

Article

A Cluster-Based Approach for Analysis of Injury Severity in Interstate Crashes Involving Large Trucks

Syed As-Sadeq Tahfim  and Yan Chen

School of Maritime Economics and Management, Dalian Maritime University, No. 1 Linghai Road, Dalian 116026, China

* Correspondence: tahfim1963@dlmu.edu.cn

Abstract: The significance of large trucks for the expansion and well-being of the economy is a well-established fact. However, crashes involving large trucks significantly threaten the overall safety on the roads. Moreover, a significant proportion of fatal crashes involving large trucks occurs on interstate roadways in the United States. However, not many studies have focused on the heterogeneous effects of the contributory factors on injury outcomes of interstate crashes involving large trucks. The current study explores the application of a k-prototypes clustering-based mixed logit model to identify and analyze the heterogeneous effects of contributory factors on injury outcomes in different scenarios of interstate crashes involving large trucks. Data from six years of crashes involving large trucks that occurred on interstate roadways in the state of Pennsylvania, US, were used in this study. The list of contributory factors included the following: drivers' demographics and behaviors; crash characteristics; vehicle-related factors; location and roadway attributes; and environmental factors. The results indicated that some of the contributory factors were significant for all scenarios of interstate crashes involving large trucks. However, the magnitude of those factors' effects varied across scenarios. Moreover, some of the contributory factors were exclusive to certain scenarios of interstate crashes involving large trucks. Lastly, the identification of random parameters in the cluster-based models indicated that a cluster-based mixed logit model is a more effective approach for accurately estimating the effects of contributory factors on injury outcomes in large-truck interstate crashes. The empirical findings of this study can be used to develop more robust traffic laws and safety measures to reduce the frequency and severity of injury in different scenarios of interstate crashes involving large trucks.

Keywords: large trucks; interstate; injury severity; heterogeneity; k-prototypes; clustering; mixed logit model



Citation: Tahfim, S.A.-S.; Chen, Y. A Cluster-Based Approach for Analysis of Injury Severity in Interstate Crashes Involving Large Trucks. *Sustainability* **2022**, *14*, 14342. <https://doi.org/10.3390/su142114342>

Academic Editor: Elżbieta Macioszek

Received: 2 August 2022

Accepted: 22 October 2022

Published: 2 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Large trucks play a critical role for the freight-moving industry and the economy. Moreover, large trucks are still the most used mode of transportation for moving freight by land in the United States (US). According to the US Bureau of Transportation Statistics, the value of freight moved by trucks across the US for the year 2017 was USD 12,017 billion, which was significantly higher than the value of freight moved by other modes of transportation. The use of large trucks for freight movement will increase as consumer demand increases and the economy expands. In the US, the number of registered medium/heavy trucks increased by 16% between 2015 and 2019 [1]. However, the risk of a crash involving large-truck increases with the growing number of large trucks.

According to a 2007 report published by the FMCSA (Federal Motor Carrier Safety Administration), the average cost of fatal crashes involving large trucks is around \$3.6 million per fatal crash. This value could be higher if adjusted for inflation from 2007 until now. Moreover, several studies have indicated that the long vehicle length, heavyweight, slow

braking mechanism, and unique operating characteristics of large trucks increase the likelihood of severe and fatal injuries in crashes [2,3]. The economic and social impacts of crashes involving large trucks have captured the attention of many researchers.

In recent years, a lot of studies have analyzed the contributory factors of injury severity in crashes involving large trucks under different scenarios and conditions. A few studies have analyzed the contributory factors to injury outcomes of different types of crashes (e.g., rear-end, run-off-road, and rollover) involving large trucks [4–6]. A couple of studies have indicated that the set of significant factors for injury severity in single-vehicle large-truck crashes differs from those in multiple-vehicle large-truck crashes [7,8]. According to Uddin and Huynh [9], the contributory factors have heterogeneous effects on injury severity under different weather conditions. A few studies explored the contributory factors to crashes involving large trucks under different lighting conditions [10,11]. Anderson and Dong [12] and Behnood and Mannering [13] analyzed the heterogeneous effects of the contributory factors on injury outcomes in crashes involving large trucks for different days of the week, and time of day, respectively. Some studies have researched the injury severities of crashes involving commercial and hazmat trucks [14,15]. The literature also includes studies that focused on crashes involving large trucks in different locations. A few previous studies have indicated that the set of contributory factors that influence the severity of injury in urban crashes involving large trucks is different from those that influence the severity of injury in rural crashes involving large trucks [8,11,16]. Osman et al. [17] analyzed the injury severity of large truck crashes in work zones. Song and Fan [18] explored the heterogeneities in truck-involved severities at cross- and T-intersections.

In 2019, 25% of fatal crashes involving large trucks occurred on interstate roadways in the US [19]. However, only a few studies have analyzed crashes involving large trucks on interstate roadways. Islam and Hernandez [20] analyzed the fatality rates for crashes involving large trucks on interstate highways. Teoh et al. [21] explored the crash risk factors for interstate large trucks in North Carolina. Anderson and Hernandez [22] investigated the statistically significant contributory factors of driver injury severity for crashes on principal arterials, major collectors, interstates and other principal arterials. Though a few studies have analyzed the contributory factors to interstate crashes involving large trucks, not many studies have explored the application of a cluster-based approach to identify and analyze the contributory factors to injury outcomes in crashes involving large trucks. Understanding the varying effects of the contributory factors that influence the severity of injury in different scenarios of crashes involving large trucks on interstate roadways can aid in developing more robust traffic laws and safety measures.

Many previous studies have indicated that considering the heterogeneous effects of the contributory factors on the severity of injury in crashes can reveal valuable and hidden insights [23–26]. The mixed logit model has been a popular approach to account for the unobserved heterogeneity in the crash data since it allows the parameters to vary across observations [7,10,27]. Several clustering techniques, such as latent class analysis (LCA) [23,28], k-means [29,30], k-modes [24], and hierarchical clustering [25], have been used to account for the heterogeneous nature of the crash data. Since the final class solution by the LCA is user-defined, it may not be reproducible [31]. On the other hand, k-means is appropriate for data sets with only numerical variables, and k-modes for only categorical variables. However, the crash data usually include both numerical and categorical variables. Most of the clustering techniques create dummy variables to represent different levels of the categorical variable. A dummy variable uses numerical values such as 0 or 1 to code the absence or presence of a level of the categorical variable. According to Huang [32], the numerical representation of the categorical variables, which are not ordered, does not produce meaningful results. In this study, the k-prototypes clustering technique was used to segment the crash data into homogeneous groups. The k-prototypes clustering is capable of segmenting data sets with both numerical and categorical variables.

Though a cluster-based approach can account for heterogeneity within the aggregated data, some sort of heterogeneity is likely to remain within the subgroups. The current

study has proposed a k-prototypes clustering-based mixed logit model approach to identify and analyze the heterogeneous effects of the contributory factors on injury severities in interstate crashes involving large trucks. Such an effort has not been previously employed for the analysis of interstate crashes involving large trucks. The next section of the study describes the methodologies and data used in this study. The following section shows the results and describes them. In the last part, the findings of the study were discussed from the perspective of the study's objective and past studies.

2. Methodologies and Materials

2.1. The K-Prototypes Clustering

The current study has used the k-prototypes clustering technique to segment the collected data set into more homogeneous groups. The k-prototypes clustering technique is an extension of the k-means clustering technique [32]. A major advantage of the k-prototypes clustering technique is that it is capable of working with categorical variables. The traditional k-means clustering technique accepts data sets only in numerical forms. When the k-means clustering is applied to the data set involving categorical variables, it creates a dummy variable (0, 1) for each level of the categorical variable. Such an approach is less informative. In light of the limitations of the traditional k-means clustering technique, Huang [32] proposed the k-prototypes clustering technique to segment data sets with both numerical and categorical variables. The k-prototypes clustering uses a simple matching coefficient to measure the distances between the observations with categorical variables. Equation (1) represents the cost function that measures the distance between an observation and a cluster-prototype.

$$E = \sum_{l=1}^k \sum_{i=1}^n y_{il} d(X_i, Q_l) \quad (1)$$

Here, the objective of the k-prototypes clustering technique is to minimize the cost function (E) and segment the data set X into k number of clusters. In Equation (1), Q_l is the center of prototype l , y_{il} is the dummy variable that equals to 0 when data object i is assigned to prototype l , and $d(X_i, Q_l)$ is the distance measure for both numerical and categorical variables in brief. The equation below describes the distance measure in more detail.

