CAIT-UTC-014

Railroad Operations Research and Training

FINAL REPORT
July 2014

Submitted by:

Trefor Williams
Professor

Christie Nelson
Post Doctoral Associate

William Pottenger
Associate Research Professor

Department of Civil & Environmental Engineering, and
Department of Computer Science
Rutgers, The State University of New Jersey
Piscataway NJ 08854

External Project Manager
Dr John Betak,
Collaborative Solutions LLC

In cooperation with

Rutgers, The State University of New Jersey
And
State of New Jersey
Department of Transportation
And
U.S. Department of Transportation
Federal Highway Administration

Disclaimer Statement

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CAIT-UTC-014		

4. Title and Subtitle	5. Report Date
Railroad Operations Research and Training	July 2014
	6. Performing Organization Code CAIT/Rutgers, The State University of New Jersey

7. Author(s)	8. Performing Organization Report No.
Trefor Williams, Christie Nelson, John Betak, William Pottenger	CAIT-UTC-014

9. Performing Organization Name and Address	10. Work Unit No.
Department of Civil & Environmental Engineering, and Department of Computer Science Rutgers, The State University of New Jersey Piscataway NJ 08854	11. Contract or Grant No. DTRT12-G-UTC16

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Center for Advanced Infrastructure and Transportation Rutgers, The State University of New Jersey 100 Brett Road Piscataway, NJ 08854	Final Report 8/1/12-9/30/13
	14. Sponsoring Agency Code

15. Supplementary Notes
U.S. Department of Transportation/Research and Innovative Technology Administration
1200 New Jersey Avenue, SE
Washington, DC 20590-0001

16. Abstract

This research is necessary to address training and research needs for railroads. Very few institutions provide instruction in railroad engineering, operations or management. With increasing government regulation there is a need for Class I railroads, short lines, and mass transit systems to provide training in a variety of areas including asset management, and maintenance systems to maintain a state of good repair. In this research, we propose to explore the possibilities for railroad education leading to the implementation of short courses for professionals and ultimately to develop courses for undergraduate Civil Engineering curriculums. The major railroads have recognized the need for additional training and are actively promoting the establishment of additional educational offerings. The goal is to seek funding from railroads, mass transit systems, and governmental agencies to implement a training program that promotes the improved rail operations and maintenance.

17. Key Words	18. Distribution Statement
Pavement; Computer-Visions; GIS; Pavement Monitoring; Bidimensional Empirical Mode Decomposition	

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	26	

Contents

Introduction

Grade crossing accidents are a major problem for the U.S. railroad industry. There were 232 fatalities and 943 injuries at railroad grade crossings in 2013 (Operation Lifesaver, 2014). Considerable progress has been made in reducing grade crossing accidents with collisions between trains and automobiles reduced from a high of approximately 12,000 in 1972 to about 2,087 incidents in 2013 (Operation Lifesaver, 2014). Research in grade crossing accidents has concentrated on methods to prioritize grade crossings using statistical and probabilistic methods (Chadwick et al., 2014, Ogden 2007). Recent advances in computer science in the areas of data mining, data visualization, and text mining have made it possible to consider other techniques to better understand the factors involved in grade crossing accidents, and to develop more accurate methods of identifying grade crossings requiring safety upgrades. In addition, new text mining algorithms can extract information and trends from text fields in a database that was previously difficult to comprehend (Blei, 2012). For example, text mining can be used to identify accidents where pedestrian trespassing is discussed in the accident form comment fields.

The Federal Railroad Administration (FRA) Research Needs Workshop on Highway-Rail Grade Crossing Safety and Trespass Prevention identified the 33 top research needs for grade crossing research (Carroll et al., 2010). Included in the list are cost/benefit analyses of grade crossing improvements, review and improvement of prediction formulae, and collecting and analyzing pedestrian trespassing data. These are all areas where emerging techniques in data mining and visualization can be used to improve existing statistical methods. It is the purpose of this paper to suggest and demonstrate new computer-based methods to better understand and visualize grade crossing data. The paper will discuss computer models that can be implemented to produce insights that can augment the existing prioritization techniques, and stand-alone models that can be used to identify potentially dangerous crossings.

Considerable data are collected about railroad grade crossings. The FRA maintains a database of railroad grade crossing accidents that provides extensive details about the type of track where accidents occurred, the installed warning devices, and the

severity of the accident. This database also contains text in the form of comment fields about the accident. This information is reported to the FRA by railroads on standard accident reporting forms. The FRA also maintains a second database, which is the inventory of all of the grade crossings in the United States. This database is useful because it contains information on both highway traffic and the frequency of trains at each grade crossing. Each grade crossing has a unique identification number assigned by the FRA. Therefore, it is possible to merge data from the two databases to build accident prediction models and visualizations.

New advances in computer science allow new relationships in the grade crossing databases to be found. Data and text can be mined for information that can highlight previously unknown relationships in the data. Grade crossings can potentially be classified, using data mining techniques, based on data in the grade crossing databases. With new text mining techniques it is now possible to extract useful information from the text fields in the accident database to relate specific words and terms to accident characteristics.

In particular, novel Higher Order Learning techniques can be utilized to potentially gain more understanding about these grade crossing accidents, especially if there are a small number of instances per class (e.g. per type of accident). Traditional machine learning techniques assume that instances are IID (independent and identically distributed). These traditional methods can be thought of as "zero-order" as they do not leverage relationships between attributes. However, Higher Order Learning utilizes these relationships between attributes and features across instances. One of the thrusts of this research is to evaluate if these Higher Order techniques can be successfully applied to the grade crossing data domain. It has already been shown to work well in several other applications in the statistical relational learning field. In particular, we will focus on a variant of the commonly used Naïve Bayes, called Higher Order Naïve Bayes (HONB), and a variant of Latent Dirichlet Allocation (LDA), called Higher Order Latent Dirichlet Allocation (HO-LDA) (Ganiz et al., 2011; Nelson et al., 2013).

