Accident Analysis and Prevention 85 (2015) 219–228

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: [www.elsevier.com/locate/aap](http://www.elsevier.com/locate/aap)

Exploring the application of latent class cluster analysis for  investigating pedestrian crash injury severities in Switzerland

Lekshmi Sasidharana*,*∗, Kun-Feng Wub, Monica Menendeza

a *Swiss Federal Institute of Technology, ETH Zurich, Stefano-Franscini-Platz 5, CH-8093 Zurich, Switzerland*

b *Department of Civil Engineering, National Chiao Tung University, Taiwan*

a r t i c l e i n f o a b s t r a c t

*Article history:*

Received 22 January 2015 Received in revised form 24 September 2015

Accepted 24 September 2015

*Keywords:*

Latent class Cluster analysis Pedestrian Severity

Binary logit

Receiver operating characteristic (ROC) curve

Switzerland

One of the major challenges in trafﬁc safety analyses is the heterogeneous nature of safety data, due to the sundry factors involved in it. This heterogeneity often leads to difﬁculties in interpreting results and conclusions due to unrevealed relationships. Understanding the underlying relationship between injury severities and inﬂuential factors is critical for the selection of appropriate safety countermea- sures. A method commonly employed to address systematic heterogeneity is to focus on any subgroup of data based on the research purpose. However, this need not ensure homogeneity in the data. In this paper, latent class cluster analysis is applied to identify homogenous subgroups for a speciﬁc crash type- pedestrian crashes. The manuscript employs data from police reported pedestrian (2009–2012) crashes in Switzerland. The analyses demonstrate that dividing pedestrian severity data into seven clusters helps in reducing the systematic heterogeneity of the data and to understand the hidden relationships between crash severity levels and socio-demographic, environmental, vehicle, temporal, trafﬁc factors, and main reason for the crash. The pedestrian crash injury severity models were developed for the whole data and individual clusters, and were compared using receiver operating characteristics curve, for which results favored clustering. Overall, the study suggests that latent class clustered regression approach is suitable for reducing heterogeneity and revealing important hidden relationships in trafﬁc safety analyses.

© 2015 Elsevier Ltd. All rights reserved.

# Introduction

Identiﬁcation of possible risk factors and implementation of appropriate countermeasures to reduce the risk of crashes is one of the most conventional and practically adopted trafﬁc safety improvement strategies. Researchers depend on trafﬁc safety data analyses to identify the alleviating and aggravating factors inﬂu- encing the frequency and severity of crashes. Nevertheless, these crashes might have occurred under different conditions, which make the trafﬁc safety data highly heterogeneous in nature, thereby making it difﬁcult to identify some hidden relationships. These rela- tionships may include different effects of the same factor under different conditions. As such, researchers focus on narrow crash variables – crashes due to speciﬁc movements (left turn or right turn movements at signalized intersections; Wang and Abdel-Aty, 2008), crashes associated with speciﬁc vehicle type (sports util- ity vehicle, pickups, minivan and passenger cars; Ulfarsson and Mannering, 2004), motorcycles (Shankar and Mannering, 1996;

∗ Corresponding author.

*E-mail addresses:* [lekshmi.sasidharan@ivt.baug.ethz.ch,](mailto:lekshmi.sasidharan@ivt.baug.ethz.ch) [lechu18181@yahoo.com](mailto:lechu18181@yahoo.com) (L. Sasidharan).

Quddus et al., 2002), crashes involving particular age group of peo- ple or gender (Zhang et al., 2000; Ulfarsson and Mannering, 2004; Sasidharan and Menendez, 2014), speciﬁc accident type (Islam and Mannering, 2006; Ulfarsson and Mannering, 2004; Savolainen and Mannering, 2007) and so on. Although this approach helps to reduce heterogeneity, it does not guarantee homogenous group of crashes in the dataset. Some previous safety studies (Yau, 2004; Depaire et al., 2008; de On˜a et al., 2013) suggest that data mining techniques such as cluster analysis aids in reducing the hetero- geneity in the data. Unlike these studies in which the main focus was on vehicle–vehicle crashes and information on different crash types (rear end, angle, head-on, side-swipe, etc.) were available and used in segmentation, the current study focuses on a spe- ciﬁc crash type–pedestrian crashes. Another approach that takes into account of the unobserved heterogeneity in analyzing pedes- trian injury severities in pedestrian–vehicle crashes includes using a mixed logit model (Kim et al., 2010). Previous pedestrian safety studies have utilized the narratives obtained from police accident report to classify pedestrian crashes into different groups such mid- block dart crashes, crashes due to pedestrian error, turning vehicle crashes, and crashes involving driver’s failure to grant right of way for pedestrians (e.g. Fontaine and Gourlet, 1997; Preusser et al., 2002). These studies suggested that pedestrian crash typologies are

<http://dx.doi.org/10.1016/j.aap.2015.09.020>

0001-4575/© 2015 Elsevier Ltd. All rights reserved.

associated with crash severities and using the whole data involving all crashes without distinguishing the typology makes the data het- erogeneous. One of the implications is that safety countermeasures needed for different pedestrian crash types are different. Accurate estimation of the effects of different factors inﬂuencing pedes- trian injury severities under different conditions is vital as trafﬁc engineers, policy makers and planners rely on this information for identifying appropriate safety countermeasures which includes geometric improvement, trafﬁc control measures, dedicated pedes- trian facilities, modifying land use, educational and enforcement actions. Therefore, an attempt is made to accommodate the sys- tematic heterogeneity in pedestrian crashes in Switzerland using a latent class clustering approach. Subsequently, a binary logit model is used for each of the identiﬁed latent clusters to identify the effect of different crash contributing factors.

* 1. *The heterogeneity in crash data*

As discussed above, heterogeneity in the crash data is unavoidable and highly undesirable. The three main problems of heterogeneous data include: (1) certain crash contributing factors will remain hidden, e.g. a crash contributing factor that is highly inﬂuential for crashes involving speciﬁc vehicle types may not be signiﬁcant in the whole data analysis (Valent et al., 2002; Yau, 2004; Depaire et al., 2008); (2) the magnitude of the effect of certain crash contributing factors may be different for different conditions,

e.g. different effects of injury severities for males and females in different age groups (Ulfarsson and Mannering, 2004; Islam and Mannering, 2006); and (3) the increase or decrease in severity levels for a crash contributing factor may be different for different crash types, e.g. an increase in the probability of less injury accidents for male drivers and increase in probability of severe injury accidents for female drivers for crashes involving guardrails (Ulfarsson and Mannering, 2004).