$$d(X_i, Q_l) = \sum_{j=1}^p (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=p+1}^m \delta(x_{ij}^c, q_{lj}^c) \quad (2)$$

In Equation (2), the first part is the squared Euclidian distance measure on the numerical variables and the second term is the simple matching dissimilarity measure on the categorical variables. In the second term of the equation, γ_l represents the weight to avoid favoring either type of variable. The readers can follow Huang [32] to read more about the influence of γ_l in the clustering process. In Equation (2), superscript "r" represents the numerical variable, and superscript "c" represents the categorical variables. The complete cost function for prototype l is computed using the equation below.

$$E_l = \sum_{i=1}^n y_{il} \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{i=1}^n y_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) = E_l^r + E_l^c \quad (3)$$

The second term in Equation (3) is explained in Equation (4). In Equation (4), C_j is the set of all the discrete values of the categorical variable j , and $p(c_j \in C_j | l)$ is the probability of the discrete value q_j from C_j being in prototype l . Detailed explanations for the equations can be found in [32,33].

$$E_l^c = \gamma_l \sum_{j=1}^{m_c} n_l (1 - p(q_{lj}^c \in C_j | l)) \quad (4)$$

2.2. Mixed Logit Model (MXL Model)

Past studies have used different types of ordered or unordered logit/probit models such as generalized ordered response models [34], random parameter ordered logit model [6], heteroskedastic ordered probit model [35], mixed logit model [7,10,27], and multinomial logit model [16,36] to analyze the injury severities of road crashes. In order to

capture the heterogeneous effects of the contributory factors, the current study has used the mixed logit model. In the mixed model, the outcomes of the dependent variable are assumed to be unordered. There are significant trade-offs for choosing an ordered logit/probit model over an unordered logit/probit model. Mannering and Bhat [37] revealed that it is impossible for a contributory factor to simultaneously increase or decrease both of the extreme severities (no injury and fatality) in ordered probability models. According to the mixed logit modelling approach presented in [38,39], the relationship between the contributory factors and any of the injury severity levels ($i = 2, \dots, I$) for observation n can be expressed using the following function.

$$U_{in} = \beta_i X_{in} + \varepsilon_{in} \quad (5)$$

Here, U_{in} is the function that determines the probability of the injury severity i for a crash event n , and X_{in} indicates the vector of a contributory factor for the n crash event. β_i is a vector of estimable parameters, and ε_{in} indicates the error term for random noise [40]. It can be shown that if ε_{in} are assumed to be extreme distributed values, then the standard multinomial logit model takes the following formation:

$$P_n(i) = \frac{\exp[\beta_i X_{in}]}{\sum_{\forall I} \exp[\beta_I X_{In}]} \quad (6)$$

The probability of crash event n leading to injury severity i is denoted by $P_n(i)$. Additionally, I is the set of all possible injury outcomes. Now to capture the variation across the observations, the following mixed model (a model with mixing distribution) can be employed.

$$P_n^m(i) = \int_x P_n(i) f(\beta|\varphi) d\beta \quad (7)$$

where $P_n^m(i)$ indicates the outcome probabilities and $f(\beta|\varphi)$ is the density function of β with φ referring to a vector of parameters of the density function (mean and variance). Substituting Equation (6) into Equation (7), we can obtain the mixed logit model [40].

$$P_n^m(i) = \int_x \frac{\exp[\beta_i X_{in}]}{\sum_{\forall I} \exp[\beta_I X_{In}]} f(\beta|\varphi) d\beta \quad (8)$$

The above equation shows that the mixed logit probabilities $P_n^m(i)$ are the weighted average of the standard multinomial logit probabilities $P_n(i)$ with the weights determined by the density function $f(\beta|\varphi)$. The distribution of β is defined by the user. Generally, the distribution is specified to normal, but several distributions such as log-normal and uniform are tested for statistical significance in some occasions. The density function $f(\beta|\varphi)$ allows parameters to vary across observations and β accounts for the observation-specific variations in $P_n(i)$ based on the effect of X [40]. Equation (8) can calculate the coefficients of the contributory factors. However, the coefficients alone are not sufficient to calculate the effects of the contributory factors on the probabilities of the injury outcomes. The effects of the contributory factors on injury outcomes were further explored using average pseudo-elasticities. The following equation is used to compute the elasticities for the numerical variables from the partial derivative for each crash event n (n subscripting omitted).

$$E_{X_{ki}}^{P(i)} = \frac{\partial P(i)}{\partial X_{ki}} \frac{X_{ki}}{P(i)} \quad (9)$$

In Equation (9), $P(i)$ is the probability of injury outcome i , and X_{ki} is the value of k th variable for injury outcome i . After taking the partial derivative of model, Equation (8) becomes

$$E_{X_{ki}}^{P(i)} = [1 - P(i)] \beta_{ki} X_{ki} \quad (10)$$

In Equation (10), β_{ki} is the estimated coefficient associated with the value X_{ki} of k th variable. In the context of the current study, elasticities are interpreted as the percentage change in the probability of injury outcome i for a unit change in the X_{ki} . The above equation is not suitable for dummy or indicator variables, which take the value 0 or 1. For dummy variables, we compute the pseudo-elasticity using the following equation.

$$E_{X_{ki}}^{P(i)} = \frac{\text{EXP}[\Delta(\beta_i X_i)] \sum_{\forall I} \text{EXP}(\beta_{ki} X_{ki})}{\text{EXP}[\Delta(\beta_i X_i)] \sum_{\forall I} \text{EXP}(\beta_{ki} X_{ki}) + \sum_{\forall I \neq I_n} \text{EXP}(\beta_{ki} X_{ki})} - 1. \quad (11)$$

where I_n is the set of all alternate injury outcomes and X_k determines the injury outcome in the function. Readers can follow Washington et al. [40] to obtain more details about the aforementioned equations.

2.3. Data Description

The crash data was obtained from the Open Data Portal of the Pennsylvania Department of Transportation. According to the Pennsylvania Department of Transportation, a reportable crash is an incident that occurred on a public highway or traffic way and involved at least one motor vehicle in transport. For the current study, the crash data included only crashes involving large trucks that occurred on the interstate roadways of Pennsylvania from 2014 to 2019. According to the US Department of Transportation, any medium or heavy truck with a gross vehicle rating of over 10,000 pounds is considered a large truck. In the database, the crash records are distributed into several data tables. Each data table includes different categories of contributory factors to the crash. The data tables are linkable using the crash record number (CRN). For some crash events, multiple records were tabulated in the database since multiple vehicles and people were involved in many cases. The severity of a crash event is defined by the most severely injured person among all those involved in the crash event. The crash record indicated the most severely injured person was kept, and others were removed. Redundant factors, factors with a lot of null or unknown values, and duplicated crash records were removed from the data set. Some of the levels of the categorical variables had very low proportions in the data set, and they were put together into one level as “others”. The final data set included 10,547 interstate crashes involving large trucks and 22 independent contributory factors. The list of contributory factors included driver demographics and behavior; vehicle-related factors; crash characteristics; roadway and traffic attributes; and environmental factors. Table 1 displays the frequency and percentage of the contributory factors.

The Pennsylvania Department of Transportation follows the KABCO scale to measure the severity of injuries in crashes. Here, K indicates fatal injury, A indicates suspected serious injury, B indicates suspected minor injury, C indicates possible injury, and O indicates no apparent injury. To make the injury severity levels more discrete, the fatal injury and suspected serious injury were put into one category as “serious injuries” (A + K); the suspected minor injury and possible injury were put into another category as “Non-severe Injuries” (B + C), and the no apparent injury was labeled as “No-injury”. A similar strategy for categorizing the injury outcomes has been used in some previous studies as well [5,10,41].

Table 1. Statistics for the Contributory Factors.

