
Ahmed Lasisi, Doctoral Researcher,
Pengyu Li, Graduate Research Assistant
Nii Attoh-Okine, Professor of Civil Engineering

University of Delaware
127 The Green
301 DuPont Hall
Newark, DE, 19716

NUMBER OF WORDS: 4123

ABSTRACT
Railways are essential components of effective transportation in most advanced countries. It has a significant influence on national, economic and industrial activities. Despite these benefits, a railway disaster is often very catastrophic. In 2017 alone, the United States recorded an approximate loss of $300 million to train accidents. This loss is accompanied by over 500 casualties with approximately 50 people killed. The state of California has an estimated population of 40 million people with over $2.7 trillion Gross Domestic Product in 2017. Meanwhile, the rich landforms and complex environmental systems create even more complexities for train services and operations, including freight and passenger. From 2008 to 2017, train accidents in California resulted in a loss of $150 million, 563 injuries and 64 fatalities. In this work, the risks associated with trains are examined from several perspectives. California being a large state with big counties, accident data from 2008 to 2017 is examined in the light of probability of occurrence, financial losses, safety implications and risk criticality. The targeted stakeholders for this study include: the Department of Public Safety, the Federal Highway Administration, the Department of Transportation, and multiple railroad corporations. The study identified the significant types and major causes of train accidents in California with Failure Mode Effect analysis; builds a risk assessment framework using appropriate risk methods like spatial Monte Carlo simulations, event tree analysis amongst others to analyze severe accident causes. The analysis is concluded by recommendations to improve safety practices for identified failure modes.

INTRODUCTION
Railways are essential components of an effective transportation system in most advanced nations. It has a significant influence on national, economic and industrial activities. The United States currently boasts of at least 140,000 miles of main track which is significantly more than all train miles in most European countries if not all of Europe [1]. Railroads and development are often much correlated. This is obvious from the Gross Domestic Product (GDP) of advanced countries when compared to developing or underdeveloped nations. Despite these benefits, a railway disaster is often very ruinous and catastrophic. In 2017, the United States recorded an approximate loss of $300 million to train accidents. This loss is accompanied by over 500 casualties with approximately 50 people killed [2]. Independent sources however report higher figures for injuries (8500), accidents (11600) and fatalities (860) for the same year [3]. While number of accidents and injuries have been steadily decreasing for the past half-decade, the same cannot be said of fatalities. The United States has consistently struggled with very fatal train accidents and derailments for the past decade. The state of New York leads the list of states with the highest number of rail incidents followed by Illinois, Texas, California, New Jersey and Pennsylvania[4]. It is argued that majority of these accidents can be avoided with the implementation of positive train control [5].

However, the emphasis of positive train control is to prevent train accidents mainly due to speeding out of control and lack of awareness which are often due to human errors. While human accidents have been the highest cause of accidents according to primary cause category [2], we are able to show in this study that
Highway-rail grade crossings (HRGCs) are critical spatial locations for transportation safety where the derailments are more severe in terms of financial damage.

In this paper, authors investigate the risks associated with Highway-rail grade crossings (HRGCs) and train derailment in the United States while taking the state of California as a case study. Learning from FRA’s 10-year data in the attached Figure 1 and Table 1, it was found that HRGC accidents account for more than half of the fatalities/casualties (although most are highway users), train derailment cost more than half of the total financial loss.

![FIGURE 1. Derailment loss proportion per year (FRA, 2018b).](image)

