Addressing Unobserved Heterogeneity in Crash Severity

A Review of Traditional and Machine Learning Methods used to Realize the Unobserved Variation in Crash Severity Analysis

Term Paper

CVEN 626

Submitted by

Pranik Koirala

Dec 2021

1 ABSTRACT

Unobserved heterogeneity is an inherent part of crash severity in crash severity data. This observation is well established in the safety research community for a while. Unobserved heterogeneity is the missing of variables from a dataset that could have explained some characteristics of the factors contributing to the event. This study is focused on the crash severity analysis and the methodology used in dealing with unobserved heterogeneity in crash severity analysis. The study aims at examining the different traditional and machine learning methodology used in literature in dealing with unobserved heterogeneity in crash data. In doing so the study also investigates the different types of data being used in crash severity analysis and the inherent unobserved heterogeneity in crash data. The study finds that the use of random parameter models coupled with clustering has the most success in addressing unobserved heterogeneity in crash severity analysis. It also infers that the use of more advanced unsupervised learning methods of clustering and machine learning can be useful in dealing with this issue. The study also suggests that with the advent of big data and numerous non-traditional sources of data, the use of machine learning and data mining can have a huge role to play in dealing with unobserved heterogeneity in crash severity analysis.

2 INTRODUCTION

Americans drove less in 2020 due to the pandemic but the rate of fatal motor vehicle related crashes increased drastically in this year. It was a surprising projection form the US department of Transportation's National Highway Traffic Safety Administration on June 2021 (NHTSA, 2021). This finding says that 38,680 people died in motor vehicle traffic crashes which is the largest projected crashes since 2007. Traffic crashes are the cause of tremendous economic, physical and emotional suffering to not only the people involved in the crash but also to many people linked to the person involved and the community as a whole. Reducing the number of fatalities and injuries, and decreasing the level of severity occurred in crashes is a huge challenge for any society. To achieve these goals, the effects of risk factors, such as road features, vehicle design, driver characteristics, weather, etc. on crash severity levels need to be carefully studied using appropriate models.

This number does not reveal the whole picture of the crashes occurring in American highways. Fatal crashes number are just a part of the picture. Victim of a crash might be injured in varying level, top being fatal. There are number of ways of defining a crash severity level which is talked about later in the paper. For many years motor vehicle crash severity modeling and proper analysis has been an important topic for highway safety research. The analysis of crash severity 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. The severity of injury is an important factor in examining crashes as it helps analyst distinguish factors contributing to different level of severity. Being an important factor, it is also a difficult task to analyze with complications for Transportation Engineers. The problem of imbalanced data or biased data due to improper reporting or under reporting of less injury crashes.

The data used in crash severity analysis is not always complete. Traditionally police reports are being used for analyzing and modeling crash severity which might not be accurate in accessing the severity (Ivan & Konduri, 2019). Linking hospital data with police reports or using other sources can be seen in the literature. The use of these new sources with traditional sources are not able to completely incorporate all the factors affecting a crash. These omittance of explanatory variables can cause biases or faulty inferences in crash severity data analysis. To address these issues inherent in crash severity data, advanced analytical methods such as random parameter models, clustering, hybrid of clustering and random parameter models are used. Recently more advanced machine learning and unsupervised learning techniques are also introduced in cash severity analysis domain. This paper aims to review these techniques used to solve the inherent issue of unobserved heterogeneity in crash severity data.

Unobserved heterogeneity is when some explanatory variables are missing in a set of data. In this condition some characteristic of the data could be explained by some variable or explanation. When relevant variables that are correlated with the independent variable observed in the data set are present in a data set it is known as unobserved heterogeneity. The importance of addressing this unobserved heterogeneity is more than important in today's scenario. With the advent of autonomous vehicle and self-driving cars it is more than essential to find the causes of crashes and

the factors effecting or influencing different level of crash severity. As (Lord & Washington, 2018) has pointed out designing and managing a safe and livable city is an important aspect in keeping our cities and surrounding sustainable and safe for future generation.

2.1 Objective

The main objective of this study is to access the state of crash severity analyses in terms of analysis methodology. More specifically, this study examines the methodological tools used to address the unobserved heterogeneity issue in the crash severity data. This study also examines the use of machine learning models in dealing with unobserved heterogeneity issues and the use hybrid model approach in dealing with unobserved heterogeneity. More advance and new analysis tools such as machine learning and artificial intelligent tools are identified that deals with the unobserved heterogeneity concern. The implication of big data and the issue is unobserved related variable in big data is also investigated.

2.2 Organization

The literature review paper titled "Addressing Unobserved Heterogeneity in Crash Severity: A Review of Traditional and Machine Learning Methods used to Realize the Unobserved Variation in Crash Severity Analysis" is divided in eight sections. Section 1 is the abstract. **Section 2** navigates through the overall background, objective and scope of the study. Likewise, **Section 3** gives a background review of what crash severity is and how and why is it studied. **Section 4** talks about the type of data used in analyzing crash severity. This section also looks into different forms of data being used for crash severity analysis and the use of big data. **Section 5** explains unobserved heterogeneity in crash severity data with examples. The methods used to analyze these data for crash severity is talked about in **Section 6** which also includes a critical review of papers using hybrid methods for overcoming unobserved heterogeneity. Then **Section 6** presents the summary and concluding remarks.