Factors	Freq. (%)	Factors	Freq. (%)
Driver Demographics and Behavior		Land use	
Drivers aged between 16 to 20 years		Rural *	6470 (61.34)
No	9834 (93.24)	Urban	4077 (38.66)
Yes	713 (6.76)	Roadway attributes	
Drivers aged between 50 to 64 years		Speed limit (kmh)	$\bar{x} = 96.59, \sigma = 12.45$
No	5566 (52.77)	Traffic control	
Yes	4981 (47.23)	Absent *	9984 (94.66)
Drivers aged over 65 years		Present	563 (5.34)
No	9193 (87.16)	Curved road	
Yes	1354 (12.84)	No	8779 (83.24)
Speeding related		Yes	1768 (16.76)
No	9596 (90.98)	Intersection	
Yes	951 (9.02)	No	9845 (93.34)
Driving under influence of alcohol or drug		Yes	702 (6.66)
No	10144 (96.18)	Lane count	$\bar{x} = 2.23, \sigma = 0.55$
Yes	403 (3.82)	Environmental Factors	
Seat belt status		Weather	
Belted	9638 (91.38)	Clear *	7326 (69.46)
Unbelted	909 (8.62)	Rain	1591 (15.08)
Driver distracted or fatigue-asleep		Snow	1348 (12.78)
No	9596 (90.98)	Others (i.e., windy, cloudy, and fog etc.)	282 (2.67)
Yes	951 (9.02)	Lighting	
Tailgating		Daylight *	6604 (62.61)
No	9495 (90.03)	Dark-not lighted	2719 (25.78)
Yes	1052 (9.97)	Dark-lighted	766 (7.26)
Crash Characteristics		Others (i.e., dusk, dawn, dark-unknown street lighting)	458 (4.34)
Collision types		Wet road	
Rear-end *	3529	No	8541 (80.98)
Hitting fixed object with collision	2268	Yes	2006 (19.02)
Sideswipe	2207	Crash Hour	
Angle	1415	Off-peak (10 a.m. to 4 p.m.) *	3608 (34.21)
No-collision	980	Midnight to morning (0 to 6 a.m.)	2416 (22.91)
Head-on	106	Am-peak (6 to 10 a.m.)	2192 (20.78)
Others (i.e., backing, hitting pedestrian)	42	Pm-peak (4 to 8 p.m.)	1530 (14.51)
Overturned		Evening (8 to 11:59 p.m.)	801 (7.59)
No	9366 (88.8)	Dependent Variable	
Yes	1181 (11.2)	Injury Severity	
Vehicle related factors		No injury *	6062 (57.48)
Commercial truck involved		Non-severe	4011 (38.03)
Yes	6407 (60.75)	Serious	474 (4.49)
No	4140 (39.25)		
Vehicle count	$\bar{x} = 2, \sigma = 0.81$		

Note: (1) $\bar{x}$ = Mean; (2) σ = Standard Deviation; (3) Freq. = Frequency (4) The reference levels are indicated by *.

3. Results

3.1. Cluster Analysis

The k-prototypes clustering algorithm was employed using the “clustMixType” package for R. To determine the optimal number of clusters, a simple and easy to interpret elbow method was used in this study. In the elbow method, the Within-Cluster Sum of Squared Errors (WSSE) is visualized in a line chart. When the line takes the shape of an elbow, the

k-value associated with the elbow is used as a suggestion for the optimal number of clusters. In most cases, the elbow occurs when segmenting the data set does not significantly reduce the WSSE. Here, the k-prototypes clustering was employed for a range of values (e.g., 2 to k values). Figure 1 shows the WSSE for different k values. Figure 1 suggested that 3 or 4 is the optimal number of clusters. After evaluating the cluster prototypes and their distribution in the clusters, the EDS was segmented into three homogeneous clusters.

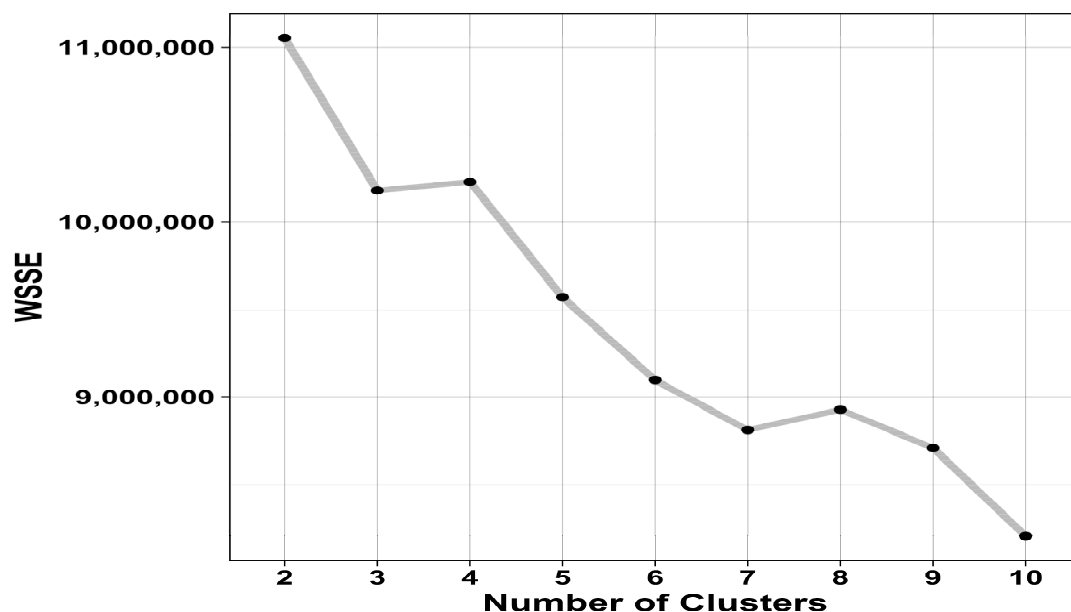


Figure 1. Optimal number of clusters.

Table 2 displays the number of observations and the distribution of the cluster prototypes in each of the clusters. The skewed distribution of cluster-prototypes can be used to characterize the clusters. Commercial truck involvement, land use, drivers aged between 50 and 64 years, lighting conditions, and speed limits were the cluster prototypes or contributing factors that distinguished the clusters from one another. The majority of the crashes in CL2 and CL3 involved commercial trucks and occurred in rural areas. CL2 and CL3 are also similar in terms of their speed limit. However, nearly 69% of the crashes in CL2 occurred under dark-not-lighted conditions, which distinguishes CL2 from CL3. On the other hand, almost 80% of the crashes in CL1 occurred in urban areas. Additionally, a good proportion of crashes in CL1 involved drivers aged between 50 and 64 years. Therefore, CL1 can be characterized based on drivers aged between 50 and 64 years old and urban areas. The distribution of the cluster prototypes indicates that the clusters are more homogeneous than the EDS.

Table 2. Cluster Description.

Variable	Cluster 1 (CL1)	Cluster 2 (CL2)	Cluster 3 (CL3)
Number of observations	3941	2998	3608
Commercial truck involved	38.06%	78.22%	71.01%
Driver aged between 50 and 64 years	63%	36.56%	38.86%
Rural	20.43%	85.26%	86.17
Urban	79.57%	14.74%	13.83%
Dark not lighted	11.22%	68.98%	5.79%
Speed limit (88–113 kmh)		90%	90%
Speed limit (72–105 kmh)	90%	-	-

3.2. Mixed Logit Model Results

3.2.1. Log-Likelihood Ratio Tests

In the current study, four mixed logit models were developed, one on the EDS and three more on the three clusters. The mixed logit models were developed using the “mlogit” package for R. Two log-likelihood ratio tests were conducted for the four models. The first log-likelihood ratio test was conducted to ensure that the models with random parameters are of more statistical significance than the models with fixed parameters. For this log-likelihood test, the null hypothesis states that there is no difference between the model with random parameters and the model with fixed parameters. Equation (11) was used for the first log-likelihood ratio test.

$$X^2 = -2[LL(\beta_{Fixed}) - LL(\beta_{Random})] \quad (12)$$

In Equation (12), $LL(\beta_{Fixed})$ is the log-likelihood of the model with fixed parameters at convergence and $LL(\beta_{Random})$ is the log-likelihood of the model with random parameters at convergence. The χ^2 (chi-squared) statistics are used to accept or reject the null hypothesis. Here, the χ^2 is normally distributed with degrees of freedom equal to the number of estimated random parameters. The random parameters were also estimated using the normal distribution. Table 3 shows the log-likelihoods, degrees of freedom (DF), chi-square statistics, and p -values for each of the four models with random and fixed parameters. The DF indicates the number of random parameters in each model. At the 5% significance level, all of the models with random parameters were found to be more statistically significant than the models with fixed parameters.

Table 3. Log-likelihood Ratio Test.