**TABLE 1: Summary of Train Accidents in the United States over the past decade (FRA, 2018b).**

<table>
<thead>
<tr>
<th>S/No.</th>
<th>Major Cause Category</th>
<th>Likelihood</th>
<th>Damage ($)</th>
<th>% $ loss</th>
<th>% Derailment</th>
<th>Fatalities</th>
<th>Injuries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Equipment</td>
<td>0.119</td>
<td>454,726,164</td>
<td>15.7</td>
<td>70.7</td>
<td>0</td>
<td>266</td>
</tr>
<tr>
<td>2.</td>
<td>Highway Rail</td>
<td>0.087</td>
<td>164,862,648</td>
<td>5.7</td>
<td>N/A</td>
<td>343</td>
<td>2318</td>
</tr>
<tr>
<td>3.</td>
<td>Human</td>
<td>0.340</td>
<td>710,106,552</td>
<td>24.5</td>
<td>58.66</td>
<td>66</td>
<td>1536</td>
</tr>
<tr>
<td>4.</td>
<td>Miscellaneous</td>
<td>0.139</td>
<td>442,361,251</td>
<td>15.2</td>
<td>56.9</td>
<td>22</td>
<td>487</td>
</tr>
<tr>
<td>5.</td>
<td>Signal</td>
<td>0.023</td>
<td>26,854,560</td>
<td>0.9</td>
<td>49.6</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>6.</td>
<td>Track</td>
<td>0.291</td>
<td>1,105,297,448</td>
<td>38.1</td>
<td>95.07</td>
<td>7</td>
<td>715</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1</td>
<td>2,904,208,623</td>
<td>1</td>
<td>N/A</td>
<td>438</td>
<td>5329</td>
</tr>
</tbody>
</table>

Train Accidents by Cause from FRA Form F 6180.54

As a result, this study included two systematic approaches to address the safety and economic issues in separate facets. The first part of the study is to predict the potential severity if an accident happens at any given HRGC using accident records and existing inventory data extracted from multiple sources. The second part of this paper examines risk analysis in the broadest sense of transportation before diving into railway risks associated with infrastructure and operations from a probabilistic point of view.

**HIGHWAY-RAIL GRADE CROSSING ACCIDENT ANALYSIS AND PREDICTION**

Highway-rail grade crossings (HRGC) are critical spatial locations for transportation safety where the
highway and railroad tracks meet at the same elevation. It is an important source of safety concern to railway, highway authorities and the public at large. The accidents at HRGCs are often very catastrophic with serious consequences, such as fatalities, injuries, environmental disasters and extensive property damage [6]. From 2008 to 2017, 80.2% of fatal and 43.4% of injured casualties due to railroad accidents in the US are caused by HRGCs [7]. Due to safety reasons, HRGC accidents analysis has drawn considerable attention for decades. Effective accident analysis and prediction methodology is strongly desired to help transportation agencies and other stakeholders accurately predict and significantly reduce the casualties due to HRGC accidents.

Hybrid Machine Learning and GIS Model

In the broadest possible sense, there are two (2) types of machine learning: supervised and unsupervised (James et al., 2013). Unsupervised learning is indicative of a situation where observations in a sample space $\omega = s_1, ..., s_n$, take properties $x_1, ..., x_p$ (e.g. lane width, train speed, highway speed limit, visibility, gates, etc.) without any targeted response $y_i$ (e.g. HRGC accident (binary) or number of casualties). This type of learning is difficult to monitor despite some expertise about the subject matter. At best, some cluster-based learning or pattern recognition in similar groups can be identified without any definitive group assignment or labeling.

Supervised learning in HRGC accident analysis describes a problem whereby each element in the sample space $\omega = s_1, ..., s_n$ is assigned an associated label or measurement $y_i$. The learning goal is to create a model that attempts to characterize the response $y_i$ in terms of the explanatory variables. Mathematically, this is shown as:

$$y_i = f(X)$$

Where:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$x = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

This study emphasizes mostly on supervised learning with both classification and regression problems discussed in the next subsection.

Machine Learning and GIS

Geographic information systems (GIS) in transportation engineering need no detail introduction as can be found in relevant cited studies [8]. However, the profuse use of the GIS tool has taken the center stage in most spatial analytical studies. Because prediction problems are hard to visualize in spatial analytics, the need for geospatial data science or machine learning cannot be over-emphasized. Although, recent breakthroughs have been made towards integrating popular data science tools with GIS [9] & [10], this study aims to expand the implementation to HRGC accident analysis and transportation engineering by extension. The framework for this implementation is provided in the next section.

Research Framework

Formulation

Firstly, the analysis starts by collecting grade crossing inventory data from Caltrans, ten-year HRGC accident data from FRA and; Highway traffic and geometry data from FHWA and Caltrans respectively. Using data interpolation, the GIS layer that contains all possible parameters from multiple sources was created. The merged data is then spatially presented in GIS to identify hot-spots and possible leads. The next step is geospatial data analytics. In this phase, there are four sub-analyses; exploratory data analysis,
classification and regression problems on merged accident data, as well as prediction or forecasting casualties on future network data. The last phase of the analysis examined what the new predictions portends for relevant stakeholders through visual analytics in GIS.