The emergence of powerful data visualization tools now allows complex data sets to be examined and analyzed. Data visualizations also have an advantage in that they

allow people without extensive statistical or computer backgrounds to examine complex relationships between data.

Potentially, data mining, text mining and text visualization can be used to improve and augment existing hazard models. The use of these techniques can aid in the identification of "hotspots" that are prone to accidents, and to better allocation of grade crossing improvement funding. It is the purpose of this paper to discuss some preliminary models and results that have been produced using text mining, data mining and data visualization to analyze the grade crossing accident database. The paper will also suggest ways these initial analyses can be expanded to produce additional insights.

Survey of Current Grade Crossing Research

Research in grade crossing accidents has focused on methods of prioritizing grade crossings for improvement. Several statistical models have been developed to select grade crossings for upgrades. The most well known models assign a probability of an accident to individual crossings. The crossings with the highest probability of an accident are selected for further study and possible safety improvements. Many statistical methods have been applied to grade crossing accident prediction. They include hazard indexes, liner regression methods, Poisson regression, binomial regression, and logit models (Ogden 2007, Chadwick et al. 2014).

The methods available to rank grade crossings are relative formulae use crossing data to rank the relative hazards at each crossing, so that improvements can be prioritized from most dangerous to least dangerous crossings. Absolute formulae use statistics and probability to predict the number of collisions expected to occur at each crossing over a certain time period, allowing for estimation of the number of lives saved by upgrading a crossing (Chadwick et al. 2014). These statistical methods include consideration of the volume of highway and railroad crossing at the crossing. A heavily used crossing with the same features as a low use crossing would tend to rank higher in these models.

U.S. DOT Accident Prediction Model

The most widely used accident prediction model is the U.S. DOT accident prediction model (Chadwick et al. 2014). This formula consists of three primary equations:

$$a = (K)(EI)(DT)(MS)(MT)(HP)(HL)(HT) \quad (1)$$

$$B = \frac{T_o}{T_o + T}(a) + \frac{T}{T_o + T}\left(\frac{N}{T}\right) \quad (2)$$

$$A = \{0.7159B\} \quad \text{For passive devices}$$
$$A = \{0.5292B\} \quad \text{For flashing lights}$$
$$A = \{0.4921B\} \quad \text{For gates (FRA, 1987; FRA, 1999)} \quad (3)$$

where a = the initial collision prediction, collisions per year at the crossing; K = formula constant; EI = factor for exposure index based on product of highway and train traffic; MT = factor for number of main tracks; DT = factor for number of through trains per day during daylight hours; HP = factor for paved highways (yes or no); MS = factor for maximum timetable speed; HT = factor for highway type; HL = factor for number of highway lanes; B = adjusted accident frequency value; T_0 = formula weighting factor; =1.0/(0.05 + a); N = number of observed accidents in T years at a crossing and A = normalized accident frequency value.

A table provides each of the factors, for crossings with passive controls, flashing lights, and gates (Austin and Carson 2002, Chadwick et al. 2014, Farr 1987, Ogden 2007). The multiplicative factors in Equation (1) are crossing characteristics maintained in the grade crossing inventory. These factors were found to be statistically significant in the prediction of accidents at highway-rail crossings using nonlinear multiple regression. Note that some important characteristics, such as sight distance, are not included in Equation 1. Factors like sight distance are unavailable in FRA's highway-rail crossing inventory.

Other Models in the Grade Crossing Literature

Considerable research has gone into statistical models to provide better predictions than models like the U.S. DOT method due to problems of prediction accuracy with the U.S. DOT method. In this section some example are discussed. Saccamanno et al (2007) have used the results of previous studies to develop a statistical method to estimate the effects of countermeasures to reduce collisions at grade crossings. They studied Canadian grade crossing data and found that lifting whistle bans reduced collisions by a larger percentage then by upgrading flashing lights to gates. Konur et al. (2013) have discussed the use of knapsack models from operations research to allocate resources for grade crossing improvements. Austin and Carson (2002) developed an alternate highway-rail crossing accident prediction model, using negative binomial regression. This technique allows for the interpretation of both the magnitude and direction of the effect of the factors significantly influencing highway-rail crossing accident frequencies. Oh et al (2006) applied the gamma probability model to model accident probabilities.

Literature About Causes of Grade Crossing Accidents

Some research has focused on factors that cause grade crossing accidents. Schartung et al (2011) have performed an interesting study of trends in grade crossing safety. Their analysis found that, between 2006 and 2011, there is an increase in the number of open crossings, an increase in the rate of accidents for HAZMAT carrying vehicles, and an increase in Amtrak injury and fatality rates. Analysis of accidents also indicates that accidents are not distributed equally geographically and that five of the top 20 states for accidents accounted for 65 percent of the accidents in that group. The study also notes that the severity of accidents has increased. They recommend that individual "hotspots" be identified for remediation.