One way to account for the heterogeneous nature of the data is to divide the data into different homogenous subgroups based on exogenous variables (crash location, crash type, speed limit, roadway geometry, trafﬁc conditions, cause of accidents and so on) and analyze each of the subgroups separately to identify the effect of inﬂuential factors for that subgroup. However, analysis involv- ing ‘dividing the dataset based on all the exogenous variables’ is unrealistic as the number of subgroups can be very large and the sample size in some of the subgroups can be very low, thereby restricting the application of severity models. For example, consid- ering all possible subgroups in a study with 8 binary variables will

results in 28 = 256 subgroups, which deﬁnitely is not practically fea-

sible to analyze and interpret separately. Previous studies show that researchers generally try to divide the data into subgroups based on the objective of research, methodologies used, or on expert domain knowledge (Ulfarsson and Mannering, 2004; Islam and Mannering, 2006; Savolainen and Mannering, 2007). However, some studies suggests that even though the above said factors can result in a workable segmentation of the crash data, one cannot guarantee that each of the subgroups comprise of homogeneous group of crashes (Depaire et al., 2008; de On˜a et al., 2013).

* 1. *The latent class clustering analysis*

To address the heterogeneity issue, a data mining technique, such as cluster analysis can be used to aid in the crash seg- mentation process (Depaire et al., 2008; de On˜a et al., 2013; Kim and Yamashita, 2007; Mohamed et al., 2013). Cluster anal- ysis is a descriptive data mining technique, which can divide a heterogeneous data set into homogenous subgroups or clusters (Berry and Linoff, 1997). It is an unsupervised learning technique that divides data into subgroups or clusters with the goal to

maximize both the homogeneity of elements within the clus- ter and heterogeneity between clusters (Hair et al., 1998). Trafﬁc safety researchers have used different types of cluster analysis in the past to meet different research objectives. Some have used a partitioning method called k-means clustering, a distance based clustering algorithm to identify homogenous crash clusters (Kim and Yamashita, 2007; Mohamed et al., 2013; Hamzehei et al., 2014). Another approach used is hierarchical clustering, which uses methods like Ward’s linkage, Single linkage, average link- age, median linkage, and centroid linkage (Depaire et al., 2008). However, the statistical properties of these methods are mostly unknown (Fraley and Raftery, 1998; Vermunt and Magidson, 2002; Depaire et al., 2008). These methods also require the researcher to specify the number of clusters in advance and assign individual observations to one cluster or the other. To avoid these problems, a probability-model based clustering technique known as latent class clustering analysis (LCA) is proposed to identify homogenous subgroups.

LCA posits that there exists an unobserved or latent categorical

variable that divides the data into mutually exclusive and exhaus- tive latent classes (Collins and Lanza, 2010; Goodman, 1974; Lanza and Rhoades, 2013). Even though the class memberships of indi- vidual crashes are unknown, it can be inferred from the observed variables. In LCA, the probabilities of each crash to be in different clusters are estimated based on different models developed for dif- ferent values of clusters speciﬁed. LCA do not put one crash into any cluster based on any one property, instead it assigns probability of that crash to be in different clusters and assigns a best index cluster for the cluster with the highest probability of accommodating that crash.

* 1. *Pedestrian safety in Switzerland*

Pedestrian safety is very important in Switzerland because of the large share of walking commuters and the high injury severity levels associated with crashes involving this group. In Switzerland, people chose to walk for more than 40% of their total trip time as part of their daily routine (Microcensus, 2010), which points to the importance of ensuring the safety of pedestrians in this coun- try. In addition, Switzerland has a very unique unexplored crash database. The year 2010 in Switzerland witnessed 2418 pedestrian crashes of which 69 were fatal crashes (Swiss council for acci- dent prevention, 2013). A pedestrian safety study in Switzerland showed that the factors inﬂuencing the injury severity levels in old and young pedestrian crashes in Switzerland and its effect are very different when compared to all pedestrian crashes, which clearly points to the need to segment the pedestrian data before analysis (Sasidharan and Menendez, 2014).

Previous studies have identiﬁed that safety countermeasures

needed for different pedestrian crash types are different. A study conducted in France suggests that a typology based on pedestrian crashes can assist in an in-depth analysis to identify inﬂuential factors and determined four groups based on a correspondence analysis for pedestrian fatal injury crashes – elderly pedestrians who were crossing a road in an urban area; children involved in day- time crashes in urban areas while playing or running; pedestrians under inﬂuence in nighttime crashes while walking on road; and pedestrians involved in secondary crashes (Fontaine and Gourlet, 1997). Similarly, another study (Preusser et al., 2002) based on pedestrian crashes in two urban areas in the United States identiﬁed crash groups for mid-block dart crashes, crashes due to pedestrian error, turning vehicle crashes, and crashes involving driver’s failure to grant right of way for pedestrians. Although there is a need to distinguish different types of pedestrian crashes for more accurate crash data analysis, such information is often not directly docu- mented and can only be assembled using the narratives in the

police accident reports. This process is time-consuming and even not feasible in some occasions, and hence there is a need to develop another approach to mitigate this problem.

study. The probability of observing a particular vector of responses is (Lanza and Rhoades, 2013).

*C S Rs*

* 1. *The objective of this study*

*P*(*Z* = *z*) = L*yc* rrrr*pI*(*ys* =*rs* )

(1)

*c*=1 *s*=1 *rs* =1

*s,r* |*cs*

The objective of this study is to explore the application of LCA in reducing the systematic heterogeneity in pedestrian crashes reported in Switzerland and to determine the factors inﬂuencing the pedestrian injury severity levels using binary logit models.

The parameters are estimated using an expectation maximi- zation (EM) algorithm. Using Bayes’s theorem, usually stated as Eq. (2), the posterior membership probability that a crash can then be obtained using Eq. (3).

The study estimated and compared the effects of different crash contributing factors for the whole data and individual clusters. To evaluate the beneﬁts of individual cluster-based analysis compared to whole data analysis, receiver operating characteristic (ROC)

The next section of this paper explains the methodology used

*P*(*A*|*B*) =

*P*(*B*|*A*)*P*(*A*) *P*(*B*)

(IT*S*

*P*(*L* = *c*|*Z* = *z*) =

*s*=1

*c*=1

IT*Rs*

*c*

*pI*(*ys* =*rs* ))*yc*

(2)

curve analysis is also included in the study.

*s*=1

*s,rs* |*c*

*C y* IT*S*

IT*Rs*

*pI*(*ys* =*rs* )

in this study. Section 3 describes the details of the pedestrian

*rs* =1

*s,rs* |*c*

*.* (3)

injury severity data. The results of the analyses are described in the ﬁnal section of the paper, followed by summary and discussions regarding actions that can reduce pedestrian injury severities in Switzerland.

# Methodology

This section of the paper describes the latent class cluster anal- ysis method and the binary logit model used for modeling crash severities. It also includes the estimation of sensitivity and speci- ﬁcity and ROC curve.

