Accident Analysis and Prevention 108 (2017) 27–36

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Comparison of four statistical and machine learning methods for crash severity prediction

Amirfarrokh Iranitalab[a,*], Aemal Khattak[b]

[a] Department of Civil Engineering and Nebraska Transportation Center, University of Nebraska-Lincoln, 330P Prem S. Paul Research Center at Whittier School, Lincoln, NE, 68583-0851, United States

[b] Department of Civil Engineering and Nebraska Transportation Center, University of Nebraska-Lincoln, 330E Prem S. Paul Research Center at Whittier School, Lincoln, NE, 68583-0851, United States

ARTICLE INFO

Keywords:
Traffic crash severity prediction
Multinomial logit
Nearest neighbor classification
Support vector machines
Random forests
Crash costs

ABSTRACT

Crash severity prediction models enable different agencies to predict the severity of a reported crash with unknown severity or the severity of crashes that may be expected to occur sometime in the future. This paper had three main objectives: comparison of the performance of four statistical and machine learning methods including Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM) and Random Forests (RF), in predicting traffic crash severity; developing a crash costs-based approach for comparison of crash severity prediction methods; and investigating the effects of data clustering methods comprising K-means Clustering (KC) and Latent Class Clustering (LCC), on the performance of crash severity prediction models. The 2012–2015 reported crash data from Nebraska, United States was obtained and two-vehicle crashes were extracted as the analysis data. The dataset was split into training/estimation (2012–2014) and validation (2015) subsets. The four prediction methods were trained/estimated using the training/estimation dataset and the correct prediction rates for each crash severity level, overall correct prediction rate and a proposed crash costs-based accuracy measure were obtained for the validation dataset. The correct prediction rates and the proposed approach showed NNC had the best prediction performance in overall and in more severe crashes. RF and SVM had the next two sufficient performances and MNL was the weakest method. Data clustering did not affect the prediction results of SVM, but KC improved the prediction performance of MNL, NNC and RF, while LCC caused improvement in MNL and RF but weakened the performance of NNC. Overall correct prediction rate had almost the exact opposite results compared to the proposed approach, showing that neglecting the crash costs can lead to misjudgment in choosing the right prediction method.

1. Introduction

Motor vehicle crash severity modeling has been an important topic of traffic safety research for the past many years. It involves the use of statistical techniques (and recently data mining/machine learning methods) to gain insights into factors that affect or are associated with crash severity, along with predicting the severity outcome of crashes with unknown severity levels. This study is focused on crash severity prediction and its application using four statistical and machine learning methods.

Different agencies may benefit from the ability to predict the severity of a reported crash with unknown severity or the severity of crashes that may be expected to occur sometime in the future. In this study, three types of these stakeholder agencies are mainly considered: transportation safety planners who are interested in predicting the crash costs imposed upon a community in a future year; the hospitals and agencies that provide emergency care that need to predict the injury severity of people involved in traffic crashes using out-of-hospital variables to be able to provide them with proper medical care as fast as possible; and insurance companies that determine their costumers premiums and perform their economic analysis based on a lot of factors, including the costs of the possible crashes they might be involved in, which depends on crash severity.

This paper has three main objectives: 1) comparison of the performance of four statistical and machine learning methods including Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM) and Random Forests (RF), in predicting traffic crash severity; 2) developing a crash costs-based approach for comparison of crash severity prediction methods; and 3) investigating the effects of data clustering methods comprising K-means

* Corresponding author.
E-mail addresses: airanitalab2@unl.edu (A. Iranitalab), khattak@unl.edu (A. Khattak).

http://dx.doi.org/10.1016/j.aap.2017.08.008
Received 24 February 2017; Received in revised form 3 August 2017; Accepted 6 August 2017
Available online 06 September 2017
0001-4575/

Clustering (KC) and Latent Class Clustering (LCC), on the performance of crash severity prediction. A literature review is presented in the second section of this paper while the methods mentioned above along with the proposed comparison method are introduced in the third section. The fourth section of this paper presents a crash dataset that the modeling was based on. The fifth section presents a comparison of the prediction results using the proposed approach. Conclusions and discussion about the performance of these prediction methods, the effects of clustering on prediction and the application of the proposed comparison approach complete the paper.

2. Literature review

This literature review is focused on crash severity modeling and prediction and use of prediction and clustering methods in traffic safety research. A significant body of safety research is focused on crash severity modeling. A common approach in these studies is using a statistical modeling tool with crash severity as the dependent variable and characteristics of crash, driver, roadway, weather, etc. as independent variables. Some research papers reported using binary discrete outcome models (e.g., binary logit or probit) with two levels of crash severity (Fan et al., 2016; Khattak et al., 1998; Shibata and Fukuda, 1994) or multinomial models (e.g. Multinomial Logit (MNL) and Probit) with multiple levels of crash severity (Malyshkina and Mannering, 2010; Ye and Lord, 2014). A number of studies used generalized extreme value models, such as nested logit, to address the correlation of unobserved portion of utility of crash severity levels (Shankar et al., 1996; Yasmin and Eluru, 2013), while ordered probit and logit models were used to account for the ordinal nature of crash severity variable in many other studies (Khattak et al., 2002; Kockelman and Kweon, 2002). The prediction performance of crash severity models is mostly used as a validation tool for the estimated/trained models, rather than being the major focus of papers. Data mining/machine learning algorithms are more frequently observed as prediction tools in the literature, rather than statistical methods. For example, Artificial Neural Networks (ANN) (Abdel-aty and Abdelwahab, 2004; Abdelwahab and Abdel-Aty, 2001; Delen et al., 2006), and Decision Trees (Abellán et al., 2013; Chong et al., 2005) were used in a number of traffic safety-related papers. (Abdelwahab and Abdel-Aty, 2001) compared the prediction performance of two well-known ANN paradigms, the multilayer perceptron (MLP) and fuzzy adaptive resonance theory (ART) neural networks on crash severity data with ordered logit models and found out that MLP had the best overall correct prediction.

Literature pertaining to the machine learning methods used in this paper was reviewed. The use of Support Vector Machines (SVM) was reported in some traffic safety-related studies (Chong et al., 2005; Li et al., 2012). (Li et al., 2012) compared the performance of SVM with ordered probit on modeling crash injury severity, and found it is better in terms of prediction and comparable regarding investigating the impacts of variables on crash injury severity. While not very frequent, Random Forests (RF) was used in crash severity modeling and prediction and its performance was reported satisfying (Das et al., 2009; Harb et al., 2009). Further, (Lv et al., 2009) used Nearest Neighbor Classification (NNC) for real-time highway traffic accident prediction; however, this review of traffic crash severity literature did not uncover any research that used NNC.

To account for heterogeneity in crash data (presence of unobserved factors that are correlated with observed variables), a number of crash severity studies utilized data clustering methods. While some papers preferred Latent Class Clustering (LCC) as a model-based clustering method for clustering crash data to obtain homogeneity prior to crash severity modeling (De Oña et al., 2013; Eluru et al., 2012; Kaplan and Prato, 2013; Zhao et al., 2016), others used K-means Clustering (KC) in traffic crash analysis (Anderson, 2009; Mauro et al., 2013; Zhang et al., 2007). (Mohamed et al., 2013) showed that LCC provided better results in a crash dataset compared to KC, while KC had better performance in another crash dataset, relative to LCC.

In summary, the reviewed literature on crash injury severity showed that significant attention has been on crash severity modeling but injury outcome prediction was not a common major focus. Statistical models were more frequently used in crash severity modeling compared to machine learning methods, while machine learning methods were mostly used as prediction tools. MNL, SVM, and RF were found to be used in crash severity modeling with varying popularity. Clustering methods LLC and KC were also found reported in crash analysis in general and crash severity modeling in particular, with varying levels of success.