3 CRASH SEVERITY

3.1 Concept

Crash severity is a way to categorize an incident based on the amount of damage done with most severe being death of a driver or passenger. Severity of a crash can be defined taking into account different criteria but usually the level of severity of injury incurred by persons involved in the crash is a standard criterion. It is important is understand the severity of injury in any crash as it can give us understanding of the crash. If crash happened and crash did not happen was the only criteria policy makers or operators had that it would be difficult to allocate resources and prevent crashes in the future. It is also essential to understand the level of influence of different factors in crash injury level which is importuning in prevention of crash injury in the future.

3.2 Defining crash severity

KABCO is the scale usually used in defining crash severity. KABCO scale was defined by the National Safety Council and includes five levels defined by the National Highway Traffic Safety Association in the 4th (current) edition of the Model Minimum Uniform Crash Criteria (MMUCC Guideline, 2012). The guideline defines K as killed, A as Suspected Serious Injury, B as Suspected Minot Injury, C as possible injury and O as No injury. Previously the guideline suggested K as killed, A as Incapacitating Injury, B as Non-Incapacitating Injury, C as Possible Injury and O as No Injury. As the KABCO scale is used by in site police officer to code crash severity, this can sometimes not account for or create confusion in coding invisible injuries or later showing injuries. To overcome this another scale was created by the Association for the Advancement of Automotive Medicine (Gennarelli & Wodzin, 2006) known as the Abbreviated Injury Scale (AIS). The AIS coding is done by public health professionals and epidemiologists in-hospital. This in-hospital clinic assessment can be regarded as more through and reliable as it is done by professional in this field. The AIS addresses all major body region in accessing a crash involved person. Another more popular scale to measure injury severity is Injury severity score (ISS). The ISS is calculated as the sum of square of the most severe AIS scores in three different body regions. The highest possible value of ISS is 75 ad the cutoff serious injury is usually defined as ISS 16. Other measures such as MAIS is also used. In spite of these different measures, most of the data on injury severity are on KABCO scale. Now with data fusion and new rich data sources more and more data are available to complement the police report data using the KABCO score.

4 DATA

To examine the factors influencing the crash severity a lot of data is required. Usually, the crash severity data are discrete in nature; it represents discrete categories such as fatal injury or killed, incapacitating injury, non-incapacitating, possible injury, and property damage, etc. Traditionally crash reports are the major source of raw data for crash severity analysis. There are multiple shortcomings with the use of traditional data set, underreporting of crashes being just one of them. Underreporting is a frequently occurring phenomenon in crash severity dataset as most lesser injury crashes are not reported (Ye & Lord, 2011). Drivers tend to avoid minor crashes and injury mainly to not have to go the process of police system and to avoid the penalties charged by the insurance companies. The reason of underreporting can also be due to some policy of the state or local authority. Some states or local authority have laws in place that only reports a crash if a certain threshold of damage is done or exceeded by the accident.

The other limitation can be the lack of explanatory variable: the traditional crash report dataset used for crash severity omits a lot of variables that can explain a lot about the crash and the severity of injury. The data are mostly observed crash data and might overrepresent over-risk taking population in the dataset. It doesn't account for the driver's risk-taking nature, the distraction parameters during driving and so on. The traditional police report also do not report a fatal crash as fatal crash if the driver or passenger dies after a few weeks due to injury incurred during the crash.

To overcome the biases incurred by these limitations some researchers have utilized hospital record data, which is hard to find, to link them with crash data to overcome underreported crashes and later stage fatal and injury crashes. (Tarko & Azam, 2011) uses linked police-hospital data to analyze pedestrian injury to overcome the underreporting bias (selectivity bias) in data. Linking hospital data to the police report data can compensate for missing data on minor injury crashes or unreported major injuries if the driver or injured was taken to the hospital. Of course, this won't solve the whole problem of under reporting but as shown by many researchers it does fill the gap of some unreported crashes and sometime give a better understanding of the injury severity. Along with police and hospital data, a direct questioner approach can also be adopted to get a full picture of the crash severity and other factors. (Lin, Hwang, & Kuo, 2001) used direct survey method to

gather information in Taiwan's college student population. In which they four stages of assessment of sociodemographic and time-dependent characteristics of the cohort were carried out over 20-month study period with a 92% average response rate.

Big data or non-traditional data has been proposed as the solution to overcome the shortcoming of overrepresentation of over-risk-taking population in traditional data and making use of non-traditional parameters and datapoints to uncover more features and influencing factors. (Xie, Ozbay, Kurkcu, & Yang, 2017) in their study makes use Crash, transportation, land use data from satellite images and sociodemographic features, and social media data of New-York to analyze the contributing factor in crashes in those area. This kind of new data sources can be implemented in analysis crash severity and find new insights.

With the advent of more and more sources of data such as mobile phone data and road surveillance and more clear satellite images coupled with increasing computational power, it is possible of monitor and analyze each individual's behavior and the factors affecting a crash. Moreover, with the advent of autonomous driving technology vehicles are having more and more sensors in them which can be used to collect crash data. It can be possible to narrow the gap of unobserved variables with these rich sources of data. Despite the increasing amount of data, it can be said that all factors and variables are not yet possible of get on any crash which will leave some room for unobserved variable bias in crash analysis. To overcome this new and advanced machine learning and data mining techniques are being tested like (Li, et al., 2018), (Yu & Abdel-Aty, 2013) to overcome the issue of unobserved heterogeneity in traditional as well as big data.

In summary, the use of traditional police report data can have a lot of short comings like underreporting, selectivity bias, overrepresentation of risk-taking demographic, omitting important explanatory variable etc. These shortcomings can be overcome with the use of other sources of data and used in combination with police report data or just using new sources of data. Linking police-hospital data can sometimes overcome underreporting problems, use of direct survey can provide much more explanatory variables and finally the use of big data can open up a lot of opportunity for being able to understand crash severity and its impacting factors.