* 1. *Latent class cluster analysis*

The LCA is a probability model based cluster analysis method (Depaire et al., 2008; Vermunt and Magidson, 2002; Collins and Lanza, 2010). Like all cluster analysis methods, LCA identiﬁes homogenous clusters of data from the heterogeneous trafﬁc safety data in such a way that the similarity within the cluster is maxi- mized and similarity between cluster elements is minimized. LCA assumes that the data is from a mixture model of different prob- ability distributions (Mohamed et al., 2013). It assumes that there is a latent variable that divides the data into mutually exclusive homogenous subgroups. LCA is reported to have several notable advantages over the conventional cluster analyses methods (Hair et al., 1998; Vermunt and Magidson, 2002; de On˜a et al., 2013) which includes (a) different statistical criteria are available in LCA output, which can be used to identify the most appropriate num- ber of clusters; (b) different types of variables (for example, counts, continuous, categorical, nominal) can be used in LCA directly with- out additional standardization process (Mohamed et al., 2013; de On˜a et al., 2013). In this analysis, the LCA plugin for Stata developed by the Penn State methodology center is used to identify latent classes (Lanza et al., 2014).

To identify latent classes based on crash characteristics, we

follow the work of Lanza and Rhoades (2013). Let *yc* represents latent class membership probability for latent class cluster *c* (*c* = 1,

1. *. . .*, *C*, the number of clusters). Suppose each crash *i* can be

The most appropriate number of clusters “*C*” that explains the

maximum out of the data in hand is unknown in LCA. The value “*C*” can be determined by trying multiple models using differ- ent number of clusters to ﬁnd the best model. The rationale of determining the number of clusters is to minimize assignment error when assigning individual crash to latent class based on their maximum posterior probability of the latent class membership conditional on the characteristics of a crash, *P*(*L* = *c Z* = *z*) (Collins and Lanza, 2010). There have been a few diagnostics proposed to measure the improvement over chance in a model’s assignment accuracy, including the odds of correct classiﬁcation (OCC) (Nagin, 2005), Bayesian Information Criteria (BIC), Akaike Information Cri- terion (AIC), Consistent Akaike Information Criterion (CAIC), and entropy-based measures. The OCC diagnostic tracks whether all latent classes’ assignment is better than chance individually (an OCC of 5 or larger for all latent class separation is considered desir- able); instead, the AIC, BIC, CAIC, and the entropy-based measures are single-number summary of a particular model. In terms of AIC, BIC, and CAIC, the number of clusters that minimizes these criteria will be the most appropriate one. Some researchers suggest that BIC is better than AIC and CAIC in deciding on the number of clusters (Biernacki and Govaert, 1999). However, some others suggested that when analyzing large samples (like trafﬁc safety databases), increasing the number of clusters might not always reach a min- imum value (Bijmolt et al., 2004). They suggest using percentage reduction in BIC between competing models.

|

*rs* =1

With respect to the entropy-based measure, the current study

adopts the entropy measure proposed by Ramaswamy et al. (1993), which is one of the most popular entropy measures that are cur- rently used (Collins and Lanza, 2010). An entropy measure is essentially a weighted average of individual’s posterior probabil- ities of membership, ranging between 0 and 1. Although larger values of entropy measures indicate better latent class separation (Celeux and Soromenho, 1996), latent class assignment error can increase simply as a function of the number of latent classes, so to ﬁnd a balance between the number of clusters and the entropy value, and past research suggested an entropy value of 0.9 is con- sidered as satisfactory. The entropy *R*2 can be estimated using Eq.

(4) (Ramaswamy et al., 1993).

−

characterized with S attributes, *Zi* represents crash *i*’s attribute of characteristic *s*, and *Zi* is a categorical variable from 1, *. . .*, *rs*. *p* indi-

Entropy *R*2 = 1 −

*n i*=1

*C*

*c*=1

*pic* log *pic*

(4)

cates the probability that crash *i* has the attributes in terms of all the S characteristics conditional on latent class membership. There- fore, *pI*(*zs* =*rs* ) represents the probability that a crash has attribute

*n* log *C*

where *pic* – posterior probability that crash *i* belong to cluster *c*; *n*

– number of crashes; *C* – number of clusters.

*r s,rs* |*c*

*s*, conditional on membership in latent class

The entropy *R*2 varies between 0 and 1. An entropy *R*2 value

*s* of characteristic

*c*. The indicator function, *I*(*Zs* = *rs*), equals 1 when the attribute of the characteristic *s* equals *rs*, and equals zero, otherwise. Indepen- dent variables (can be continuous or dichotomous) can be included when modeling *pI*(*zs* =*rs* ), if needed, but it is not considered in this

*s,rs* |*c*

greater than 0.90 indicates clear deﬁnition of clusters (McLachlan and Peel, 2000). After identifying appropriate clusters, crash sever- ity analysis was carried out for both whole dataset and individual clusters.

**Table 1**

Sensitivity and speciﬁcity.

Fatal and severe injury

Minor injury

driver characteristics, roadway characteristics, vehicle character- istics, environmental conditions, main cause of crash, temporal characteristics, and trafﬁc characteristics were included in the study (Table 2).

Test as fatal and severe injury True positive

(sensitivity)

Test as minor injury False negative (missed)

* 1. *Severity modeling*

False positive (false alarm) True negative (speciﬁcity)

There were a total of 9659 reported crashes for which injury severity levels were reported, of which 31.9% involved fatality or severe injury severity level. The remaining crashes involved only minor injuries. From the proportion of fatal and severe injury crashes in different age groups, it is evident that the injury sever- ity levels of pedestrians in different age groups differ very much. It is interesting to note that the injury severity levels of drivers

This section focuses on the method to identify the factors inﬂu- encing the pedestrian crashes in Switzerland. It was noticed that there are some issues associated with the underreporting of PDO crashes in some cantons of Switzerland. To assure the uniformity in data, PDO crashes are excluded from the current study. The number of fatal crashes is few when compared to severe injury and minor injury crashes. Besides, clustering of the data into subgroups will result in very low number of fatal crashes in individual clusters. Therefore, fatal crashes are combined with severe injury crashes in this analysis. As such, there are only two injury severity levels considered in this analysis – fatal/severe injury crashes and minor injury crashes. A binary logistic regression is used to determine the effect of inﬂuential factors (*X*) for the whole data set and individual clusters of pedestrian crashes and is shown in Eq. (5).

*e*(*˛*+*Xiˇ*)

*pi* = *P*(*Yi* = 1) = 1 + *e*(*˛*+*Xiˇ*) (5)

where *˛* – constant; *ˇ* – regression coefﬁcients for *X*.

* 1. *Receiver operating characteristic (ROC) curve*

One of the intentions of LCA is to improve the model precision in predicting severity outcomes. The ROC curve, a widely used medical diagnostic technique (e.g. Swets, 1988; Centor, 1991; Obuchowski, 2003; Pepe, 2003) to compare the effectiveness of different meas- ures for correctly predicting the outcome was used in this study to evaluate the precision of whole data and cluster based models. The result of a diagnostic test can be classiﬁed as a true positive, a true negative, a false positive, or a false negative as shown in Table 1.

