Proceedings of the Sixth International Conference on
Railway Technology: Research, Development and Maintenance
Edited by: J. Pombo
Civil-Comp Conferences, Volume 7, Paper 11.9
Civil-Comp Press, Edinburgh, United Kingdom, 2024
ISSN: 2753-3239, doi: 10.4203/ccc.7.11.9

Analysing Railway Accidents: A Statistical Approach to Evaluating Human Performance in Obstacle Detection

O. Lahneche[1], A. Haag[2], P. Dendorfer[2], V. Aravantinos[2], M. Guilbert[2] and M. Sallak[1]

[1]Laboratoire Heudiasyc, Université de Technologie de Compiègne, Compiègne, France
[2]Research and Development Unit, Futurail, Munich, Germany

Abstract

This paper evaluates human driver performance in obstacle detection on railway systems by statistically analysing accident data. Given the challenge of obtaining non-accident data, the study estimates obstacle frequency on tracks by examining accident rates under low visibility conditions (night and curves), where drivers cannot see obstacles in time.

Using data from a major European railway operator, the study finds that **human drivers can avoid around 28% of collisions in good visibility conditions**, but only about 12% when all conditions are considered. Autonomous trains need to meet at least an equivalent performance level. This research aims to support the safety certification of autonomous train systems by offering a benchmark of human performance in obstacle detection, emphasising the need for comprehensive data and rigorous validation of hypotheses regarding driver reactions and environmental conditions.

Keywords: obstacle detection, train driver performance, safety, railways, autonomous train, reference system

1 Introduction

In recent years, significant advancements have been made in the development of autonomous trains, exemplified by the Digital S-Bahn project in Hamburg, which operates at Grade of Automation 2 (GoA2) [1] or the SNCF Train de Fret Autonome project, that reached a GoA4 demo. These advancements are part of a broader effort to improve transportation efficiency and reduce greenhouse gas emissions, aligning with the objectives of the Paris Agreement to limit global warming to 2 degrees Celsius. Transportation is a major contributor to greenhouse gas emissions, accounting for 31% of emissions in France as of 2019. Given that trains are among the most energy-efficient modes of transport, enhancing the rail network's efficiency through increased automation is a crucial strategy for mitigating the environmental impact of transportation.

[2] indicates that *the absence of a driver in the cabin is therefore not enough to classify a train as autonomous. It is its on-board decision-making capacity that makes a train autonomous.* Train autonomy is categorised from GoA0 to GoA4, from no automation to operation of the train without personnel on board, with obstacle detection systems (ODS) being vital above GoA3. Despite the technological advancements, the safety certification of fully autonomous trains remains a significant hurdle, particularly concerning the ODS. The only close example of a fully autonomous certified train is the Rio Tinto train which runs in GoA4 [3]. However, this project is very different from mainline and suburban train projects since the Rio Tinto train only crosses desertic areas and thus doesn’t need obstacle detection. The ODS, which uses sensors such as lidar, radar, and cameras combined with machine learning algorithms to detect and classify obstacles, is not yet safety certified, posing a major barrier to full autonomy of train running in denser open environments.

According to the EN50126 standard [4], there are three primary methods to demonstrate the safety of an object: explicit risk estimation, reference systems, and standard practices. Additionally, other norms like EN50128 [5] for software certification, EN50129 [6] for electronics device certification, and UL600 [7] for autonomous product safety certification are relevant. Utilising a reference system for safety certification necessitates a clear understanding of the system and relevant statistics for comparison. Human drivers, who significantly impact obstacle detection, serve as the obvious reference system, raising questions about the appropriate statistics to use, such as biological parameters (e.g., eye range, reaction time) or operational statistics (e.g., human driver accidentology).

This paper aims to provide operational statistics of human drivers, which can be used as a reference in the argumentation for the safety certification of autonomous train systems. By analysing railway accidentology data, particularly focusing on curve and night scenarios, this study seeks to contribute valuable insights into the performance benchmarks necessary for the certification of autonomous train systems.

The next section provides a comprehensive review of related works, establishing the foundation for our analysis. Building on this, Section 3 introduces the data used for the study, ensuring clarity on the information sources. In Section 4, we formally define the problem, setting the stage for subsequent analysis. The methodology described in the annex is then applied in Section 5, enabling us to derive results on the distribution of accidents occurring during nights and on curves.

Following this, Section 6 addresses the known limitations of our study, providing a balanced perspective. Finally, Section 7 concludes the paper, summarising the key findings and implications.

2 Literature review

The focus of this paper is related to the introduction of an obstacle detection system in trains. According to the EN50126 standard [4], there are three methods to demonstrate the safety of an object before market release: conducting an explicit risk assessment, utilising state-of-the-art practices, and using a reference system. The first technique is applicable, unlike the second, as there are currently no established state-of-the-art practices for obstacle detection systems in trains. The third technique is particularly relevant as the obstacle detection system is intended to replace the conductor, who can be considered a reference system.

Many papers on railway accidentology classify accidents based on human factors. For instance, the paper "Understanding the Human Factors Contribution to Railway Accidents and Incidents in Australia" [8] studies human and external factors that have contributed to railway accidents. This study reviewed forty rail safety investigation reports and applied the Human Factors Analysis and Classification System (HFACS), developed by Dr. Scott Shappell and Dr. Doug Wiegmann [9], to identify errors associated with rail accidents and incidents in Australia. The analysis revealed that nearly half of the incidents resulted from equipment failures, mostly due to inadequate maintenance or monitoring programs. Additionally, slips of attention, associated with decreased alertness and physical fatigue, were identified as the most common unsafe acts leading to accidents and incidents. The study highlighted the significant role of inadequate equipment design and organisational influences in contributing to these incidents, suggesting that improvements in resource management, organisational climate, and processes are crucial for reducing accidents and incidents in the Australian rail system.