Models	$LL(\beta_{Random})$	$LL(\beta_{Fixed})$	X^2	DF	p -Value
EDS	−7876.865	−7914.41	56.67	50	0.0124
CL1	−2571.922	−2682.973	222.10	46	0.0000
CL2	−2036.724	−2094.512	115.58	46	0.0000
CL3	−2298.919	−2336.855	75.87	46	0.0036

The second log-likelihood ratio test was conducted to ensure that developing cluster-based models is a justifiable approach to estimate the heterogeneous effects of the contributory factors on injury outcomes in interstate crashes involving large trucks. The null hypothesis states that there is no difference between the parameters of the EDS and cluster-based models. For the second log-likelihood ratio test, Equation (13) was used.

$$X^2 = -2[LL_{full}(\beta^{full}) - \sum_{j=n}^J LL_j(\beta^j)] \quad (13)$$

In Equation (12), $LL_{full}(\beta^{full})$ is the log-likelihood of the EDS-based model at convergence and $\sum_{j=n}^J LL_j(\beta^j)$ is the aggregated log-likelihood of the cluster-based models at convergence. The degree of freedom for the χ^2 statistics is equal to the total number of contributory factors in the cluster-based model minus the number of contributory factors in the EDS-based model. The sum of log-likelihoods of the clusters was −6907.565. The chi-square statistics of 1938.6 with 228 degrees of freedom indicated that the cluster-based models are statistically significant at the 5% significance level. It means that the parameters of cluster-based models are significantly different from the parameters of EDS-based models.

3.2.2. Heterogeneous Effects of the Contributory Factors

The study’s objective was to identify and analyze the heterogeneous effects of the contributory factors on the injury outcomes of interstate crashes involving large trucks.

Therefore, we will keep our focus on the varying effects of the contributory factors on injury outcomes between the EDS and cluster-based models; the contributory factors with random parameters in all of the models; and the contributory factors that are exclusive to each of the cluster-based models. Tables 4–7 show the coefficients, z statistics, average pseudo-elasticities, and the mean and variance of the random parameters for the statistically significant contributory factors. The random parameters of the contributory factors followed the normal distribution. The average pseudo-elasticities were computed using the “marginaleffects” package for R. The contributory factors shown in the tables were statistically significant at a 5% confidence level. In all of the models, the no-injury outcome was set as the baseline or reference level in the dependent variable. The majority of the independent contributory factors were of the categorical type in this study. The coefficients of the categorical factors indicate the changes in the probability of each injury severity level relative to the reference level of the categorical factors. The ρ^2 of the models varied between 9% and around 18%. The number of observations, log-likelihood at zero, and log-likelihood at convergence are also displayed in the tables. The reference level of the binary categorical factors is “no” in all cases.

Table 4. Significant Contributory Factors in the EDS-based Mixed Logit Model.

Contributory Factors	Non-Severe		Serious		Average Pseudo-Elasticity		
	Coefficient	Z-Stats	Coefficient	Z-Stats	No-Injury	Non-Severe	Serious
Constant	−1.42		−7.90				
Angle collision	−0.36	−4.16			0.14	−0.19	−0.19
Head-on collision			1.61	3.33	−0.0906	−0.0036	1.2610
Hitting fixed object collision	−0.48	−5.74			0.1905	−0.2508	−0.3017
No collision	−0.87	−8.01	−4.47 (3.54)	−2.70 (3.22)	0.4673	−0.3594	−3.0116
			$\bar{x} = -4.47, \sigma = 3.54$				
Other collisions (i.e., backing, hitting pedestrian)			3.85	7.03	−0.1290	−0.1507	3.1076
Sideswipe collision	−0.62	−7.86	−2.21 (1.92)	−3.21 (3.34)	0.2895	−0.2835	−1.2251
			$\bar{x} = -2.21, \sigma = 1.92$				
Commercial truck involved			−0.40	−2.24	0.0090	0.0157	−0.2308
Drivers aged over 65 years	0.28	3.65	0.53	2.58	−0.1171	0.1418	0.3447
Driving under influence of alcohol or drug	0.60	4.50	1.86	6.67	−0.2792	0.2804	1.3144
Urban areas			−0.51	−2.23	−0.0133	0.0694	−0.4336
Overtaken	1.51	15.97	2.43	10.37	−0.6162	0.7742	1.5017
Speeding related	0.14	2.55			−0.0607	0.0710	0.1987
Tailgating	−0.24	−2.43	−0.81	−3.03	0.1112	−0.1044	−0.5793
Unbelted	1.32 (−1.75)	9.25 (−2.88)	2.80	10.56	−0.5535	0.6513	1.8643
	$\bar{x} = 1.32, \sigma = 1.75$						
Snowy weather	−0.41 (−0.90)	−3.44 (2.41)	−1.92	−2.92	0.2067	−0.1573	−1.2732
	$\bar{x} = 0.41, \sigma = 0.90$						
Lane count	0.16	3.31			−0.0577	0.0885	0.0013
Speed limit			0.03 (0.01)	2.95 (2.63)	−0.0007	−0.0015	0.0244
			$\bar{x} = 0.03, \sigma = 0.007$				
Vehicle count	0.36	8.80	0.83	8.36	−0.1586	0.1785	0.5778
Number of observations						10547	
Log-likelihood at Zero						−8705.52	
Log-likelihood at Convergence						−7876.87	
McFadden pseudo-R-squared						1-LL(β)/LL(0) = 0.10	

Note: (1) values in parenthesis indicate the standard deviation of the normally distributed random parameters; (2) $\bar{x}$ = mean; (3) σ = standard deviation.

Table 5. Significant Contributory Factors in the CL1-based Mixed Logit Model.

Contributory Factors	Non-Severe		Serious		Average Pseudo-Elasticity		
	Coefficient	Z-Stats	Coefficient	Z-Stats	No-Injury	Non-Severe	Serious
Constant	−174.54		−47.99				
Angle collision	−7.38	−3.90			0.5006	−0.1643	1.6060
Head-on collision			29.32	3.01	−0.0060	−0.1189	17.4495
No collision	−8.54	−2.17			0.5560	−0.1881	−0.8542
Collision type others			53.60	3.13	1.4257	−0.6790	35.7380
Sideswipe collision	−7.92 (12.00)	−4.69 (5.43)	−22.29 (−30.54)	−2.34 (−3.47)	0.7214	−0.1411	−5.4962
	$\bar{x} = -7.92, \sigma = 12$		$\bar{x} = -22.29, \sigma = 30.54$				
Commercial truck involved	30.17	12.70			−2.0233	0.6937	−7.6593
Drivers aged between 50 and 64 years	−18.32 (51.40)	−9.63 (12.79)			1.5705	−0.3293	4.4821
	$\bar{x} = -18.32, \sigma = 51.40$						
Driving under influence of alcohol or drugs			17.91	2.73	−0.2632	0.0136	10.0511
Evening hours	7.07 (−11.61)	1.97 (−2.38)	−47.36 (36.17)	−2.10 (2.69)	0.0049	0.2109	−26.8661
	$\bar{x} = 7.07, \sigma = 11.61$		$\bar{x} = -47.36, \sigma = 36.17$				
Urban areas	−36.09	−10.35			2.4761	−0.8023	2.0040
Dark-not-lighted condition	−6.22	−2.39			0.3364	−0.1532	8.8923
Overtaken	22.32	5.19	29.34	3.79	−1.6415	0.4638	13.0485
Speeding related	3.72 (8.22)	2.58 (3.94)	11.31	2.28	−0.1670	0.0651	6.3406
	$\bar{x} = 3.72, \sigma = 8.23$						
Tailgating	−4.25	−2.21			0.3682	−0.0714	−5.5764
Unbelted	15.61 (−12.58)	5.07 (−2.77)	35.64	4.17	−1.2223	0.2835	18.2522
	$\bar{x} = 15.61, \sigma = 12.58$						
Speed limit	2.22	14.70			−0.1436	0.0502	−0.3909
Vehicle count	4.77 (−5.30)	4.22 (−9.72)			−0.0378	0.1251	3.7767
	$\bar{x} = 4.77, \sigma = 5.30$						
Number of observations							3941
Log-likelihood at Zero							−3138.35
Log-likelihood at Convergence							−2571.92
McFadden pseudo-R-squared							1-LL(β)/LL(0) = 0.18

Note: (1) values in parenthesis indicate the standard deviation of the normally distributed random parameters; (2) $\bar{x}$ = mean; (3) σ = standard deviation.

Table 6. Significant Contributory Factors in the CL2-based Mixed Logit Model.

