Manuscript Details

Manuscript number	AAP_2020_307_R1
Title	An empirical assessment of strategies to model party interactions in crash severity analysis
Article type	Full Length Article

Abstract

Road crashes impose an important burden on health and the economy. Numerous efforts have been undertaken to understand the factors that affect road collisions in general, and the severity of crashes in particular. In this literature several strategies have been proposed to model interactions between parties in a crash, including the use of variables regarding the other party (or parties) in the collision, data subsetting, and estimating models with hierarchical components. Since no systematic assessment has been conducted of the performance of these strategies, they appear to be used in an ad-hoc fashion in the literature. The objective of this paper is to empirically evaluate ways to model party interactions in the context of crashes involving two parties. To this end, a series of models are estimated using data from Canada's National Collision Database. Three levels of crash severity (no injury/injury/fatality) are analyzed using ordered probit models and covariates for the parties in the crash and the conditions of the crash. The models are assessed using predicted shares and classes of outcomes, and the results highlight the importance of considering opponent effects in crash severity analysis. The study also suggests that hierarchical (i.e., multi-level) specifications and subsetting do not necessarily perform better than a relatively simple single-level model with opponent-related factors. The results of this study provide insights regarding the performance of different modelling strategies, and should be informative to researchers in the field of crash severity.

Keywords	Crash severity; opponent effects; ordered logit; model performance; Canada's National Crash Database; reproducible research
Taxonomy	Statistical Method, Motor Vehicle Accident, Risk Assessment

Corresponding Author	Antonio Paez
Corresponding Author's Institution	McMaster University
Order of Authors	Antonio Paez, Hany Hassan, Mark Ferguson, Saiedeh Razavi

Submission Files Included in this PDF

File Name [File Type]

Cover-Letter-v2.pdf [Cover Letter]

Response-to-Reviewers-v2.pdf [Response to Reviewers]

Highlights-v2.pdf [Highlights]

Interactions-Crash-Severity-v2.pdf [Manuscript File]

Author-Statement-v2.pdf [Author Statement]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Research Data Related to this Submission

Data set	https://github.com/paezha/Modelling-Participant-Interactions-in-Crash-Severity

Modelling Participant Interactions in Crash Severity

Data and code used in this research.

Cover Letter

5/8/2020

Dear Editor:

We would like to express our gratitude to four anonymous reviewers for their generous reviews of our paper. Their comments have helped us to improve the paper. In this letter we detail our responses to the feedback received, and actions taken in this revision.

We trust that you and the expert reviewers will find that our responses address in a thorough and conscientious manner the points raised with respect to the previous version of the paper.

Looking forward to hearing back from you in due course, we remain sincerely yours.

The Authors

Response to reviewers

5/8/2020

Reviewer 1

The paper investigates three general modelling strategies to analyze crash data. Namely, use of opponent related factors, single and multi-level model approach, and different sampling approach.

Thank you for reading the paper and for your feedback to improve the paper.

The combined application of those strategies is not completely new because it is part of a general methodological approach (and usually crash data contain information about opponent factors). The paper, in the present form, seems more an exercise, a good exercise, but not so much sufficient as to justify a paper. Why not to use other models besides ordered probit (row 525) (or ordered logit as in the abstract)?

Thank you for this perceptive comment. Indeed, the combined application of these strategies is not completely new, or new at all, since every single one of them has been used in the past by someone or other. Our argument is not that the strategies are new, but rather that they are used in a haphazard fashion (we document the different types of applications in the paper). Moreover, we argue that there is no systematic examination of how each of these strategies perform, and therefore little guidance in the literature in terms of which one to choose in applications. We believe that our paper provides new information that is not available elsewhere in the literature. With respect to the use of other models, this has been investigated by several others, who have compared multinomial to ordered models, ordered models to ordered models with random components, etc. We would like to draw your attention to this paragraph in the paper (p. 26, lines 497-510):

> *"We do not report results regarding other modelling strategies. On the one hand, more sophisticated modelling frameworks are generally capable of improving the performance of a model. On the other hand, there are well-known challenges in the estimation of more sophisticated models (see Lenguerrand et al., 2006, p. 47, for a discussion of convergence issues in models with mixed effects; Mannering et al., 2016, p. 13, for some considerations regarding the complexity and cost of estimating more complex models; and Bogue et al., 2017, p. 27, on computational demands of models with random components). The additional cost and complexity of more sophisticated modelling approaches would, in our view, have greatly complicated our empirical assessment, particularly considering the large size of the sample involved in this research (a data set with over 164,000 records in the case of the full sample models). That said, we experimented with a model with random components using monthly subsets of data to find that, indeed, estimation takes considerably longer, is more demanding in terms of fixing potential estimation quirks, and in the end resulted in variance components that could not be reliably estimated as different from zero (results can be consulted in the source R Notebook). For this reason, we choose to leave the application of more sophisticated models as a matter for future research."*

We did *try other models, but there is no squeezing out unobserved heterogeneity where none seems to exist.*

To support this argument, tables 9 and 10 (confusion matrices) show a very poor performance of the models, especially for fatality; Models Ensemble improve it a little.

If you compare the average prediction errors (APEs) and weighted average prediction errors

(WAPEs) of our models to those reported by Bogue, Paleti, and Balam (2017), you will notice that our models actually perform well according to these criteria. However, part of our argument is that a more complete comparison of the performance of models requires the use of other performance criteria, hence all the verification statistics.

It is curious that its 'percent correct by class' is generally very high (close to or greater than 90%) when numbers are so bad; it is curious too that the 'percent correct' is always lower than the lowest value of each 'percent correct by class'.

It is not that curious, really. Percent correct by class is heavily influenced by the dominant class (see note in Table 8). It is calculated as the sum of total hits and correct non-events divided by the number of cases. Accordingly (see for example Table 9, Model 1):

Outcome	Pred. No Injury	Pred. Injury	Pred. Fatality
No Injury	50652	22503	150
Injury	28232	62121	797
Fatality	2	51	3

The percent correct by class (No Injury) is calculated as follows:

$$PC_{No-Injury} = \frac{50652 + 62121 + 797 + 51 + 3}{164511} = 0.6907 = 69.07\%$$

In the above, 50652 *is the number of total hits, and* $62121 + 797 + 51 + 3$ *is the number of correct non-events. Compare to the percent correct by class (Fatality):*

$$PC_{Fatality} = \frac{3 + 50652 + 22503 + 28232 + 62121}{164511} = 0.9939 = 99.39\%$$

In the above, 3 *is the number of total hits, and* $50652 + 22503 + 28232 + 62121$ *is the number of correct non-events. It is not surprising that the value is so high: in fact, the model correctly predicts non-fatalities most of the time! However, when we look at bias (B) we see that fatalities are severely underpredicted. This is despite APE and WAPE statistics that are similar or better than those reported in the literature. Our argument is that the use of verification statistics provides a more nuanced view of the performance of the models. Again, to the best of our knowledge, this more nuanced perspective is not currently available, and that it provides new insights into the performance of models.*

The availability of crash data at a repository on the web to reproduce the research or to make other seems very interesting (not only for exercise) and should be encouraged.

Thank you. We also think that this should be encouraged. We did not find in our literature review a single paper on crash severity that was reproducible.

Title is a little misleading since the role of 'opponent' variables is not really the main focus of the paper.

We have changed the title based on this and a related comment by Reviewer 2.

Minor comments

Reorganize section 5.3.1 by listing at the beginning the considered measures for verification statistics. Furthermore, all the section is difficult to read.

We followed this recommendation in the revision and reorganized the section to describe the verification statistics at the beginning. We also edited it, hopefully improving its readability.

At row 271 in which unit is 4536?

It should be kg. We added the units.

At row 302, “in two different” what?

In two different ways. This was corrected.

At row 435, CSI requires a note.

This paragraph was deleted as part of reorganizing the section.

At row 533, there is an s too many in ‘s(e.g. . . .)’

Typo corrected.

Acknowledgment: It is quite strange that authors did not add now their acknowledgments but they reserve to do it in the ‘final’ version.

This was an oversight caused by the practice of anonymizing papers. The revision includes the acknowledgments.

Thanks again for your feedback and the time invested in helping us improve this paper. We hope that with these revisions the contributions of the paper will be more easily appreciated.

Reviewer 2

An empirical assessment of strategies to model opponent effects in crash severity analysis – the paper examines the injury severity of crashes employing ordered model.

"Accident" should be avoided. Crash/collision should be used.

> *Thank you for this comment. We thought that since the name of the journal is Accident Analysis and Prevention the term "accident" was appropriate to refer to stochastic events with undesirable outcomes. We have changed most "accidents" to "collisions" and "crashes" to reflect the nuanced differences between them.*

"ways to deal with the way the different parties in a crash" – this sentence is confusing. Does it refer to methodological ways to accommodate for such correlation, if present? Please clarify.