3. Methodology

In this research, several prediction methods (also known as classification methods) were used for crash severity prediction along with two methods for clustering the crash data. This section presents the prediction and clustering methods utilized in this study. Also, an approach for comparison of the crash severity prediction accuracy, based on the monetary costs of traffic crashes, is proposed and discussed later.

3.1. Prediction methods

In statistics and machine learning, prediction (classification) methods are used for predicting the class (category or population) of an observation, based on information extracted from a dataset consisting of observations with known classes (also known as training data). This study used prediction methods Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM), and Random Forests (RF) for classifying crash injury outcomes.

3.1.1. Multinomial logit (MNL)

Multinomial logit is the most widely used discrete choice model (Train, 2002) and has a long history of use in crash severity literature. Assume that the n^{th} crash observation with severity level i has the severity utility function V_{in}:

$$V_{in} = \boldsymbol{\alpha_i} + \boldsymbol{\beta_i X_{in}} + \varepsilon_{in} \tag{1}$$

In the above equation, V_{in} is a function of independent variables that determines the severity, $\boldsymbol{\alpha_i}$ is a constant parameter, $\boldsymbol{\beta_i}$ is a vector of model parameters (coefficients), $\boldsymbol{X_{in}}$ is a vector of observable characteristics (independent variables) that influence severity, and ε_{in} is an error term that accounts for unobserved effects, which are assumed independently and identically distributed with a generalized extreme value distribution. If $P_n(i)$ is the probability of occurrence of crash severity level i for observation n, then:

$$P_n(i) = \frac{e^{\alpha_i + \beta_i X_{in}}}{\sum_i e^{\alpha_i + \beta_i X_{in}}} \tag{2}$$

So, the probability of occurrence of each crash severity level for a crash with unknown severity level can be calculated after the model is estimated and the affiliated severity level can be predicted based on the calculated probabilities (Savolainen et al., 2011; Train, 2002).

3.1.2. Nearest neighbor classification (NNC)

NNC, also known as k-nearest neighbors, is a prediction method that classifies an observation of interest by looking at the closest k observations. The class that the majority of the k closest observations belong to, is predicted as the class of the observation of interest. In other words, nearest neighbor decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points (Cover and Hart, 1967). In the implementation of NNC two decisions are needed: the value of k, and the distance function to use (Elkan, 2011). In this paper the value of k was determined by trying different values for this quantity and finding the best prediction

accuracy and the Euclidean distance was used as the distance function. This distance can be interpreted as the physical distance between two p-dimensional points (Hewson, 2009). If d_{ij} denotes the distance between observations i and j and x_{ik} and x_{jk} represent the values of the kth variable for observations i and j, respectively, the Euclidean distance can be defined as:

$$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right)^{\frac{1}{2}} \quad (3)$$

3.1.3. Support vector machines (SVM)

SVM is a machine learning approach originally developed by Vapnik et al. (Boser et al., 1992; Vapnik, 1998). It is a system for efficiently training the linear learning machines in the kernel-induced feature spaces, while respecting the insights provided by the generalization theory, and exploiting the optimization theory (Cristiani and Taylor, 2000). It includes a set of related supervised learning methods used for prediction and regression. The statistical learning theory and structural risk minimization are the theoretical foundations for the learning algorithms of SVMs. It has been found that SVMs show comparable or better results than the outcomes estimated by other statistical and machine learning methods (Kecman, 2005). In this research, the C-classification SVM (also known as C-SVM) was utilized, while the value of γ (a parameter of nonlinear SVM with a Gaussian radial basis kernel function) for the kernel function and C, an upper bound for constraints of the minimization problem (that needs to be solved during the training process) were to determine. C-classification and Gaussian radial basis kernel function were selected to use, as they are the most appropriate combination for predicting discrete outcomes with SVM (Chang and Lin, 2011). For more details on C-SVM and the affiliated kernel function and its parameters the readers may refer to (Kecman, 2005) and (Chang and Lin, 2011).

3.1.4. Random forests (RF)

In the literature of machine learning, RF is known as a prediction method categorized as ensemble learning (methods that generate many classifiers and aggregate their results) (Liaw and Wiener, 2002). (Breiman, 2001) proposed this method as a prediction tool (classifier) consisting of a collection of tree-structured classifiers with independent identically distributed random vectors, while each tree casts a unit vote for the most popular class at input. RFs are a combination of tree predictors (in the context of decision tree learning which is a popular machine learning method) such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). In RF, each node is split using the best among a subset of predictors randomly chosen at that node. This method has performed very well compared to many other commonly-used classifiers, and is robust against overfitting (Breiman, 2001). Implementing this method requires determination of the models' parameters, including number of trees to grow and number of variables randomly sampled as candidates at each split.

3.2. Clustering methods

Clustering analysis divides a dataset into subsets called clusters in a manner that maximizes homogeneity and heterogeneity of elements within and between clusters, respectively (Shaheed and Gkritza, 2014). In this paper, assuming that homogeneity of crash data can help increase prediction accuracy, two methods were utilized to cluster the data and implement the prediction methods on each cluster, separately. Aggregated results were considered as the final prediction output of each joint cluster-prediction approach. The two clustering methods used in this study were K-means Clustering (KC) and Latent Class Clustering (LCC); each is described next.

3.2.1. K-means clustering (KC)

KC is a non-hierarchical clustering method, (an initial number of clusters is pre-specified and is not calculated through the algorithm) that groups the observations of a dataset in k clusters, based on the physical distance of the observations from clusters' means. If the dataset is comprised of p variables, this algorithm assigns the observations to the clusters with the nearest mean in a p-dimensional space. Given a number of k random starting points, the data are clustered, the means recalculated and the process iterates until stable (Hewson, 2009). The KC calculations in this study were done using the "Hartigan-Wong" algorithm (Hartigan and Wong, 1979). Similar to NNC, a distance function is selected prior to the algorithm implementation, along with the value of k. In this study, the Euclidean distance (Eq. (3)) was used as the distance function. The value of k was determined by trying different values and comparing the results in terms of the amount of within-cluster homogeneity that is accounted for by k clusters.

3.2.2. Latent class clustering (LCC)

As a model-based clustering method, LCC is probabilistic and assumes data comes from a mixture of several probability densities (Mohamed et al., 2013). The basic LCC model has the form (Hagenaars and McCutcheon, 2002):

$$f(\boldsymbol{y_i}|\theta) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y_i}|\theta_k) \quad (4)$$

in which, $\boldsymbol{y_i}$ denotes the i^{th}observation on the manifest variables (observed variables, as opposed to the latent variables), K is the number of clusters, π_k is the prior probability of belonging to cluster k and θ is the cluster specific parameter. The distribution of $\boldsymbol{y_i}$ given the model parameters θ, $f(\boldsymbol{y_i}|\theta)$, is a mixture of class-specific densities, $f_k(\boldsymbol{y_i}|\theta_k)$. Maximum-likelihood (ML) and maximum-posterior (MAP) are two estimation methods for LCC models, and an expectation-maximization (EM) algorithm is usually used by software packages to find ML or MAP estimates (Hagenaars and McCutcheon, 2002). The appropriate number of clusters in this paper was decided based on Bayesian Information Criteria (BIC) (Fraley and Raftery, 1998; Hagenaars and McCutcheon, 2002).

3.3. The prediction accuracy comparison approach

Assume that there are S levels of crash severity in a dataset and different prediction models are estimated/trained based on the dataset. To examine the accuracy of the prediction tools a validation dataset (with known crash severity levels) can be used and the prediction results for each method can be summarized in a confusion matrix (also known as an error matrix) (Stehman, 1997), as shown in Table 1. In this confusion matrix, p_{ij} is defined as the number of crashes with severity level i, predicted as severity level j and r_{ij} is the ratio of crashes with severity level i, predicted as j. In other words, $r_{ij} = p_{ij}/N_i$, while N_i denotes the actual number of crashes of severity level i in the validation

Table 1
Confusion Matrix (Error Matrix) of a crash severity prediction method.

