Dimensionality Reduction and Machine Learning-based Crash Severity Prediction using Surrogate Safety Measures

Pushkin Kachroo, Anamika Yadav, Ankit Kathuria, Shaurya Agarwal, and Md Mahmudul Islam,

Abstract This paper presents a mechanism for using Machine Learning (ML) methods to extract the information efficiently from the input data collected in the field for predicting crash severity predictions. The output from the EVT engine subsequently can be used for approximately predicting traffic crashes. In our study, we use Principal Component Analysis (PCA) and Multidimensional Scaling techniques to obtain reduced dimensionality of the information obtained from multiple variables. Logisitic and Poisson regressions are used on the reduced number of variables which are the principal components, for crash severity predictions.

1 Introduction

There are many traffic Surrogate Safety Measures (SSM) that have been used in the past ([Arun et al.(2021)Arun, Haque, Bhaskar, Washington, and Sayed]), such as Time to Crash (TTC) and its various modifications namely Time exposed time to collision (TET) and Time integrated time to collision (TIT), Post-Encroachment Time (PET), delta-V (ΔV), relative speeds, or accelerations, etc. These SSMs have been used in the models for prediction of traffic crashes as well as their severities.

Since there are multiple SSMs, it makes sense to reduce the dimensionality of this data, and then use the least number of variables with most information for further

Pushkin Kachroo
Electrical & Comp. Eng., Univ. of Nevada Las Vegas (UNLV) e-mail: pushkin.kachroo@gmail.com

Shaurya Agarwal, Md Mahmudul
Civil Engineering, University of Central Florida (UCF) e-mail: iitg.shaurya@gmail.com, mahmudulroman@gmail.com

Anamika Yadav, Ankit Kathuria
Civil Engineering, Indian Institute of Technology Jammu e-mail: 2020rce2049@iitjammu.ac.in, ankit.kathuria@iitjammu.ac.in

K. R. Rao et al. (eds.), *Traffic and Granular Flow '22* , Lecture Notes in Civil Engineering 443, https://doi.org/10.1007/978-981-99-7976-9_56

processing such as predicting crashes using extreme value theory or for crash severity prediction.

In our study, we use Principal Component Analysis (PCA) and Multidimensional Scaling techniques for dimensionality reduction.

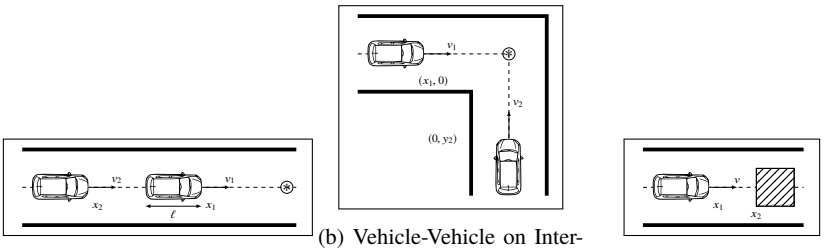
2 Surrogate Safety Measures

The following are some of the surrogate safety measures that have been proposed and used.

Time Based:

Time to Collision (TTC): Time to collision (TTC) is a classic original SSM proposed in a thesis [Hayward(1971)] with the formula for vehicle to vehicle collision, and vehicle to fixed obstacle collision given respectively by

$$TTC(t) = \frac{x_1(t) - x_2(t) - \ell}{v_2(t) - v_1(t)}, \text{ and } TTC(t) = \frac{x_2(t) - x_1(t)}{v(t)} \tag{1}$$



(a) Vehicle-Vehicle Longitudinal (b) Vehicle-Vehicle on Intersection (c) Vehicle-Obstacle

Fig. 1: TTC: Vehicle-Vehicle, and Vehicle-Obstacle

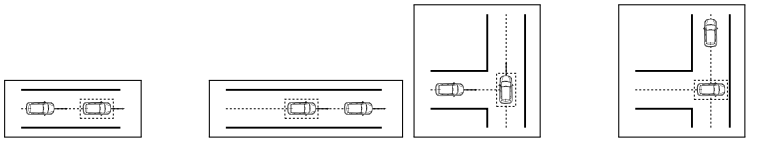
Modified Time to Collision (MTTC): A modified version of TTC is called MTTC (Modified TTC) that uses deceleration values and current speed values to assess TTC ([Ozbay et al.(2008)Ozbay, Yang, Bartin, and Mudigonda]).