> *The sentence was wordy and unclear, and it was edited for clarity. Now it reads:*
>
> > *Numerous efforts have been undertaken to understand the factors that affect road collisions in general, and the severity of crashes in particular. In this literature several strategies have been proposed to model interactions between parties in a crash, including the use of variables regarding the other party (or parties) in the collision, data subsetting, and estimating models with hierarchical components.*

"opponent effects" – this is misleading. What effects are you referring to and does opponent refer to 2 vehicles involved in a crash? After reading the paper it was possible to infer it, but, these terms should be clearly defined to avoid confusions and to improve readability.

> *"Opponent effects" are the effects on outcome severity of the attributes of the opposite party in the accident. For clarity, we now define these terms earlier in the paper.*

"participants in the crash"- how can someone participate in a crash?

> *We changed "participants" to "parties" and "individuals" to distinguish between groups of people who were party to a crash, and the individuals who composed those parties.*

Please clearly state what are the objectives of the study and how it is contributing to existing road safety research. It is not at all clear from introduction/background section. It was not even clear whether authors are focusing on motor vehicle crashes only.

> *The introduction was rewritten to remove some background information that was not strictly necessary, and to more clearly state the objectives and contributions of the paper.*

The four strategies identified in the methodology section; how different approaches contribute towards the empirical context? Please provide a detailed discussion first.

> *We identify three basic strategies, which can be combined. This is explained in sections 3.2 to 3.4.*

Line 125: the second part collect – what does collect represent here?

> *To bring together, as in, several items that are collected in the second summation in the formula. We changed this to "gather".*

Page 2, line 29 – "shamsunnahar and Eluru" should be "Yasmin and Eluru".

> *Thank you for this clarification. This has been corrected.*

Line 224: Were pedestrian/bicycle involved crashes also considered in the severity analysis along with motor vehicle crashes? The mechanisms of injury severity for these two groups are different. How these could be modeled in one model?

> *Yes, they were, unless they were modelled separately. For example, Models 1 and 2 were re-estimated after subsetting the sample, so that only pedestrians were considered, or bicyclists, or light vehicles, etc. How could several be modelled in one model? By using the strategies described in the methods section.*

4.1.2 and 4.1.3: these procedures are standard in preparing crash severity data. Not sure why these are discussed in detail. For the data cleaning, the fundamental question is, other than deleting the cases with unknown information, how the statistical property of the dependent variable was maintained?

> *Data preprocessing is not as sexy as a new ranked ordered logit model with mixtures of copulas, etc. In fact, data preprocessing is seldom report in detail: everybody assumes that it is standard. As we were working on this research, what we found is that data preprocessing is as complex as data analysis and modelling, and we could not find good, complete discussions of this topic. Since data preprocessing is an essential component of two of the strategies discussed in the paper (data subsetting and opponent-effects), we would rather maintain this as part of the paper, unless the editor indicates otherwise.*

Line 262: virtually all crashes were fatal when the opponent is a driver – this statement is fundamentally wrong.

> *That is not what we said. This is the actual statement:*
>
> > *"In terms of outcomes, we observe that virtually all fatalities occur when the opponent is a driver, and only very rarely when the opponent is a motorcyclist."*
>
> *As seen in Table 2, where we split the outcomes by the roles of the opposite parties, the only fatalities recorded outside of Opponent: Driver where for Opponent: Motorcycle (7 fatalities). In other words, virtually all fatalities occur when the opponent is a driver. This is very different from "people virtually always die when the opponent is a driver", which seems to be your odd reading of our statement.*

The data sample is high. How is it ensured that the model is not over fitted with high number of records? Number of co-efficient in the models are an indication of over fitting.

> *No. The data sample is not "high", it is large. In the present case, we have a sample of size 164,511, and the models have at most 111 coefficients: this leaves tens of thousands of degrees of freedom. In fact, thinking that the number of coefficients is an indication of overfitting is overly-simplistic, since overfitting is a combination of sample size and the number of parameters. Evidence of overfitting only can be gleaned from evaluating the performance of the model when "new" information is used. That is the point of backcasting: to evaluate how the models behave in an out-of-sample situation.*

The study design is quite confusing. Isn't it obvious that Model 3 and Model 4 are generalized version of the Model 1 and 2. More importantly, Model 3 and 4 can be easily combined to estimate a rather genralized version of severity model. What is the point of this modelling exercise? What is new in this exercise and why one should consider it?

> *It might be obvious to you that Models 3 and 4 can be easily combined, but we regret to inform you that there would be perfect multicollinearity if you tried to do so. The point is that even an expert such as yourself may fail to see this, and we consider that presenting a framework that explains the various strategies and compares them might be worthwhile for readers who eventually read the paper.*

The paper is merely a modelling exercise without any significant contribution to the state-of-the-art of severity model. The wordings of the paper are significantly misleading and confusing. If we are considering number of occupants in different vehicle units involved in a crash, the outcome is likely to be correlated. How such correlations were accommodated? What is the purpose of this modelling exercise rather than just estimating a set of models?

We regret that you were not persuaded by the paper. We have taken into account those suggestions in your report that make sense, and hope that you will be persuaded to re-read the paper with attention to maybe see its value.

Reviewer 3

The reviewer found the paper to be exceptionally well researched. It is quite long and so full of details, some quite nuanced, that reading the paper and digesting its many remarks and information is time consuming. However, the authors are praised on giving such detail, and especially on opening the code and data for others. It makes this paper highly useful as a base for further research on this important topic. The reviewer points out that Mannering has co-authored earlier work than Wang and Kockelman, which used direct opponent effects and classified models based on opponent types, in 2004 in AAP, although that particular paper did not focus on the opponent effects.

> *Thank you for your positive assessment of our paper. We realize that it is information-rich, and appreciate the time and effort you put into this review.*

One small item to improve would be to name the four models with descriptive titles rather than only model 1, model 2, model 3, and model 4, which requires readers to remember which is which and makes the tables unreadable for readers browsing the paper.

> *Thank you for this suggestion. We have labelled the models in the tables, which we think makes it easier to recall the main characteristics of each model, i.e., single-level, hierarchical, etc.*
>
> *Again, thank you for your thoughtful comments and suggestions.*

Reviewer 4

This study aims to evaluate approaches to model opponent effects in crashes involving two parties. A series of models are established and evaluated. The paper is well written and organized. Besides, authors make the data and code publicly available, which is great. Here are my comments:

Many thanks for your positive assessment of the paper.

1. The data preprocessing section could be simplified to make the paper more concise.

   Thank you for this comment. We collapsed three subsections of data preprocessing into one, and edited it to make it somewhat more concise. We are convinced that data preprocessing is a fundamental part of the process to implement the strategy of opponent-related variables. The process, as the paper probably shows, is not necessarily straightforward. And although not as sexy as presenting a new model with yet more unobserved heterogeneity terms, is as important. Our preference is to keep this aspect of the process well documented and transparent, but if the editor insists, we could reduce it even more.

2. Ln 159, for any given coefficient q, you can have equation (7). I don't quite understand it. Please clarify.

   Thanks for catching this. The sentence should say "for any given coefficient m", since the nesting is of individual l in traffic unit m. This was corrected.

3. Table 3 is mentioned in section 4 but shown in section 5. It's better to adjust the location of Table 3.

   Done.

4. Ln 376-377, as mentioned by authors, models in Table 6 are clearly overfitted to the 2017 dataset. They don't generalize well to new data. To this end, the analysis based on these models might be biased.

   It is true, and there is the perennial trade-off between bias and variance. That said, overfitting is not necessarily a bad thing, and an analyst might prefer somewhat biased predictions in exchange for lower variance. Here we do not take a strong stand on this issue because it really depends on the specific use of the models.

5. As shown in Table 8, authors evaluated the outcomes of models by lots of different indicators. Normally, it's difficult to compare model performances based on so many indicators. How did you manage that? BTW, what's the difference between Percent Correct by Class and Probability of Detection?

   We believe that using a fuller battery of indicators is a bit more challenging (it makes it somewhat more difficult to find the absolute best *model), but on the other hand it does give numerous new insights into how specific models perform well or not. As an example, Model 1 produces individual-level predictions that are less biased, but also has a higher Probability of False Detection. Does the analyst care more about reducing bias, or about reducing false detection? That might guide the choice of a model. The difference between the two verification statistics is that Percent Correct by Class is calculated with the hits and correct non-events, whereas the Probability of Detection is calculated using only the hits.*

   Thank you for your suggestions. We hope that you will find the revised version of the paper much improved.