The sensitivity (probability of detecting true positive) and speci- ﬁcity (probability of detecting true negative) of a given test for accurately identifying the injury severity, is shown in Eqs. (6) and (7).

Sensitivity (*c*) = True Positive Rate (*c*) = *P*(*Z* ≥ *c*|*Y* = 1) (6) Speciﬁcity (*c*) = 1−False Positive Rate (*c*) = 1−*P*(*Z* ≥ *c*|*Y* = 0) (7)

where a threshold *c* is deﬁned in terms of predicted probability of a severe outcome, *Z*. For every possible *c*, the test is considered as positive if *Z* > *c*, or as negative if *Z* < *c.* The idea is to ﬁrst plot a ROC plot based on all combinations of sensitivity and speciﬁcity in terms of all possible threshold *c*, and then compute the area under the ROC curve, termed as ROC area. The greater the area, the better the accuracy a model can provide.

# Data

Four years of police reported pedestrian crashes in Switzerland (2009–2012) were used in this study. The data were obtained from the national accident database maintained by the Federal Road Ofﬁce (ASTRA) from the Department of the Environment, Trans- port, Energy and Communications of the Swiss Government. The number of crashes and the observed distribution across different severity levels for many key variables including pedestrian and

involved in the crashes also increase with age and follows the same trend as in the case of pedestrian injury severity levels. Table 2 shows that both female and male pedestrians and drivers had approximately equal shares in the pedestrian crashes. The pedes- trian crashes reported in rural areas included a higher proportion of fatalities and severe injury crashes (49.5%) compared to the urban areas (30.8%). Crashes involving drunk pedestrians are found to have more fatalities and serious injuries compared to other pedes- trian crashes (38.74% compared to 31.48% of crashes are in fatal and severe injury group).

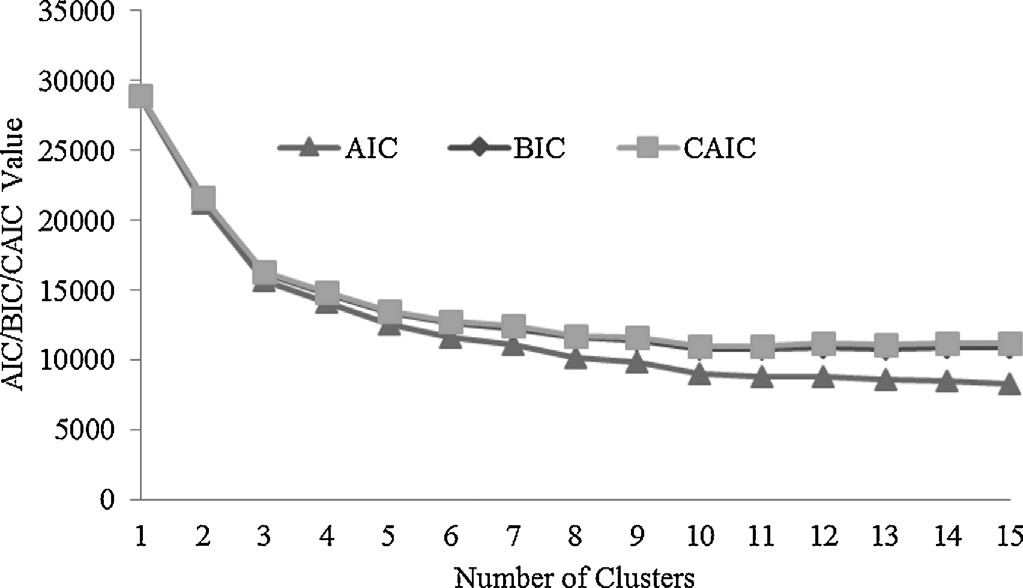
The faults of drivers have resulted in 4276 crashes. For the main reason of crash in Table 2, the authors have included only 7 cat- egories in the main reasons that had high number of reported crashes. Some categories for which the frequency of crashes were comparatively less or not categorized were not included in the table. Heavy vehicles and national highways reported high pro- portion of fatal and severe injury crashes, 43.45% and 60.32%, respectively. It is also interesting to note that the proportion of fatal and severe injury crashes increases signiﬁcantly after 60 kmph speed limit.

# Results

This section discusses the results of LCA and severity modeling for pedestrian crashes. It also includes a discussion on the different criteria and the entropy values used for the selection of appro- priate number of clusters. An ROC curve analysis comparing the whole data analysis and cluster based analysis is also included in this section.

* 1. *Cluster identiﬁcation*

Different models were estimated for pedestrian crashes using variables given in Table 2 by specifying different numbers of clus- ters (1–15). A comparison between the three different criteria (AIC, BIC and CAIC) was made to determine the most appropriate number of clusters/latent classes needed in LCA (Fig. 1).



**Fig. 1.** Identiﬁcation of number of clusters for pedestrian crash analysis.

**Table 2**

Descriptive statistics of pedestrian crashes.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Total crashes | Fatal/severe injury | Minor injury |
| Pedestrian–vehicle crashes | 9659 | 3088 | 6571 |
| Variable | Total no. of crashes | Fatal/severe injury (%) | Minor injury (%) |
| *Pedestrian characteristics* |  |  |  |
| **Gender** |  |  |  |
| Female | 4835 | 31.75 | 68.25 |
| Male | 4797 | 32.25 | 67.75 |
| **Age** |  |  |  |
| ≤15 2386 | | 26.28 | 73.72 |
| >15 and <75 | 5936 | 30.85 | 69.15 |
| 75 and above | 1283 | 48.32 | 51.68 |
| Unknown | 54 | 18.52 | 81.48 |
| **Alcohol** |  |  |  |
| 0 (no alcohol) | 9029 | 31.48 | 68.52 |
| 1 (alcohol impaired) | 604 | 38.74 | 61.26 |
| **Pedestrian behavior** |  |  |  |
| Distraction | 400 | 33.5 | 66.5 |
| Driver at fault | 4276 | 2.67 | 24.37 |
| Crossing mid-block | 8948 | 32.07 | 67.93 |
| Walking along the road | 741 | 29.01 | 70.99 |
| Turning vehicle – crossing at intersection | 711 | 30.66 | 69.34 |
| Pedestrian violating rules | 1168 | 36.9 | 63.1 |

*Vehicle characteristics*

**Vehicle type**

|  |  |  |  |
| --- | --- | --- | --- |
| Car | 7662 | 31.18 | 68.82 |
| Bicycle | 578 | 30.28 | 69.72 |
| Motorbike | 603 | 34.83 | 65.17 |
| Heavy vehicle | 527 | 43.45 | 56.55 |