5 UNOBSERVED HETEROGENITY

As discussed earlier, highway crash severity data can not fully address or draw a full picture of a crash event as a crash is affected by many factors. A crash is a complex event. An understanding of factors contributing to crash severity can be tricky as many minor and major factors can be playing a part; it is impossible to have access to all of the data that could potentially determine the likelihood of a highway accident or its resulting injury severity. It is not possible to incorporate all the factors in the data while accessing a crash scene. Working with an incomplete data for analyzing crash severity may be problematic as important explanatory variables might be missing from the data. This problem in crash data is usually known as unobserved heterogeneity (Mannering and Bhat, 2014).

Unobserved heterogeneity is an inherent occurrence in traditional crash severity data. A simple example of unobserved heterogeneity can be the type of traffic control used in observing crash on an intersection (assuming that information is absent from the dataset). Another example can be gender as an observation that affects injury severity outcomes. While there are clearly some differences between men and women there is also great differences across people of the same gender, the differences might include differences in weight, height, premedical condition, bone density, etc. which are generally unavailable to the analyst. These unobserved variables create a black spot in data which might cause biased and unbalanced analysis.

6 ANALYSIS METHODS

6.1 Fixed parameter approach

A fixed parameter model is such that returns only a mean value of the estimated coefficient. In other words, the fixed parameter model treats parameters as constant across observations. There are many traditional statistical methods that make use of fixed parameter approach such as Binary output models which include Bivariate binary probit, binary logit, etc. Fixed parameter multinomial logit model and fixed parameter ordered logit models are also some examples. There have been many studies using these fixed parameter models to analyze crash severity in the past (Savolainen, Mannering, Lord, & Quddus, 2011). These models are not very useful in addressing the unobserved heterogeneity inherent in crash data. There are other methods discussed below are

used to deal with the inherit unobserved heterogeneity in crash data. One example can be (Gkritza, R., Hallmark, & Hawkins, 2010) which uses Multinomial logit models to investigates severity outcomes of farm vehicle crashes and concluded that under insufficient lighting conditions antiquated farm vehicles are more likely to lead to serious injuries or fatalities in crashes.

6.2 Partial proportional odds

A partial proportional odds model is an alternative modeling technique which tries to overcome the two assumptions of the ordered and multinomial models. The ordered model's assumption of proportional odds limits the model to analyze different effect of independent variable in different severity level. The PPO can include both a fixed parameter as well as a random parameter. The PPO method can handle the inconsistency in level (order) distribution in an ordered model. This model can also address some unobserved heterogeneity in the data. (Sasidharan & Menéndez, 2014) makes use of the PPO method to allow the covariates that meet the proportional odds assumption to affect different crash severity levels with the same magnitude; whereas the covariates that do not meet the proportional odds assumption can have different effects on different severity levels. This flexibility can in some level address the difference issue in which predictors can only have the same effect on different levels of the dependent variable without ignoring the ordered nature of crash severity data. Also, studies such as (Song & Fan, 2020), which will be discussed thoroughly later make use of PPO in combination with other techniques to better understand and address the unobserved heterogeneity in data.

6.3 Random parameter approach

A random parameter model returns a mean and a standard deviation of the estimated parameter to account for the individual-level heterogeneity in the data. In other words, model allows parameter values to vary across the population according to some pre-specified distribution. This flexibility in the method allows for the model to incorporate for any variation in parameter within the dataset. This method has been extensively used in addressing unobserved heterogeneity in crash severity data.

Use of multilevel logistic model in crash severity is a method of random parameter approach in dealing with unobserved heterogeneity. (Lenguerrand, Martin, & Laumon, 2006) uses four years of data from the French road crash data and compares crash severity estimations calculated by

multilevel logistic models (MLM), Generalized Estimating Equation models (GEE) and logistic models (LM). In this study MLM performs the best among others. The MLM seems to be the best modelling to analyses crash data, to determine risk factors, to quantify their effects and to determine their statistical links with the outcome. A multilevel model has a hierarchical or clustered structure and it allows residuals from each level to be in the hierarchy which can be useful in analyzing crash severity determining factors as these factors might be correlated to other unobserved variables. Here in this study the authors have used three levels namely: crash level, car level and occupant level where crash level is the injured individual, the car level is the model of car in the crash and crash level is the unique crash. This analysis, with its multilevel approach has taken into account the car's model/type and its relationship to the crash severity. The study has also compared the multilevel approach to logistic models and others with the multilevel approach resulting in better performance.

Summarizing section 6.1, 6.2 and 6.3 we can say that the use of more traditional statical method such as ordered logit models, multinomial logit modes and other fixed parameter methods can not address the blind spots in crash severity data or unobserved heterogeneity inherent in dataset. Trying to address this concern in analysis safety engineering make use of more sophisticated statistical tools such as random parameter methods and partial proportional odds methods which addresses the heterogeneity in the data set to some extent. There are multiple reasons to adopt random parameter approach or multilevel model. As noted, this gives us a comprehensive framework to correctly account for complex data structures, ignoring this structure when for example using some simple single level regression or classifier will leave a substantive blind spot. Just using ordered discrete outcome models or binary outcome models cannot fully explain the crash severity data as a whole and use of more complex models are needed to make sense of these datasets.