Highlights

5/8/2020

- Strategies to model interactions between parties in crash severity modelling are investigated
- A data workflow is presented for preparing data for modelling party interactions
- Single-level and hierarchical models are compared
- Full-sample and sub-sample models are compared
- The results provide information useful to select a modelling strategy for crash severity

An empirical assessment of strategies to model party interactions in crash severity analysis

Antonio Paez[a], Hany Hassan[b], Mark Ferguson[a], Saiedeh Razavi[a]

[a] *McMaster Institute for Transportation and Logistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4K1*

[b] *Department of Civil and Environmental Engineering, Louisiana State University, Baton Rouge, Louisiana, USA 70803*

Abstract

Road crashes impose an important burden on health and the economy. Numerous efforts have been undertaken to understand the factors that affect road collisions in general, and the severity of crashes in particular. In this literature several strategies have been proposed to model interactions between parties in a crash, including the use of variables regarding the other party (or parties) in the collision, data subsetting, and estimating models with hierarchical components. Since no systematic assessment has been conducted of the performance of these strategies, they appear to be used in an ad-hoc fashion in the literature. The objective of this paper is to empirically evaluate ways to model party interactions in the context of crashes involving two parties. To this end, a series of models are estimated using data from Canada's National Collision Database. Three levels of crash severity (no injury/injury/fatality) are analyzed using ordered probit models and covariates for the parties in the crash and the conditions of the crash. The models are assessed using predicted shares and classes of outcomes, and the results highlight the importance of considering opponent effects in crash severity analysis. The study also suggests that hierarchical (i.e., multi-level) specifications and subsetting do not necessarily perform better than a relatively simple single-level model with opponent-related factors. The results of this study provide insights regarding the performance of different modelling strategies, and should be informative to researchers in the field of crash severity.

1. Introduction

Modelling the severity of injuries to victims of road crashes has been a preoccupation of transportation researchers, planners, auto insurance companies, governments, and the general public for decades. One of the earliest studies to investigate the severity of injuries conditional on a collision having occurred was by White and Clayton (1972). Kim et al. (1995) later stated that the "linkages between severity of injury and driver characteristics and behaviors have not been thoroughly investigated" (p. 470). Nowadays, there is a burgeoning literature on this subject, which often relies on multivariate analysis of crash severity to clarify the way various factors can affect the outcome of an incident, and to discriminate between various levels of injury.

Email addresses: paezha@mcmaster.ca (Antonio Paez), hassan1@lsu.edu (Hany Hassan), fergumr@mcmaster.ca (Mark Ferguson), razavi@mcmaster.ca (Saiedeh Razavi)

Crash severity is an active area of research, and one where methodological developments have aimed at improving the reliability, accuracy, and precision of models (e.g., Savolainen et al., 2011; Bogue et al., 2017; Yasmin and Eluru, 2013). Of interest in this literature is how different parties in a crash interact to influence the severity of individual outcomes. The importance of these interactions has been recognized by numerous authors (e.g., Chiou et al., 2013; Lee and Li, 2014; Li et al., 2017; Torrao et al., 2014). Lee and Li (2014) for instance, note that "[for] two-vehicle crashes, most studies only considered the effects of one vehicle on driver's injury severity or the highest injury severity of two drivers. However, it is expected that driver's injury severity is not only affected by characteristics of his/her own vehicle, but also characteristics of a partner vehicle." More generally, the severity of the outcome depends, at least in part, on the characteristics of the parties, and a crash between two heavy vehicles is likely to have very different outcomes compared to crash where a heavy vehicle hits a pedestrian.

For the purpose of this paper, we define a party as one or more individuals travelling in a traffic unit that becomes involved in a crash. Sometimes the traffic unit is a vehicle, and the party is a single individual (i.e. a single occupant vehicle); but in other cases, a party could consist of several individuals (i.e., a driver and one or more passengers). Other times, the individual *is* the traffic unit, for instance a pedestrian or a bicyclist. An opponent is a different party that is involved in the same collision. In the literature, interactions between parties in a collision are treated by means of different strategies, including the use of *data subsetting*, *opponent variables*, and estimating models with *hierarchical components*. These strategies are not new, however, a systematic comparison between them is missing from the literature. For this reason, the objective of this paper is to empirically evaluate different strategies to model party interactions in crash severity in the context of incidents involving two parties. More concretely, this research aims to:

1. Systematize the analysis of party interactions in crash severity models. Although every strategy considered here has been used in past research, in this paper they are organized in a way that clarifies their operation.
2. Present a data management workflow to prepare a dataset to implement analysis of opponent effects.
3. Provide evidence of the performance of different modelling strategies. In particular, the importance of considering opponent-level effects and the suitability of single-level models.
4. Present an example of reproducible research in crash severity analysis: all data and code are publicly available from the beginning of the peer-review process[1].

For the assessment we use data from Canada's National Collision Database, a database that collects all

[1]The source file for this paper is an 'R' Markdown document; all code and data necessary to reproduce the analysis are available from the following GitHub repository: https://github.com/paezha/Modelling-Participant-Interactions-in-Crash-Severity

police-reported collisions in the country. Using the most recent version of the data set (2017). Three levels of crash severity (no injury/injury/fatality) are analyzed using ordered logit models, and covariates for the parties in the crash and the conditions of the crash. For model assessment, we conduct an in-sample prediction exercise using the estimation sample (i.e., *nowcasting*), and also an out-of-sample prediction exercise using the data set corresponding to 2016 (i.e., *backcasting*). The models are assessed using predicted shares and predicted classes of outcomes at the individual level, using an extensive array of verification statistics.

The rest of this paper is structured as follows. In Section 2 we discuss some background matters, and follow this with a concise review of the modelling strategies used to analyze crash severity 3. Section 4 describes the data requirements, data preprocessing, and the modelling strategies, along with the results of model estimation. The results of assessing the models and the discussion of these results is found in Section 5. We then present some additional thoughts about the applicability of this approach in Section 6 before offering some concluding remarks in Section 7.

2. Background

Crash severity is often modeled using models for discrete outcomes. Analysts interested in crash severity have at their disposal an ample repertoire of models to choose from, including classification techniques from machine learning (e.g., Iranitalab and Khattak, 2017; Chang and Wang, 2006; Effati et al., 2015; Khan et al., 2015), Poisson models for counts (e.g., Ma et al., 2008), unordered logit/probit models (e.g., Tay et al., 2011), as well as ordered logit/probit models (e.g., Rifaat and Chin, 2007), with numerous variants, such as random parameters/mixed logit (e.g., Aziz et al., 2013; Haleem and Gan, 2013), partial proportional odds models (e.g., Mooradian et al., 2013; Sasidharan and Menendez, 2014), and the use of copulas (e.g., Wang et al., 2015). Recent reviews of methods include Savolainen et al. (2011), Yasmin and Eluru (2013), and Mannering et al. (2016).

Irrespective of the modelling framework employed, models of crash severity often include variables in several categories, as shown with examples in Table 1 (also see Montella et al., 2013). Many crash databases and analyses also account for the multi-event nature of many crashes. Individuals in the crash may have had different roles depending on their situation, with some acting as operators of a vehicle (i.e., drivers, bicyclists), while others were passengers. They also may differ depending on what type of traffic unit they were, for example pedestrians, or operators/passengers of a vehicle. The multiplicity of roles makes for complicated modelling decisions when trying to understand the severity of injuries; for example, what is the unit of analysis, the person, the traffic unit, or the collision? Not surprisingly, it is possible to find examples of studies that adopt different perspectives. A common simplifying strategy in model specification is to consider

only *drivers* and/or only *single-vehicle* crashes (e.g., Kim et al., 2013; Gong and Fan, 2017; Lee and Li, 2014; Osman et al., 2018). This strategy reduces the dimensions of the event, and it becomes possible, for example, to equate the traffic unit to the person for modelling purposes.

The situation becomes more complex when dealing with events that involve two traffic units (e.g., Torrao et al., 2014; Wang et al., 2015) and multi-traffic unit crashes (e.g., Wu et al., 2014; Bogue et al., 2017). Different strategies have been developed to study these, more complex events. A number of studies advocate the estimation of separate models for different parties, individuals, and/or situations. In this way, Wang and Kockelman (2005) estimated models for single-vehicle and two-vehicle crashes, while Savolainen and Mannering (2007) estimated models for single-vehicle and multi-vehicle crashes. More recently, Duddu et al. (2018) and Penmetsa et al. (2017) presented research that estimated separate models for at-fault and not-at-fault drivers. The strategy of estimating separate models relies on *subsetting* the data set, although it is possible to link the relevant models more tightly by means of a common covariance structure, as is the case of bivariate models (e.g., Chiou et al., 2013; Chen et al., 2019) or models with copulas (e.g., Rana et al., 2010; Shamsunnahar et al., 2014; Wang et al., 2015).