Similarly, "Analysis and Assessment of the Human Factor as a Cause of Occurrence of Selected Railway Accidents and Incidents" [10] utilises the HFACS to classify human factors in accidentology. This study focuses on analysing and assessing the role of human factors in selected railway accidents and incidents. By applying the HFACS framework, the authors identify various levels of human error, ranging from unsafe acts to organisational influences, providing a comprehensive understanding of how human factors contribute to railway accidents. The study emphasises the need for improvements in human factor management to enhance railway safety.

Other studies focus on the external causes of railway accidents. The paper "Statistical Analysis of the Railway Accidents Causes in Iran" [11] examines the types and frequencies of railway accidents, providing insights into the external factors contributing to these incidents. By analysing statistical data, it helps identify common accident types and potential preventive measures.

Our paper aims to propose a statistical study to assess the contribution of human conductors in avoiding railway accidents. As opposed to previous studies the idea is not to analyse the factors in human performances that could lead to an accident but more to analyse if the human driver has an impact on accident rate in operational context. That kind of study could be held by an analysis of human perception capabilities and braking distance studies. To the best of our knowledge, that specific kind of work doesn’t exist, even though some studies compare human performance with autonomous obstacle detection systems. For example, "Analysis of Human-Factor-Caused Freight Train Accidents in the United States" [12] uses human detection capabilities (e.g., detecting humans, with or without reflective vests, and vehicles) under various conditions (night and day) to benchmark the performance of the developed system. The human detection model referenced in this study originates from an internal paper by Deutsche Bahn (DB). Additionally, the paper "Automatisches Fahren" (DB's Kompass Project) [13] provides a reference for human detection range and probability but does not focus on the impact on the overall accident rate on the railway network.

There is a noticeable gap in the literature regarding the safety performance of human conductors from a railway network perspective. While existing studies provide references for human detection capabilities, they do not address the overall impact of human presence on accident rates. Our study aims to fill this gap by statistically analysing the contribution of human conductors to accident avoidance, potentially leading to a better understanding of their role in enhancing railway safety. This review highlights the importance of human factors in railway accident prevention and the need for comprehensive studies to evaluate the impact of human conductors on railway safety. By introducing a statistical approach and leveraging human perception and braking distance studies, our paper seeks to analyse the safety benefits of human conductors, contributing to the development of effective obstacle detection systems for trains.

3 Data presentation

A railway operator needs safety data in order to analyse the safety of their operations and find the weaker areas. Every year many of the European operators produce a safety report in which they use that data, like the one from SNCF [14]. Even though those reports are interesting in the safety analysis that is needed for an Obstacle Detection System (ODS), they tend to be very general and only present the conclusions. In order to be more precise, a safety analysis dedicated to ODS has to use raw data.

For this purpose, a primary railway operator provided raw accidentology data. It contains a list of every accident since January 2022 in an European region railway network. For each accident, the entries that will be used in this study are : Beginning time, End time, Line number, Kilometre, Classification, Ressources and Causes.

Unfortunately, no description was provided along with this set of data. As a consequence we will have our own definitions based on the terms that are used. A full list of those definitions can be found in the annexe of this document.

While all the values are free entries for the people who register the accidents, “Classification” is a field which is defined by a limited set of values. Three sets of incidents (Animals, Obstacle and Safety) are listed and sub-incidents are listed for them :

- Animals : hoist, hit wild animals
- Obstacle : Personal accident & Level crossing, tree, other, branch in the catenary, fall of rock, Collision with a bridge, fire on the outskirts or on the way, vehicle on the track
- Safety : stone throw & projectile shot, voluntary obstacle on the track, person on the track or suicidal

As for fields, classification categories don’t have any description, thus we will use our own.

4 Probabilistic definition

The objective of this section is to describe the world of events in a probabilistic way. The result will be a probabilistic definition of the problem and the hypotheses, as well as the expected result.

4.1 Definition of the world of events

Let us define five events :

I : There is an Incident on the railway (presence of an obstacle that requires a stop)
R : The driver detects the obstacle and Reacts soon enough to stop the train before the obstacle
A : There is an Accident (collision with the obstacle)
C : The train is driving in a Curve
N : The train is driving at Night

nX : The complementary event of X (e.g. nC = The train is not in a curve)

Moreover the set of moments of train runs can be divided in two ways : by the curve/straight-line condition and by the night/day condition.

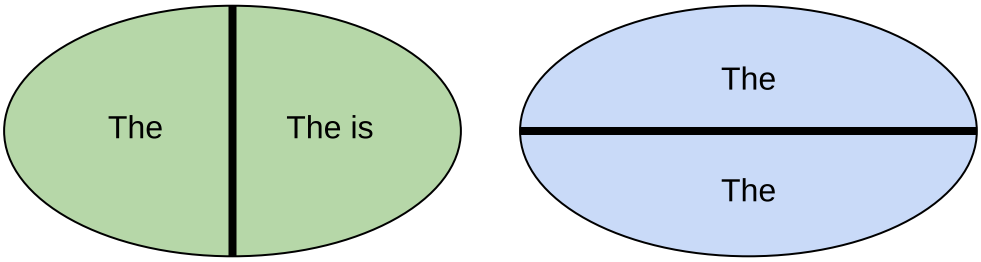

Figure 1: divisions of set of moments the train runs

In one of those two divisions, the event R, A and I are structured as such :