**Classification**

Let \( X = x^{(1)}, x^{(2)}, \ldots, x^{(n)} \), be a set of \( n \) HRGC accidents from some network or municipality, and \( y \) is a related to response \( y = y^{(1)}, y^{(2)}, \ldots, y^{(n)} \). Given that each HRGC accident or \( x^{(i)} \) has \( m \) features; \( x^{m(1)}, x^{m(2)}, \ldots, x^{m(n)} \), e.g. gates, speed limit, train speed, visibility, etc. The objective is to predict \( y^{(i)} \) for every \( x^{(i)} \).

A classification problem arises when the responses, \( y^{(i)} \) is discrete and not continuous. A binary classification problem considered in this study aimed to categorize grade crossings based on the incidence of casualties in an HRGC accident within the past ten years. In other words,

\[
y = \begin{pmatrix}
0 - no \ casualties \\
1 - casualties \\
\vdots \\
\vdots \\
0 - no \ casualties
\end{pmatrix}
\]

Example of machine learning tools used for classification in this study include: logistic or binary regression, support vector classifier, gradient boosting classifier, etc.

**Regression**

Regression problem on the other hand is very similar to classification in that they both follow similar structure of equation 4, except that \( y_i \) is a "count-continuous" variable.

**Correlation Check**

After all the selected models have been trained in classification, a correlation of their predictions is conducted to examine the differences in the model performances.

**Cross-validation**

Instead of splitting the data into two, this study employs training with cross validation wherein data is split into \( k \) folds [11]. During training, each regression or classification technique is fitted onto \( k - 1 \) folds and the isolated fold is predicted. The choice \( k \) is usually 5 or 10 but the former was used in this study. The performance over all folds is then averaged. This performance could be accuracy, ROC score, etc. (for classification) or mean absolute error, root mean squared error etc. (for regression).

**Data Description**

As mentioned earlier, the railroad/highway inventory and accidents data were obtained from multiple sources including [2], [12], and [13]. A 10-year period (2008-2017) accident data sheet was exported from FRA as an important input for machine learning. According to FRA 5.14 section, a total of 1393 accidents occurred from 2008 to 2017 [2]. In this study, 1263 cases were analyzed because the HRGC inventory gathered from Caltrans was not up-to-date, which might result in some missing inclusions of newly-added HRGCs in GIS. In this analysis, authors employed at-grade for grade type because grade separations would not have vehicle-train interference. The input variables for 6,962 at-grade HRGCs were obtained from Caltrans as well, which was updated until Oct. 31, [12]. Other variables for road network were obtained from USDOT as part of the National Transportation Atlas Database (NTAD) geospatial files [14].

A summary of the variables used in this analysis are provided in Table 2 below. The table indicates that some parameters have not been considered in the machine learning model building. This is because these parameters are difficult to specify for any future HRGC accident. The HRGC accidents in California from 2008 to 2017 can also be spatially examined from Figure 3, used as a comparison for future predictions. While California has been selected for this study, similar analysis can be done for different states or any specific rail network of interest.
Assumptions and Limitations

The following assumptions and limitations apply to this study:

- All type of trains (freight/passenger) were assumed to run within the max allowed speed on tracks.
- For some crossings that do not have recorded speed limits, a speed of 40 mph was assumed as most of the HRGC seem to be on rural roads. Also, the speed limit data for each crossing was taken from the closest road segment that has a speed limit recorded. This assumption was necessary due to lack of up-to-date accurately recorded data.
- Because third-party information sources on train traffic were questionable and not easily accessible from FRA database, a conservative uniform traffic was not assumed across all crossings to allow for randomness.
- The HRGC inventory database is not up-to-date, so number of accident data dropped from 284,139 to 1,263 and the number of at-grade HRGCs was counted as 6,962 instead of 9,145, which is listed in FRA 8.05 section as of Jan. 2019.

Table 2: Variables and Descriptions

System-Action-Management (SAM) Analysis of HRGC accidents

The concept of SAM stems from the idea of solving problems from their root-causes before making top-level decisions [15]. The concept has been applied in several fields including but not limited to offshore safety, aerospace, maritime engineering, etc. [16]. In order to make decision recommendations in HRGC accidents, a qualitative SAM study was conducted and results are also presented to complement the quantitative analysis initially presented. It is a three-step process that involves:

- Basic Events like traffic signal, driver perceptions, braking etc.
- Decisions and Actions; like identifying human decisions, behaviors and actions.
- Organizational Culture which usually is the root causes for recurrent problems like HRGC.

