Accident Analysis and Prevention 152 (2021) 105985

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Applying machine learning, text mining, and spatial analysis techniques to develop a highway-railroad grade crossing consolidation model

Samira Soleimani [a,*], Michael Leitner [a], Julius Codjoe [b]

[a] Geography &Anthropology Department, Louisiana State University, Baton Rouge, LA, 70802, United States
[b] Louisiana Transportation Research Center, Baton Rouge, LA, 70808, United States

ARTICLE INFO

Keywords:
Highway-rail grade crossing consolidation
Spatial analysis
Machine learning
Text mining

ABSTRACT

The consolidation of Highway-Railroad Grade Crossing (HRGC) is one of the effective approaches to decrease the number of crashes between trains and vehicles. From 2015–2019, there were 57 HRGC crashes at crossings in East Baton Rouge Parish (EBRP), resulting in thirteen injuries with $346,875 cost of vehicle damages. Consolidation programs help to close redundant crossings and thereby decrease the crash risks; however, it is difficult to find the best crossing in a neighborhood for closure. In our previous research working on HRGC consolidation models in 2019, from among four Machine Learning algorithms, eXtreme Gradient Boosting (XGboost) performed better in HRGC prediction models. In continuation of our previous studies on developing a HRGC prediction model, this research employed Text Mining Techniques, and Geospatial Analysis in addition to the XGboost Machine Learning algorithm. The aim was to develop a consolidation model that is customized for local implementation. The results indicated an overall accuracy of 88 % for the proposed model. The relative importance of the variables input to the model was also reported and offers an in-depth understanding of the model's behavior. Considering the different correlation threshold, a sensitivity analysis was also performed on different aggregation gain values. Subsequently, it resulted in the development of a simplified model utilizing 14 variables, with aggregated gain values of 95 % and a correlation threshold of 0.5. Based on this model, 15 % of current highway-rail grade crossings should be closed.

1. Introduction

Considering crash frequency over time, rail transportation is one of the safest modes of travel. However, according to the Federal Railroad Administration (FRA)'s Office of Safety, the severity of train-related crashes is almost twenty times greater than vehicle crashes ("Crossing Collisions & Casualties by Year", 2018; Rahimi et al., 2020). The severity of highway-railroad grade crossing (HRGC) crashes is mostly related to transporting dangerous goods, inadequate safety training, and a lack of warning devices. One way to decrease the number of crashes at HRGC is by developing crossing consolidation models that aim to close dangerous, unnecessary, and/or consecutive crossings in close proximity. FRA requires states with a high number of rail-related crashes to develop a grade crossing Safety Action Plan (SAP) to address the safety issues. Currently, such states include Alabama, California, Florida, Georgia, Illinois, Indiana, Iowa, Louisiana, Ohio, and Texas (Texas Department of Transportation, 2013; Federal Highway Administration, 2016). Crossing consolidation is considered as one of the options to address this safety issue, and generally results in the closure of redundant crossings to reduce the concentration of HRGC in an area. The crossing closure refers to the closure of roads that cross the rail.

So far, most of the current consolidation programs only rely on safety experts' knowledge to find the most suitable crossing candidates for future closures. Some studies have also used data from FRA to develop a rating formula that ranks existing crossings to find the most suitable candidates for closure. However, the traditional approach treats all crossings just based on experts' judgment by using a decision support system (Ćirović and Pamučar, 2013). This raises the question of how appropriately selected variables, used in the rating formulae, are and how much experts' judgment can be trusted.

In 2019a, Soleimani et al. created a consolidation model for Louisiana using machine learning techniques. Though the accuracy of the model was good, the model was not suitable enough for practical use. That is because the model was created for a large area and no spatial parameters, as well as crash history, were explored to account for the uniqueness of every crossing. To make the model practical for

* Corresponding author.
E-mail address: ssolei1@lsu.edu (S. Soleimani).

https://doi.org/10.1016/j.aap.2021.105985
Received 25 September 2020; Received in revised form 7 December 2020; Accepted 11 January 2021
Available online 22 January 2021

implementation in a locality, it is important to explore other aspects of crossing closures in that area. There are other features that may increase the risk of crashes at crossings and we can get access to them from FRA crash history data. Also, the accessibility of each crossing to critical places such as hospitals, fire stations, and schools is not available in FRA data, however, safety experts believe that proximity to these places may affect the crossing closures (Soleimani, 2020). These proximities can be accessed using spatial analysis if the spatial data is available for the implementation area.

Machine Learning approaches return models with higher accuracy by increasing the number of sample data (Namakian et al., 2020). By narrowing down to a parish level, the sample size would decrease, however, we can get more spatial data. Therefore, on one hand, there is a need to have as much data as possible to create a Machine Learning model, and on the other hand, we wanted to explore as many features as possible to consider all potential aspects when it comes to the implementation phase. Since not all areas have available spatial data, we first developed a general model using all FRA crossing data in our first study (Soleimani et al., 2019a). Then we narrowed down to a smaller area where the needed spatial data is available to be explored in the final consolidation model.

In addressing the model limitation of Soleimani et al. (2019a), this research aims to narrow the model down from a state scale (Louisiana) to a parish scale (East Baton Rouge parish). Moreover, this research explores not only FRA data, but it also uses the crash narrative (or crash history) data and several spatial characteristics of every crossing in East Baton Rouge Parish (EBRP). The final model will highlight the best crossing candidates for closure programs in the parish. Sometimes mining through a different set of data can open some aspects that human beings' ability cannot understand from such overwhelming data. That is why a combination of machine learning, spatial analysis, and safety experts' judgment works better.

While the presented approach in this study can aid in selecting potential HRGCs for future closures, it should be noted that machines learn what is fed to them by recognizing patterns among variables in the dataset. Microscopic and detailed analysis of each crossing before consolidation is required before selecting crossings for closure. The HRGC data and the HRGC crash data, which are publicly available on the FRA website, were used in this study as secondary data. The challenge of using existing secondary data from an available database rather than using primary data is two-fold: first, it is required to be preprocessed before running any analysis, including removing unneeded variables, and handling missing values; second, a feature selection analysis should be applied on data to find variables that may be significantly related to the outcome of interest in the research (Mousavian and Chen, 2018). Several primary variables were also created using spatial analysis to be explored in the research. Finally, the results of the proposed consolidation models were compared and evaluated with other potential models, based on statistical accuracy measures that suit models' predictive performance.

The rest of this study was organized into six more sections. Section 2 presents a brief literature review. Section 3 discusses approaches by which the data was created. Section 4 describes the methodology used in this study to explore the data and create a final prediction model. Lastly, Section 6 and 7 discuss the evaluation and conclusion of this study.

2. Literature review

The safety of HRGC is a concern in both urban and suburban areas. Grade crossings may have less effect on the traffic flow in suburban neighborhoods than in an urban area. Moreover, the higher the traffic flow, the higher the probability of crashes between trains and vehicles for high traffic volume areas (Retallack and Ostendorf, 2019). To improve safety at HRGCs, Executive Order 12866 and DOT Order was issued in 1979 to reduce the number of accidents at crossings. Almost nineteen years later, in 2008, ten HRGC safety-challenged states, Louisiana included, were required to submit their state action plan by the Rail Safety Improvement Act of 2008. In 2019, a proposed rule by FRA was issued that required all states to develop their state action plan reporting how to reduce the safety risks at grade crossings ("State Highway-Rail Grade Crossing Action Plans", 2019). One of the least expensive approaches to improve safety at crossings is referred to as consolidation programs. Removing the unnecessary crossings in a neighborhood helps to reduce the crash probability. Roads with grade crossings try to connect separated parts of neighborhoods, while they still act as barriers when trains are passing. So before making any decision on which crossing to close, a comprehensive analysis should be done considering various potential alternatives.

So far, several approaches have been employed by different transportation departments and agencies to find the best candidates for crossing closures. For instance, the Union Pacific railroad agency requests multiple public crossing closures before agreeing to allow and create a new grade crossing in a neighborhood (Union Pacific, 2020). The North Carolina DOT (NCDOT) considers several variables to select the best crossing closures candidates including where traffic congestion could be safely redirected, where crossings are within a quarter-mile of one another, where crossings have a high number of crashes, and where crossings have a limited sight distance (Texas Department of Transportation, 2013). The Florida DOT considers a set of variables including the acute angle between rail and road, the amount of nearby alternative crossings (within 1300 ft), the number of vehicles per day (less than 2000), the number of trains per day (more than 2), and crossings on routes not commonly used by emergency vehicles (Texas Department of Transportation, 2013). All these approaches either rely only on expert's knowledge or use a rating formula with a limited number of variables. In a recent research, Soleimani et al. (2019a) presented an approach employing machine learning techniques to develop a prediction model for future closures. This study explored more than fifty variables to create the consolidation model; however, the crash history and spatial characteristics of the crossings were not considered. The crash narrative data holds all the information about why, how, and when the crash happened. In another study, Soleimani et al. (2019b) introduced a method to easily recognize the reason behind crashes from crash narrative reports. This information can help to identify the crossing problems that cause crashes between trains and vehicles.