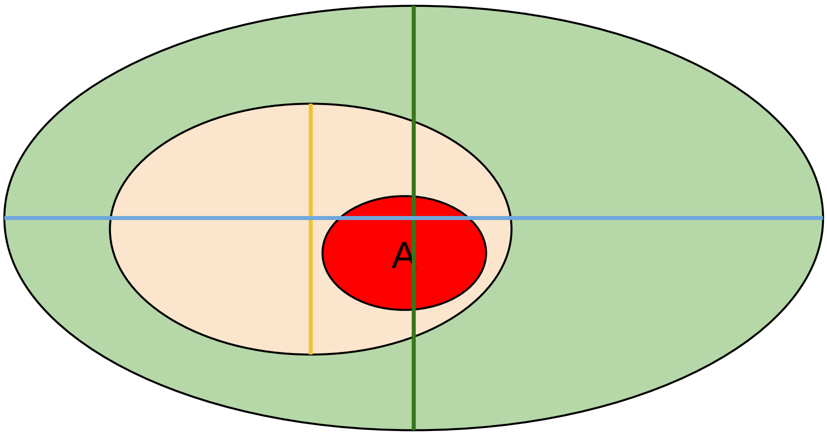

Figure 2: description of event inside the set of moment the train runs

At any moment the train runs, whether it's in a curve or not, at night or not, an incident on the railway can happen (I). In those incidents, the driver can react in time to brake (R) or not (nR), and when he doesn't react in time an accident can occur (A).

According to the DB work mentioned in the introduction, it seems that human abilities often don't provide a vision range long enough to brake in time to avoid any collision. **The goal of this paper is to quantify the impact of human drivers on accident rate.**

Let's define V as the event, "the driver has good visibility conditions", which means the train is not in a curve and not driving at night.

$$P(V) \ = \ P(nC \ and \ nN) \qquad (1a)$$

$$P(nV) \ = \ P((C \ and \ N) \ or \ (C \ and \ nN) \ or \ (nC \ and \ N)) \qquad (1b)$$

4.2 Hypotheses

The first hypothesis states that the incidents are homogeneously distributed.

$$P(I) \ = \ P(I|V) = P(I|nV) \qquad \textbf{(Hypothesis 1)}$$

The second hypothesis states that the only events that lead to an accident (A) are an incident (I) and the late reaction of the driver (nR) :

$P(A) = P(I \cap nR) = P(I) \times P(nR)$ (**Hypothesis 2**)

The third hypothesis applies to the drivers' ability to perceive the obstacles. It's considered that at night and in curves, it's impossible for the driver to brake before the obstacle in time. In curves, the obstacles is obstructed while at night the lights of the train don't enable the driver to see the obstacles more than 100-150m :

$P(nR|nV) = 1$ (**Hypothesis 3**)

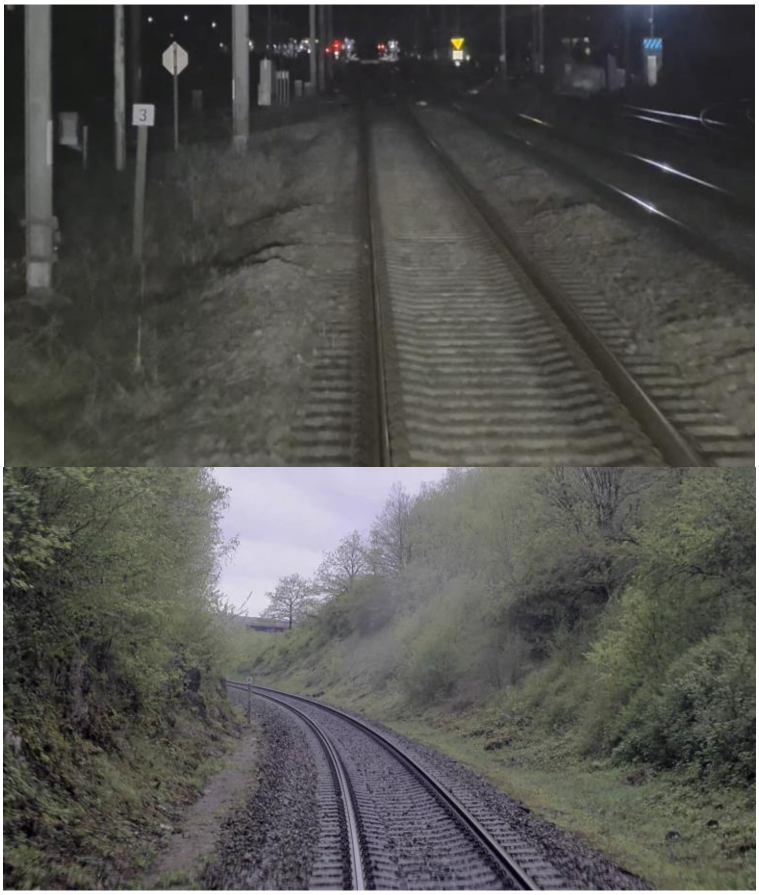

Figure 3 : Visualisation of the driver's perception at night and in curves, limited to 100-150m

4.3 Probabilistic usage of the results

The lack of solid numbers regarding human driver's capability of avoiding accidents is due to the difficulty to estimate the frequency of obstacles on tracks. While all accidents are well recorded, most of the non-accidents (near-accidents) go unrecorded. But to estimate human driver performance both sets of data would be needed.