Clarke and Loeb (2004) have developed econometric models of the determinants of grade crossings using seemingly unrelated regression. They include macroeconomic variables in their studies. They found that accidents involving trespassers, employees, and passengers were significantly correlated with alcohol consumption. Heavy passenger and freight train usage are also significant factors in grade crossing accidents. Crossings with

passenger train traffic were found to have a significant increase in fatalities. The authors also found that expenditures on crossing safety were statistically significant effect on reducing grade crossing and trespassing fatalities. Hu and Lin (2012) have studied grade crossing accidents in Taiwan. They have found that accident rates are highly correlated with the sight distance provided to the locomotive engineer. Oh et al. (2006) found that crashes in Korea were observed to increase with total traffic volume and average daily train volumes. The proximity of crossings to commercial areas and the distance of the train detector from crossings are associated with larger numbers of accidents, as is the time duration between the activation of warning signals and gates.

Limitations of Statistical Models

Statistical models for grade crossing accidents have some limitations. It has been noted in the literature (Austin and Carson 2002) that the U.S. DOT model may not always yield accurate rankings. Other models described in the research literature require extensive statistical knowledge to implement and execute. Several examples are described below that show the potential for data mining and data visualization to provide alternative techniques to analyze grade crossing safety. In particular, data visualization provides a way to understand the multi-faceted data in the accident database without needing to employ statistical models.

Proposed Models with Applications to the Grade Crossing Data

Several models have been developed to show the possibilities of using data mining, text mining and data visualization. The data used were all grade crossing accidents in the period from 2009-2013.

In addition to the results shown, several other machine-learning algorithms are also proposed. In particular, early results show that topic modeling performs well on grade crossing accident data. Latent Dirichlet Allocation (LDA) is a machine-learning algorithm primarily designed to build a model on text data. In some cases, Higher Order LDA (HO-LDA) has been shown to outperform LDA, especially when there is small

class representation. In addition, another algorithm, Higher Order Naïve Bayes (HONB) can build a model on the data.

Text Mining

With the rapidly increasing power of computers, new possibilities exist for analyzing complex data sets like grade crossing accident data, to better understand relationships between grade crossing safety treatments, highway characteristic, railroad characteristics, and accidents. In particular, it has been difficult to analyze textual data, such as comment fields in accident reports. New computer algorithms now make it possible to automatically extract meaningful information from these texts.

Basic Text Analysis

The purpose of text mining is to transform text into numeric attributes that can then be used in data mining algorithms. Text mining is often defined in the context of discovering previously unknown information that is implicit in the text but not immediately obvious. Using the RapidMiner data mining software, the text fields of the grade crossing accident database have been processed to determine the frequency of words and word pairs for each FRA track grade. The text is processed in the following way:

1. The texts were modified so that there were only lower cases letters
2. The unstructured text is transformed into a sequence of tokens. Tokens can take different forms; however in this model the tokens were equivalent to single words.
3. Common words such as “and” and “but” are removed.
4. Words that are less then five and longer then 35 characters long were removed from the word list.
6. “Generate n-gram terms” was used to allow for word pairs to also be considered.

Table 1 shows the frequency of some important words from the grade crossing data for different FRA track grades. The table shows that the word “pedestrian” occurs in accident narratives much more frequently for high-speed track. The word pair “Failed_Stop” is more evenly spread across all track grades. It indicates that the “Failed to Stop” accident type occurs on both low speed and high-speed tracks.

Table 1. Word Occurrences by FRA Track Grade

Word	Total Occurences	Document Occurences	Track Grade 1	Track Grade 2	Track Grade 3	Track Grade 4	Track Grade 5	Track Grade 6	Track Grade X
Tractor_Trailer	178	158	24	21	44	77	8	0	4
Locomotive	421	374	61	60	92	174	8	2	23
Locomotives	139	136	6	9	13	108	3	0	0
Automobile	208	185	19	20	47	110	9	1	2
Track	518	421	73	82	136	204	15	0	8
Tracks	283	257	25	37	68	141	6	0	6
Pedestrian	160	135	3	16	33	94	14	0	0
Main_Track	145	139	8	25	32	72	8	0	0
Failed_Stop	239	239	41	47	69	69	3	0	9
Crossing_Gates	137	136	20	4	30	67	16	0	0

This basic analysis of the text yields some useful insights. An extension of this basic analysis would be to combine the text with other available data in the accident database to produce predictive data mining models.

Topic Modeling

Topic modeling algorithms are statistical methods that analyze the words of original texts to automatically discover the themes that run through them (Blei, 2012). A topic model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics.

Analysis of grade crossing accident comment fields with LDA

A frequently used topic-modeling algorithm is Latent Dirichlet Analysis (LDA). In this example, a labeled LDA model has been constructed to show the words associated with one, two, three and four crossing user fatality accidents at grade crossings (No accidents with more then four user fatalities were found in the database). This analysis yields interesting results showing single fatality accidents are frequently associated with words indicating that pedestrians and bicyclists are often involved in single fatality accidents and that trespassing near the crossing by pedestrians is a serious problem. Figure 1 shows the main words associated with the different numbers of fatalities with single user fatality accidents highlighted in red. This initial example shows how labeled LDA can be used to extract useful information about the nature of grade crossing accidents from the text fields of the database.

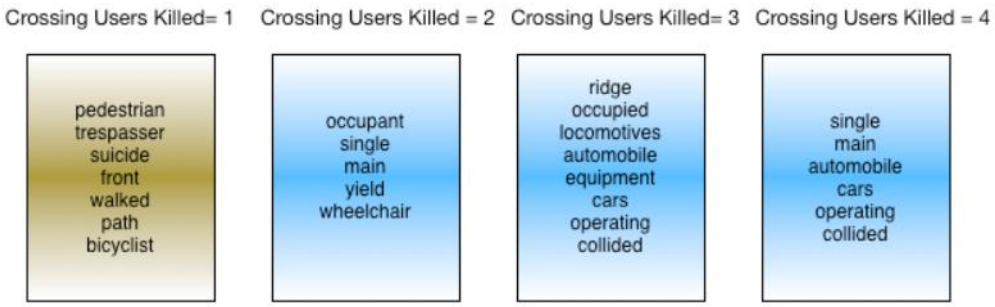

Figure 1. Labeled LDA Analysis of User Fatalities

Higher Order LDA

In addition to using LDA to build a model on the data, HO-LDA can also be utilized. HO-LDA utilizes relationships between attributes across instances, and has been shown in other applications to perform well when there is small class representation.