	Crash Severity Level	Classified					Actual Number of Crashes
		1	2	3	…	S	
Original	1	p_{11}	p_{12}	p_{13}	…	p_{1S}	N_1
		r_{11}	r_{12}	r_{13}		r_{1S}	
	2	p_{21}	p_{22}	p_{23}	…	p_{2S}	N_2
		r_{21}	r_{22}	r_{23}		r_{2S}	
	3	p_{31}	p_{32}	p_{33}	…	p_{3S}	N_3
		r_{31}	r_{32}	r_{33}		r_{33}	
	⋮	⋮	⋮	⋮	…	⋮	⋮
	S	p_{S1}	p_{S2}	p_{S3}		p_{SS}	N_s
		r_{S1}	r_{S2}	r_{S3}		r_{SS}	

Table 2
Roadway Variables and their Statistics in Model Estimation and Validation Datasets.

Category	Variable	Variable Name	Modeling		Validation	
			Freq.	Ratio	Freq.	Ratio
Road Classification	Highway	highway	14081	28.70%	5498	28.37%
	Highway Ramp	highwayramp	248	0.51%	100	0.52%
	Highway Rest Area/Scale	highwayrest	60	0.12%	1	0.01%
	Interstate Mainline	interstatemain	1886	3.84%	840	4.33%
	Interstate Ramp	interstateramp	371	0.76%	152	0.78%
	Interstate Rest Area/Scale	interstaterest	13	0.03%	7	0.04%
	Local Road or Street	local	30790	62.75%	12166	62.78%
	Recreation Road	rec	37	0.08%	19	0.10%
	Others	(base)	1582	3.22%	597	3.08%
Road Characteristics	Curved and Level	Curvedlevel	1347	2.75%	571	2.95%
	Curved and on Hilltop	Curvedhilltop	60	0.12%	25	0.13%
	Curved and on Slope	Curvedslope	853	1.74%	333	1.72%
	Straight and Level	Straightlevel	36890	75.18%	14416	74.39%
	Straight and on Hilltop	Straighthilltop	906	1.85%	383	1.98%
	Straight and on Slope	(base)	9012	18.37%	3652	18.84%
Road Surface Type	Asphalt	Asphalt	16837	34.31%	6629	34.21%
	Brick	Brick	361	0.74%	118	0.61%
	Concrete	Concrete	30873	62.92%	12249	63.20%
	Dirt	Dirt	59	0.12%	29	0.15%
	Gravel	Gravel	925	1.89%	348	1.80%
	Other	(base)	13	0.03%	7	0.04%
Road Surface Condition	Dry	Dry	39924	81.36%	15312	79.01%
	Ice	Ice	1377	2.81%	585	3.02%
	Sand, Mud	Sandmud	107	0.22%	33	0.17%
	Slush	Slush	259	0.53%	139	0.72%
	Snow	Snow	2135	4.35%	756	3.90%
	Water	Water	44	0.09%	27	0.14%
	Wet	Wet	5184	10.56%	2510	12.95%
	Other	(base)	38	0.08%	18	0.09%
Number of Lanes (as dummy variables)	One	one	2075	4.23%	726	3.75%
	Two	two	22798	46.46%	8979	46.33%
	Three	three	2928	5.97%	1239	6.39%
	Four	four	16664	33.96%	6475	33.41%
	Five	five	1398	2.85%	707	3.65%
	Six or More	(base)	3205	6.53%	1254	6.47%
Median Type	Barrier	barrier	2547	5.19%	1016	5.24%
	Grass	grass	2349	4.79%	1002	5.17%
	None	(base)	22843	46.55%	8688	44.83%
	Painted	painted	8082	16.47%	3368	17.38%
	Raised	raised	13247	27.00%	5306	27.38%
Rut, Holes, Bumps	Yes	rut	39	0.08%	10	0.05%
Shoulders (None, Low, Soft, High)	Yes	shoulders	15	0.03%	7	0.04%
Traffic Control Device Inoperative, Missing, etc	Yes	controldevice	84	0.17%	24	0.12%
Work zone (Construction/Maintenance/Utility)	Yes	workzone	451	0.92%	127	0.66%
Worn, Travel-polished Surface	Yes	worn	10	0.02%	2	0.01%
Intersection Involved	Yes	intersection	12610	74.30%	4860	74.92%

dataset. Therefore, the calculated value of r_{ii} shows the correct prediction ratio for crash severity level i and $R_{overall} = \sum_{i=1}^{S} p_{ii} / \sum_{i=1}^{S} N_i$ is the overall correct prediction rate or overall prediction accuracy.

The majority of the crash severity prediction studies have used $R_{overall}$ for comparing prediction models, e.g. in (Abdelwahab and Abdel-Aty, 2001; Zhao and Khattak, 2015). However, under certain conditions it may not provide reliable results for two reasons: first, assume a validation dataset in which a large portion of crashes are from one specific severity level (which is common in many crash datasets). A completely-insensitive prediction model that predicts all the crashes as that specific frequent level, regardless of the independent variables will have a high $R_{overall}$, while a sensitive model that does predict other severity levels as well as the frequent level, might have a lower $R_{overall}$; second, using $R_{overall}$ implies that the predictions of all crash severity levels are equally valuable, or in other words it assumes the costs of all the crashes with different severity levels equal.

In this study an alternative approach was utilized to avoid the potential issues associated with the use of $R_{overall}$. If C_i denotes the average monetary costs of each crash with severity level i, then the actual and predicted costs of crashes in the validation dataset (or the future dataset) can be defined as:

$$Actual\ \ Overall\ \ Costs\ \ of\ \ Crashes(AOCC)(\$) = \sum_{i=1}^{S} N_i C_i \tag{5}$$

$$Predicted\ \ Overall\ \ Costs\ \ of\ \ Crashes(POCC)(\$) = \sum_{i=1}^{S} \sum_{j=1}^{S} p_{ij} C_j \tag{6}$$

So, the overall prediction error and the specific prediction error can be defined as in Eqs. (7) and (8). It should be noted that OPE represents the ratio of overall costs that the prediction model is predicting relative to the actual overall costs, while SPE captures the amount of cost of each crash on average, that is being accounted for by the prediction model. In terms of comparison of different prediction models OPE and SPE lead to similar conclusions (as the denominators are similar for different models) but, they denote two different concepts.

$$Overall\ \ Prediction\ \ Error(OPE)\left(\frac{\$}{\$}\right) = \frac{|POCC - AOCC|}{AOCC} \tag{7}$$

$$Specific \quad Prediction \quad Error\,(SPE)(\$) = \frac{POCC - AOCC}{\sum_{i=1}^{S} N_i} \tag{8}$$

The use of these measures addresses the issues associated with the use of $R_{overall}$ which considers the costs of all crashes equal. The relative amount of the crash costs that each method predicts can be used as a comparison tool among different methods and is given by 1-OPE. To account for the other shortcoming of $R_{overall}$, along with using OPE and SPE for comparison, in this approach the values of r_{ii} were considered in comparison for each crash severity level. The importance of each r_{ii}, depends on the use of the prediction model. Since, usually the most severe crashes are the least frequent (harder to predict) and at the same time are the most costly, the ratio of the correct prediction of these crashes is more important. But depending on the specific use of the prediction, other severity levels may be more important.