2.1. Machine learning

Another underexplored method is integrating machine learning and/or data-driven approaches with consolidation models to facilitate solving time-consuming and costly problems. Machine learning can be used to determine the relationships of several entities using various classifiers, including Decision Trees (DT), Logistic Regressions (LR), Random Forest (RF), and eXtreme Gradient Boosting (XGboost)(Williams et al., 2020)(Ebrahimi et al., 2020; Namakian et al., 2014). XGboost is a highly effective technique that commonly displays a high order of accuracy (Bort Escabias, 2017; Omar, 2018; Chen and Guestrin, 2016, August). Using XGboost, the importance of the applied variables can also be obtained as a gain value (relative importance) (Friedman et al., 2000; Chen and Guestrin, 2016, August). Machine learning is a perfect tool to solve complex problems in the severity of crossing crashes (Keramati et al., 2020b; Lee et al., 2019; Zhou et al., 2020)(Nejad et al., 2015) and crossing consolidations studies (Soleimani et al., 2019a, 2019b). There are a variety of ML algorithms that can be used to solve complex problems. However, it is very important to choose the right one that fits your data well. For consolidation problems, Soleimani et al. (2019a) found the XGboost algorithm to be the most accurate, among others evaluated, when applied to HRGC data from an FRA dataset. The other tested algorithms include RF, XGBoost, LR, and DT.

2.2. Spatial analysis

Generally, most of the previous railroad safety studies used logistic regression to find significant variables in railroad crash studies, which only captures global relationships between variables and the outcome of interest (Liu and Khattak, 2017; Liu et al., 2016; Iranitalab et al., 2018; Yıldız and Ateş, 2020). However, recent studies investigated local relationships between variables and the outcome of interest, as well (Grauers, 2019; Keramati et al., 2020a; Zheng et al., 2016). The effective approach for capturing the geographical distribution of crash characteristics is using the Geographically Weighted Regression (GWR) modeling technique. The GWR assigns weights to each observation according to its spatial distance to other nearby observations. Liu et al. (2016) employed local analysis using spatial analysis to investigate relationships between safety outcomes and crossing characteristics. In another study, Liu and Khattak (2017) studied gate-violation behavior using GWR (as the local spatial modeling approach) and binary logistic regression analysis (as the global modeling approach).

Spatial analysis can also help to create new spatial parameters as test variables to be investigated in consolidation models. These variables could represent and evaluate the complexity of crossings. The crossing complexity could depend on several variables, including the crossing road/railroad design, accessibility of crossings to roads, community coherence at the crossing's neighborhood, number of nearby intersections, distance to nearby intersections, number of nearby crossings, distance to nearby crossings, and the difference between the grades of the road and the rail. Spatial variables do not act uniformly in different spatial contexts; that is why it would be better to investigate these variables in a smaller area (Liu et al., 2016). Also, it is sometimes impossible to generate spatial variables for a larger area due to a lack of existing spatial data.

Some research considered and explored "distance" as a spatial variable in railroad safety research (Haleem, 2016; Ma et al., 2018; Lu et al., 2020; Keramati et al., 2020a). In a very early study, Schrader and Hoffpauer, 2001a,b calculated the difference in distance between a route using an at-grade crossing and another route that uses a nearby grade-separated crossing. They employed a distance variable along with other variables, such as noise, delay, and geographic distribution, to evaluate and prioritize the location of potential HRGC separation. In 2010, Kim et al. studied how the accessibility of crossings and intersections in a neighborhood affects the safety of crossings in that area. They considered the distance between crossings and intersections as a spatial variable to explore railroad safety. It was hypothesized that when the accessibility between roads and railroads rises, traffic congestion increases, and so does the risk of a crash (Kim et al., 2010). Kim et al. (2010) found several factors for neighborhood accessibility measurements, such as road length, availability of bus stops, length of bus routes, number of intersections, and number of dead ends. Land-use diversity is another factor that affects the accessibility of a neighborhood (Vale et al., 2016). The higher the land-use diversity in a neighborhood, the more facilities in the surrounding area would be accessed, and the more accessible a neighborhood would be (Vale et al., 2016). In very recent studies, a combination of spatial variables was used to define the crossings geometric factors including acute crossing angle, the number of highway lanes, and the roadway distance between the grade crossing and the signalized intersection (Keramati et al., 2020a; Lu et al., 2020).

However, the number of passing school buses is one of the factors that FRA considers in safety research, but none of the previous studies used the distance from crossings to nearby schools. Moreover, the availability and proximity of crossings around hospitals are not yet sufficiently explored. Most of the previous crossing consolidation research used the approximate 150 m (~500 ft) as the FRA standard threshold to search the availability of nearby intersections. Haleem (2016) found that the distance between crossings and their nearby intersections has significant results on crash severity and injuries at crossings. Moreover, Keramati et al. (2020) found crossing geometric factors as one of the significant contributions in crash severity and occurrence. Therefore, according to the crash severity levels, as well as the geometric characteristics of crossings, other ranges of the distance between intersections and crossings may need to be explored in consolidation research as well.

The spatial variables suggested in this paper are either based on safety experts' recommendations, or by reference from previous research. For example, the distance to nearby school variable has a shared concept with a variable collected in FRA dataset named "average school buses on a school day". Also, the distance to a fire station and hospitals were both suggested by safety experts to be explored in our model. They believed that the crossings that are very closed to hospitals or fire stations should not be selected as the best candidates for closure, as closures may increase the commute time to hospitals and fire stations.

This paper attempts to address the existing gap in previous research efforts regarding using different datasets of FRA and crash narrative, as well as creating and exploring new spatial parameters. This study employs a machine learning algorithm, text mining technique, and spatial analysis to create a reliable prediction model for the consolidation program in EBRP.

3. Data acquisition

3.1. Study area

As of 2018, the EBRP was the most populated Louisiana parish followed by Jefferson and Orleans Parishes. Three primary railroad operators are currently running in the state, with 235 currently open HRGCs based on FRA crossing data (FRA Safety Data & Reporting, 2019). These companies are Baton Rouge Southern Railroad (BRS), Illinois Central Railroad Company (IC), and Kansas City Southern Railway Company (KCS) (RIMS, 2020. EBRP had already closed approximately 86 (27 %) of its crossings as of April 2020 (Table 1) referencing from 1979. Most of its private crossings are located on industrial land use for private travels only. Therefore, this research sought to develop a consolidation model only based on public crossings data.

3.2. Data collection

Several datasets are used in this research, including the HRGC data from FRA, the FRA crash narrative data, and spatial data. These data are discussed below:

3.2.1. The FRA HRGC data

All the variables in this FRA dataset were explored. However, only twenty-six variables were finally selected including 17 numeric variables and 9 nominal variables. Not all variables were selected because some of the variables had more than 70 % missing values (such as "Crossing Surface Width ft", and "Crossing Surface Length ft), and some of them contained the same value for all records. For example, the values in the "does track run down a street" variable for all the crossing records contained "NO"; and variable for "traffic lane" contained "two-way" for all crossings in EBRP. The descriptive statistics of the study variables from HRGC data are shown in Table 2 (numeric variables) and Table 3 (nominal variables).

Table 1
The number of crossings in East Baton Rouge Parish.

Crossing Type	Open		Closed	All
	Before 1979	After 1979		
Public	111	26	38	175
Private	70	28	48	146
All	235		86	321

Table 2
Table Descriptive information of numeric variables.

	Numeric Variables	Total	Missing	Min	Max	Mean	SD
1	number of day-thru train movements 6am to 6 pm	175	0	0	4	1.703	1.370
2	number of night-thru train movements 6 pm to 6am	175	0	0	4	1.039	1.307
3	total switching trains	175	0	0	9	1.257	2.539
4	maximum timetable speed mph	175	0	0	49	27.09	15.25
5	typical minimum speed mph	175	0	0	25	2.709	3.644
6	typical maximum speed mph	175	0	0	49	26.62	15.48
7	annual average daily traffic	175	0	10	50100	5248	7874
8	total count of flashing light pair	175	0	0	15	2.749	3.915
9	number of crossbuck assemblies	175	0	0	4	1.1086	0.9969
10	estimated percent trucks	175	0	3	90	9.126	9.770
11	avg school buses on a school day	162	13	0	8	0.2286	1.2614
12	cumulative narrative	175	0	0	43	1.606	4.162
13	number of accidents at crossing	175	0	0	22	1.160	2.714
14	total trains passing per day	175	0	0	16	4.354	3.417
15	number of traffic lanes crossing track	170	0	1	7	2.5314	1.0817
16	number of advance warning signs w10-1	170	5	0	4	0.9588	0.9317
17	posted highway speed mph	174	1	10	55	33.540	9.692

3.2.2. Crash narrative data

In a very recent study, Soleimani et al. (2019b) applied text mining and topic modeling on the narrative in US crash reports to identify the most frequently used words that correlate with crossing closures across all US states. Building upon the outcome of the study, a new variable was created that counts the number of frequent words in the police narratives of crashes at crossings. According to Soleimani et al. (2019b), the most frequent words for Louisiana were: truck, tractor, trailer, light, farm, gate, bumper, box, marking, bayou, drug, alcohol, stop, sign, grass, pickup, speed, male, pavement, fail, warning, main, fouling, shove, industry, device, private, and work. Based on the total number of times these words occurred in crash narratives at a specific crossing, a new column, called "cumulative narrative", was developed as another potential variable to predict future crossing closures. The accuracy of the new consolidation model with the added cumulative narrative column was improved. However, from among 175 public crossings, 111 crossings had no crash narrative, so this new variable only influenced a reduced subset.

