Capturing Spatial Heterogeneity in Cycling Accidents using a Latent Class Discrete Outcome Model

Miguel Costa[1,2], Carlos Lima Azevedo[3], Felix Wilhelm Siebert[3], Manuel Marques[2], and Filipe Moura[1]

[1]Civil Engineering Research and Innovation for Sustainability, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, Lisboa, Portugal
[2]Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, Lisboa, Portugal
[3]Department of Technology, Management and Economics, Technical University of Denmark, Kgs. Lyngby, 2800, Denmark

Short summary

Cities are striving for sustainable transportation, with cycling playing a pivotal role. Despite its health benefits, cyclists face various hazards, leading to accidents and injuries. To enhance cyclist safety, understanding the factors contributing to accidents is crucial. This study introduces a novel modeling framework employing a latent class discrete outcome model, integrating machine learning and econometric approaches. Results reveal the effectiveness of our approach in identifying risk factors based on accident location and their distinctive contributions. Complex relationships between built environments and accident characteristics are unveiled, illustrating variations in the impact of specific factors across environment typologies. These findings emphasize the capability to capture accident heterogeneity and its correlation with the built environment, enabling the targeted design of effective countermeasures and policies for specific risk scenarios.

Keywords: Accident Severity Model, Built Environment Typology, Cycling Safety, Latent Class Discrete Outcome Model.

1 Introduction

Cycling numbers have increased recently in many cities (Pucher & Buehler, 2017). Yet, despite a decrease in the number of fatalities from accidents in Europe in the past decade, fatalities resulting from cycling accidents have increased, representing about 10% of all accident fatalities in the European Union in 2020 (European Commission, 2022).

Researchers have tried to pinpoint and analyze what factors impact accidents to increase cyclists' safety. A particular interest has been placed on the influence of the built environment since it can help planners design current and future safer cycling infrastructures and avoid severe accidents or all accidents, desirably. The impact of intersections, building densities, mixed land use, dedicated cycling lanes, and road hierarchies all affect cycling accidents (Bi et al., 2023; Branion-Calles et al., 2020; P. Chen & Shen, 2016; Hu et al., 2018; Labetski & Chum, 2020; Zahabi et al., 2011).

Yet, conducting such studies is not easy. Data on the built environment is often missing or not captured in accident records (Costa et al., 2022), albeit being essential for such analysis. This data is particularly important to injury severity analysis to quantify danger and analyze how particular factors increase accident severities. More, while classic past research uses discrete outcome models (C. Chen et al., 2017; Kaplan et al., 2014), or spatial models (C. Chen et al., 2017; Osama & Sayed, 2017), newer approaches started to explore machine learning models due to their usual higher predicting power, albeit lower explainable power (for models typically seen as black boxes). Yet, to the best of our knowledge, little research on cycling safety has tried combining machine learning and econometric approaches, exploiting the good predictive power of the former with the conventional explainability of discrete outcome models. More, we postulate that the built environment's holistic nature may impact accident severity that individual components may not independently possess. The approach used here allows for more complex representations to capture heterogeneity in urban scenarios and better estimate interactions between built environment typologies and accidents'

contributing factors.

With this in mind, this paper's main objectives are threefold:

- Apply a new two-part framework to analyze cycling accidents using latent class discrete outcome models (LCDOM), a joint machine learning and econometric methodological tool.
- Use authoritative accident records augmented with volunteered geographic information to understand the impact of built environment typologies and accident contributing factors on cycling accident severity outcomes.
- Understand how risk factors can be directly indexed to distinct built environment typologies.

2 Methodology

This section describes the methodological framework applied to analyze cycling injury severity. We begin by describing the data used in this study. Next, we describe the Gaussian-Bernoulli Mixture Latent Class Discrete Outcome Model (GBM-LCDOM) used to model cycling accident outcomes, first introduced by Sfeir et al. (2021).