Figure 2 illustrates Higher Order paths. In particular, Figure 2 shows three sample instances, V_1, V_2, and V_3. Instance V_1 has two attributes, attributes x_1, and x_2, instance V_2 has two attributes (x_2 and x_3), and instance V_3 has two attributes (x_3 and x_4). Traditional machine learning does not leverage the latent higher order paths. However, Higher Order techniques use these higher order paths to create a link between attributes. In this example, attributes x_1 and x_4 are linked by leveraging the higher-order paths between attributes V_1, V_2, and V_3.

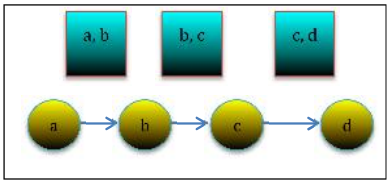

Figure 2. Illustration of a higher-order path (red boxes are documents and letters represent words)
(Nelson, C., Keiler, H. and Pottenger, W. M. (2013))

Unlike approaches that assume data instances are independent, HO-LDA leverages relations between feature values across different instances. Additionally, this framework can be generalized using a data-driven space transformation that allows any supervised discriminative classifier operating in vector spaces to take advantage of these Higher

Order relations. The utility of this transform has also been established in supervised generative algorithms including Higher-order Naïve Bayes.

The objective of this aspect of the proposed effort is to incorporate Higher Order information into the framework of LDA. For the HO-LDA algorithm that will be used from (Nelson et al., 2013), the Gibbs-sampling formula of LDA replaced feature frequencies in topics with their Higher Order frequencies. In other words, in equation (4) these counts are replaced with Higher path counts, $c_{i,j}$, for the feature *w* in topic *j*. $c_{i,j}$ is computed as follows, assuming that the input to this algorithm is a set of documents, each being labeled with a topic index:

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}.$$

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}.$$

(4)

1. Partition each instance into sets of entities E_1, E_2, …, E_k according to the topics that are assigned.
2. Each topic *j* now has a corresponding set of partial communications. The Higher Order path counts $c_{i,j}$ are computed exactly the way they are computed for Higher Order Naïve Bayes, the classes corresponding to topics.

Higher Order Naïve Bayes

As shown in Ganiz, et al., (2011), HONB has been shown effective in other studies to outperform Naïve Bayes due to its leveraging of Higher Order information. Naïve Bayes is a traditional machine learning classifier, which is based on Bayes Theorem:

$$P(A|B) = P(B|A)P(A) / P(B) \quad (5)$$

Naïve Bayes assumes strong "naïve" independence, and that the absence (or presence) of a particular attribute is unrelated to the absence (or presence) of any other attribute. Assuming independence, as Naïve Bayes does, sometimes makes a good estimation challenging, especially when there is a small amount of training data, as seen

in Nelson, C., Keiler, H. and Pottenger, W. M. (2013). Similar to HO-LDA, HONB utilizes relationships between attributes values across instances.

To utilize HONB for text classification, the text is pre-processed in a similar manner to the RapidMiner (upper and lowercase text not distinguished; words were stemmed by removing common endings such as “s”, “ing”, etc.; numeric values removed along with stem words of a “too short” length; then words were indexed and each word given an index number, using these index numbers to transform each instance).

Data Visualization

Visualization is the graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents. User understanding may involve detection, measurement and comparison, and is enhanced via interactive techniques. Data visualization aids user understanding by providing the information in multiple views. The FRA grade crossing accident data contain numerous interrelated variables. Understanding of how the variables are interrelated can be enhanced using modern visualization techniques. Data visualization can allow managers from railroads and government agencies to understand the complex relationships between variables without the need for complex statistical analyses.

Application of Data Visualization to Grade Crossing Accidents

This section shows several visualizations that we have developed using the grade crossing accident database. We used the *Weave* data visualization software to construct several examples (Institute for Visualization and Perception Research, 2014). A dashboard of linked histograms of data from the grade crossing accident database was constructed. The bar in each histogram can be selected and the linked factors in the other histograms are highlighted.

Figure 3 shows a data visualization where single fatality user accidents are analyzed. Each bar chart shows the number of cases for a data field from the database. In the bottom right hand histogram, Type of crossing user K, Pedestrian, has been selected and the other bar charts now show highlighted bars that show the distribution of single user fatalities, the type of train, the train speed, and types of crossing protection.

The visualization in Figure 4 shows accidents where the crossing user did not stop. These accidents are predominately at crossings where there are only cross bucks or stop signs. In the visualization, the "Action of Highway Users" with an entry of 3 is selected. In the grade crossing database, 3 designates a "Did Not Stop" user action. It can be seen from the visualization that this type of accident occurs most frequently at crossings with only cross bucks or crossings with both cross bucks and stop signs. These are Types 7 or 8 in the crossing protection fields. It can also be seen that truck-trailers and pick up trucks are frequently involved in this type of accident.

Figure 5 shows that most passenger train accidents occur at high speeds at crossings with flashers and gates. This may indicate that there may need to be additional traffic-control measures beyond traditional gates and flashers on lines with high-speed passenger trains.