As TTC is a value obtained at a given instant of time, it can be used to evaluate the safety level of a vehicle trajectory over a given time period. Three values that provide such an estimate are TET (Time Exposed TTC), TIT (Time Integrated TTC), and TTCm (Minimum TTC) ([Minderhoud and Bovy(2001)]).

TET which measures the time the vehicle is below a threshold TTC value is given by $TET = \int_{t_0}^{t_f} \mathbb{1}(TTC^* - TTC(t))dt$, where TTC^* is the threshold TTC value, t_0 and t_f are initial time and the final time on the vehicle trajectory, and the $\mathbb{1}(x)$ is the indicator function. TIT which measures the integral of the TTC function when the TTC value is below a threshold TTC value is given by $TIT = \int_{t_0}^{t_f} TTC(t)\mathbb{1}(TTC^* - TTC(t))dt$.

Finally, TTCm measures the minimum value of the TTC over a time period in a block which is between the values t_o and t_f, and$TTCm = \min_{t\in(t_o,t_f)} TTC(t)$,

Post Encroachment Time (PET): PET is the difference in time between when a vehicle occupies a spot and the following vehicle then occupies the same spot, and hence is given by $PET = t_2 - t_1$.



(a) Time t_1 (b) Time t_2 (c) Intersection:t_1 (d) Time t_2

Fig. 2: PET

Decelertion Based:
Deceleration Rate to Avoid the Crash (DRAC): The minimum deceleration needed to avoid a crash ([Cooper and Ferguson(1976)]) between vehicles can be computed by $DRAC(t) = [v_2^2(t) - v_1^2(t)]/[2(x_1(t) - x_2(t) - \ell)]$, and between a vehicle and a stationary obstacle as $DRAC(t) = v_1^2(t)/[2x_1(t)]$.

As DRAC is an instantaneous value it can also be integrated over time. We propose three measures for this extending the time based versions. Time Exposed DRAC (TED) is given by $TED = \int_{t_0}^{t_f} \mathbb{1}(TTD(t) - TTD^*)dt$, Time Integrated DRAC (TID) is given by $TID = \int_{t_0}^{t_f} TTD(t)\mathbb{1}(TTD(t) - TTD^*)dt$, whereas, Minimum DRAC (DRACm) is given by $DRACm = \min_{t\in(t_o,t_f)} DRAC(t)$.

Energy Based:
DeltaV Method uses the masses of the vehicles and their change in speeds during collision to assess the severity of a potential collision to develop a severity index. The details also depend on the angle of collision as well ([Shelby et al.(2011)]). The method depends on the calculation of $\Delta v_1 = [m_2(v_2 - v_1)]/(m_1 + m_2)$, and $\Delta v_2 = [m_1(v_2 - v_1)]/(m_1 + m_2)$, where m indicates mass, and the v terms indicate speeds as before. Crash index (CAI) [Ozbay et al.(2008)Ozbay, Yang, Bartin, and Mudigonda] uses kinetic energy concept without using DeltaV to develop a crash severity index.

3 Dimensionality Reduction

We use two methods from the multivariate analysis and Machine Learning techniques for reducing the dimension of the input SSM data. The two methods are Principal Component Analysis (PCA) and Multidimensional Scaling (MDS).

Principal Component Analysis (PCA):
PCA involves the rotation of axes in the input n-dimensional space of variables $x_1, x_2, \cdots, x_n$ such that the new variables $z_1, z_2, \cdots, z_n$ are such that they ac-

count for maximum variation in the data in the descending order ([Jolliffe(2002), Everitt and Hothorn(2011)]). We find variable $z_1 = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ to have maximum variance while satisfying $\alpha'\alpha = 1$. Formally, the problem to solve is Maximize Variance $(\alpha'\mathbf{x}) = \alpha'\Sigma\alpha$, where $\alpha'\alpha = 1$. Using Lagrange multiplier technique, the problem can be reduced to Maximize $\alpha'\Sigma\alpha - \lambda(\alpha'\alpha - 1)$. Differentiating with respect to α yields $\Sigma\alpha = \lambda\alpha$.