*Main reason for the crash*

**Driver**

|  |  |  |  |
| --- | --- | --- | --- |
| Violation of right of way (DVROW) | 2524 | 29.04 | 70.96 |
| Violation of rules | 59 | 20.34 | 79.66 |
| Careless reverse | 456 | 29.39 | 70.61 |
| Distraction | 518 | 29.92 | 70.08 |
| Visibility | 51 | 39.22 | 60.78 |
| External inﬂuence | 82 | 28.05 | 71.95 |
| Speeding | 174 | 41.38 | 58.62 |
| Familiar route | 4690 | 31.35 | 68.55 |
| *Roadway characteristics*  **Road category** |  |  |  |
| National road | 63 | 60.32 | 39.68 |
| Canton road | 3700 | 33.16 | 66.84 |
| Township road | 4878 | 29.97 | 70.03 |
| **Geometry** |  |  |  |
| Flat | 6954 | 30.67 | 69.33 |
| Upgrade | 1099 | 35.3 | 64.7 |
| Downgrade | 1582 | 35.15 | 64.85 |
| **Speed limit (kmph)** |  |  |  |
| ≤30  >30 and <60 | 1050  7848  761 | 27.33  30.89  49.41 | 72.67  69.11  50.59 |
| **Area type**  Urban | 9055 | 30.8 | 69.2 |
| Rural | 604 | 49.5 | 50.5 |
| **Accident location** |  |  |  |
| Intersection | 700 | 29.43 | 70.57 |
| Straight | 6066 | 32.25 | 67.75 |
| Parking | 592 | 26.69 | 73.31 |
| Curve | 712 | 34.83 | 65.17 |
| Driveway | 1058 | 32.33 | 67.67 |
| Roundabout | 136 | 33.82 | 66.67 |
| Place | 244 | 33.2 | 66.8 |
| Zebra crossing | 1919 | 30.2 | 69.8 |

≥60

*Temporal characteristics*

**Time**

|  |  |  |  |
| --- | --- | --- | --- |
| 0.01–3.00 | 200 | 37 | 63 |
| 3.01–6.00 | 185 | 32.97 | 67.03 |
| 6.01–9.00 | 1408 | 31.25 | 68.75 |
| 9.01–12.00 | 1606 | 32.25 | 67.75 |
| 12.01–15.00 | 1770 | 30.06 | 69.94 |
| 15.01–18.00 | 2640 | 31.89 | 68.11 |
| 18.01–21.00 | 1433 | 33.5 | 66.5 |
| 21.01–24.00 | 417 | 33.81 | 66.19 |

Table 2 (*Continued)*

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Total no. of crashes | Fatal/severe injury (%) | Minor injury (%) |
| **Day** |  |  |  |
| Weekdays | 7907 | 31.57 | 68.43 |
| *Environmental characteristics*  **Light condition** |  |  |  |
| Daylight | 6657 | 30.73 | 69.27 |
| Dark unlighted | 778 | 34.19 | 65.81 |
| Dark lighted | 2664 | 33.6 | 66.4 |
| Twilight | 744 | 29.44 | 70.56 |
| **Weather** |  |  |  |
| Dry | 6904 | 30.98 | 69.02 |
| Wet | 2455 | 34.99 | 65.01 |
| Snow | 275 | 30.55 | 69.45 |
| **Season** |  |  |  |
| Winter | 2536 | 34.07 | 65.93 |
| Spring | 2217 | 30.81 | 69.19 |
| Summer | 2127 | 30.04 | 69.96 |
| Fall | 2779 | 32.46 | 67.54 |
| **Population density** |  |  |  |
| <100 persons/km2 | 734 | 38.96 | 61.04 |
| 100–200 persons/km2 | 3048 | 31.0 | 69.0 |
| 200–500 persons/km2 | 2734 | 29.55 | 70.45 |
| >500 persons/km2 | 3143 | 33.38 | 66.2 |

It is clear that as the number of clusters increases the values of all three criteria decreases (Fig. 1). The percentage difference in BIC values was computed for different models and the results show that the percentage decrease in BIC drops to less than 1% from 7 clusters onwards. In addition, the entropy value estimated for 7 clusters was 0.92, which indicates a clear separation between the clusters identiﬁed. Also, the AIC and CAIC values supports dividing the data into 7 clusters. Therefore, the pedestrian crash data are divided into 7 clusters for further analysis.

The LCA successfully separates the whole sample into seven clusters (Table 3), and the characteristics of each cluster are consistent with the types of pedestrian crashes identiﬁed in pre- vious studies (e.g. Fontaine and Gourlet, 1997; Preusser et al., 2002). Some of the important variables that shows clear separa- tion among the different clusters are included in Table 3. As an example, the characteristics of cluster two are over-represented in terms of driver’s violation of right-of-way and old pedestrian intersection crashes with turning vehicles. All of the crashes in this cluster involved driver’s violation of right-of-way and 23% of crashes involved old pedestrian at intersections struck by tur- ning vehicles. Similarly, cluster six is over-represented in term of pedestrian’s violation of rules, which accounts for more than 88% of all crashes in this cluster and 80.2% crashes occurred at midblock sections. Moreover, children crossing midblock sections makes up 32% of all crashes in this cluster, which is one of the types of pedestrian crashes identiﬁed in some previous studies (Preusser et al., 2002). The seven clusters identiﬁed are given below.

**Cluster 1 –** Nighttime midblock crashes (Nigmid)

**Cluster 2 –** Driver violating right of way of pedestrians during daytime (PedROWDay)

**Cluster 3 –** Driver violating right of way of pedestrians during nighttime (PedROWNig)

**Cluster 4 –** Weekend night time crashes under inﬂuence (WENigUI)

**Cluster 5 –** Crashes involving old pedestrians at driveways (OldDW)

**Cluster 6 –** Pedestrian violation of rules, children jay walking at midblock (PedVORmid)

**Cluster 7 –** Crashes involving old pedestrians, careless reverse, on driveways (OldrevDW)

* 1. *Severity models*

The results of the binary logit models developed for the whole data and the individual clusters are included in this section. The binary logit models were estimated for the whole data and the indi- vidual clusters using the maximization of log-likelihood method (Table 4).

The predictors with positive coefﬁcients indicate an increase in the probability of occurrence of fatal/severe injury crashes associated with those predictors compared to minor injuries. The following variables including national roads, high speed lim- its ( 60 kmph), alcohol, mid-block crossings, heavy vehicles, and pedestrian violation of rules are found to increase the probabil- ity of fatal and severe injury crashes. However, the magnitude of the effects of predictors is very different in whole data analysis and individual clusters. For example, the odds ratio estimated for

≥

speed limit 60 kmph was 1.734 (e0.551) for whole data analysis

≥

where as it was 1.976 (e0.681) and 2.784 (e1.024) for clusters 1 and 6, respectively. The whole data analysis suggests that the odds of a pedestrian to get fatal/severe injuries on a high speed limit road is 73.4% higher than the base line condition. However, the high speed limit odds ratio estimated for cluster 6 indicates that pedestrians violating rules or children jay walking on high speed limit roads are 178.4% more likely to receive fatal/severe injuries compared to baseline, which is very different from the whole data analysis and very vital information for educational and enforcement purposes. This information was hidden in whole data analysis. Similarly, the odds ratio for heavy vehicle is 1.57 (e0.451) and 3.364 (e1.213) for whole data analysis and cluster 2 (DVROWZD), respectively. The whole analysis suggests that pedestrians when involved in crash with heavy vehicles are 57% more likely to receive fatal/severe injuries compared to baseline. However, cluster 2 analysis suggest that if a heavy vehicle involve in a crash with a pedestrian at zebra crossing due to driver not giving right of way to the pedestrian, the pedestrians are 236.4% more likely to be fatally/seriously injured. This makes sense as the pedestrians would expect the vehicle to stop at the zebra crossing and hence, would not make an attempt to avoid collision. Similarly, nighttime midblock crashes (cluster 1) are more likely to be fatal/serious injury crash compared to whole data analysis.