4. Data and variables

The Nebraska Department of Transportation supplied the 2012–2015 reported traffic crash data which were then limited to crashes involving two vehicles only; pedestrian or bicyclist involved crashes were excluded, along with single-vehicle crashes and those that involved more than two vehicles. This was due to the intention of having a relatively more homogenous crash dataset (desirable for a data that new approaches are being tested on), compared to including other types of crashes and transportation users and modes in the dataset. The original dataset included 68,448 crashes that involved two vehicles. The 2012–2014 crash data were used as the modeling/training dataset including 49,068 crashes, and the 2015 crash data were used for validation (validation dataset with 19,380 observations).

The original dataset provided five levels of crash severity based on the most serious injury outcome in each crash: 1. Property Damage Only (PDO) 2. Possible Injury 3. Visible Injury 4. Disabling Injury (2012–2014)/Suspicious Serious Injury (2015) 5. Fatal. As there were few observations in the severity level 5, it was aggregated with severity level 4 (i.e., disabling injury + fatal). So, the new severity levels were PDO, Possible Injury, Visible Injury and Disabling/Fatal Injury. The frequencies of these severity levels in the modeling/training set were 31487 (64.17%), 12079 (24.62%), 4049 (8.25%) and 1453 (2.96%) and in the validation set were 12436 (64.17%), 4703 (24.27%), 1682 (8.68%) and 559 (2.88%), respectivel..

The crash data variables were divided into six categories: roadway, drivers, vehicles, land-use, crash and environment (presented in Tables 2 and 3) Since prediction of the models were the objective of this analysis and the significance of effects of variables on crash severity were not to be examined, to avoid overfitting, the variables were selected to use in this study based on previous crash severity literature and ratio of missing values in the dataset. Inclusion of variables that were reported not statistically significant in previous studies and variables that would decrease the sample size due to missing values were avoided. Factors such as speed and crash type were found significant in some of the previous studies, but due to high ratio of missing values they were not considered in this study. The effect of speed could be partially captured by driver characteristics, and some of roadway-related variables, such as number of lanes, median type and curved/straight roads could play the same role for crash type.

The estimated models in this paper were at crash level (also known as crash macro-analysis), meaning each observation (or row) in the dataset represented a reported crash. In this type of modeling, independent variables belong to equal or higher levels of spatial distribution of entities (Huang and Abdel-aty, 2010; Savolainen et al., 2011). For example, in this case roadway, crash, land-use and environment variables. However, since driver behavior and vehicle characteristics are reported as effective factors on crash severity in the literature, in this study the lower-level variables, such as driver or vehicle related attributes were defined in an alternative way to be used in the modeling. For example, a MALE variable and a FEMALE variable were defined, which showed the involvement of a male or a female driver in each crash. If both variables were equal to one, one driver was male and the other was female. If MALE was one and FEMALE was zero, at least one driver was male, and if MALE was zero and FEMALE was one, at least one driver was female. Since the most serious injury outcome in each crash was considered as an indicator of severity of each crash, using the MALE (or FEMALE) variable in the model was equivalent to testing the hypothesis “involvement of a male (or female) driver increases the probability of a higher crash severity level”. This holds for all other driver and vehicle related independent variables in the study’s dataset and modeling procedure. Also, unlike the common crash datasets, the way these variables were defined made the simultaneous inclusion of opposite variables, e.g. MALE and FEMALE, possible in the models.

5. Modeling results

The four prediction models were estimated using the modeling/training dataset along with the clustered datasets, and the validation dataset was used for investigation of the methods’ prediction accuracies and comparison. Details of the clustering and prediction processes are presented in this section. Coding and execution of the statistical and machine learning calculations were accomplished using the computer programming language R version 3.2.2 (The R Foundation for Statistical Computing).

5.1. Clustering

The existence of an unobserved heterogeneity in crash datasets that might adversely affect the crash severity models is possible. Unobserved heterogeneity refers to the presence of unobserved relevant variables that are correlated with the observed variables in a model (Mannering and Bhat, 2014) and its presence in models might cause biased estimation results, incorrect inferences and weak prediction performance. To address this possibility, along with the original crash dataset, two clustering methods were utilized to split crashes into clusters that were supposedly homogeneous within but heterogeneous between clusters. The prediction methods were estimated on original and clustered datasets to examine the existence of these adverse effects.

As mentioned before, K-means Clustering (KC) as a non-hierarchical method and Latent Class Clustering (LCC) as a model-based method were used to cluster the data in this paper, in order to address the assumption that homogeneity can improve crash severity prediction accuracy. It should be noted that, since parts of both KC and LCC algorithms are based on randomly generated numbers, each algorithm was executed ten times and the best result was found. The crash severity variable was not used as a variable that affects clustering.

The amount of the within cluster sum of squares that is accounted for by the clustering, relative to the total some of squares was used to decide on the number of clusters in KC, k. Ten values (1–10) for this quantity were considered and the results were compared using a scree-plot (Fig. 1). This plot shows the number of clusters versus the within cluster sum of squares relative to the total some of squares. Using the scree-plot, 6 was chosen as the value of k as a leveling off in the plot at this point can be observed (showing that 7 or more clusters do not account for a larger amount of variability in the data compared to 6 clusters).

Roadway characteristics variables were chosen as the manifest variables in the LCC (variables that were used for clustering). This means that the resulted clusters were supposed to be homogenous in terms of roadway characteristics. The Bayesian Information Criterion (BIC) was used to determine the appropriate number of clusters. Different number of clusters, 1–10, were tested and the results were demonstrated in a scree-plot (Fig. 2). A model with a lower value of BIC is preferred (Fraley and Raftery, 1998 Hagenaars and McCutcheon,

Table 3
Drivers, Vehicles, Crash, Land-Use and Environment Variables and Their Statistics in Model Estimation and Validation Datasets.

Category	Variable	Variable Name	Modeling		Validation	
			Freq.	Ratio	Freq.	Ratio
Drivers						
Alcohol Related	Yes	alcohol	1264	2.58%	501	2.59%
Driver Less Than 25	Yes	less25	22673	46.21%	8793	45.37%
Driver Between 13 and 19	Yes	bet1319	11868	24.19%	4539	23.42%
Male Driver Involved	Yes	MALE	37477	76.38%	14891	76.84%
Female Driver Involved	Yes	FEMALE	34299	69.90%	13354	68.91%
Vehicles						
Double Bottom Trailer Involved	Yes	double	81	0.17%	35	0.18%
Tractor Trailer Involved	Yes	tractor	1636	3.33%	700	3.61%
Farm Equipment Involved	Yes	farm	145	0.30%	59	0.30%
School Bus Involved	Yes	school	229	0.47%	98	0.51%
Total Truck/Buses	0	total.truck.buses	45632	93.00%	17959	92.67%
	1		3128	6.37%	1298	6.70%
	2		308	0.63%	123	0.63%
Crash Characteristics						
Accident in Traffic	Yes	accidentintraffic	44407	90.50%	17611	90.87%
Land-use Characteristics						
Public/Private Property	Private	(base)	47475	96.75%	18781	96.91%
	Public	public	1593	3.25%	599	3.09%
Population Group	Municipal	municipal	2795	5.70%	1059	5.46%
	Rural	rural	7761	15.82%	2650	13.67%
	Urban	(base)	38512	78.49%	15671	80.86%
Environment						
Light Condition	Dark	dark	7293	14.86%	3138	16.19%
	Dawn/Dusk	dawn.dusk	2045	4.17%	893	4.61%
	Daylight	day	39706	80.92%	15337	79.14%
	Other	(base)	24	0.05%	12	0.06%
Weather Condition	Blowing Sand, Soil, Dirt, Snow	blowing	318	0.65%	155	0.80%
	Clear	clear	34888	71.10%	13415	69.22%
	Cloudy	cloudy	10686	21.78%	4106	21.19%
	Fog, Smog, Smoke	fog	294	0.60%	58	0.30%
	Rain	rain	879	1.79%	265	1.37%
	Severe Crosswinds	severe	137	0.28%	35	0.18%
	Sleet, Hail, Freezing Rain/Drizzle	sleet	153	0.31%	66	0.34%
	Snow	snow	424	0.86%	128	0.66%
	Other	(base)	1289	2.63%	1152	5.94%
Animal in Roadway	Yes	animal	148	0.30%	75	0.39%
Glare	Yes	glare	728	1.48%	276	1.42%
Vision Obstruction	Yes	visionobs	861	1.75%	303	1.56%
Weather (cause of crash)	Yes	weathercond	2777	5.66%	1210	6.24%
Debris	Yes	debris	87	0.18%	38	0.20%
Non-highway Work	Yes	nonhighwaywork	34	0.07%	12	0.06%
Obstruction in Roadway	Yes	obsroadway	163	0.33%	73	0.38%

