9th International Cycling Safety Conference
4–6th November 2020, Lund, Sweden

Changes in interpretation and empirical findings when fully specifying latent classes in a Latent Class Choice Model

M. S. Myhrmann*, **G. Sfeir#**

* Institute of Economics, Management and Technology
Danish University of Technology
Bygningstorvet 116B, 2800 Kgs. Lyngby, Denmark
e-mail: mskyum@dtu.dk

Department of Civil and Environmental Engineering
American University of Beirut
Riad el Solh 1107 2020 Beirut, Lebanon
e-mail: gms12@mail.aub.edu

ABSTRACT

This study investigates injury severity outcomes of single-bicycle accidents. It applies a Latent Class Choice Model (LCCM) with a flexible class membership component where latent classes are formulated as a Gaussian-Bernoulli mixture model while the class-specific injury severity outcomes are formulated by means of random-utility specifications. Using the aforementioned modelling framework, this study estimates and compares two different models, the socio-demographic and the fully specified. The former clusters the individuals, who had been in accidents, using only socio-demographic variables and defines the class-specific utilities of injury severity outcomes using the remaining explanatory variables (e.g. road characteristics) while the latter makes use of all available explanatory variables including the socio-demographic ones to describe the class assignment and defines the class-specific utilities of injury severity outcomes using only alternative-specific constants. A dataset of about 1,720 cyclists who have been injured in Aarhus, the second most populous municipality in Denmark, is used to illustrate the two modelling approaches and results are compared on the basis of parameters' estimates, goodness-of-fit measures, latent classes' interpretability, and the average marginal effects of all variables. Results show that the socio-demographic model outperforms the fully specified model in terms of goodness-of-fit and prediction accuracy measures. However, the fully specified model is capable of identifying a much larger number of classes and as such provides better interpretability and more in-depth understanding of the nature of accidents.

Keywords: single-bicycle accidents, injury severity, latent class choice models.

Word count: 5,460 + 8 Tables + 1 Figure

1 INTRODUCTION

With the increased endorsement that the bicycle is receiving due to it being more sustainable than vehicles powered by combustion engines, understanding factors related to the injury severity outcome of bicycle crashes. Previous safety analysis studies primarily focused on bicycle crash frequency (Christiansen & Warnecke, 2018; Dozza, 2017; Fournier et al., 2017; Saha et al., 2018). While there has been an increase in the literature regarding injury severity outcomes of road traffic crashes, these are still comparatively few. Recent studies on the injury severity following bicycle crashes primarily focus on cyclist crashes that involve collisions with other road users, such as motorised vehicles (Behnood & Mannering, 2017; Kim, Kim et al., 2007) and other road users (Kaplan et al., 2014; Prati et al., 2017). Meanwhile, single-bicycle crashes, the most frequent type, are generally overlooked. Single-bicycle accidents have been reported to make up around 50% of all bicycle crashes (Beck et al., 2016; Dozza, 2017; Møller et al., 2018). Therefore, the factors associated to the injury severity outcome of single-bicycle accidents should be considered especially important.

When modelling the injury severity outcome of road traffic crashes the standard approach is to treat accidents as discrete outcomes and to apply variations of Multinomial Logistic Regressions (MNLs). The statistical frontier of traffic safety analysis focuses on addressing the heterogeneous nature of safety data. Not accounting for possible heterogeneity in the data might lead to erroneous inferences and biased parameter estimates (Mannering & Bhat, 2014). This in turn might lead to the wrong mitigating actions being undertaken. With regard modelling injury severity outcomes, an increasingly popular method to address heterogeneity is the finite mixture model approach known as the Latent Class Model (LCM). It is based on the assumption that a population consists of a finite number of heterogeneous sub-groups that in turn are homogeneous based on the characteristics within the groups. This type of method has previously been used to model the injury severity outcome of single-car crashes (Fountas, Anastasopoulos, & Mannering, 2018; Li et al., 2018). Extending the framework further by letting the class-assignment vary as a function of explanatory variables, further addresses the possible heterogeneity of probabilistic class-assignment of the accidents.

In this approach the explanatory variables used for the class-assignment function of the latent class model are often chosen to be socio-demographic variables, to potentially allow for meaningful interpretation of, or insight in, the sub-classes, it is still very little that one is able to conclude on the sub-classes in the LCM approach. If one is more interested in segmenting the group into interpretable sub-groups, one can apply Latent Class Clustering (LCC) (Depaire et al., 2008). This method is often subsequently paired with discrete choice models to assess the contributing factors to injury severity specific to the clusters (Liu & Fan, 2020; Sasidharan et al., 2015). One potential problem of the LCC approach however, is that while yielding more interpretable latent classes/ or clusters, the clusters themselves are designed to segment the data to group similar accidents not with the injury severity outcome in mind . In contrast to this, the Latent Class Choice Model (LCCM) approach, attempts to segment the accidents into groups to make them homogeneous but with the aim of optimising the model's predictive ability. As such, practitioners are faced with the dilemma of having to choose between finding the overall factors associated to aggravating injuries or identify segments of the population to address first.