Data

This work explores how GBM-LCCM can be used to understand and analyze cycling accidents in Berlin, Germany. We start by filtering the data from the city of Berlin available in the CYCLANDS cycling accident collection (Costa et al., 2022). Accidents include vehicular traffic accidents where personal injury has occurred. Accident outcomes are split between 3 levels depending on the accident's outcome: fatalities, serious injuries, and light injuries. This analysis focuses on cycling accidents from 2018 and 2019, totaling 7516 cycling accidents. However, despite containing information about accident characteristics, no data concerning the built environment is available. Given the importance of understanding the built environment's impacts on cycling accidents, we seek to remedy this by augmenting cycling accident characteristics. Consequently, we add built environment data to each observation using mapping data and street view imagery data.

We use accident locations (geographic coordinates) as a starting point to extract information about the built environment where accidents happen. Since we are interested in the immediate surroundings of cycling accidents, we extract all available information around accident locations from OpenStreetMap (OSM, https://www.openstreetmap.org). OSM data represent physical characteristics (e.g., roads, trees, buildings, or traffic signals).

Finally, we add data extracted from street view imagery (SVI) to each cycling accident observation. Following recent trends using images to evaluate urban features and perceptions (e.g., Song et al. (2020); Ye et al. (2021)), we use SVI from accident locations as another way to add contextual built environment data to our analysis. We use images from Mapillary (https://www.mapillary.com/), which we fetch using accident locations. Images are then processed using semantic segmentation. For this, we use OCRNet (Yuan et al., 2020) and, for each image, extract the presence of different urban element classes (e.g., bike lanes, sidewalks, poles, fences).

Gaussian-Bernoulli Mixture Latent Class Discrete Outcome Model

We use a Gaussian-Bernoulli Mixture Latent Class Discrete Outcome Model (GBM-LCDOM) to model cycling accident outcomes, first introduced by Sfeir et al. (2021). Latent class discrete outcome models (LCDOM) are random utility outcome models that expand the traditional multinomial discrete outcome model by employing the concept of the latent class formulation. LCDOM are divided into two sub-modules: the class membership model, responsible for categorizing observations into classes, and the class-specific component, responsible for explaining outcomes for each class. Using latent classes, LCDOM captures heterogeneity in the severity outcome process by allocating groups of observations to a set of defined classes distinct from each other. Observations are implicitly categorized into a set of K classes, therefore assuming that accidents are modeled by discrete types of built environments, which are latent (unobserved). In turn, these latent typologies are characterized by impacting accident outcomes differently and depending on accident

characteristics. Figure 1 presents an outline of GBM-LCDOM.

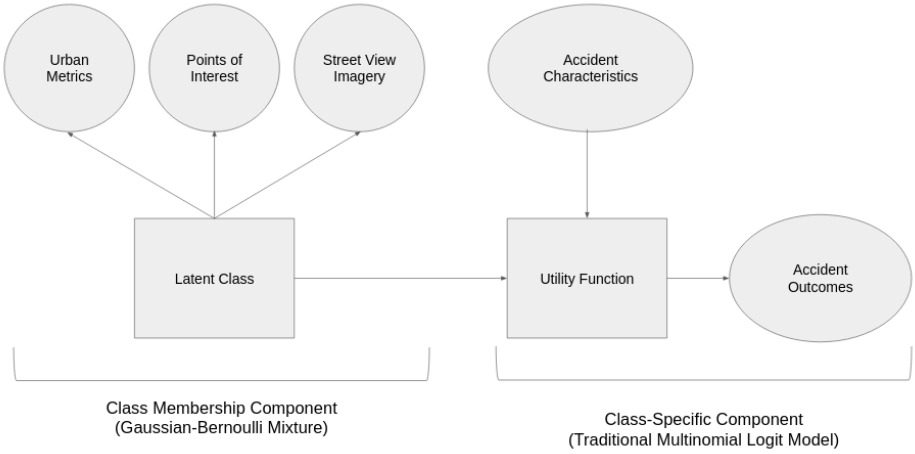

Figure 1: Gaussian-Bernoulli mixture latent class choice model for cycling accident outcome estimation.

The membership class component is a Gaussian-Bernoulli Mixture, a probabilistic machine-learning approach for unsupervised clustering, proving a useful tool to estimate built environment classes for our problem. We assume these typologies are latent classes representing different subgroups of built environments that show similar outward characteristics, which we draw from street view images, mapping points of interest, and urban metrics.