3.2.3. Spatial data

In addition to crossing variables and crash reports, this research aims to suggest, generate, and test new spatial variables that may affect crossing closures projects. The suggested variables are the distance to nearby schools, distance to nearby hospitals, type of land use in crossing's surrounding area, flood zone areas around the crossing, and the number of intersections in different distance thresholds. The crossing dataset has several similar variables, such as "is there any intersection within 500 ft" and/or "average number of school buses passing the crossing every weekday"; however, these variables either had many missing values or did not have enough information. There is also a feature in the crossing data that describes the land-use type of a crossing. However, since there could be a combination of land-use types in the neighborhood of a crossing, it may be misleading to focus on a single land-use type at the site of the crossing. Therefore, new variables that described a combination of land-use types within a distance from the crossing location were developed. This is further explained in the Methodology section.

4. Methodology

In this section, the steps for developing the crossing consolidation model for EBRP are discussed. The proposed model aims to identify the best crossing candidates for future HRGC closures in EBRP. Data preparation is crucial. The FRA data was ready to use; however, the crash report data and spatial data needed to be created.

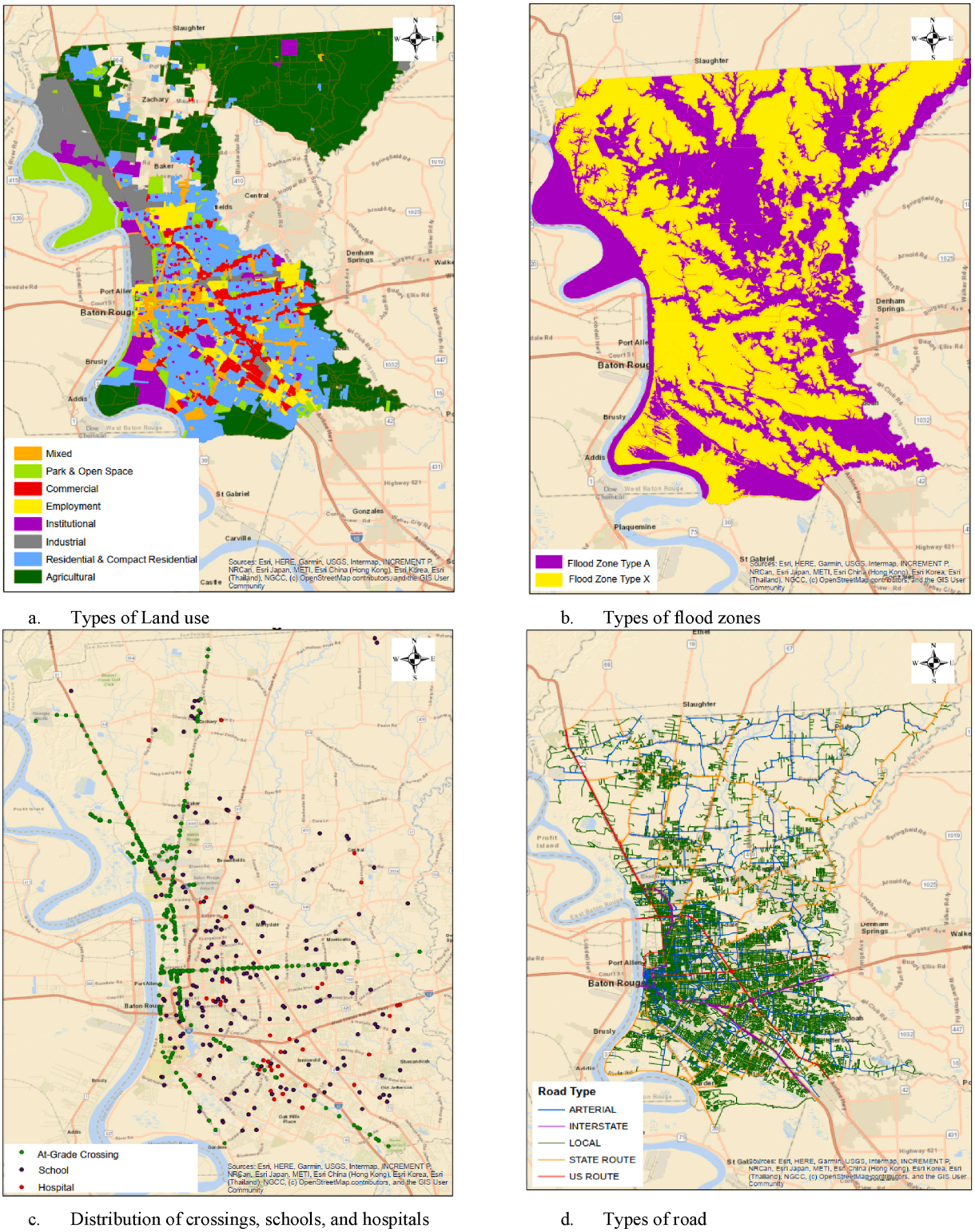
4.1. Spatial analysis

Since the case study was narrowed down to a parish and based on the

Table 3
Descriptive information of nominal variables.

#	Variable	Type	Frequency	#	Variable	Type	Frequency
1	Reason ID	closed	38	6	Type of Land Use	residential	29
		open after 1979	26			institutional	11
		open before 1979	111			commercial	53
2	In or Near City	Near City	51			industrial	70
		In City	124			farm/open space/ RR yard	12
3	Intersecting Roadway Within 500 ft	No	46	7	Pavement Markings (6 rows has missing pavement marking information)	No mark	87
		Yes	129			RR mark	15
4	Main Tracks	No	49			stop line	5
		Yes	126			combination	62
5	Crossing Surface Main Track	Asphalt	32	8	Smallest Crossing Angle (degrees)	0 to 29	15
		Asphalt & Timber	3			30 to 59	24
		Timber	64			59 to 90	136
		Concrete	55	9	Functional Classification Road Function	local	108
		Other	21			collector	20
						arterial	46

availability of spatial layers, it was possible to investigate the spatial condition of every crossing within the extent of the parish. The spatial analysis techniques are powerful tools to create or extract new insight and information from the raw spatial data. Using these techniques, new variables were created to account for a combination of land-use types, and flood zone types, around each crossing. The existence of schools and hospitals near crossings also matters for crossing closure programs; therefore, distances from each crossing to the nearby school and hospital were calculated. Fig. 1 shows the shapefiles that were used for this study. The school and hospital data were collected from ArcGIS Hub for EBRP and the various types used as input variables (EBR GIS Open Data, 2021). For example, for the hospital data, the types used were urgent care, mental health, primary care, and hospitals.



a. Types of Land use

b. Types of flood zones

c. Distribution of crossings, schools, and hospitals

d. Types of road

Fig. 1. The shapefiles used in this study.
Source: EBR GIS Open Data.

4.1.1. Polygon to point

One of the objectives of this study is to explore the distribution of land-use types, as well as flood zone types, around each crossing. Land use polygon feature includes eight different land-use types: mixed, park and open-space, commercial, employment, institutional, industrial, residential, and agricultural (Fig. 1a). The flood zone polygon feature contains two flood zone types: flood zone type A and flood zone type X (Fig. 1b). Type A indicates all areas with a high risk of flood hazard and type X indicates zones with the lowest flood risk. One way to calculate the distribution of these polygons around a crossing is by converting the polygons to points and then count the points and join them with its corresponding crossing. The Arctoolbox in ArcGIS 10.7 software has "polygon to raster" and "raster to point" tools that convert polygon features to a raster dataset and then a raster dataset to points. Using spatial join analysis within a distance threshold of 650 m (2000 ft), the distribution of each polygon can be easily merged with FRA data for each crossing based on their locations. This threshold is equal to the maximum threshold used to count the number of intersections around a crossing. The aim is to get the percentage of land-use types around a crossing and not simply relate one land-use type to a crossing based on where a crossing is located. For example, for the crossing shown in Fig. 2, there are 28 red points for flood zone type A, and 100 blue points for flood zone type X falling inside this threshold distance. The processes discussed above were applied to each land use type polygon and flood zone type polygon. Altogether, a total of ten spatial variables were added to the existing data on the 175 crossing records for further exploration. These variables are the number of cells (100m*100 m) for mixed, park and open-space, commercial, employment, institutional, industrial, residential, and agricultural land uses, as well as the number of cells (100m*100 m) for type X and type A flood zones.