Research in text and data mining applications applied to railroad issues

The initial results described above indicate that new techniques from computer science can be applied to the grade crossing accident data to yield new insights about the nature of the accidents. Many different types of analyses are possible to better understand grade crossing accidents. From the examples above it is clear that data visualization and text mining can be used to high light complex relationships in the multidimensional accident database. Several areas of grade crossing problems can be explored further using data mining and data visualization. Current topics of interest that can be more deeply explored include:

- Hazardous material spills at crossings.
- Analysis of differences in accidents between public and private crossings.
- Explore differences in accidents geographically.
- Study accidents on Class 1, regional and short line railroads
- Accidents in locations with whistle prohibitions.

Figure 3. Visualization of single fatality accidents.

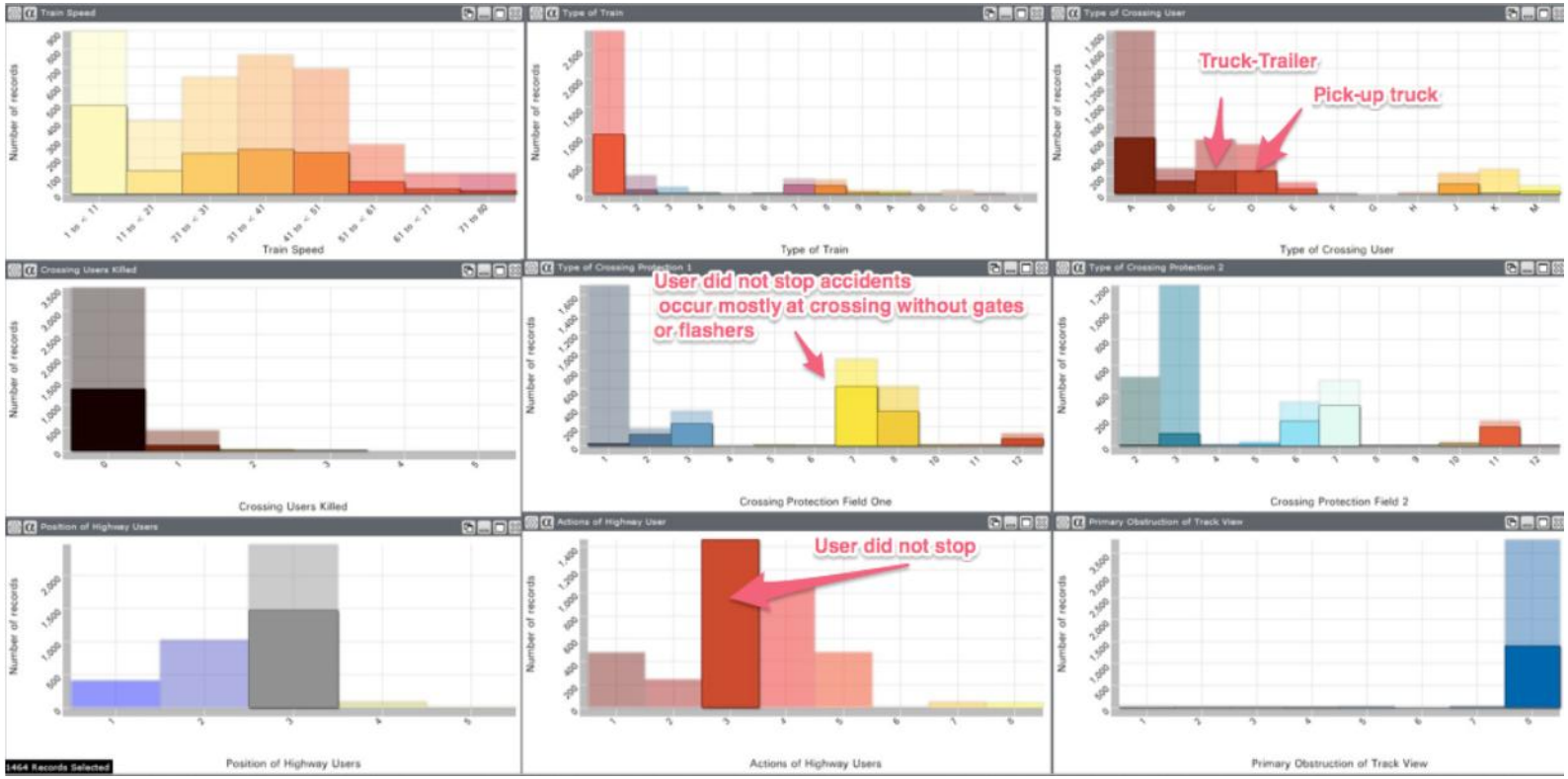

Figure 4. Did not stop accidents.