Contributory Factors	Non-severe		Serious		Average Pseudo-Elasticity		
	Coefficient	Z-Stats	Coefficient	Z-Stats	No-Injury	Non-Severe	Serious
Constant	−78.79		−293.20				
Angle collision	−36.05 (−97.44)	−3.37 (−8.37)			0.1750	−0.4560	0.1845
	$\bar{x} = -36.05, \sigma = 97.44$						
Head-on collision			83.74	2.61	−0.4815	−0.0484	29.1038
Hitting fixed object collision	−25.33 (−12.41)	−3.43 (−3.50)	−177.47 (170.84)	−3.93 (5.08)	0.9786	−0.3802	−15.5270
	$\bar{x} = -25.33, \sigma = 12.41$		$\bar{x} = -177.47, \sigma = 170.84$				
No collision	−39.37 (−30.28)	−4.57 (−4.09)	−85.47	−2.69	0.6310	−0.7903	−25.9228
	$\bar{x} = -39.37, \sigma = 30.28$						
Sideswipe collision	−38.72	−5.05	−85.13 (82.70)	−2.16 (3.02)	0.6089	−0.6941	−24.3463
	$\bar{x} = -85.13, \sigma = 82.70$						
Curved road	−15.58 (−80.16)	−2.68 (−8.12)			0.3162	−0.0866	−4.2387
	$\bar{x} = -15.58, \sigma = 80.16$						
Driving under influence of alcohol or drugs	29.22	3.86	58.82	2.38	−0.4389	0.5607	18.8383
Urban areas	−52.98	−6.65			0.4925	−0.9736	−10.0475
Dark-lighted	−42.41 (70.71)	−2.78 (4.68)			0.4869	−0.7081	−27.8177
	$\bar{x} = -42.41, \sigma = 70.71$						

Table 6. Cont.

Contributory Factors	Non-severe		Serious		Average Pseudo-Elasticity		
	Coefficient	Z-Stats	Coefficient	Z-Stats	No-Injury	Non-Severe	Serious
Dark-not-lighted	−19.98 (−18.19)	−2.05 (−5.57)			0.5135	−0.3560	−3.3331
Overtaken	74.23	13.18	106.95	4.61	−1.0829	1.3205	33.4838
Traffic control present	−49.00	−3.12			0.1734	−1.0218	12.7655
Unbelted	65.69 (92.00)	6.43 (6.60)	102.57	4.40	−0.8172	1.1044	33.2106
	$\bar{x} = 65.69, \sigma = 92$						
Weather others	−15.11	−2.20			0.3318	−0.2936	−11.9732
Rainy weather	−26.76 (54.15)	−3.16 (6.57)			0.1722	−0.2955	4.3999
	$\bar{x} = -26.76, \sigma = 54.15$						
Snowy weather	−30.05 (−44.46)	−5.06 (−6.61)	−72.79	−3.07	0.4784	−0.3294	−23.1643
	$\bar{x} = -30.05, \sigma = 44.27$						
Vehicle count	18.44 (−29.64)	4.47 (−11.84)	56.17 (15.00)	4.95 (3.62)	−0.3588	0.5961	19.1112
	$\bar{x} = 18.44, \sigma = 29.64$		$\bar{x} = 56.17, \sigma = 15$				
Number of observations							2998
Log-likelihood at Zero							−2421.06
Log-likelihood at Convergence							−2036.72
McFadden pseudo-R-squared							1-LL(β)/LL(0) = 0.16

Note: (1) values in parenthesis indicate the standard deviation of the normally distributed random parameters; (2) $\bar{x}$ = mean; (3) σ = standard deviation.

Table 7. Significant Contributory Factors in the CL3-based Mixed Logit Model.

Contributory Factors	Non-Severe		Serious		Average Pseudo-Elasticity		
	Coefficient	Z-Stats	Coefficient	Z-Stats	No-Injury	Non-Severe	Serious
Constant	−20.16		−13.30				
Head-on collision			2.48	2.03	0.0168	−0.3715	1.7205
No collision	−1.91 (2.33)	−2.96 (2.20)			0.3829	−0.6243	−2.4317
	$\bar{x} = -1.91, \sigma = 2.33$						
Collision type others			3.60	2.56	−0.2290	0.2609	2.2831
Sideswipe collision	−1.33	−3.56	−2.59 (2.02)	−2.40 (2.33)	0.2484	−0.4519	−1.2366
			$\bar{x} = -2.59, \sigma = 2.02$				
Involvement of commercial trucks	0.95 (2.59)	3.82 (3.91)			−0.1397	0.7911	−0.5781
	$\bar{x} = -0.95, \sigma = 2.59$						
Driver aged between 50 and 64 years	−1.40	−4.32			0.1773	−0.5504	0.2093
Drivers aged over 65 years	0.60	1.99			−0.0864	0.2755	0.0671
Drivers distracted or fatigued asleep	0.54	2.07			−0.0579	0.2190	−0.2576
Driving under influence of alcohol or drugs	1.58	2.90	2.47	2.99	−0.2862	0.5795	1.4593
Urban areas	−3.39	−4.45			0.4313	−1.3385	0.3424
Overtaken	2.53	4.87	3.09	4.37	−0.4286	0.9432	1.7693
Tailgating	−0.99 (−2.17)	−2.06 (−2.06)	−4.07 (3.53)	−2.20 (2.77)	0.2478	−0.2671	−2.2379
	$\bar{x} = -0.99, \sigma = 2.17$		$\bar{x} = -4.07, \sigma = 3.53$				
Unbelted	2.00	4.64	3.23	4.30	−0.3672	0.7271	1.9354
Snowy weather	−1.46 (−3.69)	−2.15 (−2.71)			0.2575	−0.4537	−1.4486
	$\bar{x} = -1.46, \sigma = 3.69$						
Speed limit	0.16	5.17	0.05	2.24	−0.0228	0.0644	0.0161
Vehicle count	0.55	2.82	1.33	3.64	−0.1175	0.2010	0.8461
Number of observations							3608
Log-likelihood at Zero							−2768.39
Log-likelihood at Convergence							−2298.92
McFadden pseudo-R-squared							1-LL(β)/LL(0) = 0.17

Note: (1) values in parenthesis indicate the standard deviation of the normally distributed random parameters; (2) $\bar{x}$ = mean; (3) σ = standard deviation.

Tables 4–7 indicated that some of the statistically significant contributory factors were common between the EDS and the cluster-based models. The following factors were significant in both the EDS-based and cluster-based models: head-on collision, sideswipe collision, no collision, driving under the influence of alcohol or drugs (DUI), urban areas, overturning, not wearing a seatbelt, and vehicle count. However, the degree of their effects on injury outcomes differed across the models.

Head-on collisions are likely to increase the chances of serious injury in almost all scenarios of interstate crashes involving large trucks. The parameters of head-on collisions were fixed in all of the models. On the other hand, the positive impacts of head-on collisions on serious injuries were substantially higher in the CL1- and CL2-based models. Several studies have also indicated that head-on collisions are more likely to increase the chance of severe injuries in crashes involving large trucks [2,42,43].

Sideswipe collisions are likely to reduce the chances of both non-severe and serious injuries under conditions such as those in the EDS and the clusters. However, the negative effects of sideswipe collisions on both non-severe and serious injuries were comparatively low in the CL2-based model. Behnood and Mannering [13] have also reported that sideswipe collisions involving large trucks are likely to reduce the chances of severe injuries. The results also indicated that the parameters of sideswipe collisions for serious injury function varied significantly in all the models. In the majority of cases, sideswipe collisions are less likely to lead to serious injuries. However, the mean and standard deviation of the parameters of sideswipe collisions for serious injury function indicated that 12.51%, 23.28%, 15.17%, and 10.05% of the sideswipe collisions involving large trucks on interstate roadways are more likely to experience serious injuries under conditions similar to EDS, CL1, CL2, and CL3, respectively. The parameters of sideswipe collisions for non-severe injury function in the CL1-based model also varied across the observations. A total of 25.47% of the sideswipe collisions in CL2 are more likely to sustain non-severe injuries. One reasonable explanation for sideswipe collisions having random parameters is that they can occur between vehicles that are traveling in the opposite direction to each other. Sideswipes from the opposite direction can increase the magnitude of the impact, increasing the likelihood of more severe injuries.