4.1.2. Closest facility

It is important to consider the location of schools and hospitals around a crossing while working on crossing closure programs. So, the distances to the nearest schools and hospitals were determined for each crossing. Both distance values can be counted as two spatial variables for the crossings dataset that is to be explored for the consolidation model for EBRP. Two types of distances can be calculated for each: first, the Euclidean distance (the straight-line between two points on a surface), and secondly, the actual travel distance (distance calculated over the road network between the two points). From the objectives of this study, the travel distance approach is preferred. Therefore, a road network was generated using the "new network dataset" tool in the network analysis toolbox. The process included creating network features, setting up the connectivity, and assigning values for defining any road attributes. After the creation of the road network, the next step involved the application of the "closet facility" tool from the network analysis toolbox. The desired facilities (schools and hospitals) and crossings were defined in advance of running the analysis for the existing 175 crossings in the dataset. After running the tool on each facility, the closest facilities were assigned to each crossing, and the actual travel distances between them were calculated. The distances from the crossings to the closest hospital and the closest school were also added as two new variables (fields) to the crossing dataset.

4.1.3. Service area

Another spatial variable to be investigated is the number of road intersections within different distances from each crossing. In a recent research in Houston, Texas, Hu, and Shelton (Hu and Shelton, 2017) found that the presence of an intersection around a crossing increases the crash risk. Also, the number of intersections in a neighborhood

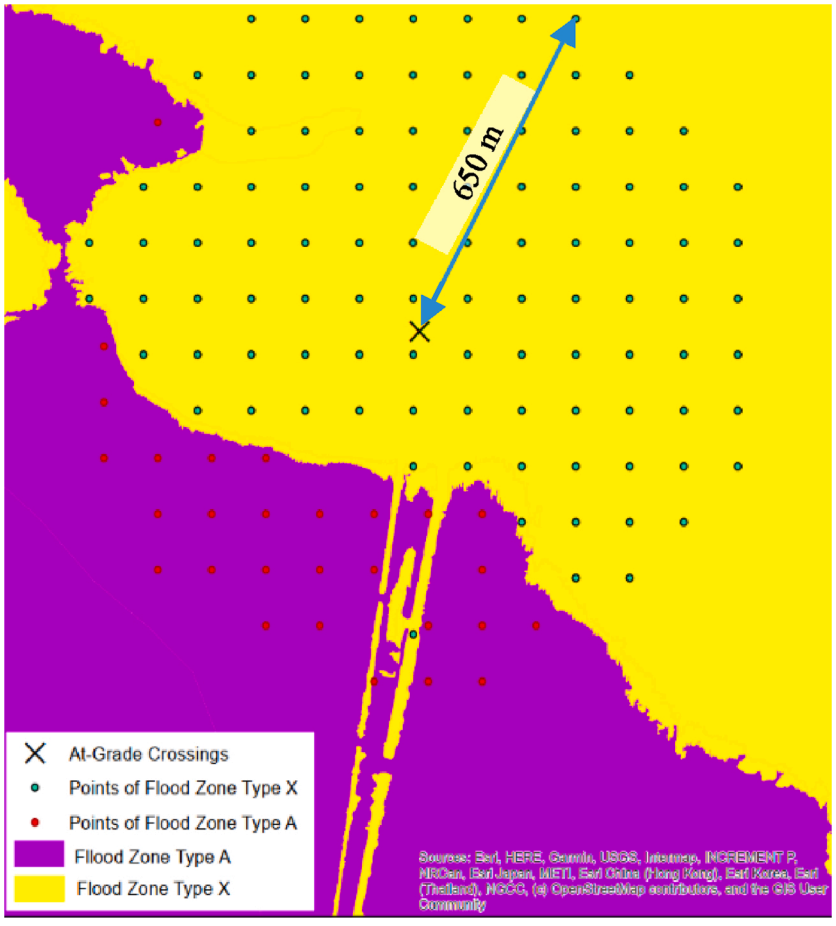


Fig. 2. Number of points in flood zone types A and X falling inside a 650 m (2,000 ft) threshold distance around one crossing.

creates traffic congestion. Traffic congestion plays a significant role in causing car accidents since it imposes significant responsibility for drivers and pedestrians to make successful judgments (Chen et al., 2012). Therefore, this paper assumes that the higher the number of intersections around a crossing, the higher the risk of crashes between trains and vehicles.

The original crossing data has a variable called “is there an intersection within 500 ft?”. For this, the number of intersections within travel distances of 150 m (~500 ft), 300 m (~1000 ft), 450 m (~1500 ft), and 600 m (~2000 ft) were counted (Fig. 3). Altogether, four variables are created and added to the dataset by counting the number of intersections falling into the four different distance thresholds around each crossing.

The final descriptive information of the new sixteen spatial variables is shown in Table 4.

4.2. Text mining

The crash reports collect valuable information about the reason behind crashes. Rather than only using the crash frequency on crossings, the narrative reports were read to find any potential reasons for crashes. There are different approaches to explore the word importance in a collection of documents, including Term Frequency (TF) and Term Frequency Inverse Document Frequency (TF-IDF). In TF, each word is mapped to a number showing its number of occurrences in the document divided by the total number of words in each document to normalize the data, as shown in Eq. (1) (Heidarysafa et al., 2018; Zhang et al., 2011).

$$TF_{t,d} = F_{t,d} / T \quad (1)$$

where “$FT_{t,d}$” is the total frequency of term “t” in the document“d”, and "T" is the total number of existing terms in the document “d”. The top frequent term within the whole crash data may have a relationship with the main reason behind the majority of crashes. While TF calculates the frequency of a word, TF-IDF computes the importance (weight) of a word. Using Eq. (2), TF-IDF tends to count the frequency of each word in the whole data along with the number of documents containing that specific word (Jones, 1972).

$$W_{t,d} = TF_{t,d} * (\log\left(\frac{D+1}{D_t+1}\right) + 1) \quad (2)$$

where "D" is the total number of documents in the collection of crash reports, and "D_t" is the number of documents in the collection that the term “t” occurs. This method dampens the impacts of common words in each document while scales up to the impact of unique words (Heidarysafa et al., 2018). Considering the objectives of this research, which is finding the most frequent, important, and related words for highway-rail crashes, the combination of both the TF and TF-IDF works well. In our latest research (Soleimani et al., 2019b), we applied a text mining technique on the FRA crash report and compared the crash reports of

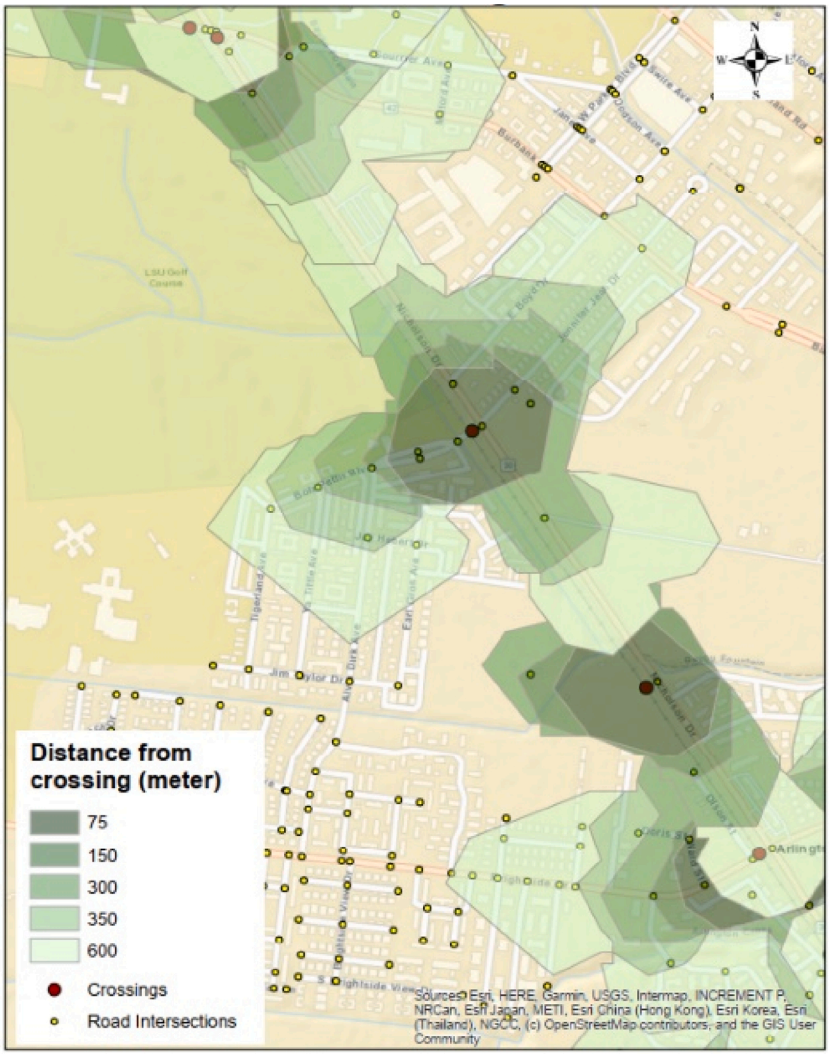


Fig. 3. Counting the number of nearby intersections within four different threshold values of 150 m, 300 m, 450 m, and 600 m of crossings.