In this study we propose a potential solution to the above problem. We use the LCCM framework, to assess the injury severity outcome of single-bicycle crashes. The difference in this specific approach is that we intend to use all the variables in the class-assignment function, similar to clustering part of Latent Class Clustering, however this way we still attempt to optimise the groups for predictive ability. We compare the output of the suggested model with the classic socio-demographic LCCM and compare results, as well as the inference drawn from them. The intention is that this approach might yield interpretation of the assigned classes similar to the clusters from LCC while providing better overall performance with regard to predictive ability.

2 METHODOLOGY

The specific approach we adopt in this study is a recently developed LCCM framework called Gaussian-Bernoulli Mixture Latent Class Choice Model (GBMLCCM)(Sfeir et al., 2020). It is a hybrid framework that combines Gaussian-Bernoulli Mixture Model and random utility models (e.g. logit models). The Gaussian-Bernoulli Mixture Model, a model-based clustering technique, is used to cluster data (e.g. people involved in accidents) into homogenous groups by using Gaussian mixtures for continuous variables and Bernoulli mixture for discrete variables, while the random utility models are used to develop class-specific injury severity models (Figure 1).

2.1 Gaussian-Bernoulli Mixture Model

The Gaussian-Bernoulli Mixture Model is a probabilistic clustering approach used to assign data points to different clusters (components of the mixture). The Gaussian Mixture Model (GMM) is a mixture of K Gaussian distributions used to account for continuous variables while the Bernoulli Mixture Model (BMM) is a mixture of the product of Bernoulli probability functions to account for discrete/binary variables.

We specify the vectors of continuous and discrete variables used for clustering of incident n as S_{cn} (with dimension D_c) and S_{dn} (with dimension D_d), respectively. As for the vector of variables that enters the random utility models, it is noted as X_n. In addition, we specify q_{nk} as the assignment variable with q_{nk} equal to 1 if incident n is assigned to cluster k and 0 otherwise.

The joint probability of S_{cn}, S_{dn}, and q_{nk} can be specified as the product of the class probability and the densities of S_{cn} and S_{dn} conditional on the class assignment as follows:

$$\mathrm{P}(\mathrm{S_{cn}}, \mathrm{S_{dn}}, \mathrm{q_{nk}}) = \mathrm{P}(\mathrm{q_{nk}}|\pi_\mathrm{k})\mathrm{P}(\mathrm{S_{cn}}|\mathrm{q_{nk}} = 1, \mu_\mathrm{ck}, \Sigma_\mathrm{ck})\mathrm{P}(\mathrm{S_{dn}}|\mathrm{q_{nk}} = 1, \mu_\mathrm{dk}) \qquad (1)$$

With:

$$P(q_{nk}|\pi_k) = \pi_k \qquad (2)$$

$$\sum_{k=1}^{K} \pi_k = 1. \qquad (3)$$

$$P(S_{cn}|q_{nk} = 1, \mu_{ck}, \Sigma_{ck}) = \mathcal{N}(S_{cn}|\mu_{ck}, \Sigma_{ck})$$

$$= \frac{1}{\sqrt{(2\pi)^{D_c}|\Sigma_{ck}|}} exp\left(-\frac{1}{2}(S_{cn} - \mu_{ck})^T \Sigma_{ck}^{-1} (S_{cn} - \mu_{ck})\right) \qquad (4)$$

$$P(S_{dn}|q_{nk} = 1, \mu_{dk}) = \prod_{i=1}^{D_d} \mu_{dk_i}^{S_{dni}} \left(1 - \mu_{dk_i}\right)^{(1-S_{dni})} \qquad (5)$$

Where μ_{ck} is the mean vector of Gaussian k, $\sum_{ck}$ is the covariance matrix of Gaussian k, $|\sum_{ck}|$ is the determinant of the covariance matrix, π_k is the mixing distribution or the probability that an accident belongs to cluster k, and μ_{dk} is the mean vector of the k Bernoulli mixture.

2.2 Class-Specific Model

Conditioned on the class membership of incident n, the class-specific model estimates the probability of having a specific injury severity outcome as a function of some exogenous variables X_n. The utility of incident n resulting in an injury severity j, conditional on the fact that incident n belongs to class k, is specified as follows:

$$\mathrm{U_{nj|k}} = \mathrm{X'_{nj}}\beta_\mathrm{k} + \varepsilon_\mathrm{nj|k} \qquad (6)$$

Where X_{nj} is a vector of exogenous variables of alternative j including an alternative-specific constant, β_k is a vector of corresponding unknown parameters that need to be estimated, and $\varepsilon_{nj|k}$ is a random disturbance term that is independently and identically distributed *(iid)* Extreme Value Type I over incidents, injury severity outcomes, and classes.

Conditional on class k, the probability of incident n resulting in an injury severity j is expressed as follows:

$$P(y_{nj}|X_{nj}, q_{nk}, \beta_k) = \frac{e^{V_{nj|k}}}{\sum_{j'=1}^{J} e^{V_{nj'|k}}} \tag{7}$$

Where J is the number of possible severity outcomes, y_{nj} is equal to 1 if incident n results in injury severity j and 0 otherwise. Conditional on class k, the probability of the actual severity outcome of incident n is expressed as follows:

$$P(y_n|X_n, q_{nk}, \beta_k) = \prod_{j=1}^{J} \left(P(y_{nj}|X_{nj}, q_{nk})\right)^{y_{nj}} \tag{8}$$

Where y_n is a vector of J injury severity outcomes y_{nj}.