Next, a traditional multinomial logit model specifically modeled each latent class, where utility functions are drawn using accident characteristics. We model the outcome of cycling accidents as a function of observed exogenous attributes using an unordered framework, a multinomial logit. Multinomial logit models are traditional discrete outcome models that consider the accident outcome severity as a linear function of unknown parameters for a set of observed attributes (i.e., accident characteristics or contributing factors).

3 Results and discussion

This section details the results of modeling cycling accidents using GBM-LCDOM. We begin by describing the built environment classes resulting from the class membership component. From the estimation procedure, we retrieve six as the most insightful number of urban classes. These are:

- **Class 1 - Cycle Haven:** Residential area with cycling infrastructure and connections to public transport, bars, cafes, and restaurants. Street furniture is present with fewer amounts of vegetation.
- **Class 2 - Green Transit Hub:** Large intersections with high presence of cycleways, sidewalks, and pedestrian crossings. Not many buildings are present, yet trees, plants, and flowers exist. Public transportation networks for buses, trams, or subways exist, with some on-street parking.
- **Class 3 - Cozy Pedal Haven:** Residential area with some vegetation, bars, cafes, and restaurants, little parking, and no heavy vehicles or primary road infrastructure.
- **Class 4 - Crossroads Transit Hub:** Main road intersections, where heavy vehicles are typically seen, together with buses and trains. There are many traffic signs and walls/fences and lower amounts of vegetation.
- **Class 5 - Industrial Artery Zone:** Commercial and predominant industrial area, highly accessible by main roads, where usually there is little infrastructure for pedestrians and cyclists. There is more vegetation and on-street parking than average and little traffic signage.

- **Class 6 - Drive-centric Commercial Zone:** Car-driven infrastructure with many commercial buildings, focused on signalized primary roads and parking spaces, with fewer opportunities for cycling.

The rationale behind the Class Membership component of LCDOM is a sort of spatial clustering submodule, grouping environments that are closely linked in terms of urban context. However, it can also serve to outline particular locations that are distinct in terms of their attributes and built environments. In practice, using this approach, we can highlight areas that are slightly different from their immediate surroundings. This may indicate that further analysis ought to be performed on these particular observations to understand the impact of that built environment on accidents and hypothesize how urban changes to that particular location may improve safety.

Together with the class membership, we jointly estimate the class-specific component, which maps the impact of different accident characteristics on accident outcomes. A multinomial logit is estimated for each urban environment class found. Table 1 shows the results for each class and some details on the model estimation.

Looking at the estimated parameters, we can notice a few key differences between the different classes. Notably, not all accident characteristics are significant in all urban environments. In fact, due to the nature of LCDOM, we can accurately identify statistically significant characteristics in some environments while remaining redundant in others (e.g., collision with a vehicle ahead). On the other hand, some features (e.g., being dark or involving a heavy vehicle) appear significant (in differing degrees) in almost all classes.

Analyzing the alternative specific constants (ASC) we notice that the average effect of each severity outcome is slightly different for each environment. Namely, Classes 2 and 5 show an ASC closer to the reference value than the remaining classes, indicating that the pre-defined/uncaptured risk is slightly higher in these environments. Unnacounting for the remaining accident contributing factors, this means that for accidents that happen in locations of classes 2 and 5, there is a higher probability of accidents resulting in fatalities.

Another set of takeaways can be drawn by analyzing the different classes. First, an accident where there was a collision with a vehicle in front is only statistically significant for Class 5. This makes sense since Class 5 environments show the highest decrease in the presence of cycleways, making cyclists ride on the road. This points to accidents involving bikes hitting a vehicle in front or being run over from behind due to the lack of dedicated cycling lanes in these locations for cyclists to ride on much more significant.

Overall, this approach brings forth a pivotal difference to traditional discrete outcome modeling as contributing factors can be directly linked to being significant in some cities' areas or whether contributing factors are impactful regardless of where the accident has happened. In practice, such understanding can help planners prioritize and target a specific set of dangerous elements in a group of known locations.

