a .geojson file used to display maps of the Netherlands [15]. This file contains all geographical locations of municipalities in the Netherlands, which was used alongside other data in datasets from table 1 to compute maps. More datasets were merged in the complete dataset, however these are not specifically mentioned in the table as these were not used in further analysis. Examples of these are crosswalks, road width and highway entrances.

Combining all datasets, the complete set explodes to a total size of 25.4GB. The combining of the dataset was done using the parameter 'WEGVAK_ID', which is an indicator in the datasets to which specific part of which road the value belongs. The program that merged the datasets used ten executors and ran for about six minutes.

2.2 Data preprocessing

To be able to perform the analysis of this research, the different datasets needed to be merged together. As noted in the previous section, a lot of the datasets can be combined based on the location to which the respective data applies to, presented by the road section ID. Before performing the join of the datasets, general preprocessing techniques are used to improve the data.

One main library is used for the preprocessing and combining of the data. This is the pyspark.sql and its sub library pyspark.sql.functions. From this library, some basic functionality is used, such as the col(), count() and row_number().
Other techniques and functions that were used were dataframe isNotNull filtering, column renaming, join operation (inner join) on wvk id, grouping by, select(), orderby(), and drop().

In certain cases, road coordinates were provided in the dataset as RD coordinates. To make them compatible with Google Maps, they were converted using gpscoordinaten.nl.

2.3 Data analysis

Many big data techniques have been used to perform the data analysis. Again the pyspark.sql with its sublibrary of pyspark.sql.functions has been greatly utilized. Other dataframe methods that have been used are select(), filter(), aggregate(), orderBy(), count(), and small data plotting methods with the matplotlib library. The pyspark machine learning libary pyspark.ml has been used for which the sublibraries feature, regression and stat were imported. These were used to perform a Random Forest machine learning algorithm.

3 Results

With all the techniques mentioned in the previous section, several features of the dataset were investigated to determine the impact on the number and outcome of road accidents. The following subsections will go through each of these results.

3.1 Trees and accidents

Included in the data was the number of trees next to the roads and their distance from the road. The hypothesis was that more trees would imply more accidents, as drivers could more easily drive into a tree. Figure 1 shows a scatter plot of the number of trees on the x-axis and the number of accidents on the y-axis.

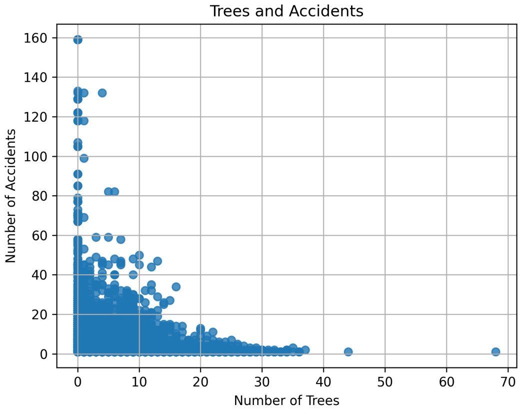


Figure 1: Scatter plot of trees and accidents

A linear trend was expected, but it can be seen that this is not the case. From figure 1 it can be seen that

Table 1: Overview of datasets that were used and their uncompressed sizes.

Dataset	Size	Nr of Entries	Description	Source
Car Accidents	357MB	1,193,484	All car accidents in the Netherlands, from 2014 up until 2023	[16]
Parties in Car Accidents	213MB	1,801,034	All parties in car accidents in the Netherlands, from 2014 to 2023	[16]
Road Section Coordinates	109MB	2,946,264	All location coordinates of road sections, from 2014 to 2023	[16]
Road Section Information	964MB	1,573,304	Road section information per ID	[17]
Speed Limits	278MB	1,590,525	Speed limits in the Netherlands	[18]
Trees Next to Roads	72MB	777,184	Trees next to roads, including distances from road	[19]
Trees Next to Roads	72MB	777,184	Trees next to roads, including distances from road	[19]
Municipalities Locations	11MB	342	Locations of municipalities in the Netherlands in geojson format	[14]
Population Density per Municipality	6KB	342	Amount of inhabitants per squared kilometer per municipality	[15]

there is no plausible connection between the number of trees and the number of accidents. The run time of this program was 29 seconds with five executors.