2.3 Joint Model

The joint probability of S_{cn}, S_{dn}, y_n, and q_{nk} is specified as the product of equations 1 and 8:

$$\begin{aligned} P(S_{cn}, S_{dn}, y_n, q_{nk}) \\ &= P(q_{nk}|\pi_k)P(S_{cn}|q_{nk} = 1, \mu_{ck}, \Sigma_{ck})P(S_{dn}|q_{nk} \\ &= 1, \mu_{dk})P(y_n|X_n, q_{nk}, \beta_k) \end{aligned} \tag{9}$$

Finally, the likelihood of the GBMLCCM for all incidents N can be obtained by summing equation 9 over all classes K:

$$P(S_c, S_d, y) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(S_{cn}, S_{dn}, y_n, q_{nk}) \tag{10}$$

The joint likelihood is maximized using the Expectation-Maximization (EM) algorithm. Firstly, the unknown parameters (μ_{ck}, μ_{dk}, etc.) are initialized, Secondly the expectation values of the class assignment variables q_{nk} are estimated (E-step) using the current parameter values. Thirdly, the parameters are re-estimated by setting the derivatives of the joint likelihood with respect to the parameters to zero (M-step). Finally, the likelihood is estimated, and convergence is checked. If convergence is not reached, the algorithms return to the E-step. For more details about the derivation and implementation of the EM algorithm for GBMLCCM, readers can refer to Sfeir et al., 2020.

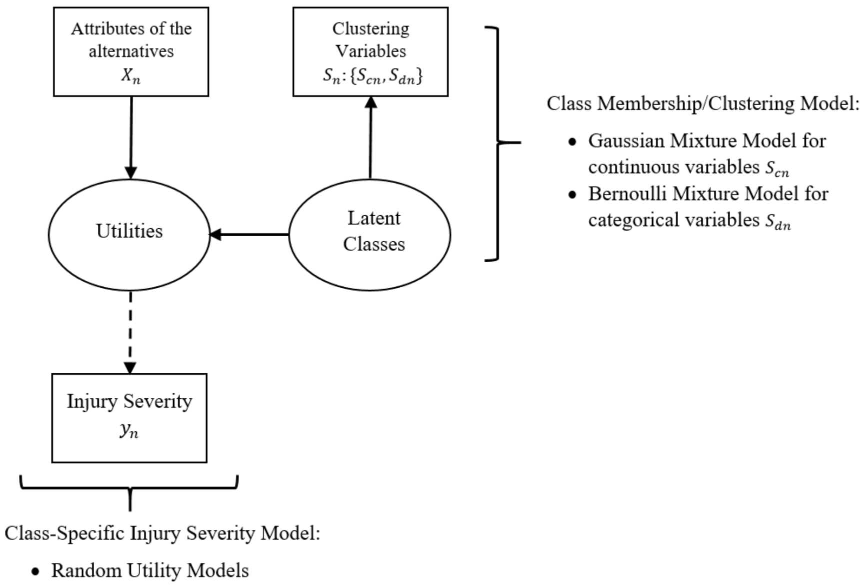

Figure 1: Gaussian-Bernoulli Mixture Latent Class Choice Model (GBMLCCM)

3 RESULTS

3.1 Data

The data for this study contain information about 1,720 injured cyclists in Aarhus (the second most populous municipality in Denmark), who visited the hospital's emergency department in the years 2010 to 2015 due to involvement in single-bicycle traffic accident. For each cyclist information on, age, gender, use of helmet, injury severity, and information on the crash (road type, surface condition (e.g. dry, slippery) and time) were recorded. The crash data were merged with road maintenance data provided by the municipality of Aarhus.

Table 1: Data set attributes

Variable Description	Categories	Frequency (%)

The Cyclist Injury Severity	Severe	19
	Slight	60
	No evident injury	21
Gender	Male	52
	Female	48
Helmet	Used	34
	Not used	66
The road AADCT	0-500	17
	501-1500	20
	1501-3000	16
	3001-5000	16
	> 5000	31
Road design	Bicycle lane	24
	Straight road	52
	Curve	3
	Intersection	16
	Roundabout	1
	Other	4
Road maintenance	Good	59
	Lacking	41
Bicycle lane condition	Good	21
	Acceptable	9
	Bad	3
	Not recorded	67
High curb stone	Yes	6
	No	94
Potholes	Yes	13
	No	87
Slippery road surface	Yes	30
	No	70
Time specific – season	Spring	23
	Summer	29
	Autumn	26
	Winter	22
Light conditions	Dark	38
	Daylight	61
	Unknown	1

The road maintenance data consist of information such as: condition of the bicycle lane and the road, as well as specific of road maintenance issues such as potholes and high curb stones and the roads condition.For the roads where single-bicycle accidents had occurred, a categorisation of the annual average daily cycling traffic (AADCT) was made. This was based on bicycle census gathered by the Aarhus municipality. This resulted in the following groups of AADCT: 1. 0-500; 2. 501-1500; 3. 1501-3000; 4. 3001-5000; and 5. more than 5001. The descriptive statistics for the variables in the study are shown in Table 1. Cyclist age is not included in Table 1 as it was considered as a continuous variable. The cyclist age follows a bi-modal distribution with the larger mode being around 23 and the other around 54 years of age and the overall average age being 36.6 years.