6.4 Clustering

Clustering is a broad term which refers to the set of techniques used to find subgroups or clusters within a given data set (James, Witten, Hastie, & Tibshirani, 2014). When clustering is done on a set of data, usually data set with some similarities are grouped together. The techniques of grouping these similar subsets of data requires a domain specific consideration. In transportation safety analysis, especially crash severity analysis sib groups are divided to account for unobserved

heterogeneity that might explain some characteristic or relationship between environmental variable, driver variable or vehicle character with crash severity. For instance, a driver's risk-taking behavior might increase in some environmental condition which can be accounted for by sub dividing the data set according to different environmental conditions.

6.4.1 Segregation

Segregation is the simplest form of clustering seen in crash severity literature. The basic idea of simple segregation is to group data into number of subsets based on domain knowledge. For example, (Wahi, Haworth, Debnath, & King, 2018) in their 2018 study investigates the difference between the crash severity in different traffic controls. The study concludes that the most important influencing factor in crash severity in no traffic control, signalized traffic and stop sign control are totally different. Here the separation of data was done on the basis of the type of intersection control only without using any model. Despite its simple nature this method is not always used by researchers as this might miss some other characteristics having correlations and this method also requires the researchers to exactly know the characteristics that might have homogenous nature in the whole dataset. To address some of these limitations other methods are implemented which are discussed below.

6.4.2 Latent class cluster

Latent class clustering has two noticeable advantages compared to more traditional clustering methods: (i) The optimized number of clusters is decided by different statistical criteria. Which means it does not have to be given in advance; (ii) standardization is not a requirement (Liu & (David), 2020). We can find the use of LCC in (Cerwick D. M., Gkritza, Shaheed, & Hans, 2014) where they compare mixed logit and latent class method for accommodate the individual unobserved heterogeneity and found the superiority of latent class cluster method over the mixed logit model alone.

6.4.3 K-mean clustering (unsupervised clustering)

A multivariate statistical method which is used to classify different observations into given K groups by their internally homogeneous and externally heterogeneous characteristics is known as K-means cluster analysis. K-mean cluster is also a popular unsupervised machine learning algorithms popularly used in segregating data into groups. There are multiple methods of determine the number of k values.

6.5 Combining Clustering and Other Models

6.5.1 Clustering and fixed parameter approach

Traditional fixed parameter models such as binary logistic regression, linear regression, logit and probit models are implemented in each cluster subdivided by clustering models. Many studies have been conducted with this hybrid approach but other more advanced methods have outshined this method so this will not be focused here. One such research is (Samerei, Aghabayk, Shiwakoti, & Mohammadi, 2021). In their study uses two methods to analyze the cyclist injury severity in motor vehicle-bicycle crashes in Victoria, Australia. The effect of bicyclist characteristics, environmental characteristics, geometry and traffic characteristics, and crash characteristics on the severity of cyclist injuries was investigated. The study used latent class clustering to overcome unobserved heterogeneity and groups the data into two clusters. The first cluster (Cluster 1) are areas with no traffic control, clear weather and dry surface condition. In this cluster high speed limit was seen to be a major factor in crash severity along with other factors. Likewise, the second cluster (Cluster 2) are areas with traffic control and unfavorable weather condition, in which wet surface was shown to increase the risk of serious crash severity. Binary Logistic Regression analysis was performed to find the critical factors in injury severity. The model (BLR) was implemented for each cluster and the significance of variables was investigated.

6.5.2 Clustering and random parameter approach

Similar to previous hybrid here random parameter approach is applied to each cluster which results in highly insightful results as seen in (Cerwick, Gkritza, Shaheed, & Hans, 2014). In their study compared two models, mixed logit and latent class methods in analysis truck crash severity with 6 years of crash data. They measured the performance of the models on the basis of fit, inference and predicted crash severity outcome probability. In this study the latent class performed slightly better than the mixed model although the probability of all the levels of injury severity were better (closer to observation) predicted by the mixed logit model. They conclude that the difference is very marginal between these two models.

Similarly, (Wahi, Haworth, Debnath, & King, 2018) in their paper examines the influence of type of traffic control on injury severity in bicycle-motor vehicle crashes at intersection. The study in its preliminary finding found that the cyclist injury severity differed significantly and depended on

whether the intersection had a stop/give way sign 4, traffic sign or no traffic control. This preliminary finding led the researchers to segment the data set according to different traffic controls and analyze the data. To analyze the different segments, they used mixed logit model to identify contributing factors to each severity category. Different contributing factors were identified for each segment. For example, it was found that it was less likely to experience minor injury at uncontrolled intersections compared to controlled. Also, older riders more likely to be fatally injured in stop/give way sign than signalized traffic which is justified by the researchers as some older cyclists have delayed perceptions, slower reaction time, and physical frailty which increase the risk and severity of injury.

6.5.3 Clustering and partial proportion

(Song & Fan, 2020) in their paper "Combined latent class and partial proportional odds model approach to exploring the heterogeneities in truck-involved severities at cross and T-intersections." Clustering is done using a latent class cluster to address the unobserved heterogeneity inherent in the crash data. The study concluded that different clusters had different significant variables. It was also noticed that same variable was differently associated with various severity levels; same variable had different parameters at varied severity levels. The use of clustering was shown to address the significant heterogeneity between classes (clusters) and the use of partial proportional odds model addressed the within class heterogeneity. Four classes were identified and the datasets were divided. The class were based on the highway characteristics and environmental character. The classes also had overlapping features but some features were dominating in one class than other. The PPO model used in the study is compared with two logit model which are usually used in crash severity analysis; the PPO model is compared with Multinomial logit and generalized ordered logit models. All the models are used to identify significant variables where the PPO model identifies more significant variables in all classes.