The correlation between the distance of the trees to the road and the outcome of the accident is shown in figure 2.

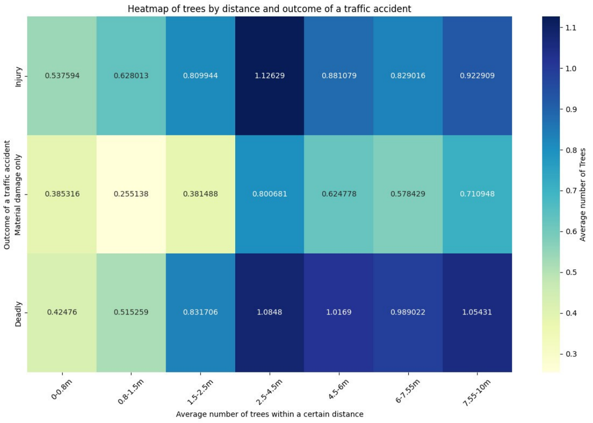


Figure 2: Heatmap distance of the trees to the road and the outcome of the accident

It can be seen that deadly accidents, sustained injuries and material damage are all more likely if there are trees between 2.5 and 4.5 meters from the road. The program ran in 53 seconds and used seven executors.

3.2 Accidents per year

The increase or decrease of the amount of accidents per year was researched. Figure 3 shows the total number of accidents over the years. To run this script, five executors were used and it took 27 seconds.

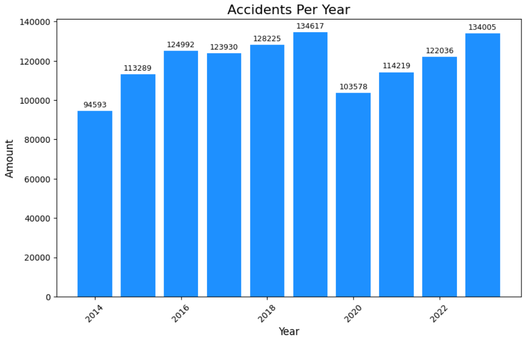


Figure 3: Total accidents over the years

A sudden decrease can be seen in 2020. This is suspected to be due to Covid19.

Figure 4 shows the amount of accidents per year for a selection of the roads.

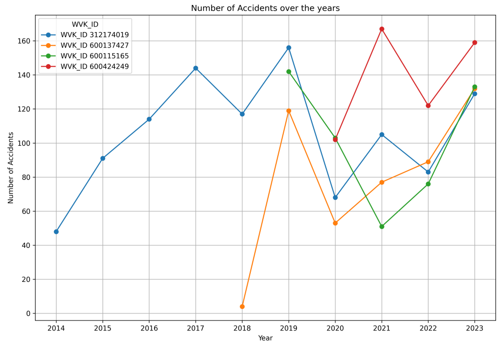


Figure 4: Number of accidents per year

The decrease in 2020 can clearly be spotted here as well.

3.3 Municipalities and accidents

The next step was to investigate where most accidents occur. This was done with the use of a heatmap, see figure 5.

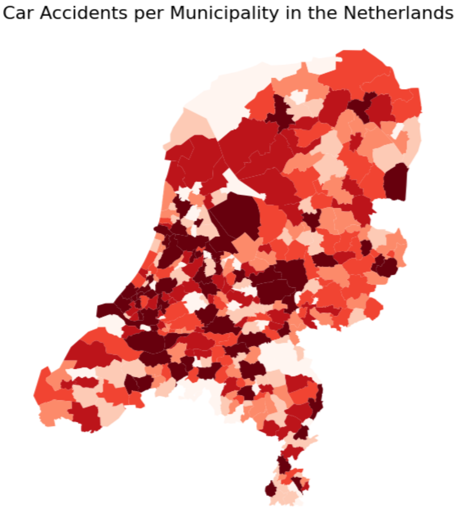


Figure 5: Car accidents per municipality in the Netherlands

Figure 6 shows the top five municipalities with the most accidents per province, and figure 7 the accidents per province.