In this work, we propose a novel approach to infer the frequency of obstacles on the track: we make the assumption that in low visibility conditions (e.g. at night and in turns), it's impossible for the driver to brake in time to avoid an accident with the obstacle (Hypothesis 3). Therefore all obstacles in low visibility result in accidents, which are recorded. This means that **the frequency of obstacles on the track is equal to the rate of accidents in low visibility conditions**. In other words

$$P(A|nV) = P(I|nV) \tag{2}$$

Given Hypothesis 1 (homogeneous distribution of Incidents), this can be generalised to:

$$\mathbf{P(I) = P(A|nV)} \tag{3}$$

Considering this, P(R|V), the probability of the driver reacting on time in good visibility conditions, can now be calculated as follow:

$$P(R|V) = 1 - P(nR|V) = 1 - \frac{P(I)*P(nR|V)}{P(I)}$$

$$= 1 - \frac{P(I|V)*P(nR|V)}{P(I)} \qquad \text{(applying hypothesis 1)}$$

$$= 1 - \frac{P(nR\cap I|V)}{P(I)} \qquad \text{(applying hypothesis 2)}$$

$$= 1 - \frac{P(A|V)}{P(I)}$$

Applying P(I)=P(A|nV), we get:

$$P(R|V) = 1 - \frac{P(A|V)}{P(A|nV)} = 1 - \frac{P(A\cap V)/P(V)}{P(A\cap nV)/P(nV)} = 1 - \frac{P(V|A)*P(A)/P(V)}{P(nV|A)*P(A)/P(nV)}$$

$$= 1 - \frac{P(V|A)/P(V)}{P(nV|A)/P(nV)}$$

We obtain :

$$P(R|V) = 1 - \frac{P(V|A)/P(V)}{P(nV|A)/P(nV)} \tag{4}$$

More precisely with this formula the distribution of accidents in good visibility is compared to the distribution of good visibility time and distance in the network. Given the three hypotheses, the difference of distribution in accidents, the impact of drivers on the accident rate will be estimated.

5 Curve and night analysis

This section aims to determine the difference of distribution of accidents with the distribution of curves and nighttime runs. To achieve this, we first calculate the proportion of curves and night runs in the studied network to obtain the proportion of conditions where the driver can’t see (P(nV)) and then the proportion of accidents occurring in those conditions (P(nV|A)). We use these proportions in the computation described in the last section. To compute night runs and curves the methodology described in the annex was used.

5.2 Analysis of the Region’s rail network

Once the conditions for every accident are determined, we estimate the distribution of these conditions in the overall train traffic in the region and during the period of the accident dataset.

We first identify the proportions of curved distance in the lines on which accidents happened (in the database). Then, this result is multiplied by the length of the line and added to the sum of those results for each line. The sum is divided by the total length of the lines studied in order to calculate the proportion of curve on all the networks. The formula used is the following :

$$\frac{\sum_{i} curveProportion(i)*lineLength(i)}{totalLineLength} \qquad (5)$$

After a visual exam, a threshold of 10 km for the radius of curvature, and 400 m for the distance between the central point and the two points taken to compute the radius of curvature, were chosen in order to conduct the analysis. The results show a curve ratio between 22.51% and 73.73% depending on the line. The weighted average curve proportion is 39.23%. With the events defined in section III : P(C) = 39.23% .

As of the night time runs, in order to have a basis of comparison the CEREMA (**C**entre for **S**tudies and **E**xpertise on **R**isks, the **E**nvironment, **M**obility and **U**rban **P**lanning) data was used [15]. Those data summarises the number of train.km per hour of the day for trains :

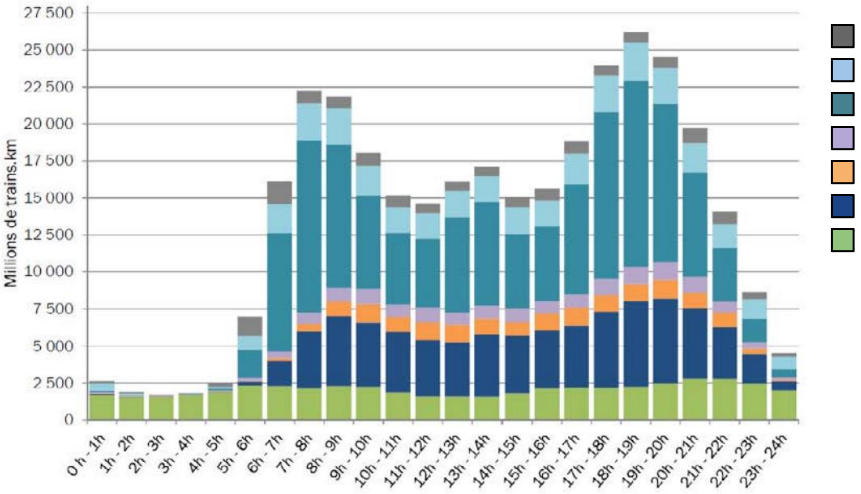

Figure 4 : Distribution of trains through the day on the national network

Using those data, the quantity of train.km before sunrise or after sunset is calculated for each day of the 2023 year. Then, the proportion of train.km by night is computed for the whole 2023 year. The result is a proportion 31.63% of train activity at night : P(N) = 31.63%. Let us consider the reasonable hypothesis that the proportion of curves travelled by trains is the same for night runs as during the day. It's then possible to compute the distance of train travelled for each condition. For example, nights runs in curved lines represent 39.23%*31.61% = 12.41% of the train's running conditions. The results for the other conditions are presented in the following table.

	Curves (C)	**Straight lines (nC)**
Night time (N)	12.41%	19.22%
Day time (nN)	26.82%	41.55%

Table 2 : Distribution of conditions for running trains

Considering that the only good visibility conditions are during day in straight lines we obtain :

$$P(V) = 41.55\%\ and\ P(nV) = 58.45\% \qquad (6)$$

5.3 Analysis of the accidents

This part intends to analyse and retrieve the accident distribution in all the conditions defined in the last part. Let us define the events :

W = accidents in Night time and Straight line
X = accidents in Day time and Curved line
Y = accidents in Day time and Straight line
Z = accidents in Night time and Curved line

For each accident those variables represent random variables following the Bernoulli law. For example, for the variable X, for one accident the variable takes 1 as a value if the accident happens during day time in a curved line, 0 either. The number of accidents available in the database is 921, unfortunately only 611 accidents are usable The remaining 310 are either lacking a kilometre or a line number, or the kilometre provided is inconsistent with the line length. We assume that these unusable points are equally distributed among the different conditions. The proportion of accidents in the different conditions is then computed.