Hence, PCA is an eigenvalue eigenvector problem for the covariance matrix Σ of the original data. Notice that $\alpha'\Sigma\alpha = \lambda\alpha'\alpha = \lambda$. The eigenvalue gives the variance. The eigenvector corresponding to the eigenvalue with the largest magnitude is the first principal component, followed by the next highest etc. We can create dimension reduction by choosing only those k eigenvectors that correspond to the variance that accounts for more than some threshold value of overall variance. The criterion used to measure variability by dimension reduction from n to k given a threshold value θ can be based on Euclidean measure as $\text{argmin}_k(\sum_{i=1}^{k}\lambda_i^2)/(\sum_{i=1}^{n}\lambda_i^2)$, given $(\sum_{i=1}^{k}\lambda_i^2)/(\sum_{i=1}^{n}\lambda_i^2) \geq \theta$.

Multidimensional Scaling (MDS):
Multidimensional scaling induces dimension reduction ([Everitt and Hothorn(2011), Härdle and Simar(2019)]) where the analysis starts with a given distance matrix **D**, which in turn may be constructed from a similarity matrix. The relationship between the distance matrix terms and the data variable terms is given by $d_{ij} = \sqrt{(x_i - x_j)(x_i - x_j)'}$.

An innerproduct semi-positive definite matrix **M** is used for dimensionality reduction, where $\mathbf{B} = \mathbf{x}'\mathbf{x}$. Once, **B** is obtained, we solve the eigenvalue-eigenvector problem $\mathbf{B}\alpha = \lambda\alpha$, and then use Euclidean measure for dimension reduction.

4 Extreme Value Theory (EVT)

The application of extreme value theory for crash prediction works by building an extreme value distribution function by estimating its parameters using data available for traffic conflicts, such as the values of PET within some time blocks. Once this distribution is built, then we can estimate the probability of a crash by checking the probability of $PET \leq 0$ from that distribution or more accurately for a negated PET, $-PET \geq 0$ as that corresponds to a crash ([Tarko(2019), Songchitruksa(2004)]).

There are two classes of distributions that are used for crash predictions using SSMs ([Zheng et al.(2014)Zheng, Ismail, and Meng]). They are Block Maxima (BM) and Peak Over Threshold (POT). In block maxima the generalized extreme value distribution $G(z)$ is used which is given by

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \quad \left\{z : 1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0\right\} \\ -\infty < \mu < \infty, \sigma > 0, \text{ and } -\infty < \xi < \infty \tag{2}$$

and for peak over threshold, the distribution turns out to be a generalized Pareto distribution given by $G(z) = 1 - [1 - (\xi z/\sigma)]^{-1/\xi}$.

5 Logistic and Poisson Regression for Severity Predictions

Both the logistic and poisson regression models are useful for predicting the crash severity ratings as the prediction variable has a categorical type value.

Logistic Regression:
Depending on a collection of independent variables, logistic regression calculates the likelihood of an event happening. A logistic function is used as the log odds or the natural logarithm of odds, and it is expressed by $p(X) = e^{\beta_0+\beta_1 X_1}/(1 + e^{\beta_0+\beta_1 X_1})$.

Poisson Regression:
In the Poisson regression model for observation i can be modelled as $\mathrm{P}(Y_i = y_i|\mathbf{X}_i, \beta) = e^{-\exp\{\mathbf{X}_i\beta\}} \exp\{\mathbf{X}_i\beta\}^{y_i}/y_i!$.

6 Application to Field Data

In this section, we discuss the process of collecting video data from the roads and then process it using DataFromSky (DFS) software to analyze the trajectories of the vehicles and classify the crash severity ratings by human observer. We also discuss how we used two dimensionality methods: Principal Component Analysis and Multi-dimensional scaling to convert the high-dimensional dataset into low-dimensional dataset. Finally, we applied logistic and poisson regression models to predict the crash severity ratings on the lower-dimensional dataset.

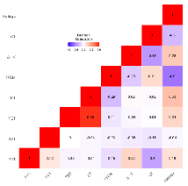
Video data was collected using drones and also by attaching cameras to high altitude infrastructure locations. The processed video was used for extracting useful data such as extraction of indicators like PET, TTC, TIT, TET, etc.