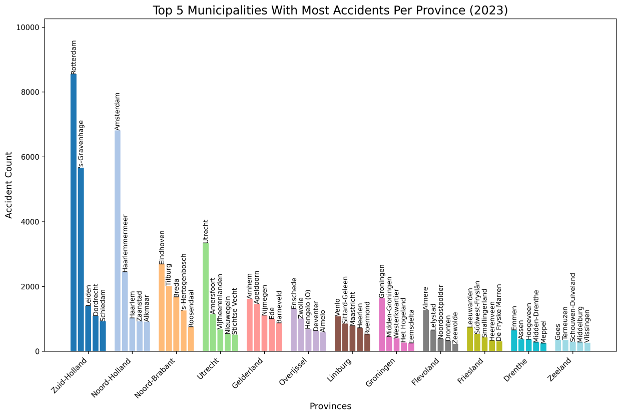


Figure 6: Top 5 municipalities per province

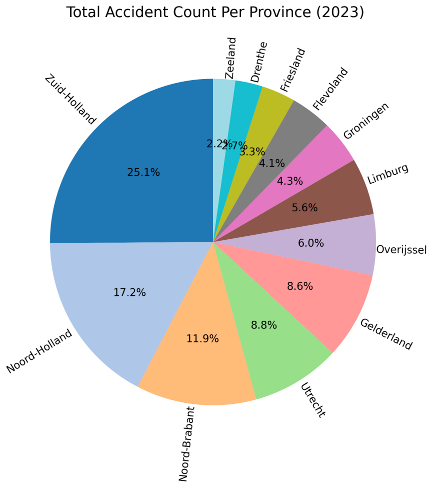


Figure 7: Accidents per province

From figure 5 and 6 it is noticeable that the largest cities regarding population count have the most accidents. Therefore normalizing the car accidents on the population count would give a better view, see figure 8.

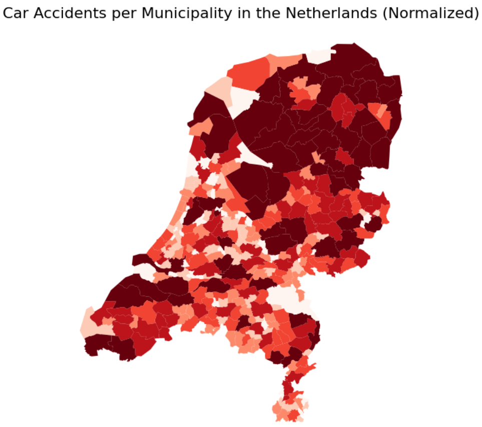

Car Accidents per Municipality in the Netherlands (Normalized)

Figure 8: Heatmap of the number of accidents per municipality normalized on the population count

It is clearly visible that when normalized, there are many accidents in the North of the Netherlands, so the provinces Drenthe, Groningen and Friesland. Computing this normalized data and projecting it onto the map of the Netherlands took 28 seconds, and five executors were used to finish this task. for the calculation of the bar and pie chart, the runtime was 30 seconds and two executors were used.

3.4 Light situation

The next feature to be investigated is the light situation, so daylight, dusk or night. Figure 9 shows a heatmap of the severity of accidents and the light situation.

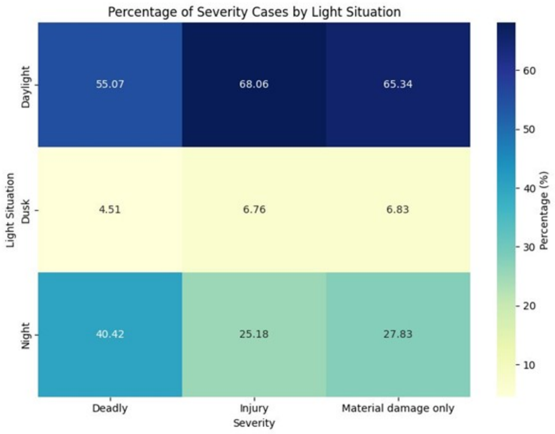
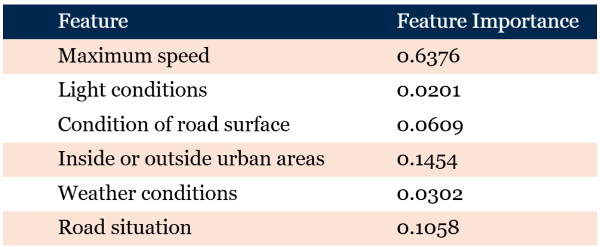


Figure 9: Heatmap of severity of accidents per light situation

It can be concluded that deadly accidents are frequent at night time. The program ran in 66 seconds and used six executors.