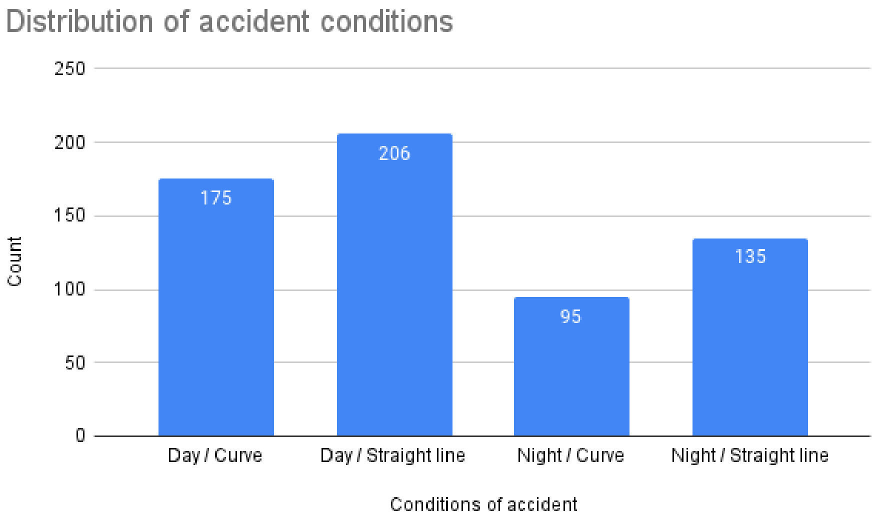

Figure 5 : Distribution of accident conditions.

Consequently, it's possible to obtain an estimation of the distribution of accident conditions.

	Curved line (C)	**Straight line (nC)**
Night time (N)	$p_Z{}' = 15.55\%$	$p_W{}' = 22.09\%$
Day time (nN)	$p_X{}' = 28.64\%$	$p_Y{}' = 33.72\%$

Table 3 : Distribution of the different accident conditions

Since X, Y, Z and W are Bernouilli random variable and the sample is big enough, it's possible to calculate confidence interval at a level of 95% for each of their population proportion p:

$p_W{}' = 22.09\% \quad p_W \in [18.80\%;\ 25.38\%]$ (7)
$p_X{}' = 28.64\% \quad p_X \in [25.06\%;\ 32.22\%]$ (8)
$p_Y{}' = 33.72\% \quad p_Y \in [29.97\%;\ 37.47\%]$ (9)
$p_Z{}' = 15.55\% \quad p_Z \in [12.68\%;\ 18.42\%]$ (10)

Since the only condition where the visibility allows the driver to see far enough to brake before the obstacle is on straight line during day times, the probability $P(V|A)$ correspond to p_Y:
$P(V|A) = 33.72\% \in [29.97\%;\ 37.47\%]$ (11)

Moreover :
$P(nV|A) = 1 - P(V|A) \Rightarrow P(nV|A) = 66.28\% \in [62.53\%;\ 70.03\%]$

We are now able to retrieve the information that is of interest for us : P(R)

5.4 Computation

In the previous sections we collected the following information :

$$P(V) = 41.55\% \; and \; P(nV) = 58.45\% \quad (12)$$
$$P(V|A) = \; 33.72\% \in [29.97\%;\; 37.47\%] \quad (13)$$
$$P(nV|A) = \; 66.28\% \in [62.53\%;\; 70.03\%] \quad (14)$$

We use (2) to compute a middle value and an interval for P(R|V) :

Mid P(R|V) :

$$mid(P(V|A)) = 33.72\% \; and \; mid(P(nV|A)) = 66.28\% \quad (15)$$
$$P(R|V) = 1 - \frac{P(V|A)/P(V)}{P(nV|A)/P(nV)} = 28.43\%$$

Max P(R|V) :

$$min(P(V|A)) = 29.97\% \; and \; max(P(nV|A)) = 70.03\% \quad (16)$$
$$P(R|V) = 1 - \frac{P(V|A)/P(V)}{P(nV|A)/P(nV)} = 39.80\%$$

Min P(R|V) :

$$max(P(V|A)) = 37.47\% \; and \; min(P(nV|A)) = 62.53\% \quad (17)$$
$$P(R|V) = 1 - \frac{P(V|A)/P(V)}{P(nV|A)/P(nV)} = 15.70\%$$

In other words, **drivers can avoid 28.43%** ∈ [15.70% to 39.80%] **of collisions** when there is an obstacle on the track in **good visibility** conditions. If we include all conditions, drivers avoid 28.43% * 41.55% = 11.81% of collisions.

6 Limitations

The analysis presented in this paper is prone to limitations.

First of all, one limitation is assuming that the visibility is equally impaired in all curves. First the visibility actually depends on the curvature. Secondly, the visibility also depends on the surrounding of the track: it will be very limited in a forest or mountain or trench area for example, but almost not affected on an open plain.

Another limitation is the descriptions of each category as well as the protocol which is used by agents in order to provide this information isn’t provided along with the database. This lack of clarity with the protocol can lead to a lack of precision for the data. Indeed, for the same accident two agents could provide different categories. Names such as “Obstacle / Personal accident & Level crossing” and “Safety / person on track" can be ambiguous and hard to differentiate in the absence of a detailed description. This ambiguity is present for other fields since there isn't any global protocol. It's possible to quote the "beginning time" of the accident which could be the time the accident happened or the time the set of actions that led to the action started.

Other low visibility conditions should be considered and not only night conditions. For example, in high intensity rain or snow or dense fog, the vision of the driver can be highly diminished. These conditions are however not that common in the studied region.

We also didn't take into account line speeds. When the train travels slower, its braking distance is shorter and the driver has more chance to avoid a collision. Lower speed lines may have more curves, but also less traffic.