Likewise, (Lin & Fan, 2021) in their paper "Exploring bicyclist injury severity in bicycle-vehicle crashes using latent class clustering analysis and partial proportional odds models" explores the use of hybrid model to overcome the unobserved heterogeneity in crash severity data. In this study they investigate the contributing factors to bicycle injury severity. Seven clusters based on different criteria were identified and the data was subdivided. Models were tested for sub-divided

group as well as the whole data and the analysis concluded that the sub-models have better goodness of fit compared to the single model developed with the whole dataset.

Similarly, (Liu & (David), 2020) explores the injury severity in head-on crashes using latent class clustering analyses coupled with mixed logit model. This study firstly segments the whole data into **xx** number of segments using latent class model to account for unobserved heterogeneity. Further the researchers use mixed logit model to account for unobserved heterogeneity within the different segments. The study confirms the existence of unobserved heterogeneity in the whole dataset as well as in homogeneous clusters and provides a better and insightful result with the use of hybrid model.

In summary, we can observe a lot of study has implemented a kind of clustering method to divide a set of data into number of groups and studied them separately. In doing so mostly Latent Class Clustering (LCC) is being implemented. The use of LCC has shown to be useful in separating a dataset into homogeneous groups and address the heterogeneity between groups. The use of clustering in (Song & Fan, 2020) (Samerei, Aghabayk, Shiwakoti, & Mohammadi, 2021) (Lin & Fan, 2021) (Liu & (David), 2020) has proven its usefulness in addressing unobserved heterogeneity which is inherent in most crash severity data with an exception of (Iranitalab & Khattak, 2017) which found not much difference with and without clustering in identifying important factors contributing in crash severity. The use of clustering has shown to address the significant heterogeneity between classes (clusters) in most of these studies and the use of partial proportional odds model and mixed logit (random parameter) seems to addressed the within class heterogeneity. The clustering has helped in understanding and addressing some forms of unobserved heterogeneity. Except for some studies, (Iranitalab & Khattak, 2017), the use of clustering has resulted in more insightful results. The use of clustering with fixed parametric method such as ordered and multinomial logit and probit methods has had better result than using just the fixed parametric method. The use of partial proportional order models resulted in more insightful findings. Many studies have also incorporated the use of multilevel or random parametric methods in different clusters to address the heterogeneity within clusters. This combined method of clustering and using random parametric model produced more rich and insightful inferences. In conclusion, most paper's use of a hybrid clustering and predicting model

has resulted in more insightful observation suggesting a superiority of the use of hybrid model in analyzing and finding significant factors affecting crash severity in different environment characteristic, highway characteristics and weather conditions.

Table 1 Hybrid Traditional Models used in Crash Severity Analysis

Study	Model A	Model B	No of data	Location	No of clusters
(Samerei, Aghabayk, Shiwakoti, & Mohammadi, 2021)	Latent class clustering	Binary logistic regression	14,759	Victoria, Australia	2 clusters
(Song & Fan, 2020)	Latent class clustering	Partial proportional odds	18,346	North Carolina (2005 2017)	4 clusters
(Lin & Fan, 2021)	Latent class cluster	Partial proportional odds	4,012	North Carolina (2007-2014)	7 clusters
(Wahi, Haworth, Debnath, & King, 2018)	segregation	Mixed logit model	5,772	Queensland, Australia (2002-2014)	3 groups
(Liu & (David), 2020)	Latent class cluster	Mixed logit model	9,153	North Carolina (2005-2013)	4 clusters

6.6 Hybrid Machine Learning models

The use of two or more machine learning models for different purposes as well as the use of hybrid machine learning and traditional statistical models is quite common in safety analysis at present. For instance, in (Najaf, Duddu, & Pulugurtha, 2018), researchers use M5' model trees method to divide a set of data into homogenous classes and models are calibrated for each class. Similarly, cluster based negative binomial regression (NBR) is applied on the same dataset to compare the results. It was seen that the use of machine learning model with traditional model was the highest

predictive model and was reliable to interpret different attribute's role on crash frequency compared to just using traditional models. Likewise, (Zhibin, Pan, Wei, & Chengcheng, 2012) compared ordered probit model and support vector model in comparing crash injury severity prediction and impact factor identification. The study found the prediction of SVM to be 4% more accurate than that done by ordered probit model. It also investigated the impacts of external factors on crash injury severity and found similar deductions from the two models with SVM resulting more reasonable outcomes in calculating several variable's impact.

Likewise, (Li, et al., 2018) in their study "Examining driver injury severity in intersection-related crashes using cluster analysis and hierarchical Bayesian models" a combined cluster analysis and hierarchical Bayesian hybrid approach model was by the researchers to examine driver injury severity patterns in intersection-related crashes based on two-year crash data in New Mexico. The k mean cluster analysis technique is used to subdivide the dataset into three clusters based on the factors such as environmental and roadway condition. This clustering was done to reveal the driver's behavior such as the risk compensating character in different environmental condition. The major differences among these clusters are variable values regarding road and environmental conditions. The three cluster along with the overall dataset were examined and hierarchical Bayesian model was implemented to find the contributing factors in multilevel driver injury outcomes. The injury outcome is divided into three classifications (levels): property injury (Level I), complaint of injury and visible injury (Level II), and incapacitating injury and fatality (Level III). The hierarchical Bayesian models (Hierarchical multinomial logistic model) performs better than the Ordinary MNL model when compared using Deviance information Criterion (DIC).
This study finds different features in different criteria that has the most impact in crash severity outcome. In time period criteria, Night is found to be a critical factor for driver injury severity. likewise, within weather, adverse weather conditions, i.e., snow and rain, lead to various influences on driver injury severity among different datasets. Looking at the light conditions Darkness is estimated to have significant say. Similarly, Area- urban, Road grade- different clusters have different impact with respect to different grade of road, Traffic controls- it has an impact in cluster 1and cluster, Vehicle action-straight driving, backing and right turn actions are found to be significantly associated with driver injury severity in all.