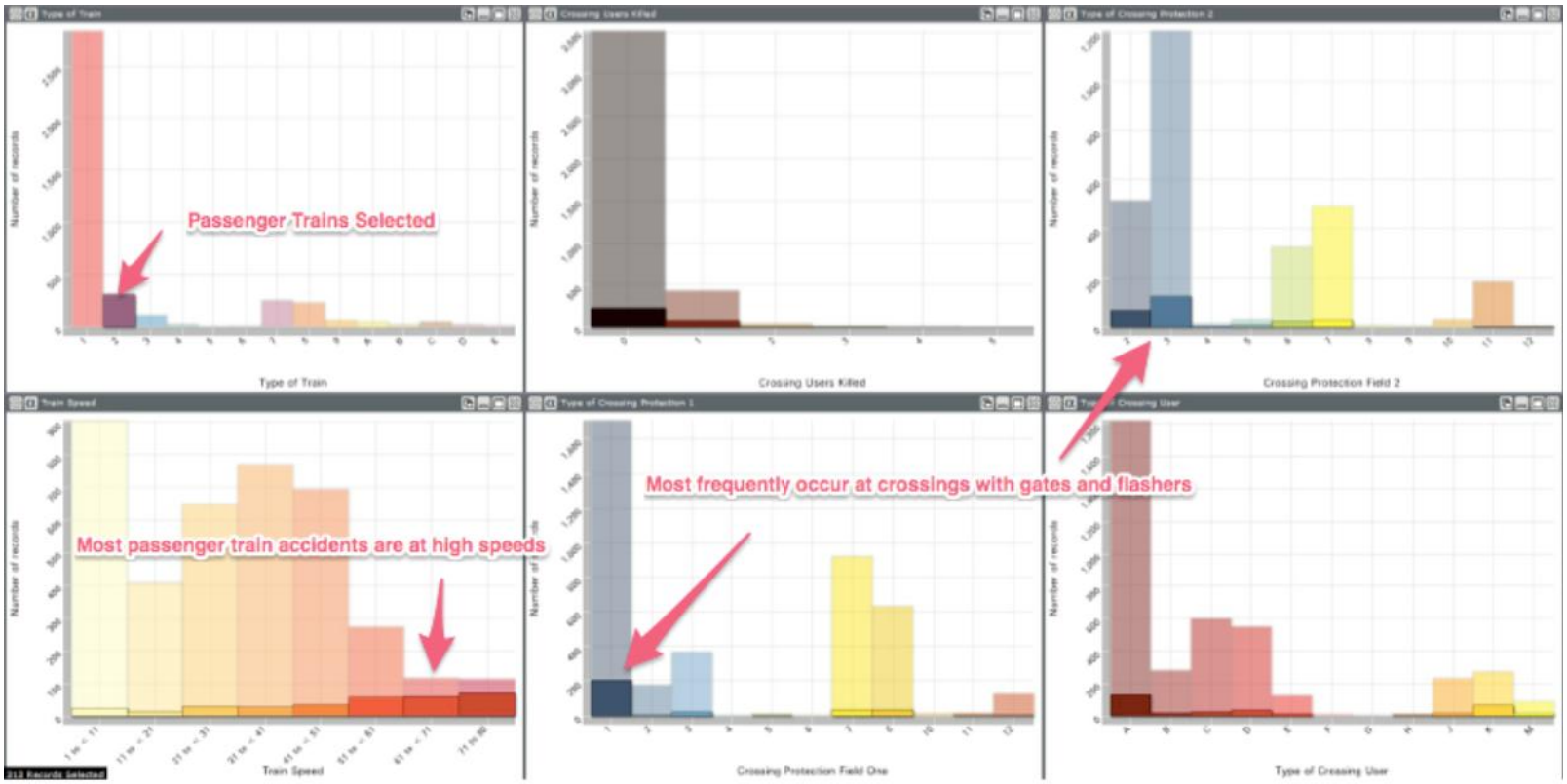

Figure 5. Passenger Train Accident

- High-speed accidents involving passenger trains.
- Differences in performance in grade crossings with only cross bucks and grade crossings with both cross bucks and stop signs.
- Tractor-Trailer accidents
- Difference between pushed train and pulled train accidents

Data and Text Mining

The labeled LDA analysis discussed above indicates that analysis of text fields in the accident reports could yield useful results. It is proposed to perform additional labeled LDA analysis on the data using different fields from the grade crossing accident database.

This analysis can yield additional insights into the nature of grade crossing accidents. In addition, HO-LDA and HONB can also be used to gain insight into the grade crossing accident data and to construct predictive models.

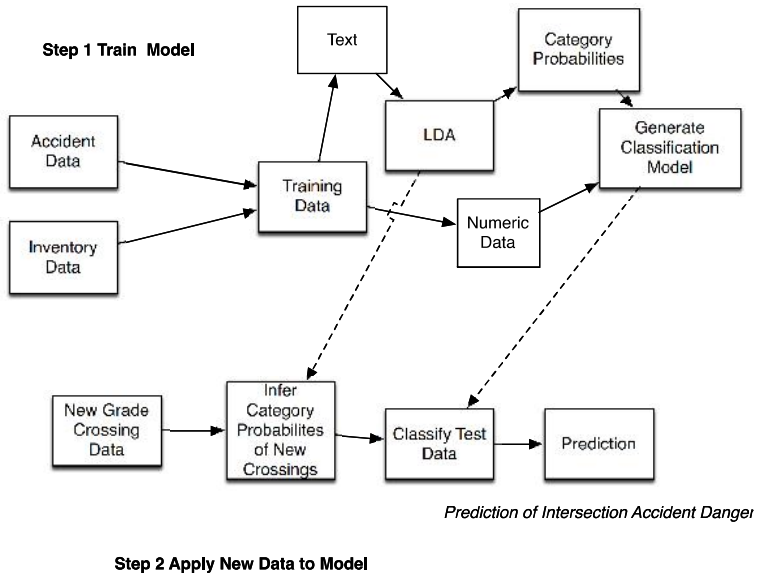

Figure 6. A predictive model using numeric and text data.