Specific Highway Rail Crossing accidents in California involving double-digits casualties for the last decade were considered. Table 4 is a summary of the major attributes in each identified event. A summary of each accident can be obtained from [17] while the following observations can be concluded from the reports:

- Most severe HRGC incidents involve trucks and trailers. Are the highway vehicles too slow to maneuver or make decisions at Highway Crossings? Should long vehicles be mandated to stop at every grade crossing just like buses?
- The railroads report may be biased against the highway users who are often blamed for either inattentiveness or deliberately violating traffic signs. Were there obscure traffic lights or limited sight distances?
- What should be done to address highway user inattentiveness and how can a traffic sign violator be stopped? Create additional physical barriers or grade separation at hot spots?
- Report analysis shows that highway users are often traveling at low speeds while trains are at high speeds during the examined HRGC accidents. Therefore, should HRGCs be considered in positive train control (PTC) for automatic slowdown?
- Is work ethic (long shifts/hours) a major cause of truck driver attentiveness at HRGCs? What should truck companies and stakeholders do to address the potential issue?
- How can automated trucks improve this process?

While the above provides a qualitative outlook to severe HRGC accidents in this case study, a holistic synthesis has been provided in the discussion section.

**DISCUSSION I**

The machine learning algorithms considered in this study are able to accurately predict HRGC accidents and the corresponding number of casualties if any. The accuracy of the prediction can be as good as 98.9% with an ROC score of 0.98. A total of 15 explanatory variables, which includes crossing attributes, highway attributes as well as both train and motor traffic features were considered.

![Feature/Variable Importance](image)

**FIGURE 2. Parameter Importance for Classification Problem**

The analysis clearly identified train speed, AheadAADT and BackAADT as the most important predictors for HRGC accidents based on considered accident data as shown in Figure 2.
FIGURE 3. California HRGC accidents from 2008 to 2017 mapped based on casualty levels.

FIGURE 4: California HRGC Casualty Prediction for the next ten years (2017-2026)
The accident prediction results are presented in GIS map above (Figure 4). HRGC accidents are marked by colored solid circle on the map. The severity of the accidents are classified into low risk, moderate risk and high risk which are represented by green, yellow and red respectively (Table 3 and Table 4). The GIS map with prediction results provide an easy and visually appealing method to identify HRGC accident locations, hot-spots and corresponding severities for transportation authorities or other stakeholders. The results from casualty predictions over the next ten-years can assist infrastructure or welfare managers to prepare insurance plans, safety/capital investment programs based on well-thought out numbers. Should long vehicles be mandated to stop at HRGCs? These are follow-up questions that can be addressed based on the results of this study. Such information allows stakeholders to evaluate the safety of each HRGC and implement appropriate plan to reduce future occurrences.

By comparing the past 10 years of HRGC accident data and prediction results, a few points can be concluded. North California HRGC casualties is likely to reduce significantly if train speed are reduced at crossings (Figure 4). One way this can be achieved is through positive train control (PTC) [5]. However, a few HRGCs’ situations did not improved perhaps due to increased traffic. In these locations, grade separation is an option that can be closely examined. In some parts Southern California however, casualty predictions are also reduced especially at the north of Santa Ana. The working attributes in these locations can be observed and implemented in other locations. In Central California e.g., at the San Joaquin Valley, accidents and casualties are reduced in general. However, there are more high risk HRGC accidents predicted for Coalinga and Delano. This can be attributed to heavy truck activity around the region. These and similar regions present opportunities for automated or self-driving trucks whose safety promises are likely examined if the recurrent "highway-user inattentiveness" cause is to be addressed from FRA reports as evident from the qualitative SAM study.