(Chen, et al., 2015) in their paper "A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes" uses logit model to investigate and identify significant contributing factors for rear-end crash driver injury severity. These injuries are classified into categories such as: no injury, injury and fatality. Bayesian network is then formulated to establish the relationship between these injury categories and the contributing factors investigated by the logit models. The data used in this study was two-year rear-end crash data records collected in New Mexico from 2010 to 2011. Various statistical model performance measures, such as F-Measure, ROC curve, AUC, and MPE, are used to quantify the BN model performance. The results shows that the trained BN model can infer interdependence y among variables and the proposed hybrid approach performs reasonably well.

Further, (Sohn & Lee, 2003) uses different approaches (Dempster-Shafer algorithm, the Bayesian procedure and logistic model) to improve the accuracy of individual classifiers (Neural network and Decision tree) models by combining them. Similarly, (Yassin & Pooja, 2020), in their study "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach" uses two different models for feature engineering and classification of the predicted crash severity. It combines unsupervised machine learning technique, K-means clustering and classification technique, Random Forest, which performs better than just using Random Forest for classification. The K-mean unsupervised learning was done to find the unobserved heterogeneity (presence of critical unseen features correlated with the observed feature in a model building) by processing/running the data set and using the clusters from k-mean as new features which represented the unobserved heterogeneity. Logistic regression, random forest models, among others, were employed to predict the crash severity. The classification model's performance using raw dataset and new features are 86.83% and 87.77% score with raw data respectively for logistic regression and random forest and 99.13% and 99.86% in the same order with the new dataset.

(Iranitalab & Khattak, 2017) their paper compares four traditional and machine learning models and methods used for crash severity analysis. They compared the predictive performance (classification ability) of Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM) and Random Forest (RF). The study concluded that the NNC

model predicted the best in terms of the overall data as well as the in more severe crashes. MNL had the weakest performance reasoning the supremacy of supervised learning over traditional statistical method.

To study if homogeneity within a cluster can improve the crash severity predictability, the paper also divided the dataset into clusters by using two different methods namely: K-mean Clustering (KC) as a non-hierarchical method and Latent Class Cluster (LCC) as a model-based method. Three datasets (original, clustered by KC and clustered by LCC) were estimated/trained by four prediction methods. The predictive accuracy of different predictive models had mixed results when clustering was implemented. In some methods clustering improved the predictability whereas in some no effect were seen. Clustering had no effect on SVM but it did improve the result of MNL slightly in different levels. Other methods (NNC and RF) also showed some improvements in predictability in some levels. The better method of clustering between KC and LCC was also not cleared by this study as the in some cases KC had better effect on prediction rates and in some LCC had better results.

In addition to using clustering and other machine learning models few research papers which has incorporated traditional statistical method of classification with new and advanced machine learning method of classification into a hybrid model quite like (Tang, Liang, Han, Li, & Huang, 2019). This study uses two-layer stacking framework which include several machine learning models as well as a logistic regression model to analyze crash severity. First layer of the two-layer stacking framework uses multiple ML models such as Random Forest, AdaBoost and Gradient Boosting Decision Tree as predictive ensemble learning and the second layer which consist of the logistic regression model completes the classification by accepting inputs form these machines learning models and predicting the final prediction. It is found that the hybrid model is better performing than any of the used machine learning or traditional statistical models. Similar advanced learning is used by (Chiu et al., 2001). where two-step clustering algorithm is used to divide crash data into similar categories. This algorithm uses a hierarchical clustering method called BIRCH.

In summary, many studies such as (Iranitalab & Khattak, 2017) , (Chen, et al., 2015) (Kumar & Toshniwal, 2017) has found supremacy of machine learning models over traditional models in

predicting crash severity. Although it is not always the case, for example (Zhibin, Pan, Wei, & Chengcheng, 2012) found not much difference in using machine learning models over traditional statistical method, here machine learning model performed slightly better in this study but the difference was minimum.

Many studies make use of k-clustering as clustering method to overcome the unobserved heterogeneity inherit in crash severity data with varying success. (Li, et al., 2018) and (Yassin & Pooja, 2020), has found the use of clustering result in far more insightful results. Likewise, (Sohn & Lee, 2003) also concludes that combining k-mean clustering with other machine learning models increase the accuracy of prediction and can provide more insightful results. Although there are some exceptions to this inference, such as in (Iranitalab & Khattak, 2017) the researchers do not find much difference in important factor affecting crash severity with and without the use of cluttering. The same study also does not find significant difference between the use of LCC and K-mean.

Likewise, some studies (Chiu et al., 2001), (Najaf, Duddu, & Pulugurtha, 2018) have also used other more advanced data clustering methods such as BRICH and Agglomerative Hierarchical Clustering method which is a hierarchical clustering algorithm used in data mining. This shows the use of big data coupled with more advanced clustering algorithms might be useful in addressing unobserved heterogeneity in crash severity data.