The effects of no-collisions on injury outcomes were not the same in all the models. The negative impacts of no-collisions on non-severe injuries were comparatively higher in the CL2-based model. The parameters of no collisions had significant variation for serious injury function in the EDS-based model and non-severe injury function in the CL2- and CL3-based models. The results indicated that the majority of the no-collisions in interstate crashes involving large trucks are less likely to increase the severity of injury. However, 10.35% of the no-collisions are likely to increase the chance of serious injuries under conditions similar to EDS. Moreover, 9.68% and 20.7% of no-collisions are likely to experience non-severe injuries under conditions similar to CL2 and CL3, respectively. The majority of the crashes in CL2 and CL3 occurred on roadways with considerably higher speed limits; driving at a higher speed is likely to increase the severity of injury.

DUI is more likely to increase the chances of both non-severe and serious injuries in crash scenarios such as those in EDS-, CL2- and CL3-based models. In the CL1-based model, DUI was a significant factor for only serious injury function. The effects of DUI on serious injury were substantially higher in the CL1- and CL2-based models. The findings about the effects of DUI in this study are consistent with previous studies [5,13,44]. Overturning and unbelted drivers were significant indicators of both non-severe and serious injuries in all the models. Overturning had similar effects on injury outcomes across the observations in all the models. Several studies have also indicated that overturning is more likely to be associated with more severe and fatal injuries [8,12,36].

On the other hand, unbelted drivers had mixed effects on non-severe injuries across the observations in the EDS-, CL1- and CL2-based models. In most cases, unbelted drivers are likely to increase the chance of non-severe injuries. However, 22.54%, 10.73%, and 23.76% of interstate crashes involving large trucks and unbelted drivers are less likely

to sustain non-severe injuries under conditions similar to those in EDS, CL1, and CL2, respectively. On such occasions, the unbelted driver probably belonged to the large truck, whose driving seat is well protected. A couple of previous studies have indicated that unbelted drivers are more likely to increase the severity of injuries [2,7].

Urban interstate crashes involving large trucks are more likely to sustain less severe injuries. Under conditions similar to those of CL3, the chances of non-severe injuries are comparatively low for urban interstate crashes involving large trucks. It is possible that the chances of severe injuries are hugely reduced for urban interstate crashes because of congestion and better traffic management.

The results have indicated that the number of vehicles in crashes is a significant indicator of both non-severe and serious injuries in the EDS-, CL2-, and CL3-based models. In the CL1-based model, the vehicle count was a significant factor for only non-severe injuries. In most cases, an increase in vehicle numbers is likely to increase the chance of both non-severe and serious injuries in interstate crashes involving large trucks. Zheng et al. [14] also indicated that an increase in the number of vehicles increases the chance of severe injuries in truck-involved crashes. In this study, the results indicated that the vehicle count can have mixed effects on non-severe injuries in interstate crashes involving large trucks. The mean and standard deviation indicated that an increase in the number of vehicles may not increase the chance of non-severe injuries in some cases, such as in CL1 and CL2. The proportion of observations that had different parameters for serious injury function in the CL2-based model was insignificant.

In addition to the aforementioned factors, some statistically significant factors were common between EDS and two cluster-based models. In crash scenarios such as in EDS, CL1, and CL2, angle collisions are more likely to reduce the chance of non-severe injuries in most cases. However, the parameters of angle collision for non-severe injury function varied across the observations in CL2. Under conditions similar to those CL2, 35.57% of the angle collisions involving large trucks on interstate roadways are more likely to experience non-severe injuries. The majority of crashes in CL2 occurred under dark-not-lighted conditions, which can reduce a driver's visibility and increase the severity of injury. Other types of collisions (i.e., hitting pedestrians and backing up), the involvement of commercial trucks, speed limits, and tailgating were significant predictors of injury outcomes in EDS-, CL1- and CL3-based models. The parameters of other types of collisions were stable, and other types of collisions are more likely to increase the chances of serious injuries in crash scenarios such as in EDS, CL1, and CL3. Interstate crashes involving commercial large trucks are likely to reduce the chances of serious injuries under conditions similar to EDS. On the other hand, interstate crashes involving commercial large trucks are more likely to sustain non-severe injuries under conditions similar to those in CL1 and CL3 in the majority of occasions. However, the parameters of commercial trucks were not the same across the observations in CL3. The mean and standard deviation indicated that nearly 36% of the interstate crashes involving large trucks are less likely to sustain non-severe injuries. In the EDS- and CL1-based models, the speed limit had a significant influence on serious and non-severe injuries and is likely to increase the chance of serious and non-severe injuries, respectively. In the CL3-based model, it increases the chance of both non-severe and serious injuries. The parameters of the speed limit in the EDS-based model varied in the EDS-based model but were insignificant. In both the EDS and CL3-based models, tailgating had negative impacts on both non-severe and serious injuries. In the CL1-based model, it had similar impacts on only non-severe injuries. However, the parameters of tailgating for both non-severe and serious injury functions were not the same across all the observations in CL3. In total, 67.58% and 87.55% of the interstate crashes involving large trucks and tailgating are less likely to lead to non-severe and serious injuries, respectively. The rest are more likely to experience non-severe and serious injuries.

Snowy weather was a significant predictor of injury outcomes in the EDS-, CL2- and CL3-based models. The effects of snowy weather on non-severe injuries were not the same across the observations in EDS, CL2 and CL3. Snowy weather is likely to reduce the chance

of non-severe injuries in most cases since drivers are usually cautious in snowy weather. However, 32.30%, 24.96%, and 34.62% of interstate crashes involving large trucks in snowy weather have a probability of experiencing non-severe injuries in crash scenarios similar to those in EDS, CL2, and CL3. Most crashes in CL2 and CL3 occurred in high speed limit zones. It is possible that some drivers were driving at a higher speed, which leads to more severe injuries.

The contributory factors, which were statistically significant in more than one model, also included drivers aged between 50 and 64 years (CL1 and CL3), drivers aged over 65 years (EDS and CL3), speeding-related (EDS and CL1), and dark-not-lighted conditions (CL1 and CL2). In the CL3-based model, drivers aged between 50 and 64 years had negative impacts on non-severe injuries, and the parameters for non-severe injury function were fixed. In the CL1-based model, the parameters for non-severe injury function varied across the observations. The results indicated that 63.92% of interstate crashes involving large trucks and drivers aged between 50 and 64 years are less likely to sustain non-severe injuries, and the rest are more likely to sustain non-severe injuries under conditions similar to those in CL1. Drivers aged between 50 and 64 years not only have years of driving experience but also have some physical limitations. This may explain the mixed effects of drivers aged between 50 and 64 years on non-severe injuries. Drivers aged over 65 years are more likely to increase the severity of injuries under conditions similar to those in EDS and CL3. Moreover, their effects on injury outcomes were stable across the observations. Speeding-related interstate crashes involving large trucks are likely to increase the chance of non-severe injuries in crash scenarios similar to EDS. Such crashes are also more likely to sustain both non-severe and serious injuries under conditions similar to those in CL1. However, the parameters of the speeding-related factor for non-severe injury function were not the same across the observations in the CL1-based model. The mean and standard deviation indicated that 32.56% of speeding-related interstate crashes involving large trucks are less likely to sustain non-severe injuries. In CL1, the majority of the crashes occurred in urban areas, where traffic rules are more strictly maintained and congestion is frequent. These issues are likely to reduce the severity of the injury. In the CL1 and CL2-based models, dark-not-lighted conditions were a significant predictor for non-severe injuries. The parameters of dark-not-lighted conditions were fixed in the CL1-based model but were random in the CL2-based model. Under conditions similar to CL2, 86.40% of the interstate crashes involving large trucks under dark-not-lighted conditions are less likely to experience non-severe injuries, but the rest are more likely to experience non-severe injuries. Usually, drivers are likely to drive more cautiously under dark-not-lighted conditions. However, significant factors such as an absence of traffic control and adverse weather conditions could increase the severity of injury.

The cluster-based approach also revealed that some contributory factors were significant only in the cluster-based models. The evening hours were significant for both non-severe and serious injury functions only in the CL1-based model. However, the parameters of evening hours for both non-severe and serious injury functions varied across the observations in CL1. In total, 72.87% of the interstate crashes involving large trucks in the evening hours are more likely to experience non-severe injury, and 90.48% are less likely to sustain serious injuries under conditions such as those in CL1. Only in the CL2-based model, curved roads, dark-lighted conditions, were the presence of traffic control, other types of weather (i.e., windy, cloudy, fog, etc.) and rainy weather significant indicators of non-severe injury. The parameters of the presence of traffic control and other types of weather were stable and had negative impacts on non-severe injuries only in the CL2-based model. On the other hand, the parameters of curved roads, dark-lighted conditions, and rainy weather for non-severe injury function were not the same across the observations. The mean and standard deviation indicated that 57.71%, 72.56%, and 68.94% of interstate crashes involving large trucks on curved roads, under dark-lighted and rainy weather conditions are less likely to experience non-severe injuries under conditions similar to CL2, respectively. Lastly, drivers being distracted or fatigued/asleep were significant predictors

of non-severe injuries only in the CL3-based model. The parameters of drivers being distracted or fatigued/asleep were stable and had positive impacts on non-severe injuries.