3.5 Feature importance

A Random Forest model was used for this particular analysis. Feature importance indicates how much each feature contributes to the severity of an accident. Figure 10 shows the results of this analysis.



Feature	Feature Importance
Maximum speed	0.6376
Light conditions	0.0201
Condition of road surface	0.0609
Inside or outside urban areas	0.1454
Weather conditions	0.0302
Road situation	0.1058

Figure 10: Feature importance

It can be seen that mostly the *Maximum speed* contributes heavily to the severity of an accident, followed by *Inside or outside urban areas*, and *Road situation.* To train this model, five executors were used and it took a little over one minute.

3.6 Nature of accident frequencies

The nature of car accidents in the Netherlands are divided in 9 different categories. They describe the cause of the accident. An analysis of the frequency of each cause can be seen in figure 11. 12 shows the results of this analysis.

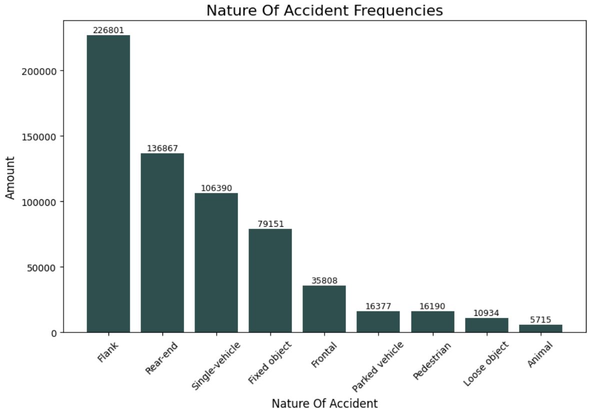
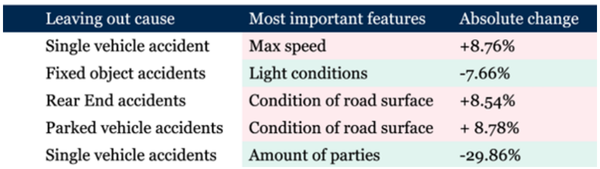


Figure 11: Leave-One-Out Feature Analysis

These causes were later used to train several models and perform feature analysis, like described in section 3.7. Running this script took 27 seconds, using five executors.

3.7 Leave-One-Class-Out Feature Analysis

The Leave-One-Class-Out Feature analysis is also done with a Random Forest model. Each accident cause is left out one by one, and for each it it is investigated which feature changes the most, indicating that they are relevant for predicting accidents with that particular cause. Figure 12 shows the results of this analysis.



Leaving out cause	Most important features	Absolute change
Single vehicle accident	Max speed	+8.76%
Fixed object accidents	Light conditions	-7.66%
Rear End accidents	Condition of road surface	+8.54%
Parked vehicle accidents	Condition of road surface	+ 8.78%
Single vehicle accidents	Amount of parties	-29.86%

Figure 12: Leave-One-Out Feature Analysis

A positive absolute change indicates that the feature is less relevant for predicting the accident with the particular feature that was left out, while a negative change indicates it is of great relevance.
For example, the maximum speed is not the most relevant for single vehicle accidents, while the light conditions are very relevant for fixed object accidents. The largest change happens for single vehicle accidents. It can be seen that the importance of the amount of parties greatly reduces, which is a very logical result.
Computing these models took some resources, since ten models (one without leaving any cause out and nine left-out cause models) needed to be trained. Therefore, ten executors were used and it took a little under ten minutes to finish the script.

3.8 Road specific analysis

An analysis was conducted to identify the causes of accidents on specific roads over the past years. The objective was to pinpoint roads with an unusually high number of accidents within particular categories. The overall share of accidents for a category (category percentage) and the proportion of that category's accidents occurring on each specific road (road percentage) was calculated. The roads were then ranked based on the most amount of accidents for that category using these percentages as tiebreakers. Coordinates for the identified roads were extracted and visualized using Google Maps. The resulting images were uploaded to gitlab. For this calculation, two executors were utilized and was the program finished in 34 seconds.