Finally, the Cerema data used to compare the night conditions proportion to accidents happening at night, include high speed trains (HST) in the study. While some of the lines that are on display here have HSTs going through them, there are not any high speed lines (HSL) in the accident database provided. Consequently, it's possible that the number of HST going through the lines of the region studied is not representative of the proportion of HST studied in the CEREMA data.

7 Conclusions and Contributions

This paper aims to analyse accident data to evaluate driver performance under operational conditions. The findings suggest that drivers can reduce collisions with obstacles when visibility is good, but only by about 30%.

One explanation is that this can be attributed to the lower speeds of trains in certain parts of the regional network, where reduced braking distances allow drivers to stop the train before reaching an obstacle.

Further research should focus on more precise data to refine these conclusions, particularly regarding driver detection capabilities in different operational contexts (e.g. high-speed vs. regional trains). Additionally, obstacle avoidance is not the only driver action that can mitigate accidents. Even a delayed emergency braking can reduce the severity of an accident by lowering the collision speed and reducing the train's post-collision speed, thereby decreasing the risk of derailment.

For the safety certification of autonomous train systems, emphasis should be on a high integrity collision detection and an obstacle detection at longer ranges that is at least as good (GAME) as a human driver, in order to reduce speed at impact. But considering that human drivers can't avoid all obstacle collisions, expectations for autonomous trains should be set accordingly reasonably.

Future work should also address the assumptions made in this study. While hypotheses 1, 2, and 3 appear reasonable, they need to be validated through rigorous research. For hypothesis 3 (P(nR|nV) = 1), a survey of drivers could provide insights into obstacle detection in curves and at night. For hypothesis 2, a comprehensive database could confirm that P(I | V) = P(I | nV). Finally, hypothesis 1 requires quantification of events leading to collisions with obstacles to determine if factors other than human vision significantly impact the results observed. Future studies will be dedicated to validating these hypotheses.

References

[1] Deutsche Bahn, “Deutsche Bahn AG Digitale S-Bahn Hamburg: Erste hochautomatisiert fahrende S-Bahn im Fahrgastbetrieb”, 2021, doi: https://digitale-schiene-deutschland.de/Digitale-S-Bahn-Hamburg (accessed on 1 June 2024)

[2] M.A. Hadded, A. Mahtani, S. Ambellouis, J. Boonaert, H. Wannous. “Application of Rail Segmentation in the Monitoring of Autonomous Train’s Frontal Environment.” ICPRAI 2022, International Conference on Pattern Recognition and Artificial Intelligence, Jun 2022, Paris, France. pp185-197, ff10.1007/978-3-031-09037-0_16ff. Ffhal-03875603f

[3] M. Yusuf, A. MacDonald, R. Stuart, H. Miyazaki, “Heavy Haul Freight Transportation System : AutoHaul”, 2020, doi : https://www.hitachi.com/rev/archive/2020/r2020_06/06a05/index.html

[4] “CENELEC EN50126 : Railway Applications - The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS)”, 2015

[5] “CENELEC EN50128 - Railway applications - Communication, signalling and processing systems - Software for railway control and protection systems”, 2011

[6] “CENELEC EN50129 - Railway applications - Communication, signalling and processing systems - Safety related electronic systems for signalling”, 2018

[7] “UL4600 - Standard for Safety for the Evaluation of Autonomous Products”, 2022

[8] Melissa T. Baysari, Andrew S. McIntosh, John R. Wilson, “Understanding the human factors contribution to railway accidents and incidents in Australia”, Accident Analysis & Prevention, Volume 40, Issue 5, Pages 1750-1757, ISSN 0001-4575, 2008, doi : https://doi.org/10.1016/j.aap.2008.06.013.

[9] S. A. Shappell, D. A. Wiegmann “The Human Factors Analysis and Classification System- HFACS”, 2000, doi : https://commons.erau.edu/publication/737

[10] K. Gawlak, “Analysis and assessment of the human factor as a cause of occurrence of selected railway accidents and incidents” Open Engineering, Vol. 13 (Issue 1), pp. 20220398., 2023, doi : https://doi.org/10.1515/eng-2022-0398

[11] I. Bargegol, V. Najafi Moghaddam Gilani, M. Abolfazlzadeh, “Statistical Analysis of the Railway Accidents Causes in Iran (TECHNICAL NOTE)”, International Journal of Engineering, 30(12), pp. 1822-1830., 2017

[12] Z. Zhang, T. Turla, X. Liu, “Analysis of human-factor-caused freight train accidents in the United States”, Journal of Transportation Safety & Security, 13(10), pp. 1157–1186. , 2019, doi: 10.1080/19439962.2019.1697774.

[13] Deutsch Bahn, “Automatisches Fahren (AF) - Lastenheft Hinderniserkennung Fahrweg", 2003, doi: https://www.dzsf.bund.de/SharedDocs/Downloads/DZSF/Veroeffentlichungen/BMBF_Lastenheft_Hinderniserkennung_Fahrweg.pdf?__blob=publicationFile&amp;v=2

[14] SNCF, “Rapport annuel sécurité 2022”, 2022, doi : https://www.sncf-reseau.com/medias-publics/2023-12/sncfreseau_rapportannuelsecurite2022.vdefweb.pdf

[15] CEREMA, “Quel avenir pour les petites ligne”, 2020, doi: https://www.cerema.fr/fr/centre-ressources/boutique/quel-avenir-petites-lignes