A DATA-DRIVEN AND MONTE-CARLO RISK ANALYSIS FRAMEWORK FOR TRAIN DERAILMENT

In the following lines, as for the second part of the study, authors propose a work-flow for a train transportation risk analysis that has been summarized in Figure 5.

\[ R_i = C_j (w_1 \times F_j + w_2 \times S_j) \]

Where: \( i = \{1,2,3, \ldots, n\} \) for each failure mode/cause

C, F, S are defined in table 3 below for likelihood, financial and safety scales respectively, where values for \( w_1 \) and \( w_2 \) will be 0.3 and 0.7 respectively.
TABLE 3. Safety, financial and likelihood levels risk scales

<table>
<thead>
<tr>
<th>Safety Levels</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.12</td>
<td>1</td>
</tr>
<tr>
<td>&lt;0.75</td>
<td>2</td>
</tr>
<tr>
<td>&lt;1.25</td>
<td>3</td>
</tr>
<tr>
<td>Else</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Financial Damage Levels</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.12</td>
<td>1</td>
</tr>
<tr>
<td>&lt;0.75</td>
<td>2</td>
</tr>
<tr>
<td>&lt;1.25</td>
<td>3</td>
</tr>
<tr>
<td>Else</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely Remote (&lt;1%)</td>
<td>1</td>
</tr>
<tr>
<td>Remote (3%)</td>
<td>2</td>
</tr>
<tr>
<td>Likely (7.5%)</td>
<td>3</td>
</tr>
<tr>
<td>Very Likely (Else)</td>
<td>4</td>
</tr>
</tbody>
</table>

Safety risks are identified by a scale from 1 to 4. An average number 16 of 10-year casualties was calculated based on the total casualties from 2008 to 2017. To assign the scales for safety risks, divide the casualties by 16. The result falls into the categories of (<0.12, <0.75, <1.25, else), which stands for 1, 2, 3, and 4. A similar calculation method was used for financial risk, where the average loss is $3,067,682 per major cause. The occurrence was measured based on the fraction of the accidents that happened in each type of specific cause, as a divisor to the total accidents. The complete Failure Mode and Effect Analysis (FMEA) table is attached in Table 4.
TABLE 4. FMEA Table for Train Accidents

<table>
<thead>
<tr>
<th>Major Cause</th>
<th>Potential Failure</th>
<th>Probability Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakdown (ECC, EEC)</td>
<td>Train handling difficulties, stability to stop main ways required</td>
<td>2.242,192,358</td>
</tr>
<tr>
<td>Car body (G21, G22C)</td>
<td>Poor train dynamics</td>
<td>2.364,192,358</td>
</tr>
<tr>
<td>Coupler (GEC, EEC)</td>
<td>Train suspension or cars</td>
<td>4,543,192,358</td>
</tr>
<tr>
<td>Track (E44C, E4C)</td>
<td>Track train dynamics</td>
<td>6,734,192,358</td>
</tr>
<tr>
<td>Bearing (E55C, E5E)</td>
<td>Potential wheel failure</td>
<td>8,925,192,358</td>
</tr>
<tr>
<td>Wheel (E55C, E5E)</td>
<td>Wheel imbalance, fluid pumps, among several others</td>
<td>1.163,192,358</td>
</tr>
<tr>
<td>Locomotives (E71, E71L)</td>
<td>Poor traction, or loss of power, etc.</td>
<td>1.234,192,358</td>
</tr>
<tr>
<td>Locomotives (E71, E71L)</td>
<td>Sprung wheel mechanism malfunction, wheel failure etc.</td>
<td>1.345,192,358</td>
</tr>
</tbody>
</table>

**Probability Simulation**

The third level of risk analysis employed in this study is the event tree analysis which characterized the distribution of major accident causes. This was a more robust approach to handling the uncertainties in the major causes of derailment rather than just doing a min/max sensitivity analysis that does not take cognizance of multi-way variable interaction. Details of Probability/Monte Carlo simulation is provided in Table 5.
A Monte Carlo based event tree analysis approach is what has been conducted in this study, where each major accident cause is considered as an uncertainty node. And they are ordered based on decreasing likelihood of derailment as follows:

- Human
- Track
- Equipment
- Miscellaneous
- Signal

The distribution of the conditional probabilities are distributed based on the observed distribution from historical data. The expected risk values and expected consequences are then outputted at the far right of the tree (Figure 6).
FIGURE 6. Schematic of Event Tree Monte Carlo Simulation
DISCUSSION II

This second part formally described the work on Monte Carlo risk analysis framework for train derailments examining a case study of California over the past decade. To plan safety policies or implement short to medium term capital programs, railroads and stakeholders require a comprehensive outlook of associated risks. While simple risk analysis provide preliminary tools to assess risk, their results can often be misleading in the case of a base event tree. This therefore necessitated a simulation approach. It should be mentioned that this analysis considered a risk neutral approach. Otherwise, appropriate utility functions would have been employed to further reflect the risk attitudes of relevant stakeholders. Also, a limitation of this study is that the time value of financial damages over the past ten years did not take cognizance of inflation or discounting rates. This is to avoid unnecessary incongruity with FRA data especially because it is a continually updated database.