The injury severity outcome, of an injured cyclist, reported by the emergency room is categorised as severe, slight or 'no evident' injury. As seen in Table 1, the majority of the accidents result in slight injury (60%). There is a slightly higher representation of male cyclists in the sample (52%) than female (48%). Most of the accidents occurred on low volume roads with an AADCT less than 1500 (37%) or on high volume roads roads with an AADCT greater than 5000 (31%). Half of the crashes occurred on straight roads without any separated bicycle lane and 24% on roads with bicycle lanes, where most of them were considered to be in good condition. Only a few roads were registered with problems such as a high curbstone (6%) and potholes (13%). Many of the single-bicycle accidents (30%) occurred on roads with slippery surfaces.

3.2 Estimation and application

As previously stated, two models with two different specifications were developed using the GBMLCCM framework and compared based on goodness-of-fit measures and cluster characteristics. The first model makes use of only socio-demographic variables to describe the class assignment and defines the class-specific utilities of injury severity outcomes using the remaining explanatory variables (e.g. road characteristics). The second model clusters the individuals who had been in accidents using all available explanatory variables including socio-demographic variables and defines the class-specific utilities using only alternative-specific constants. The variables that are used in the models are presented in Table 2.

Table 2: Explanatory variables used in the models

Variables	Socio-demographic model	Fully Specified model

	Gaussian-Bernoulli Mixture Model	Class-Specific Severity Model	Gaussian-Bernoulli Mixture Model	Class-Specific Severity Model
Alternative-specific constant (ASC)		*x*		*x*
Age (continuous)	*x*		*x*	
Male	*x*		*x*	
Helmet	*x*		*x*	
Alcohol	*x*		*x*	
Time specific season		*x*	*x*	
Road maintenance		*x*	*x*	
Light conditions		*x*	*x*	
Road design		*x*	*x*	
The road AADCT		*x*	*x*	
Bicycle lane condition		*x*	*x*	
High curb stone		*x*	*x*	
Potholes		*x*	*x*	
Slippery		*x*	*x*	

3.3 Fully Specified Model

We vary the number of latent classes from 1 to 9 and estimate the GBMLCCM model 10 times with different random initialization. Table 3 shows the Log-Likelihood (LL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) of the fully specified model as a function of the increasing number of latent classes (clusters). Results show that the model with 9 latent classes has the best LL and AIC. However, the one with 6 latent classes has the best BIC. It is known that the penalty term on the number of parameters and model complexity is higher in BIC than AIC. Therefore, we select the model with 6 classes as the optimal clustering solution using the fully specified approach.

The results of the fully specified model with 6 mixtures are shown in Table 4. They describe the fraction of characteristics present in each cluster to make up the factors experienced by the cyclists at the accident. The continuous variable, age, was standardized before applying the Gaussian Mixture Model. Therefore, a negative value means the cluster is characterized by younger cyclists and a positive value means cyclists are older than the average (36.6 years).

3.3.1 Interpreting the latent classes

K1: Elderly people – with accidents bicycle lanes and in intersections

The first cluster holds 20% of all accidents and is characterized by older cyclists from both genders, where 42% of all accidents occurred on a bicycle lane. In 11% of the cases the condition of these bicycle lanes was good, but the condition of the rest were not recorded. Furthermore, this class contains no accidents in the winter, i.e. all accidents occurred during “cycling seasons”. When the accidents did not occur on a bicycle path, generally roads were in good (63.6%) or acceptable condition (29.6%), with few high curb stones (0.7%), an average amount of potholes (11.5%), and during daylight hours (79.1%). Crashed cyclists assigned to this cluster are more likely to have had slight injuries than no injuries, and slightly less likely to suffer a severe injury as not being injured (“ASC – slight injury” is positive while “ASC – severe injury” is negative and “ASC – no injury” is fixed to 0 for identification issues).

K2: Elderly males victim to road and bicycle lane maintenance issues

The second cluster contains 15% of all accidents. The cyclists from this cluster are, on average, older than the cyclists belonging to the first cluster, and mostly male (58.1%) who were not wearing a helmet (67.7%) at the time of the accident. Most of the accidents happened during the cycling season (96.5%) and daylight hours (65.8%). The accidents generally occurred on straight road section in mixed traffic or on bicycle lanes (~75%). The roads were generally characterised by a good surface (57%) but 42% of the road sections were only deemed of acceptable condition. The conditions of all bicycle lanes where accidents occurred were recorded, and a majority of these (47%) were only of acceptable condition, and 17% were considered to be in bad condition. This group almost contains all accidents occurring on bicycle lanes of bad condition. The main part of accidents in this group also occurred on low volume roads, with cycling traffic volume equal to or below 1,500 (57.5%). On the injury severity outcome, this cluster can be considered as the cluster with the highest likelihood of severe, as well as slight injuries (see ASCs for both severe and slight injuries across all six clusters).