Annex

1 Description of the categories and fields

Name of the event	Description
Animals / hoist	An hoist animal was hit by the train
Animals / hit wild animals	A wild animal was hit by the train
Obstacle / Personal accident & Level crossing	A pedestrian or a person in a vehicle was hit at a level crossing
Obstacle / tree	The train hit a tree
Obstacle / other	The train hit an obstacle of another category than those described in this document
Obstacle / branch in the catenary	The train hit a branch in the catenary
Obstacle / fall of rock	A rock fell on the running train
Obstacle / Collision with a bridge	The train hit a bridge
Obstacle / fire on the outskirts or on the way	A fire is present on the outskirts or on the way when the train pass
Obstacle / vehicle on the track	The train hit a vehicle on the track
Safety / stone throw & projectile shot	The train was hit by stone throw or projectile shot
Safety / voluntary obstacle on the track	The train hit an obstacle outside of the categories described here, voluntary placed on the tracks
Safety / person on the track or suicidal	The train hit a person on the track which could have been suicidal

Table 4 : description of the categories of accident

Name of the field	Description
Beginning time	Time at which the accident happened
End time	Time at which the line was cleared and the operations could restart without problems
Line number	Line number of the line on which the

	accident happened
Kilometre	The kilometre of the line on which the accident happened
Classification	The category of the accident (among the one in table 2)
Ressources	The nature of the obstacle
Causes	The cause of the accident

Table 5 : Description of the fields of the accidents

2 Methodology used to compute night runs and curves

There are a lot of parameters that can influence human obstacle detection such as, level of fatigue and attention, time of reaction, other tasks or speed of the vehicle. But above all these factors, visibility is, by far, the most important. Indeed, if an obstacle cannot be seen a human driver has no chance of avoiding it. Two parameters can alter the visibility of the obstacle, its obstruction and the luminosity level of the environment. The assumption is made that by far the main cause of obstruction is curvature of the railway and the main cause of low visibility is the lack of sunlight (e.g. night vs day).

Every accident in this dataset has a time of beginning, a time of ending, a line and a kilometre point. Using this information, the purpose of this section is to find out if the accident happened at night and in a curve.

In order to know if the accident was in a curve, its position relative to the line but also the other point of the lines have to be known. The first information is given by the accident dataset and the second is coming from the railway operator’s open datasets. The one used here is the dataset called “Ligne par type” which provides a list of every line of the network, their type and their points coordinates. For every point coordinates, a radius of curvature is calculated as well as the kilometre point. If the radius of curvature is over a threshold of 10 km we consider the point in a curve. To compute the radius of curvature of a point, the point itself is considered (P2) as well as the two farthest points under 400 m in both directions (P1 and P3). The point of intersection of the perpendicular bisector of the segment P2P3 and P1P2 is localised in a mercator projection and the distance between this point and the point of interest gives the radius of curvature.

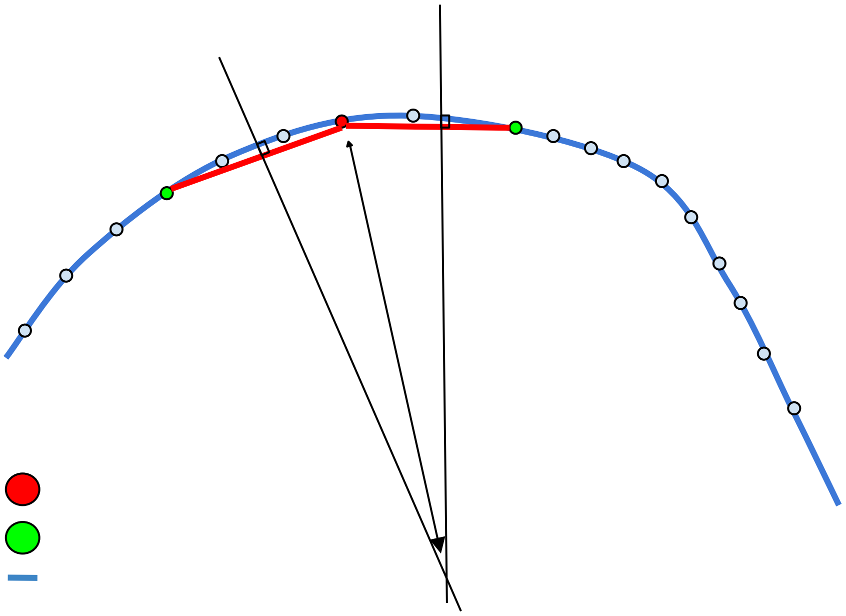

Figure 6 : Radius of curvature computation

The two thresholds of 400m and 10km were determined by a visual exam conducted on various tuples of thresholds. Once every point of the line was analysed, for every accident of the line, the farthest point before the accident is considered the point at which the accident happened and it's thus possible to know if an accident happened in a curve.

To determine if an accident happened at night, we look whether it happened before sunrise or after sunset. We get the sunrise and sunset times from the Suntime python library, using the day of the accident and the capital city of the studied region.

3 Confidence interval computation

Let's say we have a sample of n Bernouilli trial X1, X2,..., Xn and we want to build a confidence interval for the population proportion p.

Sample proportion :
The sample mean (sample proportion) p' is given by :

$$p' = \frac{1}{n} \times \Sigma Xi$$

Central limit theorem (CLT) :
For a sufficient large sample n, the sampling distribution of p' is approximately distributed with the mean p (since it's an estimator of p) and standard error $SE(p') = \sqrt{\frac{p(1-p)}{n}}$:

$$p' \sim N(p, \sqrt{\frac{p(1-p)}{n}}) \qquad (18)$$

Z-score for confidence Interval :
For a confidence level $1 - \alpha$, the corresponding z-score $z_{\alpha/2}$is used. For example, for a 95% confidence level, $z_{0.025} \simeq 1.96$

Confidence Interval :
The formula for the confidence interval for p is :

$$p' \pm z_{\alpha/2} \cdot SE(p') \qquad (19)$$

Substituting the standard error we get :

$$p' \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \qquad (20)$$

4 Frequency and proportion analysis

Without any post treatment on the lines, it's possible to use the data that was given in order to have an idea of the frequency and proportion of the different categories of accident.