4. Discussion and Conclusions

The exploration of the factors that influence the injury outcomes in interstate crashes involving large trucks is an essential step toward forming better traffic laws, roadway improvements, and safety measures. Moreover, the knowledge revealed from the analysis of interstate crashes involving large trucks can help the trucking companies as well, since the majority of the freight is carried by trucks and interstate roadways experience heavy truck traffic. The current study used k-prototypes clustering-based mixed logit models to analyze the heterogeneous effects of the contributory factors on injury outcomes of interstate crashes involving large trucks. The k-prototypes clustering technique was employed on large-truck-involved crashes that occurred between 2014 and 2019 on the interstate roadways of the state of Pennsylvania, US. The entire data set (EDS) was segmented into three clusters, where the distribution of cluster prototypes indicated that the identified clusters were more homogenous than the EDS. For example, the majority of the crashes in CL2 and CL3 occurred on rural interstate roadways and involved commercial trucks. Compared to the EDS, the proportion of crashes that occurred in rural areas in CL2 and CL3 was significantly higher. Moreover, the majority of the CL1 crashes occurred in urban areas. In terms of land use, the clusters were more homogeneous than the EDS. The clusters were more homogeneous in terms of speed limit factor as well. The dark-not-lighted condition was another cluster prototype for CL2.

Furthermore, the study developed cluster-based mixed logit models to estimate the effects of the contributory factors on injury outcomes. The k-prototypes cluster-based mixed logit models revealed some interesting and valuable insights. Foremost, the sets of significant contributory factors in the cluster-based models were different from those in the EDS-based model. This means that, under different crash scenarios, different combinations of contributory factors become significant. This indicates the existence of heterogeneity in interstate crashes involving large trucks. Traffic management authorities and road safety practitioners can rank the factors according to their average marginal effects to obtain a more comprehensive idea about the combination of factors that are important for different scenarios of interstate crashes involving large trucks.

A lot of the statistically significant contributory factors were common between EDS- and cluster-based models. However, the effects of some of the contributory factors on injury outcomes in the cluster-based models were different from the effects in the EDS-based models. For example, head-on collisions, DUI, unbelted drivers, overturning, and vehicle count are more likely to increase the chance of serious injuries under conditions such as those in CL2, when compared to the conditions of EDS and other clusters. The significantly higher positive impacts on serious injuries of the aforementioned factors can be explained by the characteristics of CL2. A good proportion of crashes in CL2 occurred under dark-not-lighted conditions, which reduces visibility and increases the chance of collisions. The magnitude of the effects of angle collisions, involvement of commercial trucks, tailgating, and snowy weather on injury outcomes was also different in the cluster-based models. The variation in the magnitude of the effects of the contributory factors should be considered while tailoring traffic rules and regulations for different scenarios of interstate crashes involving large trucks.

The cluster-based mixed logit models also identified a few contributory factors, which were statistically significant only in the cluster-based models. Evening hours, curved roads, dark-lighted, dark-not-lighted, presence of traffic control, other types of weather (i.e., windy, cloudy, fog, etc.), rainy weather, and drivers being distracted or fatigued/asleep were significant only in the cluster-based models. Due to heterogeneity, the aforementioned contributory factors remained latent in the EDS-based model. Interpreting the effects of the contributory factors on injury outcomes by considering the characteristics of the clusters, which represent different scenarios of interstate crashes, can help traffic management

authorities and road safety practitioners in customizing traffic rules and safety measures. For example, drivers being distracted or fatigued/asleep were found to increase the chances of non-severe injuries under conditions similar to CL3. In CL3, the majority of the crashes occurred on rural interstate roadways with a speed limit of between 88 and 113 km/h. It is possible to reduce the chances of non-severe injuries by warning drivers to remain alert while driving in those areas. Interpreting the effects of the contributory factors within the context of a specific crash scenario can provide more valuable insights. The mutually exclusive factors identified in the cluster-based models demonstrated the importance of cluster analysis on aggregated data of interstate crashes involving large trucks.

The results have indicated that the cluster-based approach is an efficient measure to account for heterogeneity within the aggregated data of interstate crashes involving large trucks. However, a lot of the contributory factors had random parameters not only in the EDS-based model but also in the cluster-based models. The presence of random parameters in the cluster-based models indicated that some sort of heterogeneity still remained within the clusters as well. The numbers of contributory factors with random parameters identified in the EDS-, CL1-, CL2-, and CL3-based models were five, six, eleven, and five, respectively. Based on the results, it is reasonable to state that the k-prototypes clustering-based mixed logit model is an effective approach to analyze the heterogeneous effects of the contributory factors on the injury outcomes of interstate crashes involving large trucks. Traffic management authorities and road safety practitioners should consider the mixed effects of the contributory factors on injury outcomes while tailoring safety measures for the prevention of interstate crashes involving large trucks.

Like many other studies, the current study also has some limitations. The current study did not consider the temporal instability of the data. Future studies should be cautious in applying the proposed method for forecasting. Additionally, the current study researched interstate crashes involving large trucks only in the state of Pennsylvania in the US. Future research efforts can study the severity of injury outcomes of interstate crashes involving large trucks for other states or for the whole US. Moreover, considering traffic flow in the analysis of injury severities in interstate crashes involving large trucks can reveal interesting insights. Lastly, the proportion of serious injury crashes was significantly lower when compared to no-injury crashes in the data set used for the current study. This may bias the model towards no-injury crashes while predicting the injury outcomes.

Author Contributions: Conceptualization, S.A.-S.T. and Y.C.; methodology, S.A.-S.T.; formal analysis, S.A.-S.T.; validation, S.A.-S.T. and Y.C.; writing—original draft preparation, S.A.-S.T.; writing—review and editing, S.A.-S.T. and Y.C.; supervision, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data set used in this study is publicly available. The readers can access the data set through the open data portal of Pennsylvania Department of Transportation.

Acknowledgments: The authors of the study are grateful to the Pennsylvania Department of Transportation for making the crash data publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *Highway Statistics*; Highway Statistics Series; U.S. Department of Transportation, Federal Highway Administration (FHWA): Washington, DC, USA, 2019.
2. Zhu, X.; Srinivasan, S. A Comprehensive Analysis of Factors Influencing the Injury Severity of Large-Truck Crashes. *Accid. Anal. Prev.* **2011**, *43*, 49–57. [[CrossRef](#)] [[PubMed](#)]
3. Ahmed, M.M.; Franke, R.; Ksaibati, K.; Shinstine, D.S. Effects of Truck Traffic on Crash Injury Severity on Rural Highways in Wyoming Using Bayesian Binary Logit Models. *Accid. Anal. Prev.* **2018**, *117*, 106–113. [[CrossRef](#)] [[PubMed](#)]