The use of risk analysis framework for derailment analysis employed in this study yielded the following important results: (1) Likelihood of major accident cause does not imply high severity/criticality. FMEA results show that derailment was the most critical in terms of financial loss while grade crossing accidents had the most safety implication even though human error resulted in more accidents than any other major cause category. (2) Monte Carlo simulation results revealed the distribution of expected derailment loss (financial) as a Johnson Su family with four parameters. This is a transformation of the normal distribution that is often unpopular among most statistical distribution studies [18] & [19]. This information should be considered when modeling future derailment losses.

CONCLUSION

Since over 5,000 miles of track in California is dedicated to freight, findings from this work would inform emerging markets of risk considerations for long and short term planning. This would go a long way towards improving the safety of rail operations through effective policy making. This work is a first step in incorporating Monte Carlo simulations in event tree analysis of train derailment at a large infrastructure level. Future work will analyze grade crossing accidents using machine learning integration of GIS-based spatial analysis to predict the risk profile of California High Speed Rail. Because Monte Carlo Simulation-based event tree derailment analysis has not been widely adopted, there is potential for an alternative risk management under this approach.

REFERENCES

LIST OF TABLES AND FIGURES

TABLE 1: Summary of Train Accidents in the United States over the past decade
TABLE 2: Variables and Descriptions
TABLE 3. Safety, financial and likelihood levels risk scales
TABLE 4. FMEA Table for Train Accidents
TABLE 5: Event Tree and Monte Carlo Simulation for California Train Derailments

FIGURE 1. Derailment loss proportion per year (FRA, 2018b).
FIGURE 2. Parameter Importance for Classification Problem
FIGURE 3. California HRGC accidents from 2008 to 2017 mapped based on casualty levels
FIGURE 4: California HRGC Casualty Prediction for the next ten years (2017-2026)
FIGURE 5. Proposed data-driven risk analysis framework
FIGURE 6. Schematic of Event Tree Monte Carlo Simulation
Introduction

• Railroads are good indicators of economic prosperity.

• In 2017, train accidents ≈ $300 million, while derailments ≈ $150 million (FRA, 2019)
California Highway Grade Crossing

- California Railroads operate ≈ 7000 miles of track (about 5000 miles freight)
- Approximately 7 million carloads of freight with 160 million tons (2011)
- >9000 at grade Highway Grade Crossings
Mapped Historical Data

- Most crashes happened in densely populated areas
- Multiple crashes took place in the same location during the past 10 years
California Highway Grade Crossing

- 160 per year, second to Texas
- 2000 HRGC Accidents
- ≈ 500 Trespass
- ≈ 200 HRGC

700 Fatalities Annually
California Highway Grade Crossing

≈ 500 Trespass
≈ 200 HRGC

700 Fatalities Annually

160 per year, second to Texas
2000 HRGC Accidents
Problem

• Characterize future accident risks based on data.

• Identify high safety and financial hazard causes
Problem

• Characterize future accident risks based on data.

• Identify high safety and financial hazard causes

<table>
<thead>
<tr>
<th>S/No.</th>
<th>Major Cause Category</th>
<th>Likelihood</th>
<th>Damage ($)</th>
<th>% $ loss</th>
<th>% Derailment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Equipment</td>
<td>0.119</td>
<td>454,726,164</td>
<td>15.7</td>
<td>70.7</td>
</tr>
<tr>
<td>2.</td>
<td>Highway Rail</td>
<td>0.087</td>
<td>164,862,648</td>
<td>5.7</td>
<td>N/A</td>
</tr>
<tr>
<td>3.</td>
<td>Human</td>
<td>0.340</td>
<td>710,106,552</td>
<td>24.5</td>
<td>58.66</td>
</tr>
<tr>
<td>4.</td>
<td>Miscellaneous</td>
<td>0.139</td>
<td>442,361,251</td>
<td>15.2</td>
<td>56.9</td>
</tr>
<tr>
<td>5.</td>
<td>Signal</td>
<td>0.023</td>
<td>26,854,560</td>
<td>0.9</td>
<td>49.6</td>
</tr>
<tr>
<td>6.</td>
<td>Track</td>
<td>0.291</td>
<td>1,105,297,448</td>
<td>38.1</td>
<td>95.07</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1</td>
<td>2,904,208,623</td>
<td>1</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Train Accidents by Cause from FRA Form F 6180.54
Framework