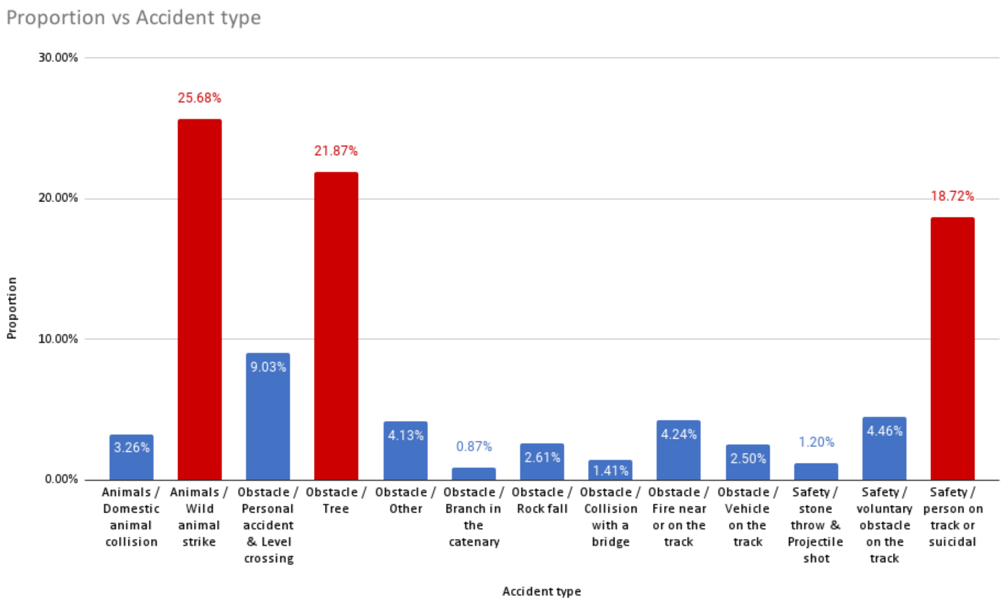

Figure 7 : Proportion vs Accident type graph

The proportion of accidents is represented for every accident class in the graph above. Three categories seem to appear more than the rest : Animals / hit wild animals, Safety /person on the track and Obstacle / Tree. Those results seem to be logical since the area crossed by trains can have a high density of humans or forest. Moreover, wild animals are not predictable and can act dangerously.

In order to compute the proportion of accidents we used the total number of them over a period of time. We can use the same logic in order to compute the frequency of accidents since we know the amount of time those data cover : from

January 2022 to January 2024 included. Besides, there are approximately 700 train trips a day in the region studied (information provided by the railway operator). We make the reasonable assumption that each train makes 5 trips a day (TrainTripPerDay) and each trip is 1.5 hours long (TrainTripDuration). With those information, we first compute the rate per year for each accident on the network :

$$AccidentRatePerYear \;= \frac{OverallAccidentCount}{YearCount} \tag{21}$$

Because there are 700 train trips a day and we made the assumption that each train makes 5 trips a day, there are approximately 140 trains (TrainCount) which are running constantly in the region's network. Let's compute the rate of accident per year per train then :

$$AccidentRatePerYearPerTrain \;= \frac{AccidentRatePerYear}{TrainCount} \tag{22}$$

The results for each accident class is presented below :

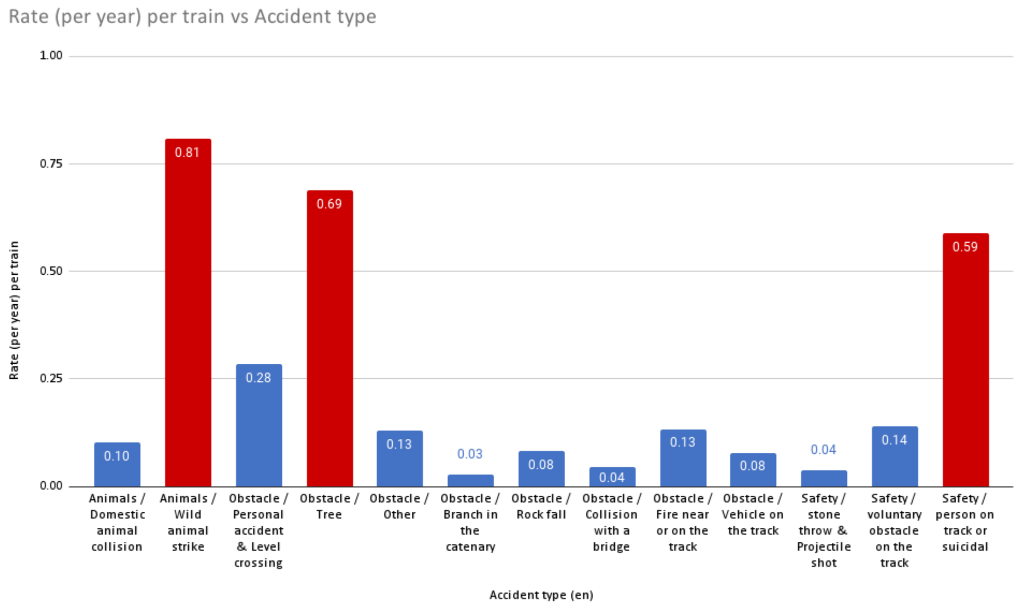

Figure 8 : Rate (per hour) on the network per train vs Accident type

Those results display that for the three most frequent accidents, the rate is between 1 accident every 1 or 2 years. In the annexe of the CENELEC EN50126 norm "Railway Applications - The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS)", this corresponds to an "occasional" accident (between 1 accident every 1 years to one accident every 10 years). If the severity of such an accident can be defined, this information can be used to measure the risk.