The dataset we created contains seven independent variables: First Vehicle Type (FVT), Second Vehicle Type (SVT), Time Exposed TTC (TET), Time Integrated TTC (TIT), TTC Minimum (TTCM), delta-V (Δ-V), and Conflict Type (CT) and one dependent variable that is the crash severity Ratings. In total, we put 493 observations in the dataset, and a glimpse of it is shown in table 3a. Some of the variables in the dataset hold numerical values and some of them hold categorical values. To find out the correlation among the variables, we first convert the categorical ones into numerical values. If we look at the Pearson Correlation matrix among the variables of the dataset shown in figure 3b, we can see that the dependent variable Ratings has some correlation with most of the independent variables.

PCA:
We fed all the independent variables into the PCA model and in return, it provided us with the seven principal components. The percentage variance of the seven components is shown in figure 4a. From the percentage variance of the PCs, we can see

FVT	SVT	TET	TIT	TTCM	ΔV	CT	Ratings
Car	Car	0.37	2.38	336	25.6	Rear-end	1
Motorcycle	Car	0.77	4.62	769	57.6	Crossing	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Bus	Car	1.1	5.55	3219	57.1	Rear-end	0

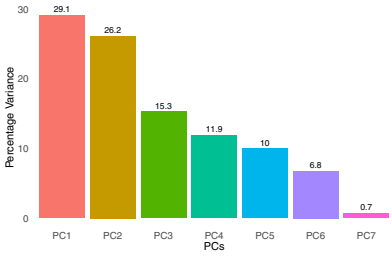
(a) Dataset before Dimensionality Reduction



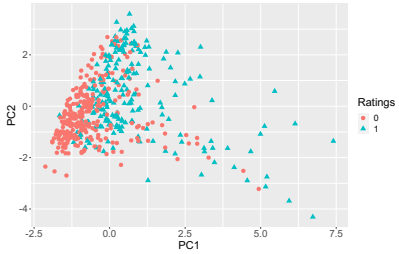
(b) Pearson Correlation of the Dataset

Fig. 3: Dataset

that the first two PCs capture 55.3 percentage data of the whole dataset. So, we chose PC1 and PC2 as the new independent variables of the lower dimensional dataset. We construct the new lower-dimensional dataset with two independent variables and one dependent variable and visualize the data samples in figure 4b.



(a) Percentage Variance of the PCs



(b) Dataset after Reducing the Dimension with PCA

Fig. 4: PCA Results

We applied logistic regression and Poisson regression to the new lower-dimensional dataset. At first, we split the dataset as a ratio of 70 and 30. Among the 493 total observations, 70 % of the data, that is 346 observations are put into the training set and the rest 147 observations which account for 30 % of the dataset are put into the testing set. The accuracy of the logistic regression model accounted for 69.36 % for the training and 78.91 % for the testing dataset, while Poisson regression achieved 69.07 % and 76.87 % on the training and testing dataset, respectively. The accuracy results of the prediction data depict that both the trained logistic and Poisson regression models are neither overfitting nor underfitting. And on the testing dataset, both the models perform well by predicting nearly 80 % of the crash severity ratings correctly.

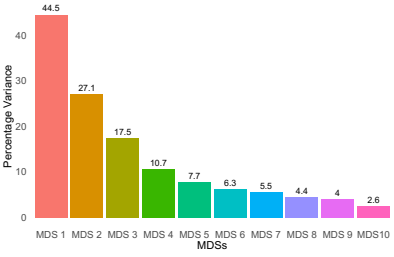
MDS:

In MDS, we utilized the average of the absolute value of the log fold changes between the observations or rows of the dataset to create the similarity matrix. We first calculates the log2 values for all the samples in the dataset. Then, we calculated

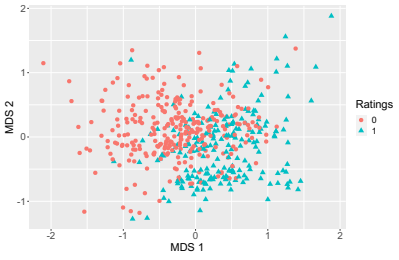
Table 1: Prediction Results of Logistic and Poisson Regression on PCA Data