Table 2 Hybrid Machine Learning Models used in Crash Severity Analysis

Study	Model A	Model B	No of data	Location	Remarks
(Iranitalab & Khattak, 2017)	LCC, K-mean clustering	MNL, Nearest Neighbor Classification, SVM, Random Forest	68448	Nebraska, USA (2012-2015)	LCC and k-mean clustering had similar effects

(Li, et al., 2018)	K-mean cluster	Hierarchical Bayesian	49073	NMDOT, New Mexico, USA	Clustering of data according to road and environment conditions., superiority to ordinary MNL model
(Yassin & Pooja, 2020)	K-mean cluster	logistic regression, Random Forest, SVM, K-Nearest Neighbors	5000	Addis Ababa, Ethiopia (2011 to 2018)	proposes using k-mean and RF
(Yu & Abdel-Aty, 2013)	Random Forest	Fixed parameter Logit model, SVM, Random parameter Logit model	670	Colorado, USA (2007- 2011)	Crash, road. Realtime weather, real-time traffic data
(Najaf, Duddu, & Pulugurtha, 2018)	Agglomerative hierarchical clustering method	NBR, M5' model trees	12,995	Charlotte, North Carolina, USA (2015)	cluster-based NBR has higher predictability than conventional general NBR
(Chen, et al., 2015)	MNL	Bayes Classification	11,383	New Mexico, USA (from 2010 to 2011)	Identifying significant contributing factor (which is then used in bayes classifier)
(Tang, Liang, Han, Li, & Huang, 2019).	Random Forest, AdaBoost, GBDT	Logistic regression	5,538	Florida, USA (2004-2006)	Advanced two-layer stacking framework

7 SUMMARY AND CONCLUSION

In summary, the reviewed literature on crash injury severity showed that significant attention has been on crash severity modeling, some literatures are also dedicated to unobserved heterogeneity. Mostly traditional methods of analysis are being used for addressing unobserved heterogeneity with good results. 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. Some new literatures are also exploring the use of machine learning techniques to address the unobserved heterogeneity in crash severity data. One such popular technique of clustering seen in many papers is k-mean clustering. K-mean clustering coupled with some classification machine learning techniques has shown some promising results. Along with the use of unsupervised learning clustering techniques such as k-mean and machine learning classification techniques, feature sensitivity analysis techniques are being used to overcome the infamous black box nature of machine learning models. Over-all the use of unsupervised clustering technique with machine learning and traditional analysis tools seems to be creating excitement (as inferred by its frequent use in recent literatures) in safety community.

The conclusion of the review is presented in bullet format below:

- Unobserved heterogeneity is an inherit characteristic of traditional and can even be observed in more recent data types used in crash severity analysis.
- Partial proportional odds model is an alternative modeling technique. It performs better than ordered model and multinomial models in addressing unobserved heterogeneity. The PPO tries to overcome the two assumptions of the ordered and multinomial models, namely the ordered model's proportional odds and multinomial model's assumption of no order.
- Use of random parameter models is preferred over fixed parameter models in addressing unobserved heterogeneity in crash severity data.
- The use of hybrid model such as clustering and some kind to model in every cluster has shown successful results. The use of clustering has shown to address the significant heterogeneity between classes (clusters) in most of these studies and the use of partial

proportional odds model and mixed logit (random parameter) seems to addressed the within class heterogeneity.

- The use of clustering technique has shown significantly insightful results except in some analysis (Iranitalab & Khattak, 2017) which indicates that the type of data can also have a significant say in if a method works or not.
- A hybrid of clustering technique and the use of machine learning model in each cluster coupled with feature (factor) importance analysis can prove more successful than more traditional method of analysis in addressing unobserved heterogeneity in crash severity data.
- The use of more advanced data clustering methods used in data mining shows the applicability of big data in crash severity analysis and in addressing unobserved heterogeneity in crash data.

8 REFERENCES

Cerwick, D. M., Gkritza, K., Shaheed, M. S., & Hans, Z. (2014). A comparison of the mixed logit and latent class methods for crash severity analysis. *Analytic Methods in Accident Research*, 11-27. doi:https://doi.org/10.1016/j.amar.2014.09.002

Cerwick, D. M., Gkritza, K., Shaheed, M. S., & Hans, Z. (2014). A comparison of the mixed logit and latent class methods for crash severity analysis. *Analytic Methods in Accident Research*, 11-27.

Chen, C., Zhang, G., Tarefder, R., Ma , J., Wei, H., & Guan , H. (2015). A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. *Accident Analysis and Prevention*, 76-88. doi:http://dx.doi.org/10.1016/j.aap.2015.03.036

Gennarelli, T., & Wodzin, E. (2006). AIS 2005: a contemporary injury scale. *Injury*, 1083-91. doi:10.1016/j.injury.2006.07.009

Gkritza, K. K., R., C., Hallmark, S., & Hawkins, N. (2010). An empirical analysis of farm vehicle crash injury severities on Iowa's public road system. *Accident Analysis & Prevention*, 1392-1397.

Granville, G. (1998). *Adding Years to Life, Life to Years.* Dublin: National Council on Ageing and Older People.

Gu, T., & Yang, S. (2019). Duration Prediction for Truck Crashes Based on the XGBoost Algorithm. *CICTP: Nanjing, China*, 5031.

Heejin Jeong, Y. J. (2018). Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. *Accident Analysis and Prevention*, 250-261.

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis and Prevention*, 27-36. doi:http://dx.doi.org/10.1016/j.aap.2017.08.008

Ivan, J. N., & Konduri, K. C. (2019). *Safe Mobility: Challenges, Methodology and Solutions.* Emerald.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning with Application in R.* Springer .