Table 4
Descriptive information of the new spatial variables.

	Numeric Variables	Total	Missing	Min	Max	Mean	SD
1	Number of intersections within 150 m	175	0	0	16	2.69	3.22
2	Number of intersections within 300 m	175	0	0	25	4.37	5.37
3	Number of intersections within 450 m	175	0	0	34	7.257	7.418
4	Number of intersections within 600 m	175	0	0	62	8.91	11.13
5	Distance to the Closest Hospital m	175	0	349	11037	4241	2021
6	Distance to the Closest School m	175	0	47	8131	1616	1546
7	Number of residential/ compact Land-use Cells (100*100 m)	175	0	0	57	6.977	10.838
8	Number of agricultural Land-use Cells (100*100 m)	175	0	0	68	2.731	10.937
9	Number of Industrial Land-use Cells (100*100 m)	175	0	0	132	5.606	13.086
10	Number of Institutional Land-use Cells (100*100 m)	175	0	0	67	3.777	10.100
11	Number of Employment Land-use Cells (100*100 m)	175	0	0	45	3.063	7.484
12	Number of Commercial Land-use Cells (100*100 m)	175	0	0	29	1.074	3.810
13	Number of Park Land-use Cells (100*100 m)	175	0	0	36	1.840	5.694
14	Number of Mixed Land-use Cells (100*100 m)	175	0	0	49	2.686	6.219
15	Number of Flood X Cells (100*100 m)	175	0	0	115.00	28.97	24.02
16	Number of Flood A Cells (100*100 m)	175	0	0	133.00	10.69	22.14

different states. Based on the obtained results, a new column "cumulative narrative" was added to this research data as well. This column was created by comparing the words with maximum TF, and TF-IDF values in Louisiana. Several words with higher TF-IDF values were specifically considered as significant ones for highway-rail crashes in Louisiana including "truck", "tractor", "trailer", "light", "farm", "gate", "bumper", "box", "marking", "bayou", "drug", "alcohol", "sign", "grass", "pick up", "speed", "pavement", "fail", "warning", "main", "industry", "private", and "work". The newly created column is called "cumulative narrative" since it is the cumulative frequency of all the selected words in the crash history at each crossing.

4.3. Machine learning

Machine learning techniques help to explore the applicability and significance of FRA variables in Tables 2 and 3, and the cumulative narrative variable created by applying TF-IDF on crash reports, as well as new spatial variables in Table 4. Among Decision Tree, Random Forest, Logistic Regression, and XGboost, Soleimani et al. (2019a) found XGboost algorithm to be the most accurate for the purpose of their crossing consolidation model. Therefore, the XGboost method was used to create a local consolidation program for EBRP. According to Table 1, for public crossings in EBRP, there were 111 open crossings before 1979, 26 open crossings after 1979, and 38 closed crossings after 1979. We assumed that the open crossings after 1979 (when the Executive Order 12866 was issued) were the best candidates for training the model regarding crossings expected to be open. The reason is that the decision for opening a new crossing after 1979 should have been based upon detailed safety considerations. Therefore, the consolidation model needs to be trained based on 38 closed and 26 lately open crossings. Then, the trained model should be applied to the 111 open crossings before 1979 to identify the best candidates for future closures.

4.3.1. Model performance measures

The data used for training was considered imbalanced because of the disproportionate number of closed crossings to open crossings after 1979. For such imbalanced models, the level of accuracy for trained models in predicting tends to be very high. To address this problem, other performance measures were also explored. The Area Under the Curve (AUC) metric, a common model performance measure, was used to evaluate the performance of the model over the training and validation data. The AUC metric refers to the areas under the Receiver Operating Characteristics (ROC) curve. The AUC value deals with the skewed sample situation to prevent overfitting. There is also always a trade-off between specificity (how correctly negative events are classified) and sensitivity (how correctly positive events are classified) in most classifiers. The AUC value attempts different thresholds to classify data and to plot specificity and sensitivity. A non-performing ROC curve has an area of no more than 0.5, while a perfect ROC curve has an AUC value of 1 (Brown and Davis, 2006). The sensitivity and accuracy values were used as the performance measures over the testing data using the equations below.

$$\text{Sensitivity} = a / A \tag{3}$$

$$\text{Specificity} = b / B \tag{4}$$

$$\text{Accuracy} = a + b / A + B \tag{5}$$

Where a is the number of correctly classified closed crossings, b is the number of correctly classified newly open crossings, A is the total number of actually closed crossings, and B is the total number of actual open crossings in the dataset. In this study, 70 % of the data was used for tuning hyper-parameters, validating, and training the models, while 30 % was used for the final test of the developed models.

4.3.2. Model training

The existing data in this research was small and from the machine learning perspective, the larger sample size would perform a better result. However, because of the objective of this research to create a customized model for EBR parish, there were no more crossings to be counted and used in this research. So we applied several techniques to improve the number of samples and to make sure the final results would be reasonable. Firstly, we used a sampling technique to increase the number of minor class. The existing imbalanced data (many closed crossings compared to the small number of opened ones after 1979), would bias the performance of the model towards the closed crossing instances. Therefore, to avoid any bias, a sampling technique was required to balance the class of data. There are two approaches to combat imbalanced data either "downsampling" the major class or "upsampling" the minor class of data set (Kuhn and Johnson, 2013). This study used the upsampling technique for training the data. The

upsampling technique randomly samples from the minor class to set a nearly uniform distribution before the classification step began in each iteration (Zhang and Schuller, 2012).

Secondly, several algorithms can be used to create a consolidation model. Based on our previous results (Soleimni et al., 2019), comparing four algorithms of the decision tree, random forest, logistic regression, and XGboost, the latter one performed better for the HRGC problem. Also, compared to other algorithms, XGBoost has the appealing properties of small sample sizes (Samat et al., 2020; Dealing with very small datasets, 2018). Therefore, we used XGboost algorithm that can control the overfitting problem by tuning the hyperparameters by selecting a low max_depth of a tree, high gamma and eta values, and high L1 and L2 regularization values. XGboost is a tree-based algorithm that requires the tuning of D (Maximum tree depth) and T (number of trees), as well as the extra regularization parameters L, γ, and λ. It is worth mentioning that, XGboost requires tuning the D value due to the sequential process of growing the trees. The γ and λ are assigned a value of 1 while tuning the hyper-parameters. The role of the L value is to avoid overfitting by decreasing the contribution of each successive tree ($0 < L < 1$). The accuracy of the model could be increased by increasing the T value, while a greater T value may also cause an overfitting problem. A combination of ten-fold and grid search techniques was applied to tune these hyper-parameters for different algorithms. Grid search, as an exhaustive search, works to define the optimal combination of hyper-parameters values. The different parameters spaces are defined as D ε [1, 2, …, 10], S ε [10 %, 20 %, 25 %, 30 %, 50 %, 75 %, 100 %], T ε [1, 2, …, 4000], and L ε [0.0001, 0.0005, 0.001, 0.005, 0.008, 0.009, 0.01, 0.02, 0.03, 0.05, 0.1, 0.5]. While the learning rate values are commonly assumed to fall between 0.1 and 0.3, this study implemented a wider range of learning rate values due to a large number of trees (1-2000). The searched learning rate values and the varying step size was determined based on sensitivity analysis and preliminary investigation using different values.

Lastly, we use k-fold cross-validation to shuffle the dataset randomly to train and test the model and find the optimal one.

5. Results

5.1. Model validation

A ten-fold cross-validation technique was applied to the XGboost model to select the optimal combination of hyper-parameters. The 70 % training/validation dataset was divided into ten subsets. Then, the model training was performed using nine subsets and the validation was carried out using the remaining subset. This process was repeated ten times by every time changing the validation subset. The AUC value, which is the area under the ROC curve, was then used to evaluate the performance of the model over the training and validation data. Regarding the AUC measurement, the closer the value is to 1, the higher the performance of the model. The final best combination for hyper-parameters was L = 0.01, D = 15, S = 20 %, T=300.

5.2. Model performance

Soleimani et al. (2019a), created a consolidation model for Louisiana that predicts crossings within the state that are the best candidates for closures. This current study generally tries to customize that model for EBRP to make sure that this model works accurately for a smaller part of Louisiana. To do so, 16 new spatial variables and one variable based on the crash report, as discussed earlier, were added to the data of the EBRP. Before running the model, the rest of the crossing characteristics from the FRA database were added to the EBRP data. Although these variables had already been tested in Soleimani et al. (2019a), it was necessary to retest their significance since the current study area was different compared to that of the previous study. Moreover, the percentage of missing values of specific variables were different for the different study areas, which makes the importance of the variables to be different in the final model. This underscores the important role "data" plays in machine learning analyses.