4 Conclusion

This research has focused on the impact of geographical features on car accidents in the Netherlands. The following research question was used: *Which features have the most significant impact on the number and outcome of road accidents?*
The research question was sought to answer by investigating several features with different techniques, including extensive analysis and the use of machine learning models.
The first conclusion is that trees within 2.5 and 6 m of the road are a relevant cause for road accidents. Therefore the advice would be to decrease the number of trees in those areas. This is also suggested by the Bomen Stiching [20]. The number of trees does not seem to influence the number of accidents on the road, which was different than expected.
Secondly, it has been found that the North of the Netherlands has relatively many accidents per inhabitant. The advice would therefore be to first focus on improving road safety in those areas, so Drenthe, Friesland and Groningen.
Moreover, deadline accidents occur more frequently at night. Therefore it would be wise to reduce instead of increase (which is the current situation) the maximum speed at night.
The machine learning models showed that the maximum speed largely impacts the severity of an accident. This again adds to the advise to not increase the maximum speed but rather decrease it on national roads, but also on smaller regional roads. Furthermore, fixed objects should be clearly marked, such that they are more visible in darker light situations.
All these factors contribute to the answer to the research question. It can be now we answered as follows: Trees in close proximity to the road, particular areas (North of the NL), light conditions, and maximum speed are important features that contribute to the number and outcome of road accidents.

5 Discussion

5.1 Implications

The findings of this research are of importance to national and regional governments, as they can use the results to improve road safety and thus reduce road accidents, to reach the UN goals.

5.2 Limitations

The data set was incomplete in a number of factors. The date and time of accidents was for instance not documented, while there was a column for this data. This prevented the use of weather data, to investigate the impact of different weather types on road accidents. Furthermore, the data in the dataset is delivered by multiple parties, and there is no way to know for sure that all accidents are registered.

5.3 Recommendations

Given the limitations of the dataset, it is crucial that the findings of this study are validated in real-life scenarios. This should be done before implementing large-scale changes to our infrastructure. Small-scale pilot projects could be conducted, like planting or removing trees in a controlled environment and monitoring the impact of road safety. Next to this, collaboration with relevant authorities is necessary to prioritize more comprehensive data collection. In this way, critical data such as timestamps or weather data can be consistently recorded. By addressing these gaps, future studies can provide more robust insights.

A Appendix

During the preparation of this work, I (and my fellow authors) used ChatGPT 3.5 to perform grammar checks and provide suggestions for proper flow of sentences. After using this tool/service, we thoroughly reviewed and edited the content as needed, taking full responsibility for the final outcome.

References

[1] United Nations General Assembly, *Resolution a/res/74/299: Improving global road safety*, Accessed: 2024-12-16, 2020. [Online]. Available: `https://undocs.org/A/RES/74/299`.

[2] European Commission, *20,400 lives lost in eu road crashes last year*, Accessed: 2024-12-16, 2024. [Online]. Available: `https://transport.ec.europa.eu/news-events/news/20400-lives-lost-eu-road-crashes-last-year-2024-10-10_en`.

[3] European Commission, Directorate-General for Mobility, and Transport, *Next steps towards 'Vision Zero' – EU road safety policy framework 2021-2030*. Publications Office, 2020. DOI: `doi/10.2832/391271`.

[4] European Commission, *European commission proposes updated requirements for driving licences and better cross-border enforcement of road traffic rules*, Accessed: 2024-12-16, 2023. [Online]. Available: https://transport.ec.europa.eu/news-events/news/european-commission-proposes-updated-requirements-driving-licences-and-better-cross-border-2023-03-01_en.

[5] Mehrnaz Asadi, Mehmet Baran Ulak, Karst T. Geurs, Wendy Weijermars, and Paul Schepers, "A comprehensive analysis of the relationships between the built environment and traffic safety in the dutch urban areas," *Accident Analysis Prevention*, vol. 172, p. 106 683, 2022, ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2022.106683. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001457522001191.