A related strategy to specify crash severity models is to organize the data in such a way that it is possible to introduce *opponent effects*. There are numerous examples of studies that consider at least some characteristics of the opposite party (or parties) in two- or multi-vehicle crashes. For example, Wang and Kockelman (2005) considered the type of the opposite vehicle in their model for two-vehicle collisions. Similarly, Torrao et al. (2014) included in their analysis the age, wheelbase, weight, and engine size of the opposite vehicle, while Bogue et al. (2017) used the body type of the opposite vehicle. Penmetsa et al. (2017) and Duddu et al. (2018) are two of the most comprehensive examples of using opponent's information, as they included attributes of individuals in the opposite party (their physical condition, sex, and age), as well as characteristics of the other party's traffic unit (the vehicle type of the opponent). The twin strategies of subsetting the sample and using the attributes of the opponent are not mutually exclusive, but neither are they used consistently together, as a scan of the literature reveals.

Another strategy is to introduce *hierarchical components* in the model, a technique widely used in the hierarchical or multi-level modelling literature. This involves considering observations as being nested at different levels: individuals nested in traffic units, which in turn are nested in accidents, as an example.

In this paper we consider three general modelling strategies, as follows:

- Strategy 1. Introducing opponent-related factors
- Strategy 2. Single-level model and multi-level (hierarchical) model specifications
- Strategy 3. Full sample and sample subsetting

Table 1: Categories of variables used in the analysis of crash severity with examples

Category	Examples
Human-factors	Attributes of individuals in the crash, e.g., injury status, age, gender, licensing status, professional driver status
Traffic unit-factors	Attributes of the traffic unit, e.g., type of traffic unit (pedestrian, car, motorcycle, etc.), maneuver, etc.
Environmental-factors	Attributes of the crash, e.g., location, weather conditions, light conditions, number of parties, etc.
Road-factors	Attributes of the road, e.g., surface condition, grade, geometry, etc.
Opponent-related factors	Attributes of the opponent, e.g., age of opponent, gender of opponent, opponent vehicle type, etc.

These are discussed in more detail in the following section.

3. Methods

3.1. Choice of model

Before describing the modelling strategies, it is important to explain our choice of model. There have already been comparisons between different models. Yasmin and Eluru (2013), for instance, conducted an extensive comparison of models for discrete outcomes in crash severity analysis, and found only small differences in the performance of unordered models and ordered models; however, ordered models are usually more parsimonious since only one latent functions needs to be estimated for all outcomes, as opposed to one for each outcome in unordered models. Bogue et al. (2017) also compared unordered and ordered models in the form of the mixed multinomial logit and a modified rank ordered logit, respectively, and found that the ordered model performed best. To keep the empirical assessment manageable, in this paper we will consider only the ordinal logit model, and will comment on potential extensions in Section 6.

The ordinal model is a latent-variable approach, whereby the severity of the crash (observed) is linked to an underlying latent variable that is a function of the variables of interest, as follows:

$$y^*_{itk} = \sum_{l=1}^{L} \alpha_l p_{itkl} + \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \epsilon_{itk} \tag{1}$$

The left-hand side of the expression above (y^*_{itk}) is a latent (unobservable) variable that is associated with the severity of crash k ($k = 1, \cdots, K$) for individual i in traffic unit t. The right-hand side of the expression is split in four parts. The first part gathers $l = 1, \cdots, L$ human-factors p for individual i in traffic unit t and crash k; these could relate to the person (e.g., age, gender, and road user class). The second part gathers $m = 1, \cdots, M$ attributes u related to traffic unit t in crash k; these could be items such as maneuver or vehicle type. The third part gathers $q = 1, \cdots, Q$ attributes c related to the crash k, including

environmental-factors and road-factors, such as weather conditions, road alignment, and type of surface. Lastly, the fourth element is a random term specific to individual i in traffic unit t and crash k. The function consists of a total of $Z = L + M + Q$ covariates and associated parameters.

For conciseness, in what follows we will abbreviate the function as follows:

$$y^*_{itk} = \sum_{z=1}^{Z} \theta_z x_{itkz} + \epsilon_{itk} \tag{2}$$

The latent variable is not observed directly, but it is possible to posit a probabilistic relationship with the outcome y_{itk} (the severity of crash k for individual i in traffic unit t). Depending on the characteristics of the data and the assumptions made about the random component of the latent function different models can be obtained. For example, if crash severity is coded as a variable with three outcomes (e.g., property damage only/injury/fatal), we can relate the latent variable to the outcome as follows:

$$y_{itk} = \begin{cases} \text{fatality} & \text{if } y^*_{itk} > \mu_2 \\ \text{injury} & \text{if } \mu_1 < y^*_{itk} < \mu_2 \\ \text{PDO} & \text{if } y^*_{itk} < \mu_1 \end{cases} \tag{3}$$

where μ_1 and μ_2 are estimable thresholds. Due to the stochastic nature of the latent function, the outcome of the crash is not fully determined. However, we can make the following probability statements:

$$\begin{aligned} P(y_{itk} = \text{PDO}) &= 1 - P(y_{itk} = \text{injury}) - P(y_{itk} = \text{fatality}) \\ P(y_{itk} = \text{injury}) &= P(\mu_1 - \textstyle\sum_{z=1} \theta_z p_{itkz} < \epsilon_{itk} < \mu_2 - \sum_{z=1} \theta_z p_{itkz}) \\ P(y_{itk} = \text{fatality}) &= P(\epsilon_{itk} < \mu_1 - \textstyle\sum_{z=1} \theta_z p_{itkz}) \end{aligned} \tag{4}$$

If the random terms are assumed to follow the logistic distribution, the ordered logit model is obtained; if the normal distribution, then the ordered probit model. Estimation methods for these models are very well-established (e.g., Maddala, 1986; Train, 2009). There are numerous variations of the basic modelling framework above, including hierarchical models, bivariate models, multinomial models, and Bayesian models, among others (see Savolainen et al., 2011 for a review of methods).

3.2. Strategy 1: opponent-related factors

When opponent-related variables are included, the function is augmented as follows:

$$y^*_{itk} = \sum_{l=1}^{L} \alpha_l p_{itkl} + \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \sum_{r=1}^{R} \delta_r o_{jvkr} + \epsilon_{itk} \tag{5}$$

This function includes one additional summation compared to Equation 3. This summation gathers $r = 1, \cdots, R$ attributes o related to individual j in traffic unit v that was an opposite party to individual i in traffic unit t in crash k. These attributes could be individual characteristics of the operator of the opposite traffic unit (such as age and gender) and/or characteristics of the opposite traffic unit (such as vehicle type or weight). To qualify as an opposite party, individual j must have been an individual in crash k but operating traffic unit $v \neq t$. Sometimes the person *is* the traffic unit, as is the case of a pedestrian. And we exclude passengers of vehicles as opponents, since they do not operate the traffic unit. In case the opponent attributes include only characteristics of the traffic unit, the condition for the traffic unit to be an opponent is that it was part of crash k and was different from t. After introducing this new set of terms, the latent function now consists of a total of $Z = L + M + Q + R$ covariates and associated parameters.

3.3. Strategy 2: hierarchical model specification

We can choose to conceptualize the event leading to the outcome as a hierarchical process. There are a few different ways of doing this. For example, the hierarchy could be based on individuals in traffic units. In this case, we can rewrite the latent function as follows:

$$y^*_{itk} = \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \sum_{r=1}^{R} \delta_r o_{jvkr} + \epsilon_{itk} \tag{6}$$

The coefficients of the traffic unit nest the individual attributes as follows. For any given coefficient m:

$$\beta_m = \sum_{l=1}^{L} \beta_{ml} p_{itkl} \tag{7}$$

Therefore, the corresponding term in the latent function becomes (assuming that $p_{itk1} = 1$, i.e., it is a constant term):

$$\begin{aligned} \beta_m u_{tkm} &= \big(\beta_{m1} + \beta_{m2} p_{itk2} + \cdots + \beta_{mL} p_{itkL}\big) u_{tkm} \\ &= \beta_{m1} u_{tkm} + \beta_{m2} p_{itk2} u_{tkm} + \cdots + \beta_{mL} p_{itkL} u_{tkm} \end{aligned} \tag{8}$$