After preparing the required data, the XGboost model was applied. To evaluate the performance of the model, several accuracy measurements were calculated, including general accuracy, sensitivity, and specificity. The accuracy of consolidation models with and without considering the newly added spatial variables and crossing characteristic variables are shown in Table 5.

As expected, the accuracy of the same model with data from different study areas may be different. The developed consolidation model in (Soleimani et al., 2019a) was 97.61 % accurate with crossing data from Louisiana. However, the accuracy of the same model with data from EBRP is 80.77 %. The confusion matrix for the EBRP wrongly classified two closed crossings as open crossings. Three newly open crossings were also wrongly classified as closed ones. Several factors can explain the difference in accuracy between Louisiana and the EBRP models. First, the Louisiana model was trained using the whole crossing data of Louisiana, including urban/rural, near-city/city, agriculture/industry/residential land use, etc. However, the crossing data for the EBRP mostly includes urban, city, and residential/industry land-use type of data. Second, the number of crossings in the Louisiana data, based on which the model was trained, is much higher than that of the EBRP data. For example, from among 2,771 open public crossings in Louisiana, only 5 % (137) is in EBRP.

In contrast, the accuracy of the new model for EBRP, based on the newly added spatial variables and other crossings' variables, is 88.46 %. This project generally aims to fit a model that truly classifies closed crossings (as a positive class), as well as newly open crossings (as a negative class). Thus, it is necessary to consider both sensitivity and specificity values, as well. The sensitivity of the new model is almost 80 % (Table 5), which indicates how accurately closed crossings were classified. The sensitivity of the new model with spatial variables is more than 6% higher than the model without spatial variables. The specificity, which shows how accurately open crossings were classified, is approximately 91 % for both models, however. Sensitivity is more important than specificity in this study since it indicates how well the classification of closed crossings can be trusted.

5.3. The significance of variables

XGboost, which was used for the machine learning analysis, can calculate the importance of every variable in the developed model. XGboost returns a value named "Gain" that indicates the relative importance of each variable. The importance of variables indicates the relative influence each variable has in the final developed model. Included in Table 6 is a listing of all the important values, without considering correlations between any two variables. Based on XGboost results, the top ten significant variables are "Average School Buses on a School Day", "Typical Minimum Speed mph", "Cumulative Narrative", "Estimated Percent Trucks", "Number of Intersections in 450 m", "Maximum Timetable Speed mph", "Number of Employment Cells (100*100)", "Number of Industrial Cells (100*100)", "Number of Flood Type A Cells (100*100)", and "Crossbuck Assemblies".

The significant contribution of the two variables "Average School Buses on a School Day" and "Estimated Percent Trucks" were expected, since the higher the number of buses and trucks passing the crossing is, the higher the probability that a crossing should be kept open. Also, the higher the number of intersections within 450 m from crossings, the higher the probability of an accident near the crossing would be. Moreover, the "Cumulative Narrative" variable was identified as one of the important variables that affect future closures.

Several variables that were significant in the model developed in (Soleimani et.al., 2019a) for Louisiana and therefore expected to be significant for this model were not identified as being significant, such as "In or Near City". However, it is to be noted that the EBRP data had most

Table 5
The performance of the XGboost consolidation model for the EBRP with and without the new spatial variable.

All FRA variables model developed by (Soleimani et al., 2019a)				All FRA variables, 16 spatial variables, and 1 text mining variable			
True Class	Predicted Class			True Class	Predicted Class		
	Closed	Open			Closed	Open	
Closed	11	4	Sensitivity 0.737	Closed	12	3	Sensitivity 0.8
Open	1	10	Specificity 0.909	Open	1	10	Specificity 0.909
Correctly classified cases: 80.77 %				Correctly classified cases: 88.46 %			
AUC: 0.927				AUC: 0.896			

Table 6
Significance of variables based on XGboost.

ID	Variable	Gain %	ID	Variable	Gain value
1	Avg School Buses on a School Day	0.239	21	Posted Highway Speed mph	0.012
2	Typical Minimum Speed mph	0.092	22	Number of intersections within 150 m	0.012
3	Cumulative Narrative	0.078	23	Day Thru Train Movements 6AM to 6 P M	0.01
4	Estimated Percent Trucks	0.065	24	Total Switching Trains	0.009
5	Number of intersections within 450 m	0.062	25	Total Trains	0.007
6	Maximum Timetable Speed mph	0.047	26	Intersecting Roadway Within 500 ft	0.006
7	Number of Employment cells	0.041	27	Closest School m	0.006
8	Number of Industrial cells	0.037	28	Night Thru Train Movements 6 P M to 6AM	0.005
9	Number of Flood A cells	0.031	29	Number of intersections within 600 m	0.004
10	Crossbuck Assemblies	0.031	30	In or Near City	0
11	Number of Residential/Compact cells	0.028	31	Crossing Surface Main Track	0
12	Total Count of Flashing Light Pairs	0.027	32	Number of Agricultural cells	0
13	Smallest Crossing Angle	0.023	33	Number of Commercial cells	0
14	Number of Flood X cells	0.022	34	Number of Park cells	0
15	Closest Hospital m	0.021	35	Number of Mixed cells	0
16	Pavement Markings	0.02	36	Type of Land Use	0
17	Number of Institutional	0.02	37	Number of Traffic Lanes Crossing Track	0
18	Typical Maximum Speed mph	0.018	38	Advance Warning Signs W10-1	0
19	Number of intersections within 300 m	0.016	39	Main Tracks	0
20	AADT	0.013	40	Functional Classification Road Function	0

crossings “In City” and that could explain why the “In or Near City” variable did not show up as a significant one.

5.4. Simplified model development

For developing a classification model, it is always important to select an algorithm that results in a high-performance model. However, it is also beneficial to develop a model that is as simple as possible. A simpler model makes the practical implementation phase easier and helps to avoid the statistical bias of randomly selected parameters/variables in the model. Also, a simpler model would have fewer variables than the original model without sacrificing too much accuracy.

To make a simpler model, first, correlated variables were detected and removed. Then, variables with a higher gain value (relative importance of the variable in the model) were selected for the final model. Four different correlation coefficient (R) thresholds were used: “larger than 0.9”, “larger than 0.8”, “larger than 0.7”, and “larger than 0.5” (Minitab Blog Editor, 2016). For each correlation threshold (0.5, 0.7, 0.8, and 0.9), six different aggregated gain values (75 %, 80 %, 85 %, 90 %, 95 %, and 97 %) were explored for the desired optimized model. The performance of the simplified model for each aggregated gain and the correlation threshold combination is included in Table 7.

Illustrated in Table 7 are the accuracy measurements of different tested models. To get the results, first, the hyperparameters, including learning rate (η), model complexity (γ), and the maximum depth of the tree, were tuned. The learning rate, ranging from 0 to 1, helps to avoid the overfitting problem; model complexity, ranging from 0 to infinite, takes care of loss reduction as the larger gamma is, the more conservative the algorithm will be; and maximum tree depth, ranging from 0 to infinite, indicates the number of features in the model, as the larger the depth is, the more likely the algorithm is overfitting (XGboost Parameters, 2020). Several hyperparameter values were tested, including learning rates (0.01, 0.05, 0.1, 0.3, 0.5), max depths (15, 12, 10, 8, 6, 4, 2), gamma values (0, 1, 5, 2), and the number of trees (100, 300, 400, 500, 1000, 2000). The optimized hyperparameter combination for this project was selected as 0.05, 0, 15, and 300 for learning rate, model complexity, tree depth, and the number of trees, respectively.

The aim is to select the simplest yet optimal and most accurate model among the eighteen models in Table 7. There is a need to trade-off between smaller correlation coefficients and aggregated gain values (to have the simplest model possible), and higher accuracy measurement values.

Sensitivity shows how accurate the “Open Crossing” class data were classified, and specificity identifies the accuracy of the “Closed Crossing” classification. Both accuracy measurements are important because of the objective of the project. Therefore, it was essential to develop a model that makes a balance between all accuracy measurements. According to Table 7, model 20 has the highest accuracy value of 88.46 %. The sensitivity, specificity, and AUC were 86.66 %, 90.9 %, and 95.15 %, respectively. This model contained only 21 variables with an aggregated gain value of 95 % and a correlation threshold of 0.5.

Accuracy values of the original model with one hundred percent of gain values (including all 40 variables) and without considering correlations were 88.46 %, 80 %, 90.9 %, and 89.6 % for accuracy, sensitivity, specificity, and the area under the curve, respectively. However, while model 20 has fewer variables than the original model, the accuracy of both models was almost the same. This is the reason why the simplification of models helps to save time and budgets when it comes to the implementation phase.