The indicators for spring, weekdays, morning and evening peak periods, sight obstruction, familiarity of route, bicycle, urban, zebra

Weekday morning peak

25.0

13.0

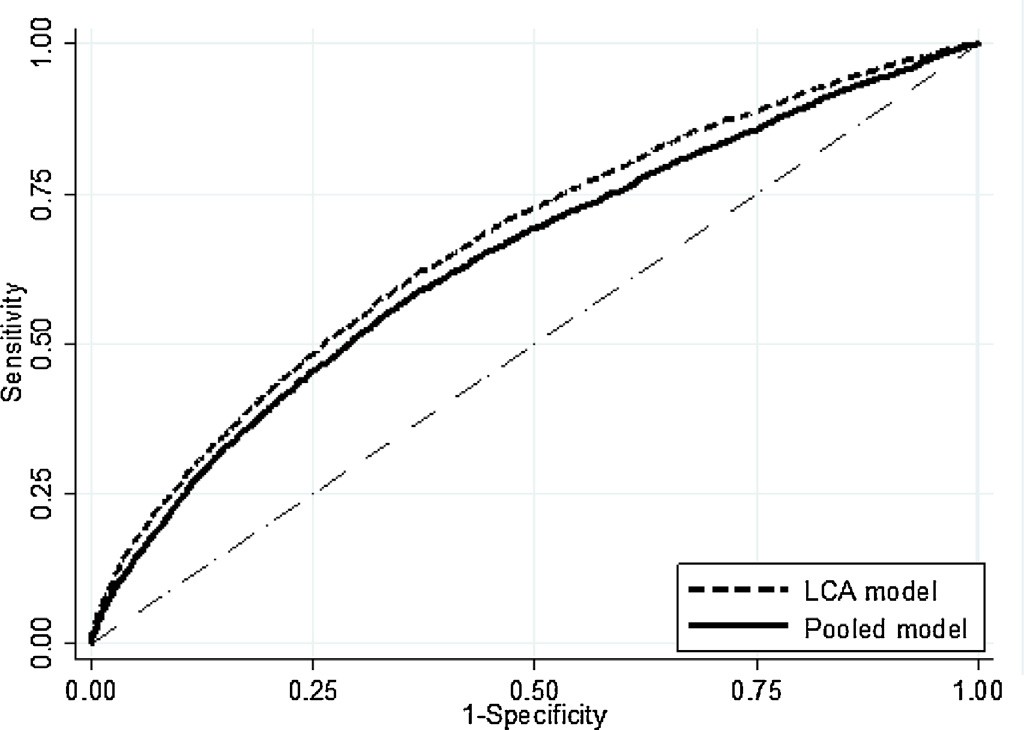
29.0

0.0

11.0

12.0

10.0

**Fig. 2.** ROC curves for pooled and reduced (the LCA model) model.

Old-intersection- turning vehicle

11.0

23.0

12.0

2.0

12.0

9.0

3.0

Young- midblock

11.0

24.0

10.0

2.0

20.0

32.0

3.0

Midblock

Weekend

69.0

88.1

74.3

35.8

55.4

80.2

10.0

9.0

16.0

9.0

62.0

16.0

15.0

21.0

crossing, careless reverse are not signiﬁcant in the whole data analysis, but are highly signiﬁcant in some clusters indicating the importance of these predictors for some speciﬁc pedestrian crash groups. For example, cluster 6 analysis shows that presence of zebra crossings helps to reduce injury severities by 42% when pedestrians are violating rules and children are walking on the road and it was not signiﬁcant in the whole data analysis.

Old pedestrian

The estimated effects of each predictor on crash severity from the pooled model (using whole data) can be considered as a weighted average of that from each reduced model (LCA model). A counterpart to Chow test is conducted to examine whether there is a structural difference between the pooled and reduced model, i.e. whether the coefﬁcients are statistically the same. The result shows that the crash severity model should be modeled separately due to a structural difference between the pooled and reduced models (chi- squared value = 244.9, critical value at 0.1% of signiﬁcance = 59.7). The ROC analysis indicates that the ROC area for the reduced model (0.67) is signiﬁcantly greater than that for the pooled model (0.64), in which the *p*-value is less than 0.00001, suggesting the predicted crash probability are more accurate when utilizing LCA as opposed to the ones predicted by the pooled model (Fig. 2). All taken, both the Chow test and ROC curve conﬁrm that there is a need to separate the whole data set into several subgroups for severity analysis, and the LCA provides a probabilistic approach to identify subgroups.

Careless reverse

0.0

0.0

2.4

3.1

0.0

0.2

29.7

Under inﬂuence

2.0

6.0

5.0

38.0

2.0

11.0

4.0

Drive-way

Dark unlighted

7.1

0.0

3.7

28.8

0.0

0.8

0.0

Night

13.7

6.1

13.9

4.4

10.6

8.8

16.6

100.0

0.0

100.0

100.0

0.0

19.6

2.4

15.3

26.1

16.0

7.1

24.4

12.1

32.5

−

# Summary and discussion

Unlike many developed countries, Swiss population prefer to walk or bike as part of their daily mode of commuting due to healthy life styles, trafﬁc policies, public transportation facilities and land use patterns. The high proportion of pedestrians and their vulnerability to get fatal/serious injuries in crashes make it criti- cal to identify factors inﬂuencing the crash injury severity levels in these crashes. The conventional trafﬁc safety analysis meth- ods, when used to analyze heterogeneous trafﬁc safety data fails to reveal some hidden relationships between injury severity lev- els and inﬂuential factors, which are very important for safety improvements and policy development. Therefore, homogenous latent class clusters for pedestrian crashes in Switzerland were identiﬁed and severity models were developed for the identiﬁed clusters to determine the effect of different inﬂuential factors on injury severity levels. Four years of pedestrian crash data obtained from Swiss Federal Roads Ofﬁce were used in the study.

**Table 3**

Percentage latent class clusters.