4. Yuan, Q.; Lu, M.; Theofilatos, A.; Li, Y.-B. Investigation on Occupant Injury Severity in Rear-End Crashes Involving Trucks as the Front Vehicle in Beijing Area, China. *Chin. J. Traumatol.* **2017**, *20*, 20–26. [\[CrossRef\]](#)
5. Al-Bdairi, N.S.S.; Hernandez, S. An Empirical Analysis of Run-off-Road Injury Severity Crashes Involving Large Trucks. *Accid. Anal. Prev.* **2017**, *102*, 93–100. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Azimi, G.; Rahimi, A.; Asgari, H.; Jin, X. Severity Analysis for Large Truck Rollover Crashes Using a Random Parameter Ordered Logit Model. *Accid. Anal. Prev.* **2020**, *135*, 105355. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Chen, F.; Chen, S. Injury Severities of Truck Drivers in Single- and Multi-Vehicle Accidents on Rural Highways. *Accid. Anal. Prev.* **2011**, *43*, 1677–1688. [\[CrossRef\]](#)
8. Islam, S.; Jones, S.L.; Dye, D. Comprehensive Analysis of Single- and Multi-Vehicle Large Truck at-Fault Crashes on Rural and Urban Roadways in Alabama. *Accid. Anal. Prev.* **2014**, *67*, 148–158. [\[CrossRef\]](#)
9. Uddin, M.; Huynh, N. Injury Severity Analysis of Truck-Involved Crashes under Different Weather Conditions. *Accid. Anal. Prev.* **2020**, *141*, 105529. [\[CrossRef\]](#)
10. Al-Bdairi, N.S.S.; Hernandez, S.; Anderson, J. Contributing Factors to Run-Off-Road Crashes Involving Large Trucks under Lighted and Dark Conditions. *J. Transp. Eng. Part A Syst.* **2018**, *144*, 04017066. [\[CrossRef\]](#)
11. Uddin, M.; Huynh, N. Truck-Involved Crashes Injury Severity Analysis for Different Lighting Conditions on Rural and Urban Roadways. *Accid. Anal. Prev.* **2017**, *108*, 44–55. [\[CrossRef\]](#)
12. Anderson, J.; Dong, S. Heavy-Vehicle Injury Severity Analysis by Time of Week: A Mixed Logit Approach Using HSIS Crash Data. *ITE J.* **2017**, *87*, 41–49.
13. Behnood, A.; Mannering, F. Time-of-Day Variations and Temporal Instability of Factors Affecting Injury Severities in Large-Truck Crashes. *Anal. Methods Accid. Res.* **2019**, *23*, 100102. [\[CrossRef\]](#)
14. Zheng, Z.; Lu, P.; Lantz, B. Commercial Truck Crash Injury Severity Analysis Using Gradient Boosting Data Mining Model. *J. Saf. Res.* **2018**, *65*, 115–124. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Uddin, M.; Huynh, N. Factors Influencing Injury Severity of Crashes Involving HAZMAT Trucks. *Int. J. Transp. Sci. Technol.* **2018**, *7*, 1–9. [\[CrossRef\]](#)
16. Khorashadi, A.; Niemeier, D.; Shankar, V.; Mannering, F. Differences in Rural and Urban Driver-Injury Severities in Accidents Involving Large-Trucks: An Exploratory Analysis. *Accid. Anal. Prev.* **2005**, *37*, 910–921. [\[CrossRef\]](#)
17. Osman, M.; Paleti, R.; Mishra, S.; Golias, M.M. Analysis of Injury Severity of Large Truck Crashes in Work Zones. *Accid. Anal. Prev.* **2016**, *97*, 261–273. [\[CrossRef\]](#)
18. Song, L.; Fan, W. Combined Latent Class and Partial Proportional Odds Model Approach to Exploring the Heterogeneities in Truck-Involved Severities at Cross and T-Intersections. *Accid. Anal. Prev.* **2020**, *144*, 105638. [\[CrossRef\]](#)
19. *Large Truck and Bus Crash Facts 2019*; U.S. Department of Transportation, Federal Motor Carrier Safety Administration Analysis Division: Washington, DC, USA, 2021; p. 118.
20. Islam, M.B.; Hernandez, S. *An Empirical Analysis of Fatality Rates for Large Truck Involved Crashes on Interstate Highways*; Transportation Research Board: Indianapolis, IN, USA, 2011.
21. Teoh, E.R.; Carter, D.L.; Smith, S.; McCartt, A.T. Crash Risk Factors for Interstate Large Trucks in North Carolina. *J. Saf. Res.* **2017**, *62*, 13–21. [\[CrossRef\]](#)
22. Anderson, J.; Hernandez, S. Roadway Classifications and the Accident Injury Severities of Heavy-Vehicle Drivers. *Anal. Methods Accid. Res.* **2017**, *15*, 17–28. [\[CrossRef\]](#)
23. De Ona, J.; López, G.; Mujalli, R.; Calvo, F.J. Analysis of Traffic Accidents on Rural Highways Using Latent Class Clustering and Bayesian Networks. *Accid. Anal. Prev.* **2013**, *51*, 1–10. [\[CrossRef\]](#)
24. Kumar, S.; Toshniwal, D. A Data Mining Framework to Analyze Road Accident Data. *J. Big Data* **2015**, *2*, 26. [\[CrossRef\]](#)
25. Taamneh, M.; Taamneh, S.; Alkheder, S. Clustering-Based Classification of Road Traffic Accidents Using Hierarchical Clustering and Artificial Neural Networks. *Int. J. Inj. Control Saf. Promot.* **2017**, *24*, 388–395. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Assi, K.; Rahman, S.M.; Mansoor, U.; Ratrou, N. Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5497. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Li, J.; Liu, J.; Liu, P.; Qi, Y. Analysis of Factors Contributing to the Severity of Large Truck Crashes. *Entropy* **2020**, *22*, 1191. [\[CrossRef\]](#)
28. Behnood, A.; Roshandeh, A.M.; Mannering, F.L. Latent Class Analysis of the Effects of Age, Gender, and Alcohol Consumption on Driver-Injury Severities. *Anal. Methods Accid. Res.* **2014**, *3–4*, 56–91. [\[CrossRef\]](#)
29. Alikhani, M.; Nedaie, A.; Ahmadvand, A. Presentation of Clustering-Classification Heuristic Method for Improvement Accuracy in Classification of Severity of Road Accidents in Iran. *Saf. Sci.* **2013**, *60*, 142–150. [\[CrossRef\]](#)
30. Iranitalab, A.; Khattak, A. Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction. *Accid. Anal. Prev.* **2017**, *108*, 27–36. [\[CrossRef\]](#)
31. Berkhin, P. A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data: Recent Advances in Clustering*; Kogan, J., Nicholas, C., Teboulle, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 25–71. ISBN 978-3-540-28349-2.
32. Huang, Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov.* **1998**, *2*, 283–304. [\[CrossRef\]](#)
33. Soria, J.; Chen, Y.; Stathopoulos, A. K-Prototype Segmentation Analysis on Large-Scale Ridesourcing Trip Data. *Transp. Res. Rec.* **2020**, *2674*, 383–394. [\[CrossRef\]](#)

-
34. Castro, M.; Paleti, R.; Bhat, C.R. A Spatial Generalized Ordered Response Model to Examine Highway Crash Injury Severity. *Accid. Anal. Prev.* **2013**, *52*, 188–203. [[CrossRef](#)]
 35. Lemp, J.D.; Kockelman, K.M.; Unnikrishnan, A. Analysis of Large Truck Crash Severity Using Heteroskedastic Ordered Probit Models. *Accid. Anal. Prev.* **2011**, *43*, 370–380. [[CrossRef](#)] [[PubMed](#)]
 36. Qin, X.; Wang, K.; Cutler, C.E. Logistic Regression Models of the Safety of Large Trucks. *Transp. Res. Rec.* **2013**, *2392*, 1–10. [[CrossRef](#)]
 37. Mannering, F.L.; Bhat, C.R. Analytic Methods in Accident Research: Methodological Frontier and Future Directions. *Anal. Methods Accid. Res.* **2014**, *1*, 1–22. [[CrossRef](#)]
 38. McFadden, D.; Train, K. Mixed MNL Models for Discrete Response. *J. Appl. Econ.* **2000**, *15*, 447–470. [[CrossRef](#)]
 39. Train, K. *Discrete Choice Method With Simulation*; Cambridge University Press: Cambridge, UK, 2003.
 40. Washington, S.; Karlaftis, M.; Mannering, F.; Anastasopoulos, P. *Statistical and Econometric Methods for Transportation Data Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2020; ISBN 0-429-24401-0.
 41. Pahukula, J.; Hernandez, S.; Unnikrishnan, A. A Time of Day Analysis of Crashes Involving Large Trucks in Urban Areas. *Accid. Anal. Prev.* **2015**, *75*, 155–163. [[CrossRef](#)] [[PubMed](#)]
 42. Chang, L.-Y.; Chien, J.-T. Analysis of Driver Injury Severity in Truck-Involved Accidents Using a Non-Parametric Classification Tree Model. *Saf. Sci.* **2013**, *51*, 17–22. [[CrossRef](#)]
 43. Chen, C.; Zhang, J. Exploring Background Risk Factors for Fatigue Crashes Involving Truck Drivers on Regional Roadway Networks: A Case Control Study in Jiangxi and Shaanxi, China. *SpringerPlus* **2016**, *5*, 582. [[CrossRef](#)]
 44. Behnood, A.; Al-Bdairi, N.S.S. Determinant of Injury Severities in Large Truck Crashes: A Weekly Instability Analysis. *Saf. Sci.* **2020**, *131*, 104911. [[CrossRef](#)]