4 Conclusions

This work explores how a Gaussian-Bernoulli Mixture Latent Class Discrete Outcome Model can be used to model and analyze cycling safety. It combines machine learning techniques with traditional discrete outcome modeling, leveraging the former's flexibility and power of dealing with complex interactions and the latter's conventional explainability capabilities. By integrating such an approach, we capture heterogeneity within built environment contexts and can better understand what key accident characteristics are impactful in specific urban environments (and not in others) and vice-versa.

Table 1: Class-specific component: outcome model for each built environment class.

	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**
ASC (SI)	4.89 (15.00)***	4.41 (14.80)***	4.72 (20.18)***	4.89 (14.18)***	4.29 (10.90)***	4.88 (22.46)***
ASC (LI)	6.32 (19.42)***	5.95 (20.07)***	6.48 (27.77)***	6.49 (18.90)***	5.93 (15.11)***	6.14 (28.21)***
Collision w/ car (SI)						
Collision w/ car (LI)						0.47 (2.23)**
Collision w/ heavy vehicle (SI)	-3.60 (-5.37)***	-2.45 (-4.93)***	-2.80 (-7.24)***	-3.37 (-6.11)***	-2.20 (-3.95)***	-2.37 (-6.47)***
Collision w/ heavy vehicle (LI)	-4.14 (-6.55)***	-3.13 (-6.29)***	-3.15 (-8.48)***	-3.95 (-7.63)***	-3.00 (-5.54)***	-2.92 (-8.08)***
Collision w/ other vehicle (SI)	-0.80 (-1.79)*	-0.60 (-1.68)*	-1.02 (-2.68)***	-1.10 (-2.10)**	-0.60	-0.57 (-1.94)**
Collision w/ other vehicle (LI)						
Wet/Slipery road (SI)						
Wet/Slipery road (LI)	0.47 (1.58)*		0.50 (2.16)**			
Darkness (SI)	-0.81 (-2.16)**	-0.47 (-2.09)**	-1.14 (-5.12)***	-0.40 (-1.63)**		-1.03 (-4.79)***
Darkness (LI)	-0.84 (-2.25)**	-0.64 (-2.82)***	-1.24 (-5.60)***	-0.61 (-2.51)***	-0.54 (-1.80)*	-0.83 (-3.89)***
Moving vehicle (SI)	1.11 (2.24)**	0.96 (2.53)***	0.77 (2.49)***	1.76 (4.02)***	1.19 (2.47)**	0.97 (3.06)***
Moving vehicle (LI)				1.08 (2.48)***		0.53 (1.67)*
Turning into road (SI)	0.72 (3.02)***	0.58 (2.93)***	0.29 (1.63)*	0.43 (1.91)**		0.24 (1.53)*
Turning into road (LI)						
Crossing the road (SI)					0.63 (1.36)**	
Crossing the road (LI)						
Stationary vehicle (SI)		0.78 (2.32)**				
Stationary vehicle (LI)		0.64 (1.91)**	0.92 (2.35)***			
Moving in carriageway (SI)	0.96 (2.26)**		0.61 (2.08)**		1.12 (2.00)**	
Moving in carriageway (LI)	1.03 (2.45)***		0.56 (1.90)**	0.72 (1.63)*	1.09 (1.97)**	
Collision w/ vehicle ahead (LI)					-1.89 (-4.37)***	
Collision w/ vehicle ahead (SI)					-2.46 (-3.62)***	
# of Observations	7516					
AIC	15181.22					
BIC	45713					
Joint Final Log-Likelihood	-219505.1					
Joint-ρ^2	0.33					

Notes: Fatal outcome was set as the reference category for all variables. SI: Serious Injury, LI: Light Injury.
***, **, *: Significance at 1%, 5%, 15% level. Parameters beyond a 15% significance level were omitted.

Acknowledgements

This work is part of the research activity partially funded by Fundação para a Ciência (FCT) e Tecnologia via grant [PD/BD/142948/2018] that was partially carried out at the Civil Engineering Research and Innovation for Sustainability (CERIS) funded by FCT in the framework of project [UIDB/04625/2020], the Associate Laboratory of Robotics and Engineering Systems (LARSyS) funded by FCT in the framework of project [UIDB/50009/2020], and the Department of Technology, Management, and Economics at the Technical University of Denmark (DTU). We also thank Georges Sfeir (DTU) for distributing the code to estimate the GBM-LCDOM.