Kumar, S., & Toshniwal, D. (2017). Severity analysis of powered two wheeler traffic accidents. *Eur Transp Res Rev. European Transport Research Review, 9*. doi:DOI 10.1007/s12544-017-0242-z

Kwon, O. H., Rhee, W., & Yoon, Y. (2014). Application of classification algorithms for analysis of road safety risk factor dependencies. *Accident Analysis and Prevention*, 1-15.

Lenguerrand, E., Martin, J., & Laumon, B. (2006). Modelling the hierarchical structure of road crash data—Application to severity analysis. *Accident Analysis & Prevention*, 43-53.

Li, Z., Chen, C., Ci, Y., Zhang, G., Wu, Q., Liu, C., & Qian, Z. (. (2018). Examining driver injury severity in intersection-related crashes using cluster analysis and hierarchical Bayesian models. *Accident Analysis and Prevention*, 139-151. doi:https://doi.org/10.1016/j.aap.2018.08.009

Lin, M.-R., Hwang, H.-F., & Kuo, N.-W. (2001). Crash Severity, Injury Patterns, and Helmet Use in Adolescent Motorcycle Riders. *The journal of Trauma and Acute Care Surgery*, 24-30.

Lin, Z., & Fan, W. (. (2021). Exploring bicyclist injury severity in bicycle-vehicle crashes using latent class clustering analysis and partial proportional odds models. *Journal of Safety Research*, 101-117. doi:https://doi.org/10.1016/j.jsr.2020.11.012

Liu, P., & (David), W. F. (2020). Exploring injury severity in head-on crashes using latent class clustering analysis and mixed logit model: A case study of North Carolina. *Accident Analysis and Prevention*. doi:https://doi.org/10.1016/j.aap.2019.105388

Lord, D., & Washington, S. (2018). *Safe Mobility: Challenges, Methodology and Solutions.* USA: Emerald Publishing Limited.

(2012). *MMUCC Guideline.*

Najaf, P., Duddu, V. R., & Pulugurtha, S. S. (2018). Predictability and interpretability of hybrid link-level crash frequency models for urban arterials compared to cluster-based and general negative binomial regression models. *INTERNATIONAL JOURNAL OF INJURY CONTROL AND SAFETY PROMOTION*, 3-13. doi:http://dx.doi.org/10.1080/17457300.2017.1285789

NHTSA. (2021, 6 3). *2020 Fatality Data Show Increased Traffic Fatalities During Pandemic*. Retrieved from NHTSA: https://www.nhtsa.gov/press-releases/2020-fatality-data-show-increased-traffic-fatalities-during-pandemic

Ortman, J., Velkoff, V., & Hogan, H. (2014). *An Aging Nation: The Older Population in the United States.* Washington DC: Census Bureau.

Samerei, S. A., Aghabayk, K., Shiwakoti, N., & Mohammadi, A. (2021). Using latent class clustering and binary logistic regression to model Australian cyclist injury severity in motor vehicle–bicycle crashes. *Journal of Safety Research*, 246-256. doi:https://doi.org/10.1016/j.jsr.2021.09.005

Sasidharan, L., & Menéndez, M. (2014). Partial proportional odds model—An alternate choice for analyzing pedestrian crash injury severities. *Accident Analysis & Prevention*, 330-340. doi:https://doi.org/10.1016/j.aap.2014.07.025

Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis and Prevention*, 1666-1676.

Sohn, S. Y., & Lee, S. H. (2003). Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. *Safety Science*, 1-14.

Song, L., & Fan, W. (. (2020). Combined latent class and partial proportional odds model approach to exploring the heterogeneities in truck-involved severities at cross and T-intersections. *Accident Analysis and Prevention*. doi:https://doi.org/10.1016/j.aap.2020.105638

Tang, J., Liang, J., Han, C., Li, Z., & Huang, H. (2019). Crash injury severity analysis using a two-layer Stacking framework. *Accident Analysis and Prevention*, 226-238.

Tarko, A., & Azam, M. S. (2011). Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data. *Accident Analysis & Prevention*, 1689-1695. doi:https://doi.org/10.1016/j.aap.2011.03.027

Wahi, R. R.-H., Haworth, N., Debnath, A., & King, M. (2018). Influence of type of traffic control on injury severity in bicycle-motor vehicle crashes at intersections. *Transportation Research Record*, 199-209. doi:10.1177/0361198118773576

Xie, K., Ozbay, K., Kurkcu, A., & Yang, H. (2017). Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots. *Risk Analysis*, 1459-1476. doi:https://doi.org/10.1111/risa.12785

Yassin, S. S., & Pooja. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*. doi:https://doi.org/10.1007/s42452-020-3125-1

Ye, F., & Lord, D. (2011). Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models: Multinomial Logit, Ordered Probit, and Mixed Logit. *Transportation Research Record: Journal of the transportation Research Board*, 51-58. doi:https://doi.org/10.3141/2241-06

Yu, R., & Abdel-Aty, M. (2013). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety Science*, 50-56.

Zhao, H., Yu, H., Li, D., Mao, T., & Zhu, H. (2019). VANETs, Vehicle Accident Risk Prediction Based on AdaBoost-SO in VANETs. *IEEE Access*.

Zhibin, L., Pan, L., Wei, W., & Chengcheng, X. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention*, 478-486. doi:https://doi.org/10.1016/j.aap.2011.08.016

Zong, F., Xu, H., & Zhang, H. (2019). Prediction for Traffic Accident Severity: Comparing the Bayesian. *Mathematical Problems in Engineering, 2013*(Article ID 475194). doi:http://dx.doi.org/10.1155/2013/475194