As an alternative, the nesting unit could be the interaction person-opponent, in which case the opponent-level attributes are nested in the following fashion:

$$y^*_{itk} = \sum_{l=1}^{L} \alpha_l p_{itkl} + \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \epsilon_{itk} \tag{9}$$

with any person-level coefficient l that we wish to expand defined as follows:

$$\alpha_l = \sum_{r=1}^{R} \alpha_{lr} o_{jvkr} \tag{10}$$

with the same conditions as before, that $j \neq i$ is the operator of traffic unit $v \neq t$. The corresponding term in the latent function is now (assuming that $o_{jvk1} = 1$, i.e., it is a constant term):

$$\begin{aligned} \alpha_l p_{itkl} &= \big(\alpha_{l1} + \alpha_{l2} o_{jvk2} + \cdots + \alpha_{lR} o_{jvkR}\big) p_{itkl} \\ &= \alpha_{l1} p_{itkl} + \alpha_{l2} o_{jvk2} p_{itkl} + \cdots + \alpha_{lR} o_{jvkR} p_{itkl} \end{aligned} \tag{11}$$

Discerning readers will identify this model specification strategy as Casetti's expansion method (Casetti, 1972; Roorda et al., 2010). This is a deterministic strategy for modelling contextual effects which, when augmented with random components, becomes the well-known multi-level modelling method (Hedeker and Gibbons, 1994, more on this in Section 6). It is worthwhile to note that higher-order hierarchical effects are possible; for instance, individual attributes nested within traffic units, which in turn are nested within collisions. We do not explore higher-level hierarchies further in the current paper.

3.4. Strategy 3: sample subsetting

The third model specification strategy that we will consider is subsetting the sample. This is applicable in conjunction with any of the other strategies discussed above. In essence, we define the latent function with restrictions as follows. Consider a continuous variable, e.g., age of person, and imagine that we wish to concentrate the analysis on older adults (e.g., Dissanayake and Lu, 2002). The latent function is defined as desired (see above), however, the following restriction is applied to the sample:

$$\text{Age of individual } i \text{ in traffic unit } t \text{ in crash } k = \begin{cases} \geq 65 & \text{use record } itk \\ < 65 & \text{do not use record } itk \end{cases} \tag{12}$$

Suppose instead that we are interested in crashes by or against a specific type of traffic unit (e.g., pedestrians, Amoh-Gyimah et al., 2017):

$$\text{Road user class of individual } i \text{ in traffic unit } t \text{ in crash } k = \begin{cases} \text{Pedestrian} & \text{use record } itk \\ \text{Not pedestrian} & \text{do not use record } itk \end{cases} \tag{13}$$

or:

$$\text{Road user class of individual } j \text{ in traffic unit } v \text{ in crash } k = \begin{cases} \text{Pedestrian} & \text{use record } jvk \\ \text{Not pedestrian} & \text{do not use record } jvk \end{cases} \tag{14}$$

More generally, for any variable x of interest:

$$x_{itk} = \begin{cases} \text{Condition: TRUE} & \text{use record } itk \\ \text{Condition: FALSE} & \text{do not use record } itk \end{cases} \tag{15}$$

Several conditions can be imposed to subset the sample in any way that the analyst deems appropriate or suitable.

4. Setting for empirical assessment

In this section we present the setting for the empirical assessment of the modelling strategies discussed in Section 3, namely matters related to data and model estimation.

4.1. Data for empirical assessment

To assess the performance of the various modelling strategies we use data from Canada's National Collision Database (NCDB). This database contains all police-reported motor vehicle collisions on public roads in Canada. Data are originally collected by provinces and territories, and shared with the federal government, that proceeds to combine, track, and analyze them for reporting deaths, injuries, and collisions in Canada at the national level. The NCDB is provided by Transport Canada, the agency of the federal government of Canada in charge of transportation policies and programs, under the Open Government License - Canada version 2.0 [https://open.canada.ca/en/open-government-licence-canada].

The NCDB is available from 1999. For the purpose of this paper, we use the data corresponding to 2017, which is the most recent year available as of this writing. Furthermore, for assessment we also use the data corresponding to 2016. Similar to databases in other jurisdictions (see Montella et al., 2013), the NCDB contains information pertaining to the collision, the traffic unit(s), and the person(s) involved in a crash. The definitions of variables in this database are shown in the Appendix at the end of this document, in Tables 11, 12, and 13. Notice that, compared to Table 1, environmental-factors variables and road-factors variables are gathered under a single variable class, namely collision-related, since they are unique for each crash.

Data are organized by person; in other words, there is one record per individual in a collision, be they drivers, passengers, pedestrians, etc. The only variable directly available with respect to opponents in a

collision is the number of vehicles involved (see models in Bogue et al., 2017). Therefore, the data needs to be processed to obtain attributes of the opposing party for each individual in a collision. The protocol to do this is described next.

4.2. Data preprocessing

To prepare the data for analysis, in particular for Strategy 1 (opponent-related factors), we apply an initial filter, whereby we scan the database to remove all records that are not a person (including parked cars and other objects) or that are missing information.

After the initial filter, the database is summarized to find the number of individual-level records that correspond to each collision (C_CASE). At this point, there are 32,298 collisions, involving only one (known) individual, there are 46,483 collisions involving two parties, 19,433 collisions with three parties, 8,250 collisions involving four parties, 3,783 collisions with five parties, 1,789 collisions with six parties, and 1,491 collisions involving seven or more parties These parties were possibly occupants in different vehicles or were otherwise their own traffic units (e.g., pedestrians). Accordingly, the sample includes 174,741 drivers, 61,403 passengers, 10,798 pedestrians, 5,286 bicyclists, and 6,564 motorcyclists.

The next step is to remove all collisions that involve only one party. This still leaves numerous cases where multiple parties could have been in a single vehicle, for instance in a collision against a stationary object. Therefore, we proceed to use the vehicle sequence number to find the number of vehicles involved in each collision. This reveals that there are 20,732 collisions that involve only one vehicle but possibly multiple individuals (i.e., driver and one or more passengers). In addition, there are 165,520 collisions involving two vehicles (and possibly multiple individuals). Finally, there are 40,242 collisions with three or more vehicles.

Once we have identified the number of vehicles in each collision, we select all cases that involve only two vehicles. The most common cases in two-vehicle collisions are those that include drivers (40,297 collisions; this is reflective of the prevalence of single-occupant vehicles). This is followed by cases with passengers (14,120 collisions), pedestrians (5,204 collisions), bicyclists (2,238 collisions), and motorcyclists (1,016 collisions). The distribution of individuals per traffic unit is as follows: 80,382 individuals are coded as being in V_ID = 1, 76,523 individuals are coded as being in V_ID = 2, and 7,932 individuals are coded as pedestrians. In addition, 683 individuals are coded as being in vehicles 3 through 9, despite our earlier filter to retain only collisions with two vehicles. At this point we select only individuals assigned to vehicles 1 or 2, as well as pedestrians. As a result of this filter a number of cases with only one known individual need to be removed.

At this point we have a complete, workable sample of individual records of parties in two-vehicle collisions. There are two possible cases for the collisions, depending on the traffic units involved: 1) vehicle vs vehicle collisions ("vehicle" is all motorized vehicles, including motorcycles/mopeds, as well as bicycles); and 2)

Table 2: Summary of opponent interactions and outcomes by road user class

	Road User Class of Opponent				Outcome			Proportion by Road User Class		
Road User Class	Driver	Pedestrian	Bicyclist	Motorcyclist	No Injury	Injury	Fatality	No Injury	Injury	Fatality
All opponents										
Driver	97582	7880	3799	2498	59180	52143	436	0.52953	0.46657	0.003901
Passenger	35359	1282	667	818	19308	18667	151	0.50643	0.48961	0.003961
Pedestrian	7880	0	0	0	145	7507	228	0.01840	0.95266	0.028934
Bicyclist	3799	1	0	40	49	3760	31	0.01276	0.97917	0.008073
Motorcyclist	2498	30	40	338	204	2598	104	0.07020	0.89401	0.035788
Opponent: Driver										
Driver	97582	0	0	0	45493	51657	432	0.46620	0.52937	0.004427
Passenger	35359	0	0	0	16672	18536	151	0.47151	0.52422	0.004270
Pedestrian	7880	0	0	0	145	7507	228	0.01840	0.95266	0.028934
Bicyclist	3799	0	0	0	43	3725	31	0.01132	0.98052	0.008160
Motorcyclist	2498	0	0	0	98	2299	101	0.03923	0.92034	0.040432
Opponent: Pedestrian										
Driver	0	7880	0	0	7693	187	0	0.97627	0.02373	0.000000
Passenger	0	1282	0	0	1246	36	0	0.97192	0.02808	0.000000
Pedestrian	0	0	0	0	0	0	0	-	-	-
Bicyclist	0	1	0	0	0	1	0	0.00000	1.00000	0.000000
Motorcyclist	0	30	0	0	11	19	0	0.36667	0.63333	0.000000
Opponent: Bicyclist										
Driver	0	0	3799	0	3706	93	0	0.97552	0.02448	0.000000
Passenger	0	0	667	0	649	18	0	0.97301	0.02699	0.000000
Pedestrian	0	0	0	0	0	0	0	-	-	-
Bicyclist	0	0	0	0	0	0	0	-	-	-
Motorcyclist	0	0	40	0	16	24	0	0.40000	0.60000	0.000000
Opponent: Motorcyclist										
Driver	0	0	0	2498	2288	206	4	0.91593	0.08247	0.001601
Passenger	0	0	0	818	741	77	0	0.90587	0.09413	0.000000
Pedestrian	0	0	0	0	0	0	0	-	-	-
Bicyclist	0	0	0	40	6	34	0	0.15000	0.85000	0.000000
Motorcyclist	0	0	0	338	79	256	3	0.23373	0.75740	0.008876