For example, the LDA, HO-LDA, or HONB models' output can be combined with the numeric and nominal data in the grade crossing accident database can be modeled to produce predictions of a grade crossing's safety performance. The output of

LDA, HO-LDA, or HONB models of the text can be combined with numeric data from the accident data to produce predictive models. Figure 6 shows a flow chart for a model of this type. In addition, the LDA or HO-LDA topic probabilities and nominal data from the accident database may also prove useful if used as input to data mining classification models like neural networks, or algorithms that automatically generate rules about the data to produce a prediction of a crossing's likelihood of accidents. The grade crossing accident database can be merged with the crossing inventory data to provide more input data for the classification model. In particular, the crossing inventory contains details about how busy the crossing is (vehicular and train traffic) that could make any predictive models more accurate. Data mining techniques can also be used to compare the characteristics of grade crossings where no accidents have occurred with the characteristics of grade crossings where accidents have occurred.

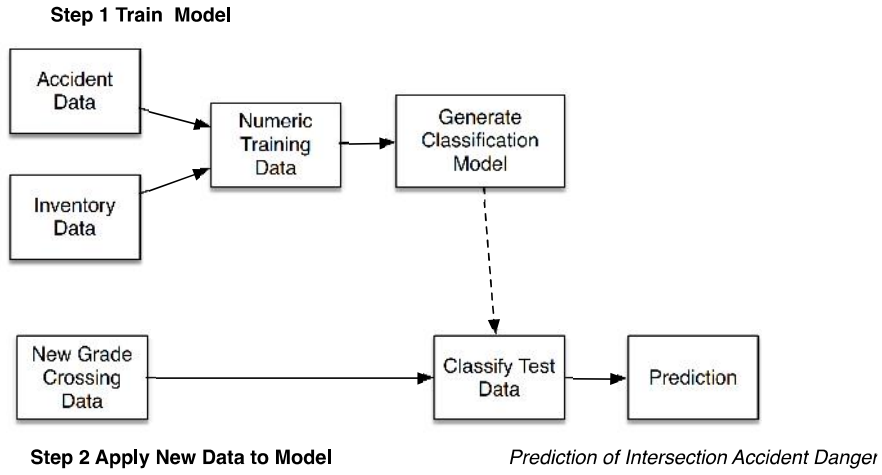

Figure 7. Predictive Model Using Numeric Data Only

Because text may not be available describing problems at grade crossings with no accident record, it is also possible to develop a model using only the numeric and nominal data that are contained in the grade crossing inventory. The training set would contain inventory data from crossings with a record of accidents (found from the grade crossing accident database) and for an equal number of crossings selected at random that are not included in the accident database. The premise of this model would be that there are some discernable differences in the crossing data that can be used to identify riskier crossings. A model could then be constructed that would accept crossing data as input

and then classify the crossings potential for accidents. Several different data mining algorithms exist to classify and group data. In particular, it is possible to automatically generate rules from the training data that will automatically classify the crossings risk of accidents. Figure 7 shows how such a model can be constructed. Other classification models that can be tested include neural networks, and ensemble classifiers.

Data Visualization

It is proposed to use data visualizations to further explore the interactions between factors contained in the accident database and the frequency of use of the crossing by both trains and automobiles. We plan to merge grade crossing accident data with grade crossing inventory data to explore the relationships of grade crossing accidents with crossing usage by trains and vehicles.

Another interesting area of research is to use data visualization to explore changes over time in the grade crossing data. The *Weave* Visualization tool we used is capable of slicing the data by date so changes in the nature of grade crossing accidents can be visualized. This may be useful in measuring the effectiveness of safety programs to update grade crossings.

To show relationships between data, many types of visualizations are possible besides bar charts. We used bar charts in our initial examples but there are many other types of visualizations that can be used to convey information. For example, Figure 8 shows a data visualization called a treemap. The treemap was developed using construction cost data from highway projects in California. Each small rectangle represents a project, and the size of each rectangle indicates the magnitude of the low bid amount. The projects are grouped first by project size and then by the number of bidders. The shading of each project rectangle indicates the percentage cost overrun on the project. Darker shading indicates projects with large cost increases. The treemap provides a method to quickly see relationships between the variables.