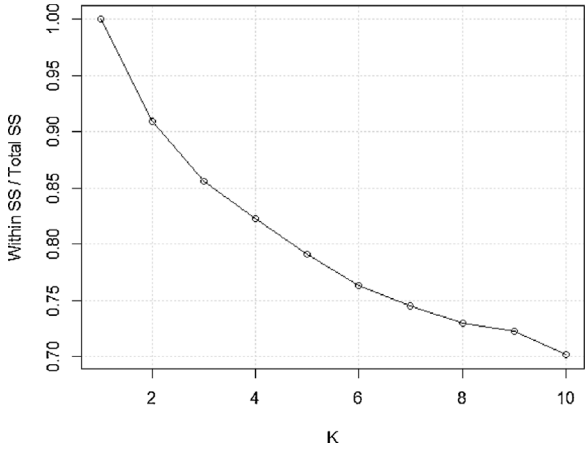

Fig. 1. Scree-plot for KC.

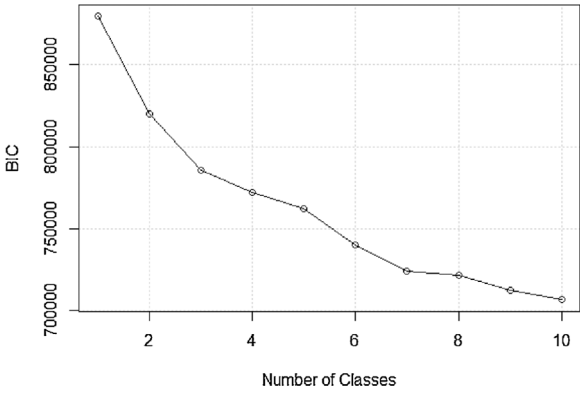

Fig. 2. Scree-plot for LCC.

2002). Similar to KC, 6 was chosen as the number of clusters.

The size ratios of the clusters (size of each cluster divided by the size of the whole dataset) for KC was 0.22, 0.17, 0.16, 0.16, 0.18 and 0.11, and for LCC was 0.23, 0.20, 0.18, 0.05, 0.28 and 0.07.

5.2. Prediction

The four prediction methods were estimated/trained using the three datasets (original, clustered by KC and clustered by LCC), and were validated with the validation dataset. Cross validation (validating models using the training/modeling datasets) was not carried out to

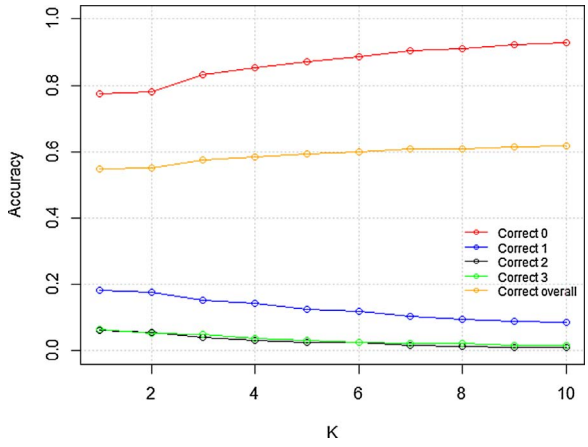

Fig. 3. Prediction Accuracy Results for the NNC method with k = 1 to 10.

avoid misjudgment due to overfitting.

The MNL model was estimated using all the independent variables, including variables with statistically significant and not significant coefficients. The reason for keeping the variables with statistically not significant coefficients in the model was that a similar model with only significant coefficients was estimated and the prediction results were weaker than the model with all the variables.

As was stated in section 3.1.2, in this paper the value of k in NNC was determined by trying different values for this quantity and finding the best prediction accuracy. This was done for values of 1–10 and the prediction results were investigated for all the crash severity levels. This is shown in Fig. 3. As can be seen, the lower values for k resulted in better correct prediction ratios for possible injury, visible injury and disabling/fatal injury severity levels, and worse correct prediction ratios for PDO and overall. Since the less frequent levels are harder to predict and they have more importance in terms of crash costs, lower values for k were more desirable and the lowest value, $k = 1$, was

Table 4
Confusion Matrix for Crash Severity Prediction Models.

Crash Severity Levels	Prediction Method	Classified							
		p_{ij}				r_{ij}			
		Level 0	Level 1	Level 2	Level 3	Level 0	Level 1	Level 2	Level 3
Level 0 N_1 = 12436	MNL	**12415**	12	2	7	**99.83%**	0.10%	0.02%	0.06%
	MNL + KMEANS	**12343**	60	15	18	**99.25%**	0.48%	0.12%	0.14%
	MNL + LCC	**12358**	59	5	14	**99.37%**	0.47%	0.04%	0.11%
	NNC	**9618**	2036	568	214	**77.34%**	16.37%	4.57%	1.72%
	NNC + KMEANS	**9530**	2075	598	233	**76.63%**	16.69%	4.81%	1.87%
	NNC + LCC	**9557**	2082	581	216	**76.85%**	16.74%	4.67%	1.74%
	SVM	**11591**	635	162	48	**93.21%**	5.11%	1.30%	0.39%
	SVM + KMEANS	**11591**	635	162	48	**93.21%**	5.11%	1.30%	0.39%
	SVM + LCC	**11591**	635	162	48	**93.21%**	5.11%	1.30%	0.39%
	RF	**10899**	1189	247	101	**87.64%**	9.56%	1.99%	0.81%
	RF + KMEANS	**10611**	1448	270	107	**85.32%**	11.64%	2.17%	0.86%
	RF + LCC	**10723**	1328	267	118	**86.23%**	10.68%	2.15%	0.95%
Level 1 N_2 = 4703	MNL	4682	**13**	0	8	99.55%	**0.28%**	0.00%	0.17%
	MNL + KMEANS	4662	**25**	6	10	99.13%	**0.53%**	0.13%	0.21%
	MNL + LCC	4655	**34**	2	12	98.98%	**0.72%**	0.04%	0.26%
	NNC	3521	**847**	251	84	74.87%	**18.01%**	5.34%	1.79%
	NNC + KMEANS	3498	**886**	233	86	74.38%	**18.84%**	4.95%	1.83%
	NNC + LCC	3536	**861**	226	80	75.19%	**18.31%**	4.81%	1.70%
	SVM	4328	**303**	61	11	92.03%	**6.44%**	1.30%	0.23%
	SVM + KMEANS	4328	**303**	61	11	92.03%	**6.44%**	1.30%	0.23%
	SVM + LCC	4328	**303**	61	11	92.03%	**6.44%**	1.30%	0.23%
	RF	4023	**542**	102	36	85.54%	**11.52%**	2.17%	0.77%
	RF + KMEANS	3900	**659**	113	31	82.93%	**14.01%**	2.40%	0.66%
	RF + LCC	3946	**591**	118	48	83.90%	**12.57%**	2.51%	1.02%
Level 2 N_3 = 1682	MNL	1671	5	**0**	6	99.35%	0.30%	**0.00%**	0.36%
	MNL + KMEANS	1657	16	**4**	5	98.51%	0.95%	**0.24%**	0.30%
	MNL + LCC	1652	20	**0**	10	98.22%	1.19%	**0.00%**	0.59%
	NNC	1226	312	**104**	40	72.89%	18.55%	**6.18%**	2.38%
	NNC + KMEANS	1207	323	**108**	44	71.76%	19.20%	**6.42%**	2.62%
	NNC + LCC	1213	322	**115**	32	72.12%	19.14%	**6.84%**	1.90%
	SVM	1527	125	**26**	4	90.78%	7.43%	**1.55%**	0.24%
	SVM + KMEANS	1527	125	**26**	4	90.78%	7.43%	**1.55%**	0.24%
	SVM + LCC	1527	125	**26**	4	90.78%	7.43%	**1.55%**	0.24%
	RF	1395	210	**52**	25	82.94%	12.49%	**3.09%**	1.49%
	RF + KMEANS	1330	258	**70**	24	79.07%	15.34%	**4.16%**	1.43%
	RF + LCC	1361	233	**62**	26	80.92%	13.85%	**3.69%**	1.55%
Level 3 N_4 = 559	MNL	547	3	0	**9**	97.85%	0.54%	0.00%	**1.61%**
	MNL + KMEANS	542	7	1	**9**	96.96%	1.25%	0.18%	**1.61%**
	MNL + LCC	542	4	1	**12**	96.96%	0.72%	0.18%	**2.15%**
	NNC	383	101	36	**39**	68.52%	18.07%	6.44%	**6.98%**
	NNC + KMEANS	387	107	36	**29**	69.23%	19.14%	6.44%	**5.19%**
	NNC + LCC	381	109	35	**34**	68.16%	19.50%	6.26%	**6.08%**
	SVM	512	35	9	**3**	91.59%	6.26%	1.61%	**0.54%**
	SVM + KMEANS	512	35	9	**3**	91.59%	6.26%	1.61%	**0.54%**
	SVM + LCC	512	35	9	**3**	91.59%	6.26%	1.61%	**0.54%**
	RF	446	71	18	**24**	79.79%	12.70%	3.22%	**4.29%**
	RF + KMEANS	423	89	20	**27**	75.67%	15.92%	3.58%	**4.83%**
	RF + LCC	423	82	24	**30**	75.67%	14.67%	4.29%	**5.37%**