vehicle vs pedestrian collisions. To identify the opposite parties in each collision it is convenient to classify collisions by pedestrian involvement. In this way, we find that the database includes 16,636 collisions that are vehicle vs pedestrian (possibly multiple pedestrians), and 147,594 collisions that involve two vehicles. After splitting the database according to pedestrian involvement, we can now extract relevant information about the different parties in the collision. This involves renaming the person-level variables so that we can distinguish each individual by their party in a given record. Notice that when working with individuals in vehicles, only drivers are considered opposites in a collision.

Once the personal attributes of opposite operators in a given collision are extracted, their information is joined to the individual records by means of the collision unique identifier. As a result of this process, a new set of variables are now available for analysis: the age, sex, and road user class of the opposite driver, as well as the type of the opposite vehicle. A summary of opponent interactions and outcomes can be found in Table 2. The information there shows that the most common type of opponent for drivers are other drivers, followed by pedestrians. The only opponents of pedestrians, on the other hand, are drivers. Bicyclists and motorcyclists are mostly opposed by drivers, but occasionally by other road users as well. In terms of outcomes, we observe that virtually all fatalities occur when the opponent is a driver, and only very rarely

when the opponent is a motorcyclist. Injuries are also more common when the opponent is a driver, whereas "no injury" is a relatively more frequent outcome when the opponent is a pedestrian or a bicyclist.

4.3. Model estimation

Before model estimation, the variables are prepared as follows. First, age is scaled from years to decades. Secondly, new variables are defined to describe the vehicle type. Three classes of vehicle types are considered: 1) light duty vehicles (which in Canada include passenger cars, passenger vans, light utility vehicles, and light duty pick up trucks); 2) light trucks (all other vehicles $\leq$ 4536 kg in gross vehicle weight rating); and heavy vehicles (all other vehicles $\geq$ 4536 kg in gross vehicle weight rating). Furthermore, this typology of vehicle is combined with the road user class of the individual to distinguish between drivers and passengers of light duty vehicles, light trucks, and heavy vehicles, in addition to pedestrians, bicyclists, and motorcyclists. This is done for both the individual and the opponent. Variable interactions are calculated to produce hierarchical variables. For example, for a hierarchical definition of traffic unit-level variables, age (and the square of age to account for possible non-monotonic effects) are interacted with gender, road user class, and vehicle type. For hierarchical opponent variables, age (and the square of age) are interacted with the age of opponent (and the corresponding square). The variables thus obtained are shown in Table 3. As seen in the table, Models 1 and 2 are single-level models, and the difference between them is that Model 2 includes opponent variables. Models 3 and 4, in contrast, are hierarchical models. Model 3 considers the hierarchy on the basis of the traffic unit, while Model 4 considers the hierarchy on the basis of the opponent.

Models 1 through 4 are estimated using the full sample. As discussed above, a related modelling strategy is to subset the sample (e.g., Islam et al., 2014; Lee and Li, 2014; Torrao et al., 2014; Wu et al., 2014). In this case we subset by a combination of traffic unit type of the individual (i.e., light duty vehicle, light truck, heavy vehicle, pedestrian, bicyclist, and motorcyclist) and vehicle type of the opponent (i.e., light duty vehicle, light truck, heavy vehicle). This leads to an ensemble of eighteen models to be estimated using subsets of data (see Table 3). By subsetting the sample, at least *some* opponent effects are incorporated implicitly. Models 1 and 2 are re-estimated using this strategy, dropping variables as necessary whenever they become irrelevant (for instance, after filtering for pedestrians, no other traffic unit types are present in the subset of data). In addition to variables that are no longer relevant in some data subsets, it is important to note that when using some data subsets a few variables had to be occasionally dropped to avoid convergence issues. This tended to happen particularly with smaller subsets where some particular combination of attributes was rare as a result of subsampling (e.g., in 2017 there were few or no collisions that involved a motorcyclist and a heavy vehicle in a bridge, or overpass, or viaduct). The process of estimation carefully paid attention to convergence issues to ensure the validity of the models reported here.

Table 3: Summary of variables and model specification

Variable	Notes	Model 1 Single-level /No opponent	Model 2 Single-level /Opponent attributes	Model 3 Hierarchical: Traffic unit	Model 4 Hierarchical: Opponent attributes
Individual-level variables					
Age	In decades	✓	✓	✓	✓
Age Squared		✓	✓	✓	✓
Sex	Reference: Female	✓	✓	✓	✓
Use of Safety Devices	7 levels; Reference: No Safety Device	✓	✓	✓	✓
Traffic unit-level variables					
Passenger	Reference: Driver	✓	✓		✓
Pedestrian	Reference: Driver	✓	✓		✓
Bicyclist	Reference: Driver	✓	✓		✓
Motorcyclist	Reference: Driver	✓	✓		✓
Light Truck	Reference: Light Duty Vehicle	✓	✓		✓
Heavy Vehicle	Reference: Light Duty Vehicle	✓	✓		✓
Opponent variables					
Age of Opponent	In decades		✓	✓	
Age of Opponent Squared			✓	✓	
Sex of Opponent	Reference: Female		✓	✓	
Opponent: Light Duty Vehicle	Reference: Pedestrian/Bicyclist/Motorcyclist		✓	✓	✓
Opponent: Light Truck	Reference: Pedestrian/Bicyclist/Motorcyclist		✓	✓	✓
Opponent: Heavy Vehicle	Reference: Pedestrian/Bicyclist/Motorcyclist		✓	✓	✓
Hierarchical traffic unit variables					
Light Truck Driver:Age				✓	
Light Truck Driver:Age Squared				✓	
Heavy Vehicle Driver:Age				✓	
Heavy Vehicle Driver:Age Squared				✓	
Light Truck Passenger:Age				✓	
Light Truck Passenger:Age Squared:				✓	
Heavy Vehicle Passenger:Age				✓	
Heavy Vehicle Passenger:Age Squared				✓	
Pedestrian:Age				✓	
Pedestrian:Age Squared				✓	
Bicyclist:Age				✓	
Bicyclist:Age Squared				✓	
Motorcyclist:Age				✓	
Motorcyclist:Age Squared				✓	
Hierarchical opponent variables					
Age:Age of Opponent					✓
Age:Age of Female Opponent					✓
Age:Age of Male Opponent Squared					✓
Age:Age of Female Opponent Squared					✓
Age Squared:Age of Male Opponent					✓
Age Squared:Age of Female Opponent					✓
Collision-level variables					
Crash Configuration	19 levels; Reference: Hit a moving object	✓	✓	✓	✓
Road Configuration	12 levels; Reference: Non-intersection	✓	✓	✓	✓
Weather	9 levels; Reference: Clear and sunny	✓	✓	✓	✓
Surface	11 levels; Reference: Dry	✓	✓	✓	✓
Road Alignment	8 levels; Reference: Straight and level	✓	✓	✓	✓
Traffic Controls	19 levels; Reference: Operational traffic signals	✓	✓	✓	✓
Month	12 levels; Reference: January	✓	✓	✓	✓

5. Model assessment

In this section we report an in-depth examination of the performance of the models. We begin by inspecting the statistical goodness of fit of the models by means of Akaike's Information Criterion (AIC). Next, we use the models to conduct in-sample predictions (i.e., *nowcasting*), using the same sample that was used to estimate the models, and out-of-sample predictions, using the data set corresponding to the year 2016 (i.e., *backcasting*). Predictions are commonly evaluated in two different ways in the literature. Some researchers analyze the outcome shares based on the predicted probabilities (e.g., Bogue et al., 2017; Yasmin and Eluru, 2013). This is a form of aggregate forecasting. Other researchers, in contrast, evaluate the classes of outcomes based on the individual-level predictions (e.g., Tang et al., 2019; Torrao et al., 2014). This is a

form of disaggregate forecasting.