Cluster

Driver-ROW

Peds VOR

Inappropriate Ped behavior

0.3

0.0

0.4

18.8

0.0

98.4

1.4

0.0

100.0

91.0

20.0

0.0

0.0

28.0

0.0

0.0

0.0

6.0

0.0

88.0

0.0

The results show that the LCA approach is promising in sepa-

rating the whole data set into meaningful subgroups, which helps to reduce heterogeneity. The ROC technique also conﬁrmed that

1

2

3

4

5

6

7

**Table 4**

Severity model for pedestrian crashes – whole data and clusters. (All variables are signiﬁcant at 0.1 signiﬁcant level.).

Fatal + severe Whole data Nigmid PedROWDay PedROWNig WENigUI OldDW PedVORmid OldrevDW Coeff. S.E. Coeff. S.E. Coeff. S.E. Coeff. S.E. Coeff. S.E. Coeff. S.E. Coeff. S.E. Coeff. S.E.

Summer −0.105 0.061 – – – – – – – – – – – – −0.261 0.155

Spring – – −0.501 0.233 – – – – – – – – – – – –

Weekdays – – – – – – – – −0.333 0.196 – – – – – –

Popdensity >500/km2 −0.191 0.090 – – – – – – – – −0.297 0.152 – – – –

|  |  |  |  |
| --- | --- | --- | --- |
| –  –  – | – – −0.508 0.151  – – −0.613 0.158  – – – – | –  –  – | –  –  – |
| – | – – – – | – | – |
| – | – – – – | – | – |
| – | – – – – | – | – |
| 0.168 | – – – – | – | – |
| – | – – −0.304 0.090 | −0.336 | 0.138 |
| – | – – | 0.310 | 0.185 |

Popdensity 200–500/km2 −0.384 0.091 – – −0.919 0.330 –

Popdensity 100–200/km2 −0.298 0.089 – – – – –

|  |  |  |  |
| --- | --- | --- | --- |
| Time 6 am–9 am | – | – | – – |
| Time 3 pm–6 pm | – | – | – – |
| Intersection | – | – | – – |
| National road | 0.572 | 0.279 | – – |

– – –

−0.335 0.163 –

– – –

– – – Township road −0.124 0.048 – – −0.331 0.145 −0.394

Flat −0.237 0.050 −0.387 0.144 – – –

Sight obstruction – – – – – – –

– –

– –

0.334 0.200

– –

– –

– –

– –

−0.325 0.134

– –

Dry −0.119 0.056 −0.283 0.142 – – – – −0.343 0.206 – – – –

Speed limit ≥60 0.551 0.130 0.681 0.318 – – – – – – – – 1.024 0.381 – –

Speed limit ≤30 −0.129 0.078 – – – – – – – – – – −0.383 0.173

Familiar route – – −0.682 0.238 – – – – – – −0.290 0.161 – – – –

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Alcohol | 0.250 | 0.093 | – | – | – – – – – – 0.555 | 0.312 | – – – – |
| Dark unlighted | 0.245 | 0.061 | 0.304 | 0.164 | – – – – – – – | – | – – – – |

Age ≤15 – – −0.393 0.207 – – – – – – – – – –

Age ≥75 0.892 0.058 0.887 0.184 0.911 0.161 1.014 0.201 1.453 0.368 0.743 0.098 1.267 0.201 0.957 0.145

Heavy vehicle 0.451 0.120 – – 1.213 0.556 – – – – 0.477 0.205 – – 0.663 0.364

Bicycle – – – – – – – – – – – – – – −1.016 0.429

Midblock crossing 0.208 0.057 0.620 0.182 – – – – – – – – 0.334 0.176 – – Urban −0.312 0.147 – – – – – – −0.858 0.480 – – – – – –

Snow −0.327 0.145 – – – – −1.618 0.582 – – – – – – – –

Driver violating ROW −0.199 0.088 – – – – −0.597 0.355 – – – – – – – –

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.082 | – | – | – – | – | – | – | – | – – |
| – | – | – | – – | – | – | – | – | – – |
| – | – | – | – – | – | – | – | – | – – |
| – | – | – | – – | – | – | – | – |  |
| 0.219 | 0.920 | 0.668 | −1.593 0.866 | 0.304 | 0.937 | 0.590 | 0.943 | 0.103 0.394 |

Pedestrian violating rules 0.299

Zebra crossing –

Old female –

Careless reverse –

Constant −0.099

– – – –

−0.540 0.282 – –

– – −0.414 0.243

– – −0.347 0.165

−0.286 0.753 −1.119 0.638

Number of observations 9593 1153 1173 914 630 3076 1265 1382

226

*L. Sasidharan et al. / Accident Analysis and Prevention 85 (2015) 219–228*

Log likelihood at zero −6019.750 −756.170 −695.120 −569.530 −412.902 −1866.310 −835.130 −860.010

Log likelihood at convergence −5740.990 −691.050 −647.760 −526.370 −367.804 −1801.460 −778.510 −807.260

‘–’ Predictors not signiﬁcant at 0.1 or dropped due to collinearity.

+ sign of coefﬁcients indicates an increase in the probability of fatal/severe injuries in crashes.

the predictive crash severity model in terms of the clusters is more accurate than the pooled model. This study revealed that LCA is promising in making clear separation between the clusters identi- ﬁed in the analysis, and hence enhancing homogeneity in the data. It was also found that clustering of data helps in identifying the effect of some factors in subgroups of the data, which were not sig- niﬁcant in the whole data analysis. In addition, the clustering helps trafﬁc safety researchers to understand the variation in the effect of a factor over different subgroups. Together, all these observations point to the importance of segmenting the data into homogenous subgroups to identify some important safety relationships. Among the major ﬁndings are:

* + - The LCA provides a probabilistic approach to separate the whole sample into clusters, and the characteristics of each cluster are consistent with the types of pedestrian crashes identiﬁed in pre- vious studies (Fontaine and Gourlet, 1997; Preusser et al., 2002).
    - A counterpart to Chow test is conducted, the result shows that the

crash severity should be modeled separately due to a structural difference between the pooled and reduced models.

* + - The ROC curve analysis indicates that the predicted crash proba-

bility is more accurate when utilizing LCA as opposed to the ones predicted by the pooled model.

* + - Some predictors are only inﬂuential in some subgroups, indi-

cating that the variables that are signiﬁcant in some particular clusters may be relevant only for the safety in that context and may not pose a safety concern in other conditions.

In this study, not all variables were found to be common in the whole dataset and individual clusters, and this is likely due to the fact that some predictors are only inﬂuential in some subgroups and should be dropped from the reduced model. For example, heavy vehicles are only allowed to operate under certain low speed limits in some residential areas. Even though pedestrian crashes involv- ing heavy vehicles would more likely lead to severe outcome, for some clusters for which most of the observations are in the residen- tial areas, because of the low operating speed, the effects of heavy vehicles on crash severity may not be statistically signiﬁcant. The current study also points to the need of educational actions and policy changes to demote the public from making midblock road crossings at locations without zebra crossings. In addition, exten- sive enforcement measures need to be undertaken in Switzerland to make sure that drivers, especially drivers of heavy vehicles, are giving right of way to the pedestrians at intersections and midblock locations with zebra crossings.