K3: Younger cyclists on high volume roads

This cluster, K3, has the youngest cyclists from both genders. In addition, K3 has the highest percentage of cyclists who were tested positive for having alcohol in their blood (25.9%), were not wearing a helmet (92.4%) at the time of the accident and were generally involved in accidents during some dark hours (62.0%). The accidents are, to some extent, equally distributed across all seasons. In addition, most of the accidents occurred on straight roads (66.1%), with good surface (65.4%), very few potholes (3.7%), unreported bicycle lane conditions (76.4%), and very high annual daily cycling traffic volume (75.5% above 5000), however often slipper conditions (24.1%). This cluster is associated with higher baseline likelihood of slight injury outcome from crashes, compared to the likelihood of no injury. Meanwhile, the cyclists in this cluster are unlikely to suffer severe injury from accidents.

K4: Middle-aged on bicycle lanes

Accidents belonging to this cluster have happened across the year with the highest percentage being registered during the fall season (33.9%). All of the accidents occurred on road sections and bicycle lanes with medium to high cyclist volume (3001-5000). The majority of the accidents occurred on bicycle lanes (51.2%) that were generally in good condition (77.7%). The rest of the accidents occurred mainly on straight roads (27.0%) and intersections (20.8%) with most of the roads sections' surface conditions (83.6%) only deemed to be of acceptable condition. K4 is the smallest cluster with only 5% of all accidents. Compared to the other clusters, accidents from K4 were the least likely to result in severe injuries since "ASC – severe injury" is the lowest (-0.603) across all clusters.

K5: Winter crash group

This cluster has the highest percentage of cyclists who tested negative for alcohol (96.5%) at the time of the accident and the highest amount of female cyclists (55.6%). Compared to the other 5 clusters, cyclists in K5 tend to be wearing a helmet at the time of the accident (44.4%). All accidents were registered during the winter which explains the high percentage accidents in slippery conditions (79.3%) and the high percentage of accidents in dark hours (60.6%) since Denmark is known for its short daylight hours during the winter. In addition, most of the accidents happened on straight roads (46.5%), bicycle lanes (24.0%), or intersections (21.1%), with good to acceptable road surface conditions (93%).

K6: Elderly cyclists who had accidents on straight roads

This is the largest among the six clusters with 25% of all accidents. This cluster includes the oldest cyclists with high helmet usage (39.5%) and low alcohol consumption (8.9%). No accidents were registered during the winter while the highest number of accidents occurred during the summer (42.4%) and in daylight hours (76.5%). All accidents happened on straight roads with good surface (62.7%), few potholes (13.2%), and no reports on bicycle lane conditions (89.2%). Cyclists in this cluster are more likely to have had slight to severe injuries since both ASCs are positive.

Table 3: Log-Likelihood, AIC, and BIC of the fully specified model

Number of Classes (K)	Log-Likelihood (LL)	AIC	BIC
2	-21,592.27	43,302.54	43,624.09
3	-21,296.29	42,770.58	43,255.64
4	-20,967.22	42,172.43	42,820.99
5	-20,785.16	41,868.32	42,680.38
6	-20,631.56	41,621.12	**42,596.69**
7	-20,539.66	41,497.32	42,636.39
8	-20,433.89	41,345.78	42,648.35
9	**-20,356.34**	**41,250.68**	42,716.75

Table 4: Estimates of the Fully Specified Approach with 6 latent Classes

		K1	K2	K3	K4	K5	K6
		Mean Matrix of the Gaussian-Bernoulli Mixture Model					
Age	Continuous	0.152	0.170	-0.737	-0.011	0.144	0.194
Gender	Male	0.545	0.581	0.473	0.561	0.444	0.551
Helmet	Yes	0.453	0.323	0.076	0.282	0.441	0.395
Alcohol	Yes	0.059	0.115	0.259	0.130	0.035	0.089
Season	Winter	0	0.036	0.178	0.194	1	0
	Spring	0.284	0.283	0.254	0.183	0	0.307
	Summer	0.378	0.359	0.227	0.284	0	0.424
	Fall	0.339	0.323	0.341	0.339	0	0.269
Road Maintenance	Good	0.636	0.570	0.654	0.164	0.559	0.627
	Acceptable	0.296	0.421	0.327	0.836	0.371	0.299
	Bad	0.068	0.009	0.019	0	0.070	0.074
Light	Daylight	0.791	0.658	0.380	0.482	0.391	0.765
	Dark	0.205	0.316	0.620	0.507	0.606	0.215
	Unknown	0.004	0.027	0	0.011	0.003	0.020
Road Design	Bicycle Lane	0.419	0.357	0.217	0.512	0.240	0
	Straight	0	0.408	0.661	0.270	0.465	1
	Intersection	0.318	0.189	0.092	0.208	0.211	0
	Curve	0.089	0.014	0.010	0	0.035	0
	Roundabout	0.025	0.006	0	0	0.013	0
	Others	0.150	0.026	0.020	0.011	0.036	0
AADT	500	0.219	0.114	0.009	0	0.225	0.266
	500-1500	0.233	0.469	0.076	0	0.194	0.172
	1501-3000	0.189	0.182	0.161	0	0.193	0.138
	3001-5000	0.113	0.007	0.236	1	0.108	0.092
	5001-	0.246	0.229	0.519	0	0.279	0.334
Bicycle Lane Condition	Good	0.111	0.371	0.223	0.777	0.115	0.108
	Acceptable	0	0.469	0.013	0.078	0.071	0
	Bad	0	0.160	0	0.000	0.039	0
	No record	0.889	0	0.764	0.146	0.775	0.892
High Curb stone	Yes	0.007	0.091	0.003	0.585	0.016	0.018
Slippery	Yes	0.202	0.205	0.241	0.354	0.793	0.136
Potholes	Yes	0.115	0.176	0.067	0.169	0.127	0.132
	Parameter Estimates of the Class-Specific Injury Severity Model						
ASC – Severe Injury		-0.0321	0.281	-0.500	-0.603	-0.112	0.0784
ASC – Slight Injury		1.148	1.188	1.137	1.037	1.041	0.983
		Mixing Coefficients (Class Probability)					
		0.20	0.15	0.17	0.05	0.17	0.25