To find variables included in model 20, first, correlation coefficients higher than 0.5 between variables were removed. According to the correlations between variables, there are no 0.5 or higher correlations between variables by excluding seventeen variables from the original dataset. The variables in model 20, their gain values, and their positive/negative effect on open crossings in the class variable are illustrated in Table 8.

As shown in Table 8, six variables in the model showed a negative correlation, implying they caused crossing closures, while the other

Table 7
Accuracy measurements for tested models.

Aggregate Gain	Model Parameters	Model performance (r > 0.9)		Model performance (r > 0.8)		Model performance(r > 0.7)		Model performance(r > 0.5)	
75 %	AUC	Model 1	0.7636	Model 2	0.8121	Model 3	0.8121	Model 4	0.8606
	Sensitivity		0.8		0.8666		0.8666		0.8
	Specificity		0.7272		0.5454		0.5454		0.8181
	Accuracy		76 %		73.08 %		73.08 %		80.77 %
80 %	AUC	Model 5	0.8787	Model 6	0.76363	Model 7	0.8686	Model 8	0.903
	Sensitivity		0.8686		0.8		0.8		0.8
	Specificity		0.8181		0.6363		0.7272		0.8181
	Accuracy		84.62 %		73.08 %		76.92 %		80.77 %
85 %	AUC	Model 9	0.8666	Model 10	0.8686	Model 11	0.8606	Model 12	0.93333
	Sensitivity		0.8		0.8		0.8		0.8666
	Specificity		0.909		0.909		0.8181		0.8181
	Accuracy		84.62 %		84.62 %		80.77 %		84.62 %
90 %	AUC	Model 13	0.8909	Model 14	0.87272	Model 15	0.8848	Model 16	0.93939
	Sensitivity		0.7333		0.7373		0.7373		0.8666
	Specificity		0.909		0.909		0.909		0.8181
	Accuracy		80.77 %		80.77 %		80.77 %		84.62 %
95 %	AUC	Model 17	0.9333	Model 18	0.87878	Model 19	0.9575	**Model 20**	**0.9515**
	Sensitivity		0.8		0.8		0.8		**0.8666**
	Specificity		0.909		0.909		0.909		**0.909**
	Accuracy		84.62 %		84.62 %		84.62 %		88.46 %
97 %	AUC	Model 21	0.8969	Model 22	0.8969	Model 23	0.9595	Model 24	0.9515
	Sensitivity		0.8		0.8686		0.8		0.8666
	Specificity		0.909		0.909		0.909		0.909
	Accuracy		84.62 %		88.46 %		84.62 %		88.46 %

Table 8
Gain value of variables and correlations between variables and open crossings.

	Variable	Gain Value (%)	Correlation
1	Avg School Buses on a School Day	31.06	Positive
2	Typical Minimum Speed mph	15.35	Negative
3	Estimated Percent of Trucks	10.16	Positive
4	Cumulative Narrative	5.4	Positive
5	Count Institutional	5.39	Negative
6	Crossbuck Assemblies	4.59	Positive
7	In or Near City	4.44	Negative
8	Pavement Markings	3.42	Positive
9	Smallest Crossing Angle	3.26	Positive
10	Closest Hospital m	2.8	Positive
11	Posted Highway Speed mph	2.62	Negative
12	Total Switching Trains	2.52	Negative
13	Typical Maximum Speed mph	2.29	Positive
14	Intersecting Roadway Within 500 ft	1.97	Negative

eight helped to keep a crossing open by showing a positive correlation. For example, the higher the average number of school buses and trucks passing a crossing, the less likely it is that the crossing will be considered for future closure. On the contrary, the higher the posted and minimum speed at a crossing, the more likely it is that the crossing is a future closure candidate.

The "Pavement Marking" variable has four categorical classes, including "1. No Marking", "2. Railroad Markings", "3. Stop Line", and "4. Combination". As expected, crossings that have a combination of different markings are not the best candidates for closures. For the "Smallest Crossing Angle" with three categories of "1. 1–29 degrees", "30 to 59 degrees", and "60 to 90 degrees", those crossings that have a tighter angle are best candidates for future closures. A tighter angle between a highway and a railroad limits the drivers' view and increases crash rates.

5.5. Simplified model prediction

In previous steps, the consolidation model was developed for EBRP using several spatial and non-spatial variables. The model was then simplified to only include 14 variables. Simplification of the model not only filters less important variables affecting closures, but it also decreases the complexity of the consolidation model for its future implementations. This simplified model predicts crossings that are the best candidates for future HRGC closure projects. So, by applying this model to the 111 currently open crossings, the candidate list of crossings for closure can be generated. Accordingly, the customized consolidation model for EBRP was developed and presented. This model consists of 14 variables, including "Avg School Buses on a School Day passing the crossing per day", "Typical Minimum Speed mph", "Estimated Percent of Trucks passing the crossing per day", "Cumulative Narrative", "Number of Institutional Land Use", "Crossbuck Assemblies", "In or Near City", "Pavement Markings", "Smallest Crossing Angle", "Distance to the Closet Hospital m", "Posted Highway Speed mph", "Total Switching Trains", "Typical Maximum Speed mph", and "Intersecting Roadway Within 500 ft". The accuracy of the consolidation model was 88.46 %, with 86.66 % sensitivity and 90.9 %, specificity. The model predicted 17 (approximately 15 %) crossings in EBRP as the best candidates for future crossing closures (Fig. 4.a).

6. Conclusion

This study explored the feasibility of applying cutting-edge machine learning and text mining techniques to develop a consolidation model capable of identifying the most suitable highway-railroad grade crossings (HRGC) to be closed within the East Baton Rouge Parish (EBRP) study area. The assumption was that a localized consolidation model for a small study area would be more accurate and reliable when making final decisions about future crossing closures. For a smaller geographic area, spatial variables could also be collected much easier than for a larger area.

Previous studies only considered a limited number of variables, but a crossing consolidation program is influenced by many variables of engineering, spatial, and economic value. This study investigated the role of crossing characteristics, the crash reports on the crossings, and spatial variables. Unlike previous studies that solely focused on decision systems based on expert judgment to select the most significant variables for the consolidation program, this study implemented Machine Learning, Text Mining Techniques, and Geospatial Analysis to retrieve information from HRGC data. Training these models with 70 % of the data and testing them with the remaining 30 % of the data revealed an overall accuracy of 88.46 % for the proposed model. Based on this simplified model, 15 % of current highway-rail grade crossings in EBRP should be either closed or undergo safety improvements.

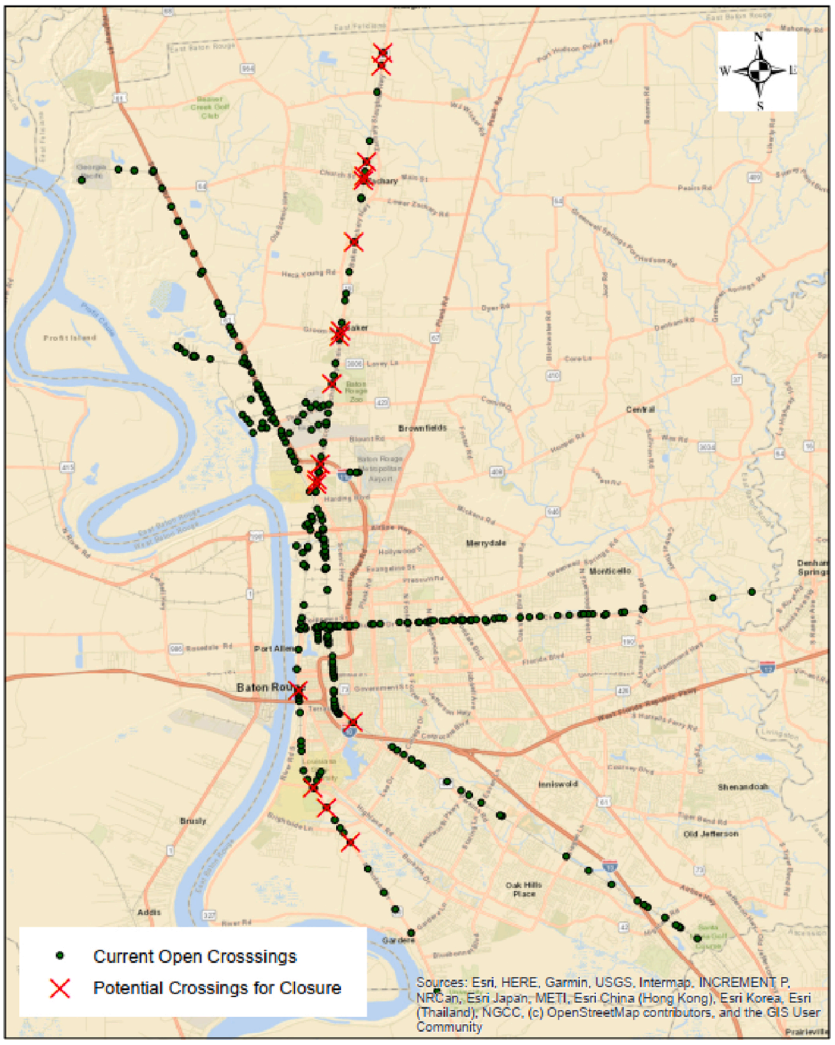


Fig. 4. The 17 potential crossing candidates for closures in EBRP based on the developed model.