selected.

For training the SVM model, the value of C, the upper bound of the constraints in the minimization problem of the c-classification algorithm and the value of γ, a parameter of the kernel function (Gaussian radial basis function) had to be determined. Literature (Chang and Lin, 2011) suggests trying small and large values for C (e.g. 1 and 1000), deciding which is better using cross validation, and then trying several γ's for the better C's. Using this procedure, the values $C = 1000$ and $\gamma = 1$ were chosen.

Training RF, needs tuning of number of trees to grow and number of variables randomly sampled as candidates at each split. To make sure that every input row gets predicted at least a few times and the results produced with different random seeds do not vary systematically (Strobl et al., 2008), different number of trees to grow was tried. While in many cases, 500 trees are sufficient, there is no penalty for having too many trees, other than longer computation time (Svetnik et al., 2003). Approaching 1000 trees, small improvement relative to the increase in computation time was started to observe. So, the value of 1000 was selected for the number of trees to grow. The number of variables randomly sampled as candidates at each split is suggested in the literature as the square root of the number of independent variables ($\sqrt{73} \cong 8$) (Svetnik et al., 2003). Again, larger values might result in better prediction performance at the price of larger computation time (Strobl et al., 2008). Larger values were tried and led to better prediction results. Finally, the value of 70 was selected for this parameter as a result of a tradeoff between prediction accuracy and computation time.

As was stated, each prediction method was executed on the original dataset and the clustered datasets (clustered by KC and LCC). This resulted in twelve sets of results, including a confusion matrix that shows the prediction frequencies and rates for each crash severity level (Table 4). Also, prediction accuracy rates are shown in Fig. 4 as bar charts, for easier comparison.

As can be observed, NNC, with and without clustering, had the best performance among these four methods in predicting crash severity levels 1, 2 and 3 (possible injury, visible injury and disabling/fatal injury), while it was the worst in prediction of level 0 (PDO). RF, also with and without clustering, was the second best in predicting levels 1,2 and 3 and the second worst in prediction of level 0. SVM was the third best in predicting levels 1 and 2, the second best in prediction of level 0 and the worst in prediction of level 3. MNL was the worst prediction model for levels 1 and 2, and the second worst in level 3, while it was the best in prediction of level 0.

Table 5
2015 Crash Costs based on Severity Level.

Crash Severity	2015 Comprehensive Crash Costs
Level 0 (PDO)	$10,000*
Level 1 (Possible Injury)	$61,700*
Level 2 (Visible Injury)	$109,400*
Level 3 (Disabling/Fatal Injury)	$913,000*

* Rounded to the nearest hundred dollars.

The effects of clustering on different methods in terms of correct crash prediction rates were not identical. While clustering did not have any effects on the results of SVM, it improved the results of MNL slightly in levels 1, 2 and 3, improved the results of NNC in levels 1 and 2, and improved the results of RF in levels 1, 2 and 3. Clustering had negative effects on correct prediction rates of the rest of the cases. The difference between the effects of KC and LCC on prediction rates was not consistent as in some cases KC had better effects and in some LCC.

As was introduced in Section 3.3, the overall prediction accuracy of the models in this paper was investigated by a proposed approach based on crash costs. To calculate, C_i, the average monetary costs of each crash of severity level i, the 2001 crash costs were extracted from the Highway Safety Manual (AASHTO, 2010), and using Consumer Price Index (CPI) and Employment Cost Index (ECI) for 2015 (the year of the validation dataset), 2001 costs were converted to 2015 costs. These costs are presented in Table 5. It should be noted that costs of level 3 (disabling/fatal injury) crashes were considered as the weighted average of costs of disabling crashes and fatal crashes (using ratios of these two severity levels as weights).

Using the proposed approach Actual Overall Crash Costs (AOCC), Predicted Overall Crash Costs (POCC) for each method, Overall Prediction Error (OPE) for each method and Specific Prediction Error (SPE) for each method, as defined in Section 3.3, are presented in Table 6. $R_{Overall}$ is also presented for each method in this table, for comparison between the common and proposed approaches. Also, $1 - \text{OPE}$ and $R_{Overall}$ are shown as bar plots in Fig. 5 for easier comparison (higher values in both plots denote a better prediction result).