Data Acquisition, Merging, Preparation, Analysis
- FRA Safety Data, Caltrans GIS, etc.

Identified Major Accident Causes
- A. Highway Grade Crossings
- B. Derailments

Risk Presentation, Discussion
- Statistical Simulation, Spatial (GIS)

Multiple Accidents and Major Causes

Failure Mode, Effect and Analysis

Event Tree, SAM, Monte Carlo

Employed Appropriate Risk Analysis/Assessment Methods
Risk Criticality

• Risk criticality for train derailments is defined as:

\[ R_i = C_j (w_1 \times F_j + w_2 \times S_j) \]

Where \( i = \{1,2,3, \ldots, n\} \) for each failure mode

• \( C, F, S \) are likelihood, Safety, Financial damage
Scales for key factors

<table>
<thead>
<tr>
<th>Safety Levels</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.12</td>
<td>1</td>
</tr>
<tr>
<td>&lt;0.75</td>
<td>2</td>
</tr>
<tr>
<td>&lt;1.25</td>
<td>3</td>
</tr>
<tr>
<td>Else</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Financial Damage Levels</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.12</td>
<td>1</td>
</tr>
<tr>
<td>&lt;0.75</td>
<td>2</td>
</tr>
<tr>
<td>&lt;1.25</td>
<td>3</td>
</tr>
<tr>
<td>Else</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely Remote (&lt;1%)</td>
<td>1</td>
</tr>
<tr>
<td>Remote (3%)</td>
<td>2</td>
</tr>
<tr>
<td>Likely (7.5%)</td>
<td>3</td>
</tr>
<tr>
<td>Very Likely (Else)</td>
<td>4</td>
</tr>
</tbody>
</table>

- Safety risk scales were calculated by normalizing the casualties by 16 (Average for each type of failure mode).
- Financial risk scales were calculated by dividing the financial loss by $3,067,682 (Average loss for each failure mode).
- Likelihood were calculated by dividing the number of derailments for each failure mode by the total number of derailments.
### Failure Mode and Effect Analysis (FMEA) table based on the risk criticality

This helped to identify the most critical failure modes.
Improvements

• Distribution of Failure modes unknown
• Past doesn’t Imply Future (Uncertainties)
• Effect of accident characteristics
Monte-Carlo Event Tree Simulation

The distribution of the conditional probabilities are distributed based on the observed distribution from historical data.

Each major accident cause is considered as an uncertainty node. And they are ordered based on decreasing likelihood of derailment as follows:

- Human
- Track
- Equipment
- Miscellaneous
- Signal
Result Indications

• Helps understand the likelihood of major accident cause, does not imply high severity/criticality

• Expected derailment loss (financial) as a Johnson Su family with four parameters.

• Transformation of the normal distribution (Johnson, 1949)
More Problems: HRGCs

• Will HRGC accidents occur in the next time step (10 years)?

• Where? And how much casualties?

• What methods are data are required to answer these questions?
Approach

• Predict potential casualties for each location for the next ten years (rather than predicting where accidents might occur) - simplifying assumption

• Based on the data from past years, updated highway-rail grade crossing GIS data will be used in machine learning process to predict future risks for the same period
Framework Steps

1. Data Merging with HRGC ID
2. Visual Inventory of Ten-year HRGC Accident
3. Geo-spatial Data Analytics
4. GIS Implementation, Forecasting