3.4 Socio-Demographic Model

For the socio-demographic specification, only 3 classes were identified. Increasing the number of classes beyond 3 resulted in large parameter estimates and standard deviations in the class-specific injury severity model. We can conclude from Table 5 that two latent classes is the best solution using the socio-demographic specification since it has the lowest BIC.

Table 5: Log-Likelihood, AIC, and BIC of the socio-demographic model

Number of Classes (K)	Log-Likelihood (LL)	AIC	BIC
2	-6,581.06	13,280.12	**13,601.67**
3	**-6,523.45**	**13,224.90**	13,709.96

Next, the two latent classes are described based on the results of the GBMLCCM (Table 6). All variables within the class-specific injury severity model have generic coefficients that are included in the utilities of severe and slight injuries while the alternative "no injury" is kept as a base.

3.4.1 Interpreting the latent classes

K1: Cautious old cyclists

Latent class 1 contains on average 68% of all accidents and is characterized by older cyclists ($\mu_{age} > 0$) who, compared to latent class 2, are more likely to be wearing a helmet and less likely to be under the influence of alcohol at the time of the accident. Gender does not seem to differ significantly between the two clusters. However, under the baseline conditions, cyclists belonging to this cluster are more likely to have had slight and severe injuries compared to both the likelihood of sustaining no injury (ASCs of severe and slight injuries are positive compared to the ASC of no injury which is fixed to 0 for identification issues), as well as when compared to the other latent cluster, K2. Accidents with slight to severe injuries are more likely to have occurred during the summer or winter, as well as in lack of daylight. Road surfaces that are not classified to be in good or adequate condition are also associated to more severe injuries and so are accidents occurring on roads with bicyclist volume less than 5000 AADCT and on roads with acceptable to bad maintenance conditions and slightly high curb stones.

Table 6: Estimates of the Socio-Demographic Approach with 2 latent Classes

		K1	K2
Mean Matrix of the Gaussian-Bernoulli Mixture Model			
Age	Continuous	0.364	-0.772
Gender	Male	0.552	0.460
Helmet	Yes	0.454	0.104
Alcohol	Yes	0.083	0.166
Parameter Estimates of the Class-Specific Model			
ASC – Severe Injury		0.556	-0.391
ASC – Slight Injury		1.496	1.345
Season	Winter	-0.085	-0.150
	Spring	-0.554	-0.035
	Fall	-0.259	0.348
Road Maintenance	Acceptable	0.337	-0.404
	Bad	0.307	-1.156
Light	Dark	0.159	0.547
	Unknown	-0.387	0.528
Road Design	Straight	-0.202	-1.135
	Intersection	-0.742	-1.314
	Curve	0.240	-0.894
	Roundabout	-0.033	-2.796
	Others	0.013	-1.293
AADT	500	0.233	0.534
	500-1500	0.212	-0.520
	1501-3000	0.151	0.111
	3001-5000	0.086	-0.126
Bicycle Lane Condition	Acceptable	0.260	0.637
	Bad	-0.263	0.882
	No record	-0.116	0.716
High Curb stone	Yes	0.067	0.653
Potholes	Yes	-0.289	-0.108
Slippery	Yes	-0.354	-0.022
Mixing Coefficients (Class Probability)			
		0.68	0.32

K2: Young, drunk, and careless

In contrast to latent class 1, the second latent class consists of younger people, who are less likely to be wearing a helmet and more likely to be under the influence of alcohol at the time of the accident. On average 32% of the accidents belong to this cluster. In contrast to initial believes regarding the segmentation of this latent class, accidents from this latent class were less likely to result in severe injuries (ASC – Severe Injury < 0) compared to no injuries and less likely to result in slight injuries compared to latent class 1, given the baseline characteristics. However, severe and slight injuries are more likely to have happened during the summer or fall, in some dark or unknown hours, and on roads with good maintenance conditions and high curb stones.

3.5 Comparison

Based on the Log-Likelihood results (Table 3 and 5), it is clear that the socio-demographic model is the superior model. However, this is the joint Log-Likelihood of the injury outcome and the variables used for clustering (equation 9). For more in-depth comparison, we do inference and find the marginal probability of observing a vector of injury outcomes y of all decision-makers N as follows:

$$\prod_{n=1}^{N}\sum_{k=1}^{K} \mathrm{P}(\mathrm{q}_{\mathrm{nk}}|\mathrm{S}_{\mathrm{cn}}, \mathrm{S}_{\mathrm{dn}}, \mu_{\mathrm{ck}}, \Sigma_{\mathrm{ck}}, \mu_{\mathrm{dk}}, \pi_{\mathrm{k}})\mathrm{P}(\mathrm{y}_{\mathrm{n}}|\mathrm{X}_{\mathrm{n}}, \mathrm{q}_{\mathrm{nk}}, \beta_{\mathrm{k}}) \tag{10}$$

Where $P(q_{nk}|S_{cn}, S_{dn}, \mu_{ck}, \Sigma_{ck}, \mu_{dk}, \pi_k)$ is the posterior probability of vectors S_{cn} and S_{dn} being generated by latent class k.