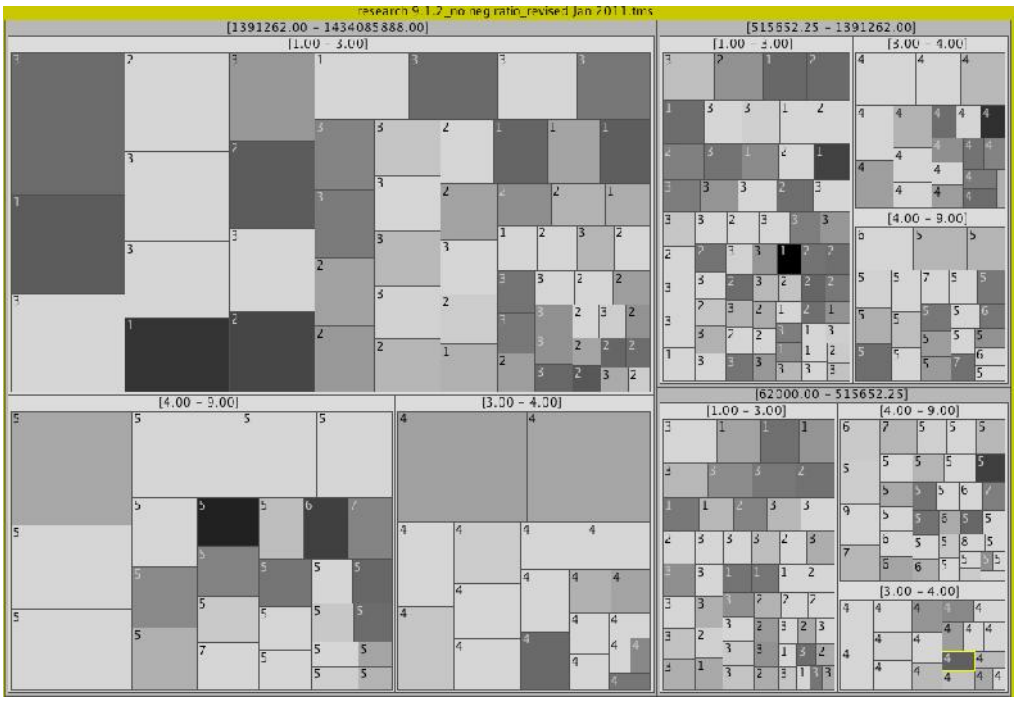

Figure 8. A Treemap Data Visualization

Customized "dashboards" of information containing different types of charts (histograms, pie charts, treemaps etc.) can be constructed to focus on the particular needs of a government agency or railroad to monitor accident causes and changes in accident type over time. The treemap in Figure 8 illustrates some of the sophisticated visualizations that are possible to allow managers to comprehend complex relationships in the data.

Proposed Research Outputs and Benefits

This research using advanced computer techniques demonstrates how emerging techniques in data and text mining can be applied to railroad grade crossing safety issues. New data visualizations can highlight previously unknown relationships in the grade crossing data. The text mining examples show how the text fields in the accident database can now be mined to produce useful insights. Information from the text fields can now be extracted and used as input for predictive models. Potentially using data mining and data visualization can lead to new methods of identifying problem crossings and to new predictive models.

Data Visualizations that show relationships in the data are more easily understood by managers without extensive backgrounds in statistics. Potentially, customized dashboards of data visualizations can be constructed that are customized to the needs of railroads and government agencies.

Conclusions

Based on the examples we have produced, there are many opportunities to apply data mining, text mining and data visualization to grade crossing databases.

1. Advanced data and text mining techniques can provide useful insights into railroad grade crossing accidents.
2. Data visualization can provide a useful tool to explore the interactions between the data collected about grade crossings.
3. Data visualization and data mining models may be easier to use by railroad managers and managers from public agencies without extensive knowledge of statistics.
4. There are opportunities to develop predictive models using data mining techniques that can prioritize problem grade crossings.

References

Austin, R. D., and L. Carson, J. (2002). "An alternative accident prediction model for highway-rail interfaces." *Accident Analysis & Prevention,* 34(1), 31-42.

Blei, D. M. (2012). "Probabilistic Topic Models." *Communications of the ACM,* 55(4), 77-7.

Carroll, A., daSilva, M., and Ngamdung, T. (2010). "USDOT Federal Railroad Administration's Third Research Needs Workshop on Highway-Rail Grade Crossing Safety and Trespass Prevention: Volume 1-Summary of Results." *Rep. No. DOT/FRA/ORD-10/01,* FHWA Office of Research and Development, Washington, DC.

Chadwick, S. G., Zhou, N., and Saat, M. R. (2014). "Review: Highway-rail grade crossing safety challenges for shared operations of high-speed passenger and heavy freight rail in the U.S." *Saf. Sci.,* 68 128-137.

Clarke, W. A., and Loeb, P. D. (2005). "The determinants of train fatalities: keeping the model on track." *Transportation Research Part E: Logistics and Transportation Review,* 41(2), 145-158.

Farr, E. H. (1987). "Rail-Highway Crossing Resource Allocation Procedure-Users Guide." *Rep. No. DOT/FRA/08-87/10,* Federal Railroad Administration, Office of Safety Analysis, Washington, DC.

Ganiz, M. C., George, C., and Pottenger, W. M. (2011). "Higher Order Naïve Bayes: A Novel Non-IID Approach to Text Classification." *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1022-1034.

Hu, S., and Lin, J. (2012). "Effect of Train Arrival Time on Crash Frequency at Highway-Railroad Grade Crossings." *Transportation Research Record: Journal of the Transportation Research Board,* 2298(1), 61-69.

Institute for Visualization and Perception Research. (2014). "Weave Web-based Analysis and Visualization Library." https://www.oicweave.org/index.php?page=about (July 26, 2014).

Konur, D., Golias, M. M., and Darks, B. (2013). "A mathematical modeling approach to resource allocation for railroad-highway crossing safety upgrades." *Accident Analysis & Prevention,* 51(0), 192-201.

Nelson, C., Pottenger, W. M. (2011). "Nuclear Detection Using Higher Order Learning." In Proc. of *IEEE Homeland Security Technologies*.

Nelson, C., Pottenger, W. M., Keiler, H., Grinberg, N. (2012). "Nuclear Detection Using Higher Order Learning." In the Proc. of *IEEE Homeland Security Technologies*.

Nelson, C., Keiler, H., Pottenger, W. M. (2013). "Modeling Microtext with Higher Order Learning." *Association for the Advancement of Artificial Intelligence*.

Ogden, B. D. (2007). "Railroad-Highway Grade Crossing Handbook." *Report No. FHWA-SA-07-010,* US Department of Transportation, Federal Highway Administration, Washington, D.C.

Oh, J., Washington, S. P., and Nam, D. (2006). "Accident prediction model for railway-highway interfaces." *Accident Analysis & Prevention,* 38(2), 346-356.

Saccomanno, F. F., Park, P. Y., and Fu, L. (2007). "Estimating countermeasure effects for reducing collisions at highway–railway grade crossings." *Accident Analysis & Prevention,* 39(2), 406-416.

Schartung, C. T., Lesales, T., Human, R. J., and Simpson, D. M. (2011). "Crossing Paths: Trend Analysis and Policy Review of Highway-Rail Grade Crossing Safety." *Journal of Homeland Security & Emergency Management,* 8(1), 1.