5.1. Goodness of fit of models

We begin our empirical assessment by examining the results of estimating the models described above. Tables 4 and 5 present some key summary statistics of the estimated models. Of interest is the goodness of fit of the models, which in the case is measured with Akaike's Information Criterion (AIC). This criterion is calculated as follows:

$$AIC = 2Z - 2\ln\hat{L} \tag{16}$$

where Z is the number of coefficients estimated by the model, and $\hat{L}$ the maximized likelihood of the model. Since AIC penalizes the model fit by means of the number of coefficients, this criterion gives preference to more parsimonious models. The objective is to minimize the AIC, and therefore smaller values of this criterion represent better model fits. Model comparison can be conducted using the relative likelihood. Suppose that we have two models, say Model a and Model b, with $AIC_a \leq AIC_b$. The relative likelihood is calculated as:

$$e^{(AIC_a - AIC_b)/2} \tag{17}$$

The relative likelihood is interpreted as the probability that Model b minimizes the information loss as well as Model a.

Turning our attention to the models estimated using the full sample (Table 4), it is possible to see that, compared to the base (single-level) model without opponent variables (Model 1), there are large and significant improvements in goodness of fit to be gained by introducing opponent effects. However, the gains are not as large when hierarchical specifications are used, even when the number of additional coefficients that need to be estimated is not substantially larger (recall that the penalty per coefficient in AIC is 2). The best model according to this measure of goodness of fit is Model 2 (single-level with opponent effects), followed by Model 4 (hierarchical opponent variables), Model 3 (hierarchical traffic unit variables with opponent effects), and finally Model 1 (single-level without opponent effects).

It is important to note that the likelihood function of a model, and therefore the value of its AIC, both depend on the size of the sample, which is why AIC is not comparable across models estimated with different sample sizes. For this reason, the full sample models cannot be compared directly to the models estimated with subsets of data. The models in the ensembles, however, can be compared to each other (Table 5). As seen in the table, introducing opponent variables leads to a better fit in the case of most, but not all models.

Table 4: Summary of model estimation results: Full sample models

Model	Number of observations	Number of coefficients	AIC
Model 1. Single-level/No opponent	164,511	102	195,215
Model 2. Single-level/Opponent attributes	164,511	108	178,943
Model 3. Hierarchical: Traffic unit	164,511	118	181,333
Model 4. Hierarchical: Opponent)	164,511	111	179,018

The simplest model (single-level without opponent effects) is clearly the best fitting candidate in the case of bicycle vs light truck collisions, bicycle vs heavy vehicle collisions, motorcyclist vs light duty vehicle collisions, and motorcyclist vs heavy vehicle collisions. Model 1 is a statistical toss for best performance with two competing models in the case of pedestrian vs heavy vehicle collisions. The relative likelihood of Model 1 compared to Models 2 and 3 in this case is 0.56, which means that these two models are 0.56 times as probable as Model 1 to minimize the information loss.

Model 2 is the best fit in the case of light truck vs heavy vehicle collisions. This model is also tied for best fit with Model 2 in the case of pedestrian vs light duty vehicle and pedestrian vs light truck collisions, and is a statistical toss with Model 4 in the case of heavy vehicle vs heavy vehicle collisions (relative likelihood is 0.592). Model 3 is the best fit in the case of light duty vehicle vs light duty vehicle collisions and heavy vehicle vs light duty vehicle. Model 4 is the best fit in the case of light duty vehicle vs light truck collisions, light duty vehicle vs heavy vehicle collisions, light truck vs light duty vehicle collisions, heavy vehicle vs light truck collisions, and motorcycle vs light truck collisions. This model is a statistical toss with Model 2 in the case of light truck vs light truck collisions, with a relative likelihood of 0.791.

These results give some preliminary ideas about the relative performance of the different modelling strategies. In the next subsection we delve more deeply into this question by examining the predictive performance of the various modelling strategies. The results up to this point indicate that different model specification strategies might work best when combined with subsampling strategies. For space reasons, from this point onwards, we will consider the ensembles of models for predictions and will not compare individual models within the ensembles; this we suggest is a matter for future research.

Table 5: Summary of model estimation results: Data Subset Models

		Model 1 Single-level/No opponent		Model 2 Single-level/Opponent attributes		Model 3 Hierarchical: Traffic unit		Model 4 Hierarchical: Opponent	
Model	Number of observations	Number of coefficients	AIC	Number of coefficients	AIC	Number of coefficients	AIC	Number of coefficients	AIC
Light duty vehicle vs light duty vehicle	114,841	94	145,390	97	143,903	100	143,896	100	144,004
Light duty vehicle vs light truck	3,237	79	3,943	82	3,927	85	3,937	85	3,922
Light duty vehicle vs heavy vehicle	5,013	88	5,895	91	5,878	94	5,888	94	5,864
Light truck vs light duty vehicle	3,121	79	3,885	82	3,877	85	3,881	85	3,875
Light truck vs light truck	809	67	1,170	70	1,156	73	1,162	73	1,155
Light truck vs heavy vehicle	198	64	288	67	281	70	287	70	286
Heavy vehicle vs light duty vehicle	4,726	79	4,326	84	4,283	86	4,268	87	4,287
Heavy vehicle vs light truck	180	64	225	65	205	67	207	66	187
Heavy vehicle vs heavy vehicle	779	74	1,147	77	1,136	80	1,141	80	1,137
Pedestrian vs light duty vehicle	7,176	88	2,826	91	2,821	91	2,821	93	2,827
Pedestrian vs light truck	328	62	202	65	200	65	200	68	206
Pedestrian vs heavy vehicle	376	64	409	67	410	67	410	70	417
Bicyclist vs light duty vehicle	3,521	80	654	83	659	83	659	85	657
Bicyclist vs light truck	116	42	84	57	114	57	114	54	108
Bicyclist vs heavy vehicle	-	-	-	-	-	-	-	-	-
Motorcyclist vs light duty vehicle	2,298	78	1,367	81	1,373	81	1,373	84	1,373
Motorcyclist vs light truck	127	56	153	59	153	59	153	47	94
Motorcyclist vs heavy vehicle	62	43	88	45	90	46	92	51	102

Note:
There are zero cases of Bicyclist vs heavy vehicle in the sample

5.2. Outcome shares based on predicted probabilities

In this, and the following section, *backcasting* refers to the prediction of probabilities and classes of outcomes using the 2016 data set. When conducting backcasting, the data set is preprocessed in identical manner as the 2017 data set. In addition, the variables used in backcasting match exactly those in the models. This means that some variables were dropped when they were present in the 2016 data set but not in the models. This tended to happen in the case of relatively rare outcomes (e.g., in 2016, there was at least one collision between a heavy vehicle and a light duty vehicle in a school crossing zone; no such event was observed in 2017).

The shares of each outcome are calculated as the sum of the estimated probabilities for each observation:

$$\begin{aligned} \hat{S}_{\text{PDO}} &= \textstyle\sum_{itk} \hat{P}(y_{itk} = \text{PDO}) \\ \hat{S}_{\text{injury}} &= \textstyle\sum_{itk} \hat{P}(y_{itk} = \text{injury}) \\ \hat{S}_{\text{fatality}} &= \textstyle\sum_{itk} \hat{P}(y_{itk} = \text{fatality}) \end{aligned} \tag{18}$$

where $\hat{P}(y_{itk} = h_w)$ is the estimated probability of outcome h_w for individual i in traffic unit t and crash k. The estimated share of outcome h is $\hat{S}_{h_w}$.