Future research is recommended to utilize the LCA approach

to better understand latent factors that lead to different types of crashes. In-depth analysis should be conducted using narra- tives and collision diagram in police accident reports. Although it is intuitive that safety countermeasures needed for different pedestrian crash types are different, it is unfortunate that unlike vehicle–vehicle crashes where different crash types are deﬁned (e.g. rear end, angle, head-on, side-swipe, etc.), such information is not available for crashes involved pedestrians. This study suggests that pedestrian crash typologies are associated with crash severi- ties and identiﬁed the factors associated with the crash typologies. It is hoped that these ﬁndings can be used to develop better def- initions for pedestrian crash typologies, and ultimately develop effective countermeasures for different types of pedestrian crashes. The results of this study provide directions for researching into different types of pedestrian crashes, especially, when this information is not available as such in the data. The clustering based on relevant variables can serve as a starting point for crash typology. Moreover, although many of the past research have suggested that a crash is a conﬂuence of human factors, vehicle design, and driving environment, this study suggests that more

could be done to further break down crashes. Even the same crash types can be grouped under different homogenous clusters, which will help in the identiﬁcation of more effective countermeasures focusing on the speciﬁc reason behind those identiﬁed clusters of crashes. Finally, the application of latent class clusters is not limited to pedestrian safety and can be extended to all types of crashes for identifying better solutions for improving safety.

# References

Berry, M., Linoff, G., 1997. Data Mining Techniques for Marketing, Sales and Customer Support. John Wiley & Sons.

Biernacki, C., Govaert, G., 1999. Choosing models in model-based clustering and discriminant analysis. J. Stat. Comput. Simul. 64, 49–71.

Bijmolt, T.H., Paas, L.J., Vermunt, J.K., 2004. Country and consumer segmentation: multi-level latent class analysis of ﬁnancial product ownership. Int. J. Res.

Market. 21, 323–340.

Celeux, G., Soromenho, G., 1996. An entropy criterion for assessing the number of clusters in a mixture model. J. Classif. 13, 195–212.

Centor, R.M., 1991. Signal detectability: the use of ROC curves and their analyses.

Med. Decis. Mak. 11 (2), 102–106.

Collins, L.M., Lanza, S.T., 2010. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. Wiley, New York.

de On˜a, J., López, G., Mujalli, R., Calvo, F.J., 2013. Analysis of trafﬁc accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accid. Anal. Prev. 51, 1–10.

Depaire, B., Wets, G., Vanhoof, K., 2008. Trafﬁc accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40 (4), 1257–1266.

Fontaine, H., Gourlet, Y., 1997. Fatal pedestrian accidents in France: a typological analysis. Accid. Anal. Prev. 29 (3), 303–312.

Fraley, C., Raftery, A.E., 1998. How many clusters? Which clustering method?

Answers via model-based cluster analysis. Comput. J. 41, 578–588.

Goodman, L.A., 1974. Exploratory latent structure analysis using both identiﬁable and unidentiﬁable models. Biometrika 61, 215–231.

Hair Jr., J.F., Anderson, R.E., Tatham, R.L., Black, W.C., 1998. Multivariate Data Analysis. Prentice Hall.

Hamzehei, A., Chung, E., Miska, M., 2014. Trafﬁc safety risks trends and patterns analysis on motorways. In: The Transportation Re-search Board (TRB) 93rd Annual Meeting, 12–16 January 2014, Washington, DC.

Islam, S., Mannering, F., 2006. Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. Accid. Anal. Prev. 37 (2), 267–276.

Kim, K., Ulfarsson, G.F., Shankar, V.N., Mannering, F.L., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accid. Anal. Prev. 42, 1751–1758.

Kim, K., Yamashita, E., 2007. Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. J. Adv. Transp. 41 (1), 69–89 (Winter 2007).

Lanza, S.T., Rhoades, B.L., 2013. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. Prev. Sci. 14, 157–168.

Lanza, S.T., Dziak, J.J., Huang, L., Wagner, A.T., Collins, L.M., 2014. LCA Stata Plugin Users’ Guide (Version 1.1). The Methodology Center, University Park, PA, Available from: methodology.psu.edu.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York. Microcensus, 2010. Federal Statistical Ofﬁce, Bern.

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

Nagin, D., 2005. Group-based Modeling of Development. Harvard University Press, Cambridge, MA.

Obuchowski, N.A., 2003. Receiver operating characteristic curves and their use in radiology. Radiology 229 (1), 3–8.

Pepe, M.S., 2003. The Statistical Evaluation of Medical Tests for Classiﬁcation and Prediction. Oxford University Press, New York, NY.

Preusser, D.F., Wells, J.K., Williams, A.F., Weinstein, H.B., 2002. Pedestrian crashes in Washington, DC and Baltimore. Accid. Anal. Prev. 34, 703–710.

Quddus, M., Noland, R., Chin, H., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. J. Saf. Res. 33 (4), 445–462.

Ramaswamy, V., DeSarbo, W., Reibstein, D., Robinson, W., 1993. An empirical pooling approach for estimating marketing mix elasticities with PIMS data. Market. Sci. 12, 103–124.

Sasidharan, L., Menendez, M., 2014. Beneﬁcial and Detrimental Factors Inﬂuencing Pedestrian Crash Injury Severities in Switzerland Using Partial Proportional Odds Model. Working paper. [http://www.ivt.ethz.ch/svt/publications.](http://www.ivt.ethz.ch/svt/publications)

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists’ injury severities in single- and multi-vehicle crashes. Accid. Anal. Prev. 39 (5), 955–963.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. J. Saf. Res. 27 (3), 183–194.

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.

228 *L. Sasidharan et al. / Accident Analysis and Prevention 85 (2015) 219–228*

Ulfarsson, G.F., Mannering, F.L., 2004. Difference in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accid. Anal. Prev. 36 (2), 135–147.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., Barbone, F., 2002. Risk factors for fatal road trafﬁc accidents in Udine, Italy. Accid. Anal. Prev. 34 (1), 71–84.

Vermunt, J.K., Magidson, J., 2002. Latent class cluster analysis. In: Hagenaars, J.A., McCutcheon, A.-L. (Eds.), Applied Latent Class Analysis. Cambridge University Press, Cambridge, pp. 89–106.

Wang, X., Abdel-Aty, M., 2008. Analysis of left-turn crash injury severity by conﬂicting pattern using partial proportional odds models. Accid. Anal. Prev. 40 (8), 1674–1682.

Yau, K.K.W., 2004. Risk factors affecting the severity of single vehicle trafﬁc accidents in Hong Kong. Accid. Anal. Prev. 36 (3), 333–340.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G., Mao, Y., 2000. Factors affecting the severity of motor vehicle trafﬁc crashes involving elderly drivers in Ontario. Accid. Anal. Prev. 32 (1), 117–125.