Table 7 presents the marginal Log-Likelihood and the corresponding AIC and BIC of both models. It is clear that the socio-demographic model performs better in terms of prediction accuracy. However, the fully specified model with 6 latent classes seems to be more interpretable and could as such enable policymakers and transportation planners to make better preventive efforts to decrease the number of crashes/injuries and increase the cycling mode share.

Table 7: Comparison

Model	Number of Classes	Number of Parameters	Marginal LL	AIC	BIC
Socio-Demographic	2	59	**-1,581.16**	**3,280.32**	**3,601.87**
Fully Specified	6	179	-1,619.09	3,596.18	4,571.74

To compare the differences in inference associated to the two different models, the average marginal effects of all variables were computed, when all other variables were held constant. The results of this are presented in Table 8 which shows respectively the changes in the probabilities of severe and slight injuries given the changes in the explanatory variables. Results show that there is generally equality in the signs with regard to the percentage point change of the probabilities of severe or slight injury outcomes from the bicycle accidents. Meanwhile, it is observed that the magnitude of the changes in the socio-demographic model tend to be larger, especially with regard to the severe injury outcomes. Lastly, we notice that while there is generally consensus of the sign of the marginal effects of explanatory variables with regard to the probabilities of slight and severe injuries, there are some discrepancies in the two models. Some of the most pronounced are given different light conditions where an increase of 8.3 percentage points in the probability of severe injury are observed given no sunlight in the socio-demographic model, while this is associated to a reduction of 3.46 percentage points in the fully specified. Meanwhile, almost the exact opposite is observed for the accidents, where the light conditions are unknown. This is also extremely evident regarding accidents of straight road sections, not being bicycle lanes, and intersections, as well as given adequate bicycle lane conditions. These discrepancies could be due to the differences in the number and composition of the latent classes (Tables 4 and 6).

Table 8: Comparison of the average marginal effects

Variables		Fully specified model $p.p.$		Socio demographic model $p.p.$	
		Severe	Slight	Severe	Slight
Age	20% reduction	0.71	-0.05	-5.32	0.3
	10% reduction	-0.15	0.07	-2.66	0.14
	10% increase	1.07	-0.24	2.98	-0.14
	20% increase	2.78	-0.59	6.38	-0.32
Gender	Male	0.6	-0.14	0.9	-0.05
Helmet	Used	4.93	-1.13	-3.56	-3.98
Alcohol	Yes	-3.26	0.85	-13.18	-13.5
Season	Winter	-8.1	2.06	-2.82	-3.06
	Spring	-0.81	0.42	-7.3	-7.36
	Autumn	-1.86	0.69	-0.53	-0.26
Road condition	Adequate	-0.59	0.12	1.87	1.51
	Bad	4.76	-0.93	-4.76	-5.62
Light condition	No sunlight	-3.46	0.9	8.33	8.85
	Unknown	9.67	-2.04	-0.06	-0.24
Road Design	Straight	4.72	-2.39	-9.46	-9.77
	Intersection	-6.94	3.09	-20.33	-20.9
	Curved road	-4.35	2.84	-3.34	-3.72
	roundabout	7.26	0.48	-22.15	-23.79
	Other road design	-7.04	3.16	-7.13	-7.54
AADCT	<= 500	10.31	-2.03	5.33	5.67
	501 - 1500	6.59	-1.05	-0.35	-0.49
	1501 - 3000	2.66	-0.29	1.53	1.81
	3001 - 5000	-7.02	1.56	-0.11	-0.2
Bike path condition	Adequate	-8.01	3.34	7.54	7.83
	Bad	22.41	-2.46	4.21	4.89
	Unknown	-3.56	0.53	5.83	6.41
High curb stones	Yes	2.6	-0.6	5.41	6.12
Slippery surface	Yes	-0.05	0.04	-5.77	-5.58
Pothole	Yes	1.75	-0.41	-5.35	-5.17

4 DISCUSSION & CONCLUSION

In light of the dilemma that practitioners could face when choosing between Latent Class Clustering analysis and Latent Class Choice models to investigate injury severity outcomes of bicycle accidents, we proposed a potential approach by fully specifying the class-membership function of an LCCM based on a Gaussian-Bernoulli-Mixture model, while letting the class specific functions associated to each latent class only be determined by alternative specific constants. By doing so the fullest description of accident classes was attained, while attaining them with the injury severity outcome in mind.

Based on an empirical case of single-bicycle accidents in Aarhus, the resulting model arrived at six latent classes as the best trade-off between parsimony and model fit. This evidenced the model's ability to address heterogeneity in the population sample. The groups in themselves also seemed meaningful as well as specific enough to perform population based mitigative efforts, as witnessed in group 5 and 6 of the model, where all accidents occurred in the winter and on straight road sections not being bicycle lanes. These groups make of 17% and 25% of the total average class-probability and addressing this would be incredibly meaningful.