The results of the proposed comparison approach showed that NNC combined with KC had the best overall performance with predicting

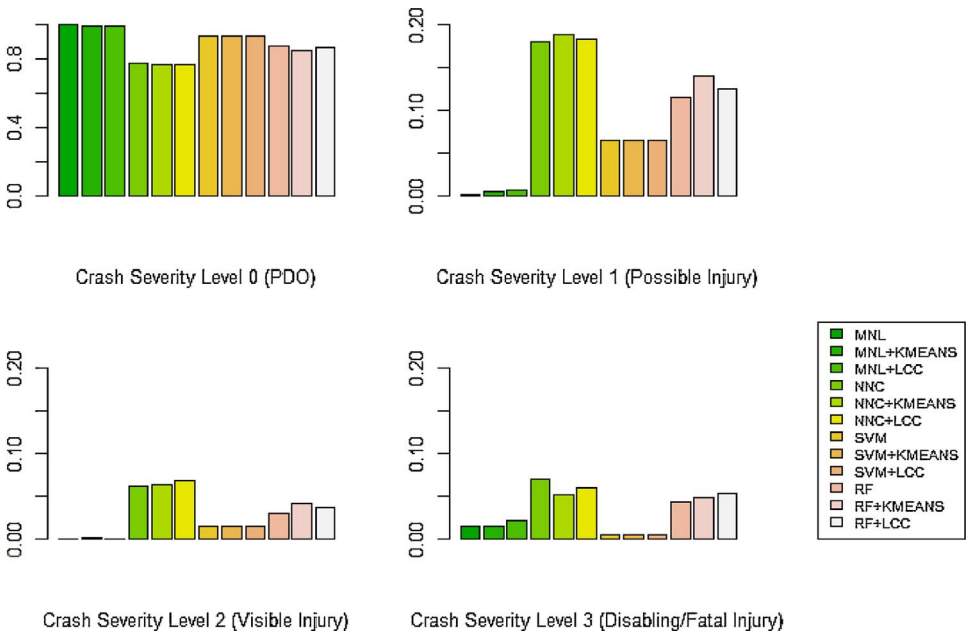

Fig. 4. Prediction Accuracy Rates.

Table 6
Overall Prediction Comparison Measures.

Method	$R_{Overall}$	POCC	OPE	1-OPE	SPE
MNL	64.17%	$222,840,600*	79.91%	20.09%	$45,700*
MNL + KMEANS	63.89%	$239,940,700*	78.36%	21.64%	$44,800*
MNL + LCC	64.00%	$244,035,200*	78.00%	22.00%	$44,600*
NNC	54.74%	$800,050,400*	27.86%	72.14%	$15,900*
NNC+KMEANS	54.45%	$820,098,700*	26.05%	73.95%	$14,900*
NNC+LCC	54.53%	$790,340,600*	28.74%	71.26%	$16,400*
SVM	61.52%	$335,871,200*	69.71%	30.29%	$39,900*
SVM + KMEANS	61.52%	$335,871,200*	69.71%	30.29%	$39,900*
SVM + LCC	61.52%	$335,871,200*	69.71%	30.29%	$39,900*
RF	59.43%	$507,502,500*	54.24%	45.76%	$31,000*
RF + KMEANS	58.65%	$538,437,600*	51.45%	48.55%	$29,400*
RF + LCC	58.85%	$556,659,900*	49.81%	50.19%	$28,500*
AOCC	$1,109,020,600*				

* Rounded to the nearest hundred dollars.

73.95% of the crash costs. NNC, in general, had the best prediction performance. RF and SVM could be sorted as next two methods, while MNL predicted only 22.00% of the crash costs in its best performance, combined with LCC, and was the weakest method. Although clustering did not affect the prediction results of SVM, KC improved the prediction accuracies of MNL, NNC and RF, and LCC improved this performance for MNL and RF, (more than KC in both cases) but weakened the performance of NNC. It should be noted that $R_{Overall}$ has almost the exact opposite results compared to the proposed approach, demonstrating that neglecting the crash costs in comparison can be completely misleading

6. Conclusions and discussion

The interpretation of the final comparison results depends on how the prediction is being used in practice. For example, safety planners that need to predict annual crash costs look at the calculated OPE's and use a method with the lowest value, which is the combination of NNC and KC. They will know that the method has underestimated the costs by 26.05%. From a hospital/emergency point of view, a model with the lowest SPE value is preferred, since prediction of costs of each crash is desirable (as a representative of the probable conditions of a crash and the people involved in it). Choosing the best method, the combination of NNC and KC, the researcher will know that the prediction is an underestimate by $14,908.25. An insurance company, similarly, will interpret both OPE and SPE.

As was mentioned, besides OPE and SPE as overall prediction methods, the values of r_{ij} can be interpreted and used, separately. As a rule of thumb it can be said that more infrequent levels of crash severity (usually more severe levels) are harder to predict. So a method that works better in predicting these levels, is usually preferable. In this paper, MNL had a good prediction performance for PDO (the most frequent level) and relatively weak performance in other injury levels. Using OPE, MNL turned out to be the weakest method overall. NNC had a very good performance in more severe (and less frequent) crash levels, and also overall, based on OPE. Looking at the POCC's and AOCC, Table 6 shows that all the models underestimated crash costs. The reason was that more severe crashes were less frequent. So, the prediction models misclassified a significant portion of the more severe crashes as less severe and consequently less costly. This led to underestimation of the overall costs of crashes, even by the best method, NNC.

There are other commonly used statistical and machine learning classification methods that were not used in this study, e.g. Discriminant Analysis and Artificial Neural Networks. Discriminant Analysis relies on a Normal distributional assumption of the data (Hewson, 2009), which is not applicable to crash datasets with a significant number of categorical variables. Artificial Neural Networks, although commonly used in crash severity literature, are not suggested with data that results in sparse matrices (data with significant portions of values of 0 among 0/1 binary variables, which is the case in this study and usual crash datasets), since it ends up in very slow convergence and poor prediction performance (Foroosh et al., 2015; Twomey and Smith, 1998).

This study investigated the performance of classification methods for crash severity prediction, based on results of the application of these methods to real crash data. Although these results do not uncover why some models outperform others, possible reasons can be identified based on mechanism of the methods. The weak performance of MNL can be attributed to its strong assumptions about data (e.g. the linear form of utility functions and distribution assumption of the error terms), which may not necessarily hold for crash data, while the three other methods are freer to learn functional forms from the training data. Among these three methods, KNN's best performance may be because this method does not make any assumptions about the functional form of data. Requirement of fewer parameters to tune and making fewer assumptions about the data may be one reason RF outperformed SVM.

Despite the efforts that were made to get the best prediction from each method, model specification in MNL, parameter tuning in NNC, SVM, RF, KC and LCC, choice of distance function in NNC and KC, and selection of type and kernel function in SVM are the factors that enter a level of uncertainty and subjectivity in the results and conclusions of this study. In other words, it is possible that different settings for the methods provide better results and change the drawn conclusions. One other limitation of this study, as was mentioned in section 4, was the imposed exclusion of some of the potentially important variables, such as speed and crash type, due to missing values from the models. In case of availability of these variables in the dataset, their inclusion in the models may lead to better prediction outcomes.

Researchers and users of crash severity prediction models may take into account crash costs in assessment of prediction methods in a similar manner as proposed herein. Policy makers and transportation agency personnel may wish to be cognizant of the level of uncertainty and inaccuracy of crash severity prediction models when making decisions. As was mentioned, all prediction methods, including the best model, underestimated the crash costs. Therefore, for future studies, other statistical models and machine learning methods for predicting crash severity may be investigated to minimize the underestimation, and their performance may be compared to the results of this study. Also, the methods used in this paper may be implemented to other crash

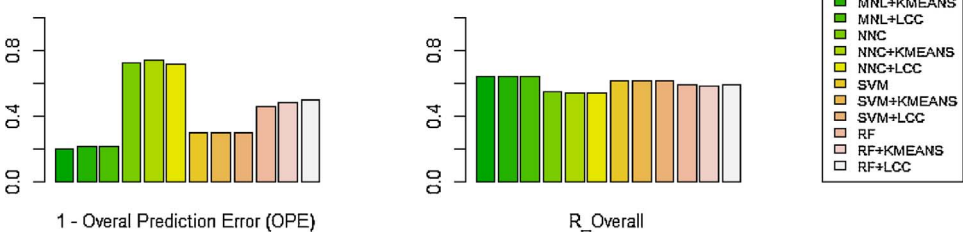

Fig. 5. 1 – OPE and $R_{Overall}$ Measures for Prediction Comparison.

datasets (e.g. multi-vehicle crashes, crashes occurred at intersection, pedestrian crashes, etc.) to investigate whether the conclusions of this paper are data-specific or generalizable. Other specifications for MNL models and different training methods for NNC, SVM and RF may be examined on crash datasets which may result in different conclusions.