[6] Marta Rodrigues Obelheiro, Alan Ricardo da Silva, Christine Tessele Nodari, Helena Beatriz Bettella Cybis, and Luis Antonio Lindau, "A new zone system to analyze the spatial relationships between the built environment and traffic safety," *Journal of Transport Geography*, vol. 84, p. 102 699, 2020, ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2020.102699. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096669231930225X.

[7] Yuan Huang, Xiaoguang Wang, and David Patton, "Examining spatial relationships between crashes and the built environment: A geographically weighted regression approach," *Journal of Transport Geography*, vol. 69, pp. 221–233, 2018, ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2018.04.027. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0966692317306373.

[8] Frits Bijleveld and Tony Churchill, "The influence of weather conditions on road safety, An assessment of the effect of precipitation and temperature," SWOV, Leidschendam, Netherlands, Tech. Rep. R-2009-9, 2009, pp. 24 + 23.

[9] Marina Klanjčić, Laetitia Gauvin, Michele Tizzoni, and Michael Szell, "Identifying urban features for vulnerable road user safety in europe," *EPJ Data Science*, vol. 11, no. 1, p. 27, 2022, ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-022-00339-5. [Online]. Available: https://doi.org/10.1140/epjds/s13688-022-00339-5.

[10] Z. Keken, J. Sedoník, T. Kušta, R. Andrášik, and M Bíl, "Roadside vegetation influences clustering of ungulate vehicle collisions," *Transportation Research Part D: Transport and Environment*, vol. 73, pp. 381–390, 2019, ISSN: 1361-9209. DOI: https://doi.org/10.1016/j.trd.2019.07.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920918311659.

[11] Li Xiong, "Exploring the influence of the environment on traffic accidents in enschede, the netherlands," Specialization: Urban Planning and Management, Master of Science Thesis, Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, The Netherlands, Feb. 2018.

[12] International Road Assessment Programme, *Case studies—road safety toolkit—irap*, https://toolkit.irap.org/case-studies/, Accessed: 2024-06-16.

[13] Richard B. Watson and Peter J. Ryan, "Geospatial factors applied to road accidents: A review," *Journal of Advances in Information Technology*, vol. 15, no. 3, pp. 451–457, 2024. DOI: 10.12720/jait.15.3.451-457. [Online]. Available: https://www.jait.us/uploadfile/2024/JAIT-V15N3-451.pdf.

[14] Centraal Bureau voor de Statistiek, *Inwoners per gemeente*, nl-NL, webpagina, Last Modified: 15-10-2024T23:03:42. [Online]. Available: https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/regionaal/inwoners (visited on 01/27/2025).

[15] *Gemeenten - Netherlands*, en-US, Publisher: Opendatasoft. [Online]. Available: https://data.opendatasoft.com/explore/dataset/georef-netherlands-gemeente@public/export/ (visited on 01/27/2025).

[16] Rijkswaterstaat, *Verkeersongevallen - bestand geregistreerde ongevallen nederland*, Accessed: 2025-01-06, 1/01/2014 - 01/01/2023. [Online]. Available: https://data.overheid.nl/dataset/9841-verkeersongevallen---bestand-geregistreerde-ongevallen-nederland#location-time.

[17] UBR—KOOP: het Kennis-en Exploitatiecentrum voor Officiële Overheidspublicaties, *NWB wegen - Wegvakken (RWS) — Data overheid*, en. [Online]. Available: https://data.overheid.nl/en/dataset/47092-nwb-wegen---wegvakken--rws-,%20https://data.overheid.nl/en/dataset/47092-nwb-wegen---wegvakken--rws- (visited on 01/27/2025).

[18] *Index of /wkd/Maximum Snelheden/*. [Online]. Available: https://downloads.rijkswaterstaatdata.nl/wkd/Maximum%20Snelheden/ (visited on 01/27/2025).

[19] Rijkswaterstaat, *Bomen database*, Accessed: 2025-01-06, 1/10/2022 - 01/01/2024. [Online]. Available: https://downloads.rijkswaterstaatdata.nl/wkd/Bomen/.

[20] Ceciel Van Iperen, Annemiek Van Loon, and Gerrit-Jan Van Prooijen, "Bomen langs n-wegen," *Bomenstichting*,