However, this research has several limitations that are difficult to overcome and need to be improved in future studies. Firstly, the number of records in EBRP crossing data was small for machine learning analysis. To overcome this challenge, data-driven techniques like K-fold training as well as up-sampling the records of a smaller group were used. For future studies, it may be worth exploring adding up the records of several adjoining cities or parishes to develop a consolidation model for a larger area with more data. However, acquiring spatial attributes for a bigger geographic area may itself present another challenge. Secondly, the quality of FRA data along with its temporal and seasonal effect needs to be further improved.

Declaration of Competing Interest

The authors report no declarations of interest.

References

Bort Escabias, C., 2017. Tree Boosting Data Competitions with XGboost. Master's Thesis. Universitat Politècnica de Catalunya.

Brown, C.D., Davis, H.T., 2006. Receiver operating characteristics curves and related decision measures: a tutorial. Chemom. Intell. Lab. Syst. 80 (1), 24–38.

Chen, T., Guestrin, C., 2016, August. XGboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

Chen, H., Cao, L., Logan, D.B., 2012. Analysis of risk factors affecting the severity of intersection crashes by logistic regression. Traffic Inj. Prev. 13 (3), 300–307.

Ćirović, G., Pamučar, D., 2013. Decision support model for prioritizing railway level crossings for safety improvements: application of the adaptive neuro-fuzzy system. Expert Syst. Appl. 40 (6), 2208–2223.

Dealing with very small datasets, 2018. Kaggel Website. Accessd from: https://www.kaggle.com/rafjaa/dealing-with-very-small-datasets. Accessed by November 2020..

EBR GIS Open Data. Accessed by August 2020, Accessed from: https://data-ebrgis.opendata.arcgis.com.

Ebrahimi, F., Tushev, M., Mahmoud, A., 2020. Mobile App Privacy in Software Engineering Research: A Systematic Mapping Study. Information and Software Technology.

Federal Highway Administration, 2016. Highway-Railway Grade Crossing Action Plan and Project Priorization. Accessed: https://safety.fhwa.dot.gov/hsip/xings/fhwasa16075/fhwasa16075.pdf.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. 28 (2), 337–407.

Grauers, H., 2019. Risk, Rail and the Region: a Spatial Analysis of Regional Differences of Infrastructural Safety and the Risk of Accidents at Swedish Level-crossings.

Haleem, K., 2016. Investigating risk factors of traffic casualties at private highway-railroad grade crossings in the United States. Accid. Anal. Prev. 95, 274–283.

Heidarysafa, M., Kowsari, K., Barnes, L., Brown, D., 2018. Analysis of railway accidents' narratives using deep learning. December.. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 1446–1453.

Hu, Y., Shelton, K., 2017. Dangerous Crossings: The Links Between Intersections and Crashes in Houston.

Iranitalab, A., Kang, Y., Khattak, A., 2018. Modeling the probability of hazardous materials release in crashes at highway–rail grade crossings. Transp. Res. Rec. 2672 (10), 28–37.

Jones, K.S., 1972. A statistical interpretation of term specificity and its application in retrieval. J. Doc.

Keramati, A., Lu, P., Tolliver, D., Wang, X., 2020a. Geometric effect analysis of highway-rail grade crossing safety performance. Accid. Anal. Prev. 138, 105470.

Keramati, A., Lu, P., Iranitalab, A., Pan, D., Huang, Y., 2020b. A crash severity analysis at highway-rail grade crossings: the random survival forest method. Accid. Anal. Prev. 144, 105683.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, Vol. 26. Springer, New York.
Lee, D., Warner, J., Morgan, C., 2019. Discovering crash severity factors of grade crossing with a machine learning approach. April 2019 Joint Rail Conference. American Society of Mechanical Engineers Digital Collection.
Liu, J., Khattak, A.J., 2017. Gate-violation behavior at highway-rail grade crossings and the consequences: using geo-spatial modeling integrated with path analysis. Accid. Anal. Prev. 109, 99–112.
Liu, J., Wang, X., Khattak, A.J., Hu, J., Cui, J., Ma, J., 2016. How big data serves for freight safety management at highway-rail grade crossings? A spatial approach fused with path analysis. Neurocomputing 181, 38–52.
Lu, P., Zheng, Z., Ren, Y., Zhou, X., Keramati, A., Tolliver, D., Huang, Y., 2020. A gradient boosting crash prediction approach for highway-rail grade crossing crash analysis. J. Adv. Transp. 2020.
Ma, C., Hao, W., Xiang, W., Yan, W., 2018. The impact of aggressive driving behavior on driver-injury severity at highway-rail grade crossings accidents. J. Adv. Transp. 2018.
Minitab Blog Editor, 2016. How to Identify the Most Important Predictor Variables in Regression Models, The Minitab Blog. Accessed: http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models. Accessed by Dec, 2018.
Mousavian, M., Chen, J., 2018. Feature Selection and Imbalanced Data Handling for Depression Detection. International Conference on Brain Informatics.
Namakian, R., Shodja, H.M., Mashayekhi, M., 2014. Fully enriched weight functions in mesh-free methods for the analysis of linear elastic fracture mechanics problems. Engineering Analysis with Boundary Elements.
Namakian, R., Voyiadjis, G.Z., Kwaśniak, P., 2020. On the slip and twinning mechanisms on first order pyramidal plane of magnesium: Molecular dynamics simulations and first principal studies. Materials & Design.
Nejad, F.M., Motekhases, F.Z., Zakeri, H., Mehrabi, A., 2015. An image processing approach to asphalt concrete feature extraction. Journal of Industrial and Intelligent Information 3.
Omar, K.B.A., 2018. XGboost and LGBM for Porto Seguro's Kaggle Challenge: a Comparison.
Rahimi, A., Azimi, G., Asgari, H., Jin, X., 2020. Injury severity of pedestrian and bicyclist crashes involving large trucks. August. In: International Conference on Transportation and Development 2020. Reston, VA: American Society of Civil Engineers., pp. 110–122.
Rail Inventory Management System (RIMS) Database. Accessed: https://rims.tavlasolutions.com. Accessed by Aug, 2020.
Retallack, A.E., Ostendorf, B., 2019. Current understanding of the effects of congestion on traffic accidents. Int. J. Environ. Res. Public Health 16 (18), 3400.
Samat, A., Li, E., Wang, W., Liu, S., Lin, C., Abuduwaili, J., 2020. Meta-XGBoost for hyperspectral image classification using extended MSER-guided morphological profiles. Remote Sens. 12 (12), 1973.
Schrader, M., Hoffpauer, J., 2001a. Methodology for evaluating highway-railway grade separations. Transp. Res. Rec.: J. Transp. Res. Board (1754), 77–80.
Schrader, M., Hoffpauer, J., 2001b. Methodology for evaluating highway–railway grade separations. Transp. Res. Rec. J. Transp. Res. Board (1754), 77–80.
Soleimani, S., 2020. Using Spatial Analysis and Machine Learning Techniques to Develop a Comprehensive Highway-Rail Grade Crossing Consolidation Model..
Soleimani, S., Mousa, S.R., Codjoe, J., Leitner, M., 2019a. A comprehensive railroad-highway grade crossing consolidation model: a machine learning approach. Accid. Anal. Prev. 128, 65–77.
Soleimani, S., Mohammadi, A., Chen, J., Leitner, M., 2019b. Mining the highway-rail grade crossing crash data: a text mining approach. December. In: 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 1063–1068.
Texas Department of Transportation, 2013. Integrated Priorization Method for Active and Passive Highway-Rail Crossings. Accessed: https://static.tti.tamu.edu/tti.tamu.edu/documents/0-6642-1.pdf.
Union Pacific, 2020. New Road Crossing Openings and Conversion of Private Crossings To Public. Accessed by August 2020, Accessed from: https://www.up.com/real_estate/roadxing/industry/new_conversion/index.htm.
Williams, M., Tushev, M., Ebrahimi, F., Mahmoud, A., 2020. Modeling user concerns in Sharing Economy: the case. Automated Software Engineering.
Yıldız, K., Ateş, A.D., 2020. Evaluation of level crossing accident factors by logistic regression method: a case study. Iran. J. Sci. Technol. Trans. Civ. Eng. 1–10.
Zhang, Z., Schuller, B., 2012. Semi-supervised learning helps in sound event classification. March. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 333–336.
Zheng, Z., Lu, P., Tolliver, D., 2016. Decision tree approach to accident prediction for highway–rail grade crossings: empirical analysis. Transp. Res. Rec. 2545 (1), 115–122.
Zhou, X., Lu, P., Zheng, Z., Tolliver, D., Keramati, A., 2020. Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. Reliab. Eng. Syst. Saf. 106931.