References

AASHTO, 2010. Highway Safety Manual. Federal Highway Administration, Washington D.C.

Abdel-aty, M.A., Abdelwahab, H.T., 2004. Predicting injury severity levels in traffic crashes: a modeling comparison. J. Transp. Eng. 130, 204–210.

Abdelwahab, H., Abdel-Aty, M., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Rec. J. Transp. Res. Board 1746, 6–13. http://dx.doi.org/10.3141/1746-02.

Abellán, J., López, G., Oña, J.De, 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Syst. Appl.: Int. J. 40, 6047–6054. http://dx.doi.org/10.1016/j.eswa.2013.05.027.

Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accid. Anal. Prev. 41, 359–364. http://dx.doi.org/10.1016/j.aap.2008.12.014.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. Proc. 5th Annu. ACM Work. Comput. Learn. Theory 144–152 (10. 1. 1. 21 3818).

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Chang, C.-C., Lin, C.-J., 2011. Libsvm. ACM Trans. Intell. Syst. Technol. 2, 1–27. http://dx.doi.org/10.1145/1961189.1961199.

Chong, M., Abraham, A., Paprzycki, M., 2005. Traffic accident analysis using machine learning paradigms. Informatica 29.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13, 21–27.

Cristiani, N., Taylor, S.J., 2000. An Introduction to Support Vector Machines.

Das, A., Abdel-Aty, M., Pande, A., 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. J. Saf. Res. 40, 317–327. http://dx.doi.org/10.1016/j.jsr.2009.05.003.

De Oña, J., López, G., Mujalli, R., Calvo, F.J., 2013. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accid. Anal. Prev. 51, 1–10. http://dx.doi.org/10.1016/j.aap.2012.10.016.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38, 434–444. http://dx.doi.org/10.1016/j.aap.2005.06.024.

Elkan, C., 2011. Nearest Neighbor Classification. pp. 16.

Eluru, N., Bagheri, M., Miranda-Moreno, L.F., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. Accid. Anal. Prev. 47, 119. http://dx.doi.org/10.1016/j.aap.2012.01.027.

Fan, W.D., Gong, L., Washing, E.M., Yu, M., Haile, E., 2016. Identifying and quantifying factors affecting vehicle crash severity at highway-rail grade crossings: models and their comparison. Transportation Research Board 95th Annual Meeting.

Foroosh, H., Tappen, M., Penksy, M., 2015. Sparse convolutional neural networks. IEEE Conf. Comput. Vis. Pattern Recognit. 806–814. http://dx.doi.org/10.1109/cvpr.2015.7298681.

Fraley, C., Raftery, a.E., 1998. How many clusters – which clustering method – answers via model-based cluster analysis. Comput. J. 41, 578–588. http://dx.doi.org/10.1093/comjnl/41.8.578.

Hagenaars, J. a., McCutcheon, A., 2002. Applied latent class analysis. Methodology.

Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accid. Anal. Prev. 41, 98–107. http://dx.doi.org/10.1016/j.aap.2008.09.009.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a k-means clustering algorithm. J. R. Stat. Soc. Ser. C Appl. Stat. 28, 100–108.

Hewson, P.J., 2009. Multivariate Statistics with R. Multivar. Stat. with R 1–189. 10.1080/09540250802213123.

Huang, H., Abdel-aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42, 1556–1565. http://dx.doi.org/10.1016/j.aap.2010.03.013.

Kaplan, S., Prato, C.G., 2013. Cyclist-motorist crash patterns in Denmark: a latent class clustering approach. Traffic Inj. Prev. 14, 725–733. http://dx.doi.org/10.1080/15389588.2012.759654.

Kecman, V., 2005. Support vector machines–an introduction. Support Vector Machines: Theory and Applications. Springerpp. 1–47.

Khattak, A., Kantor, P., Council, F., 1998. Role of adverse weather in key crash types on limited-access: roadways implications for advanced weather systems. Transp. Res. Rec. J. Transp. Res. Board 19, 10.

Khattak, A., Pawlovich, M., Souleyrette, R., Hallmark, S., 2002. Factors related to more severe older driver traffic crash injuries. J. Transp. Eng. 128, 243–249. http://dx.doi.org/10.1061/(asce)0733-947x(2002)128:3(243).

Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34, 313–321. http://dx.doi.org/10.1016/s0001-4575(01)00028-8.

Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486. http://dx.doi.org/10.1016/j.aap.2011.08.016.

Liaw, a., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22. http://dx.doi.org/10.1177/154405910408300516.

Lv, Y., Tang, S., Zhao, H., 2009. Real-time highway traffic accident prediction based on the k-nearest neighbor method. Int. Conf. Meas. Technol. Mechatron. Autom. ICMTMA 3, 547–550. http://dx.doi.org/10.1109/ICMTMA.2009.657.

Malyshkina, N.V., Mannering, F.L., 2010. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accid. Anal. Prev. 42, 131–139. http://dx.doi.org/10.1016/j.aap.2009.07.013.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22. http://dx.doi.org/10.1016/j.amar.2013.09.001.

Mauro, R., De Luca, M., Dell'Acqua, G., 2013. Using a k-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis. Mod. Appl. Sci. 7, 11–19. http://dx.doi.org/10.5539/mas.v7n10p11.

Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., Ukkusuri, S.V., 2013. A clustering regression approach: a comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada. Saf. Sci. 54, 27–37. http://dx.doi.org/10.1016/j.ssci.2012.11.001.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43, 1666–1676. http://dx.doi.org/10.1016/j.aap.2011.03.025.

Shaheed, M.S., Gkritza, K., 2014. A latent class analysis of single-vehicle motorcycle crash severity outcomes. Anal. Methods Accid. Res. 2, 30–38. http://dx.doi.org/10.1016/j.amar.2014.03.002.

Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accid. Anal. Prev. 28, 391–401. http://dx.doi.org/10.1016/0001-4575(96)00009-7.

Shibata, A., Fukuda, K., 1994. Risk factors of fatality in motor vehicle traffic accidents. Accid. Anal. Prev. 26, 391–397. http://dx.doi.org/10.1016/0001-4575(94)90013-2.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ. 62, 77–89. http://dx.doi.org/10.1016/s0034-4257(97)00083-7.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinf. 9, 307.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.

Train, K., 2002. Discrete Choice Methods with Simulation. Cambridge Univ. Presspp. 1–388. http://dx.doi.org/10.1017/CBO9780511753930.

Twomey, J.M., Smith, A.E., 1998. Bias and variance of validation methods for function approximation neural networks under conditions of sparse data. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 28, 417–430. http://dx.doi.org/10.1109/5326.704579.

Vapnik, V.N., 1998. Statistical Learning Theory. John Wiley & Sons, Wiley New York. http://dx.doi.org/10.2307/1271368.

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accid. Anal. Prev. 59, 506. http://dx.doi.org/10.1016/j.aap.2013.06.040.

Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Anal. Methods Accid. Res. 1, 72–85. http://dx.doi.org/10.1016/j.amar.2013.03.001.

Zhang, C., Ivan, J.N., Jonsson, T., 2007. Collision type categorization based on crash causality and severity analysis. 86th Annual Meeting of the Transportation Research Board.

Zhao, S., Khattak, A., 2015. Motor vehicle drivers' injuries in train–motor vehicle crashes. Accid. Anal. Prev. 74, 162–168. http://dx.doi.org/10.1016/j.aap.2014.10.022.

Zhao, S., Iranitalab, A., Khattak, A., 2016. Investigation of pedestrian injury severity in crashes at highway-rail grade crossings using latent class analysis. Transportation Research Board 95th Annual Meeting.