When comparing the fully-specified LCCM to a socio-demographic LCCM, the latter proved to have the superior predictive ability, especially when considering the number of parameters used. Following this, the average marginal effects of changing variables were investigated and compared. The models generally arrived at the same direction of effect related to the different variables, but with sizeable difference in magnitude. Also, there were some disagreements between the two models associating different directions of effect to some variables. Generally, there is no identifying which model is right given limited data. The superiority of the socio-demographic model's predictive ability leads to more trust in the individual effects of this model while the fully specified model offers better representations of heterogeneity within the data and improves the interpretability of the latent classes.

However, given the small absolute difference in the marginal likelihood and the increased information gained from the fully-specified model's latent classes, it is not impossible that the two models would be more similar in the presence of more data.

As such, the fully-specified model would be somewhat superior to the practitioner, as it does allow for the estimation of individual factor effects associated to the probability of different injury severity outcomes while also revealing more minute details about accident groups and their similarities than the socio-demographic LCCM.

5 REFERENCES

Beck, B., Stevenson, M., Newstead, S., Cameron, P., Judson, R., Edwards, E. R., … Gabbe, B. (2016). Bicycling crash characteristics: An in-depth crash investigation study. *Accident Analysis & Prevention*, *96*, 219–227. https://doi.org/10.1016/J.AAP.2016.08.012

Behnood, A., & Mannering, F. (2017). Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variances. *Analytic Methods in Accident Research*, *16*, 35–47. https://doi.org/10.1016/J.AMAR.2017.08.001

Christiansen, H., & Warnecke, M.-L. (2018). *Risiko i trafikken 2007-2016*. Retrieved from https://orbit.dtu.dk/files/146185353/Risiko_i_trafikken_2007_16.pdf

Depaire, B., Wets, G., & Vanhoof, K. (2008). Traffic accident segmentation by means of latent class clustering. *Accident Analysis and Prevention*, *40*(4), 1257–1266. https://doi.org/10.1016/j.aap.2008.01.007

Dozza, M. (2017). Crash risk: How cycling flow can help explain crash data. *Accident Analysis & Prevention*, *105*, 21–29. https://doi.org/10.1016/J.AAP.2016.04.033

Fountas, G., Anastasopoulos, P. C., & Mannering, F. L. (2018). Analysis of vehicle accident-injury severities: A comparison of segment- versus accident-based latent class ordered probit models with class-probability functions. *Analytic Methods in Accident Research*, *18*, 15–32. https://doi.org/10.1016/J.AMAR.2018.03.003

Fournier, N., Christofa, E., & Knodler, M. A. (2017). A mixed methods investigation of bicycle exposure in crash rates. *Accident Analysis & Prevention*. https://doi.org/10.1016/J.AAP.2017.02.004

Kaplan, S., Vavatsoulas, K., & Prato, C. G. (2014). Aggravating and mitigating factors associated with cyclist injury severity in Denmark. *Journal of Safety Research*, *50*, 75–82. https://doi.org/10.1016/J.JSR.2014.03.012

Kim, J.-K., Kim, S., Ulfarsson, G. F., & Porrello, L. A. (2007). Bicyclist injury severities in bicycle–motor vehicle accidents. *Accident Analysis & Prevention*, *39*(2), 238–251. https://doi.org/10.1016/J.AAP.2006.07.002

Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P. D., & Ma, D. T. (2018). Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. *Analytic Methods in Accident Research*, *20*, 1–14. https://doi.org/10.1016/J.AMAR.2018.08.001

Liu, P., & Fan, W. (2020). Exploring injury severity in head-on crashes using latent class clustering analysis and mixed logit model: A case study of North Carolina. *Accident Analysis and Prevention*, *135*, 105388. https://doi.org/10.1016/j.aap.2019.105388

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, *1*, 1–22. https://doi.org/10.1016/J.AMAR.2013.09.001

Møller, M., Janstrup, K. H., & Pilegaard, N. (2018). Bicycle accidents in Denmark – the contribution of cyclist behavior, the vehicle and the road. Retrieved from http://orbit.dtu.dk/en/publications/bicycle-accidents-in-denmark--the-contribution-of-cyclist-behavior-the-vehicle-and-the-road(2f033d56-8fa7-424c-aff4-0fb4c075c50d).html

Prati, G., Pietrantoni, L., & Fraboni, F. (2017). Using data mining techniques to predict the severity of bicycle crashes. *Accident Analysis & Prevention*, *101*, 44–54. https://doi.org/10.1016/J.AAP.2017.01.008

Saha, D., Alluri, P., Gan, A., & Wu, W. (2018). Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. *Accident Analysis & Prevention*, *118*, 166–177. https://doi.org/10.1016/J.AAP.2018.02.014

Sasidharan, L., Wu, K.-F., & Menendez, M. (2015). Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. *Accident Analysis & Prevention*, *85*, 219–228. https://doi.org/10.1016/j.aap.2015.09.020

Sfeir, G., Abou-Zeid, M., Rodrigues, F., Pereira, F., & Kaysi, I. (2020). *Semi-nonparametric Latent Class Choice Model with a Flexible Class Membership Component: A Mixture Model Approach*.