References

Bi, H., Li, A., Zhu, H., & Ye, Z. (2023). Bicycle safety outside the crosswalks: Investigating cyclists' risky street-crossing behavior and its relationship with built environment. *Journal of Transport Geography*, *108*, 103551.

Branion-Calles, M., Götschi, T., Nelson, T., Anaya-Boig, E., Avila-Palencia, I., Castro, A., ... Winters, M. (2020). Cyclist crash rates and risk factors in a prospective cohort in seven european cities. *Accident Analysis & Prevention*, *141*, 105540. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0001457519314320` doi: https://doi.org/10.1016/j.aap.2020.105540

Chen, C., Anderson, J. C., Wang, H., Wang, Y., Vogt, R., & Hernandez, S. (2017). How bicycle level of traffic stress correlate with reported cyclist accidents injury severities: A geospatial and mixed logit analysis. *Accident Analysis & Prevention*, *108*, 234-244. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0001457517303160` doi: https://doi.org/10.1016/j.aap.2017.09.001

Chen, P., & Shen, Q. (2016). Built environment effects on cyclist injury severity in automobile-involved bicycle crashes. *Accident Analysis & Prevention*, *86*, 239-246. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0001457515301184` doi: https://doi.org/10.1016/j.aap.2015.11.002

Costa, M., Marques, M., Roque, C., & Moura, F. (2022). Cyclands: Cycling geo-located accidents, their details and severities. *Scientific data*, *9*(1), 1–9.

European Commission. (2022). *Annual statistical report on road safety in the eu, 2021*. European Road Safety Observatory. Brussels, European Commission, Directorate General for Transport.

Hu, Y., Zhang, Y., & Shelton, K. S. (2018). Where are the dangerous intersections for pedestrians and cyclists: A colocation-based approach. *Transportation Research Part C: Emerging Technologies*, *95*, 431-441. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0968090X18307526` doi: https://doi.org/10.1016/j.trc.2018.07.030

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. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0022437514000437` doi: https://doi.org/10.1016/j.jsr.2014.03.012

Labetski, A., & Chum, A. (2020). Built environmental correlates of cycling accidents involving fatalities and serious injuries in london, uk. *Frontiers in Sustainable Cities*, *2*. Retrieved from `https://www.frontiersin.org/articles/10.3389/frsc.2020.599635` doi: 10.3389/frsc.2020.599635

Osama, A., & Sayed, T. (2017). Evaluating the impact of socioeconomics, land use, built environment, and road facility on cyclist safety. *Transportation Research Record*, *2659*(1), 33-42. Retrieved from `https://doi.org/10.3141/2659-04` doi: 10.3141/2659-04

Pucher, J., & Buehler, R. (2017). Cycling towards a more sustainable transport future. *Transport Reviews*, *37*. doi: 10.1080/01441647.2017.1340234

Sfeir, G., Abou-Zeid, M., Rodrigues, F., Pereira, F. C., & Kaysi, I. (2021). Latent class choice model with a flexible class membership component: A mixture model approach. *Journal of Choice Modelling*, *41*, 100320. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1755534521000531` doi: https://doi.org/10.1016/j.jocm.2021.100320

Song, X. P., Richards, D. R., & Tan, P. Y. (2020). Using social media user attributes to understand human–environment interactions at urban parks. *Scientific reports*, *10*(1), 1–11.

Ye, C., Zhang, F., Mu, L., Gao, Y., & Liu, Y. (2021). Urban function recognition by integrating social media and street-level imagery. *Environment and Planning B: Urban Analytics and City Science*, *48*(6), 1430-1444. Retrieved from `https://doi.org/10.1177/2399808320935467` doi: 10.1177/2399808320935467

Yuan, Y., Chen, X., & Wang, J. (2020). Object-contextual representations for semantic segmentation.

Zahabi, S. A. H., Strauss, J., Manaugh, K., & Miranda-Moreno, L. F. (2011). Estimating potential effect of speed limits, built environment, and other factors on severity of pedestrian and cyclist injuries in crashes. *Transportation Research Record*, *2247*(1), 81-90. Retrieved from `https://doi.org/10.3141/2247-10` doi: 10.3141/2247-10