Visual Presentation
## Data Sources & Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Input Values</th>
<th>Source</th>
<th>Use in Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casualties</td>
<td>Number of Casualties per Accident</td>
<td>Numeric</td>
<td>FRA</td>
<td>Yes</td>
</tr>
<tr>
<td>Visibility</td>
<td>Depends on daylight, time of day</td>
<td>Category</td>
<td>FRA</td>
<td>No</td>
</tr>
<tr>
<td>Track Type</td>
<td>Mainline or Yard, etc.</td>
<td>Category</td>
<td>FRA</td>
<td>No</td>
</tr>
<tr>
<td>Train Speed</td>
<td>Speed of the train in mph</td>
<td>Numeric</td>
<td>FRA</td>
<td>No</td>
</tr>
<tr>
<td>WD</td>
<td>Traffic Control Type (Gates, Flashers, or Passive)</td>
<td>Category</td>
<td>Caltrans</td>
<td>Yes</td>
</tr>
<tr>
<td>Transit_YN</td>
<td>Passenger/Freight Rail</td>
<td>Category/Binary</td>
<td>Caltrans</td>
<td>Yes</td>
</tr>
<tr>
<td>cracking_p</td>
<td>Pavement Cracking</td>
<td>Numeric</td>
<td>USDOT</td>
<td>Yes</td>
</tr>
<tr>
<td>rutting</td>
<td>International Roughness Index</td>
<td>Numeric</td>
<td>USDOT</td>
<td>Yes</td>
</tr>
<tr>
<td>speed_limit</td>
<td>Pavement Rutting</td>
<td>Numeric</td>
<td>USDOT</td>
<td>Yes</td>
</tr>
<tr>
<td>TrackClass</td>
<td>Track Class</td>
<td>Category/Numeric</td>
<td>Caltrans</td>
<td>Yes</td>
</tr>
<tr>
<td>FRT_SPEED</td>
<td>Freight Train Speed</td>
<td>Numeric</td>
<td>Caltrans</td>
<td>Yes</td>
</tr>
<tr>
<td>PASS_SPEED</td>
<td>Passenger Train Speed</td>
<td>Numeric</td>
<td>Caltrans</td>
<td>Yes</td>
</tr>
<tr>
<td>NUM_TRACK</td>
<td>Number of Tracks</td>
<td>Numeric</td>
<td>Caltrans</td>
<td>Yes</td>
</tr>
<tr>
<td>Back_AADT</td>
<td>Annual Average Daily Traffic (West/South of intersection)</td>
<td>Numeric</td>
<td>Caltrans</td>
<td>Yes</td>
</tr>
<tr>
<td>Ahead_AADT</td>
<td>Annual Average Daily Traffic (East/North of intersection)</td>
<td>Numeric</td>
<td>Caltrans</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Classification Problem

• Variables are counted as different parameters represented by unique values.

• The resulted fatalities/casualties (in this case combined as casualties as a whole because fatalities in accidents are also hard to predict) are noted by a binary outcomes.

• The classification would be used in machine learning tools such as support vector classifier and gradient boosting classifier.
Machine Learning (Spatial)

- Green, orange, and red dots represent low, moderate, and high risks respectively.
- Grade Separation in Increased Traffic Regions.
- Coalinga and Delano are likely to see high casualties.
- Thoughts on PTC.
Train speed and Annual Average Daily Traffic (AADT) are the two most important variables related to the risk associated with HRGCs.
Results Indications

• The map with hot spots of severe accidents can help transportation planner/Infrastructure managers to decide where to improve the safety of HRGCs.

• It indicates a few influential variables that could be further studied to help mitigate the increasing risks at HRGCs (i.e., freight train traffic frequency in day time peak hours, tunnels/bridges to avoid at grade interferences)
Limitations

- Some data (Speed limit, Annual Average Daily Traffic) was assumed for part of the crossings because the way how data was collected from different sources does not have a certain standard.
- All Trains run at maximum allowed speed for the analysis.
- There were 1,394 accidents as gathered from the reports; however, only 1,263 was mapped because of the outdated inventory data.
Concluding Remarks

• High Likelihood of major accident cause does not imply high severity/criticality (FMEA)

• Maximum Likelihood approach FRA WBAPS...is not spatially valid. Attribute/Feature Learning can improve future predictions
Concluding Remarks

• High Likelihood of major accident cause does not imply high severity/criticality (FMEA)

• Maximum Likelihood approach FRA WBAPS...is not spatially valid. Attribute/Feature Learning can improve future predictions
Thank You!
Any Questions?