	Prediction Results							
	Logistic Regression				Poisson Regression			
	Training		Testing		Training		Testing	
Actual\Prediction	0	1	0	1	0	1	0	1
0	153	42	75	12	165	30	79	8
1	64	87	19	41	77	74	26	34
Accuracy(%)	69.36		78.91		69.07		76.87	

the average difference between two observations and take the absolute value to fill out the similarity matrix. Once we formed the similarity matrix, we used the cmdscale function to apply the MDS. As we used the log2 values for creating the data matrix, we found more than seven MDSs, but to keep the figures simpler, we plot the first ten MDSs in figure 5a. The first two MDSs account for 71.6 % of the variation in the dataset.



(a) Percentage Variance of the MDSs



(b) Dataset after Reducing the Dimension with MDS

Fig. 5: MDS Results

Using the first two MDSs and the Ratings variable, we created the new lower-dimensional dataset, which is visualized in figure 5b. We see that MDS can distinguish between the data samples more accurately than PCA. We can see from table 2 that the accuracy on the training set is improved by around 5.5% and 3% for logistic and poisson regressions, respectively. Results on the testing dataset remained around 1% in the prediction results. Overall, logistic regression performs well over the poisson regression on the lower-dimension dataset obtained from the MDS analysis.

Table 2: Prediction Results of Logistic and Poisson Regression on MDS Data

	Prediction Results							
	Logistic Regression				Poisson Regression			
	Training		Testing		Training		Testing	
Actual\Prediction	0	1	0	1	0	1	0	1
0	154	41	71	16	168	27	75	12
1	46	105	13	46	70	81	25	35
Accuracy(%)	74.85		80.27		71.96		74.82	

7 Conclusions

This paper proposed a method for data compression for SSms into a low dimensional space. Logistic and Poisson regression were proposed for estimating collision severity. Finally, the technique was applied to a collected field data and the results presented.

References

[Arun et al.(2021)Arun, Haque, Bhaskar, Washington, and Sayed] Arun A, Haque MM, Bhaskar A, Washington S, Sayed T (2021) A systematic mapping review of surrogate safety assessment using traffic conflict techniques. Accident Analysis & Prevention 153:106016

[Cooper and Ferguson(1976)] Cooper DF, Ferguson N (1976) Traffic studies at t-junctions. 2. a conflict simulation record. Traffic Engineering & Control 17(Analytic)

[Everitt and Hothorn(2011)] Everitt B, Hothorn T (2011) An introduction to applied multivariate analysis with R. Springer Science & Business Media

[Härdle and Simar(2019)] Härdle WK, Simar L (2019) Applied multivariate statistical analysis. Springer Nature

[Hayward(1971)] Hayward JC (1971) Near misses as a measure of safety at urban intersections thesis. Dept of Civil Engineering, The Pennsylvania State University, Pennsylvania

[Jolliffe(2002)] Jolliffe IT (2002) Principal Component Analysis, Second Edition. Springer

[Minderhoud and Bovy(2001)] Minderhoud MM, Bovy PH (2001) Extended time-to-collision measures for road traffic safety assessment. Accident Analysis & Prevention 33(1):89–97

[Ozbay et al.(2008)Ozbay, Yang, Bartin, and Mudigonda] Ozbay K, Yang H, Bartin B, Mudigonda S (2008) Derivation and validation of new simulation-based surrogate safety measure. Transportation research record 2083(1):105–113

[Shelby et al.(2011)] Shelby SG, et al. (2011) Delta-v as a measure of traffic conflict severity. In: 3rd International Conference on Road Safety and Simulati. September, pp 14–16

[Songchitruksa(2004)] Songchitruksa P (2004) Innovative non-crash-based safety estimation: An extreme value theory approach. PhD thesis, Purdue University

[Tarko(2019)] Tarko A (2019) Measuring road safety with surrogate events. Elsevier

[Zheng et al.(2014)Zheng, Ismail, and Meng] Zheng L, Ismail K, Meng X (2014) Freeway safety estimation using extreme value theory approaches: A comparative study. Accident Analysis & Prevention 62:32–41