The estimated shares can be used to assess the ability of the model to forecast for the population the total number of cases of each outcome. A summary statistic useful to evaluate the performance is the Average Percentage Error, or APE (see Bogue et al., 2017, p. 31), which is calculated for each outcome as follows:

$$APE_{h_w} = \left| \frac{\hat{S}_{h_w} - S_{h_w}}{S_{h_w}} \right| \times 100 \tag{19}$$

The Weighted Average Percentage Error ($WAPE$) aggregates the APE as follows:

$$WAPE = \frac{\sum_w^W APE_{h_w} \times S_{h_w}}{\sum_w^W S_{h_w}} \tag{20}$$

The results of this exercise are reported in Table 6. Of the four full-sample models (Models 1-4), the APE of Model 2 is lowest in the nowcasting exercise for every outcome, with the exception of Fatality, where Model 4 produces a considerably lower APE. When the results are aggregated by means of the $WAPE$, Model 2 gives marginally better results than Model 4. It is interesting to see that the four ensemble models have lower APE values across the board in the nowcasting exercise, and much better $WAPE$ than the full sample models. However, once we turn to the results of the backcasting exercise, these results do not hold, and it is possible to see that the Average Percentage Errors of the ensemble worsen considerably, particularly in the case of Fatality. The Weighted Average Prediction Error of the ensemble models in the backcasting

Table 6: Predicted shares and average prediction errors (APE) by model (percentages)

	No Injury			Injury			Fatality			
Model	Observed	Predicted	APE	Observed	Predicted	APE	Observed	Predicted	APE	WAPE
In-sample (nowcasting using 2017 data set, i.e., estimation data set)										
Model 1	78886	79029.00	0.18	84675	84533.74	0.17	950	948.26	0.18	0.17
Model 2	78886	78928.98	0.05	84675	84641.94	0.04	950	940.08	1.04	0.05
Model 3	78886	79027.29	0.18	84675	84512.50	0.19	950	971.21	2.23	0.20
Model 4	78886	78939.18	0.07	84675	84622.54	0.06	950	949.28	0.08	0.06
Model 1 Ensemble	62413	62402.78	0.02	83564	83573.58	0.01	931	931.64	0.07	0.01
Model 2 Ensemble	62417	62407.00	0.02	83595	83604.14	0.01	931	931.86	0.09	0.01
Model 3 Ensemble	62411	62401.23	0.02	83596	83604.71	0.01	933	934.06	0.11	0.01
Model 4 Ensemble	62405	62395.28	0.02	83578	83586.75	0.01	932	932.97	0.10	0.01
Out-of-sample (backcasting using 2016 data set)										
Model 1	96860	96364.67	0.51	101605	102002.59	0.39	1109	1206.74	8.81	0.50
Model 2	96860	96361.41	0.51	101605	102112.08	0.50	1109	1100.51	0.77	0.51
Model 3	96860	96354.01	0.52	101605	102086.18	0.47	1109	1133.82	2.24	0.51
Model 4	96860	96325.85	0.55	101605	102136.72	0.52	1109	1111.43	0.22	0.54
Model 1 Ensemble	77457	76822.49	0.82	100013	100580.60	0.57	1072	1138.91	6.24	0.71
Model 2 Ensemble	77459	76799.11	0.85	100049	100630.48	0.58	1071	1149.41	7.32	0.74
Model 3 Ensemble	77459	76786.76	0.87	100050	100644.29	0.59	1072	1149.95	7.27	0.75
Model 4 Ensemble	77461	76766.08	0.90	100029	100630.21	0.60	1070	1163.71	8.76	0.78

Note:
Model 1. Single-level/No opponent
Model 2. Single-level/Opponent attributes
Model 3. Hierarchical: Traffic unit
Model 4. Hierarchical: Opponent

exercise is also worse than for any of the full sample models. Excellent in-sample predictions but mediocre out-of-sample predictions are often evidence of overfitting, as in the case of the ensemble models here.

In terms of backcasting, full sample Model 1 is marginally better than full sample Models 2 and 3, and better than full sample Model 4. The reason for this is the lower APE of Model 1 when predicting Injury, the most frequent outcome. However, the performance of Model 1 with respect to Fatality (the least frequent outcome) is the worst of all models. Whereas Model 4 has the best performance predicting Fatality, its performance with respect to other classes of outcomes is less impressive. Model 3 does better than Model 2 with respect to Injury, but performs relatively poorly when backcasting Fatality. Overall, Model 2 appears to be the most balanced, with good in-sample performance and competitive out-of-sample performance that is also balanced with respect to the various classes of outcomes.

5.3. Outcome frequency based on predicted classes

APE and $WAPE$ are summary measures of the performance of models at the aggregated level. Aggregate-level predictions (i.e., shares of outcomes) are of interest from a population health perspective. In other cases, an analyst might be interested in the predicted outcomes at the individual level. In this section we examine the frequency of outcomes based on predicted classes, using the same two settings as above: nowcasting and backcasting.

The individual-level outcomes are examined using an array of verification statistics. Verification statistics are widely used in the evaluation of predictive approaches were the outcomes are categorical, and are often

based on the analysis of *confusion matrices* (e.g., Provost and Kohavi, 1998; Beguería, 2006). Confusion matrices are cross-tabulations of *observed* and *predicted* classes. In a two-by-two confusion matrix there are four possible combinations of observed to predicted classes: hits, misses, false alarms, and correct non-events, as shown in Table 7. When the outcome has more than two classes, the confusion matrix is converted to a two-by-two table to calculate verification statistics.

Table 7: Example of a two-by-two confusion matrix

	Observed		
Predicted	Yes	No	Marginal Total
Yes	Hit	False Alarm	Predicted Yes
No	Miss	Correct Non-event	Predicted No
Marginal Total	Observed Yes	Observed No	

The statistics used in our assessment are summarized in Table 8, including brief descriptions of their interpretation. The statistics evaluate different aspects of the performance of a model. Some are concerned with the ability of the model to be right, others are concerned with the ability of the model to match the observed outcomes, and yet others measure the ability of the model to not be wrong. These verification statistics are discussed briefly next.

Percent correct and percent correct by class PC and PC_c are relatively simple statistics, and are calculated as the proportion of correct predictions (i.e., hits and correct non-events) relative to the number of events, for the whole table PC or for one class only PC_c.

Bias (B) measures for each outcome class the proportion of total predictions by class (e.g., hits as well as false alarms) relative to the total number of cases observed for that class. For this reason, it is possible for predictions to have low bias (values closer to 1) but still do poorly in terms of hits.

Critical Success Index (CSI) evaluates forecasting skill while assuming that the number of correct non-events is inconsequential for demonstrating skill. Accordingly, the statistic is calculated as the proportion of hits relative to the sum of hits plus false alarms plus misses.

Probability of False Detection (F) is the proportion of false alarms relative to the total number of times that the event is not observed. This statistic measures the frequency with which the model incorrectly predicts an event, but not when it incorrectly misses it. The Probability of Detection (POD), in contrast, measures the frequency with which the model correctly predicts a class, relative to the number of cases of that class.

The False Alarm Ratio (FAR) is the fraction of predictions by class that were false alarms evaluate a different way in which a model can make equivocal predictions. In this case, lower scores are better.

The last three verification statistics that we consider are skill scores that simultaneously consider different

aspects of prediction, and are therefore overall indicators of prediction skill. Heidke's Skill Score (HSS) is the fraction of correct predictions above those that could be attributed to chance. Peirce's Skill Score (PSS) combines the Probability of Detection (POD) of a model and its Probability of False Detection (F) to measure the skill of a model to discriminate the classes of outcomes. Lastly, Gerrity Score (GS) is a measure of the model's skill predicting the correct classes that tends to reward correct forecasts of the least frequent class.

We discuss the results of calculating this battery of verification statistics, first for the nowcasting case (Table 9) and subsequently for the backcasting case (Table 10).

Table 8: Verification statistics

Statistic	Description	Notes
Percent Correct (PC)	Total hits and correct non-events divided by number of cases	Strongly influenced by most common category
Percent Correct by Class (PC_c)	Same as Percent Correct but by category	Strongly influenced by most common category
Bias (B)	Total predicted by category, divided by total observed by category	$B > 1$: class is overpredicted; $B < 1$: class is underpredicted
Critical Success Index (CSI)	Total hits divided by total hits + false alarms + misses	$CSI = 1$: perfect score; $CSI = 0$: no skill
Probability of False Detection (F)	Proportion of no events forecast as yes; sensitive to false alarms but ignores misses	$F = 0$: perfect score
Probability of Detection (POD)	Total hits divided by total observed by class	$POD = 1$: perfect score
False Alarm Ratio (FAR)	Total false alarms divided by total forecast yes by class; measures fraction of predicted yes that did not occur	$FAR = 0$: perfect score
Heidke Skill Score (HSS)	Fraction of correct predictions after removing predictions attributable to chance; measures fractional improvement over random; tends to reward conservative forecasts	$HSS = 1$: perfect score; $HSS = 0$: no skill; $HSS < 0$: random is better
Peirce Skill Score (PSS)	Combines POD and F; measures ability to separate yes events from no events; tends to reward conservative forecasts	$PSS = 1$: perfect score; $PSS = 0$: no skill
Gerrity Score (GS)	Measures accuracy of predicting the correct category, relative to random; tends to reward correct forecasts of less likely category	$GS = 1$: perfect score; $GS = 0$: no skill

5.3.1. Nowcasting: verification statistics

At first glance, the results of the verification statistics (Table 9) make it clear that no model under comparison is consistently a top performer from every aspect of prediction. Recalling Box's aphorism, all models are wrong but some are useful - in this case it just so happens that some models are more wrong than others in subtly different ways. That said, it is noticeable that the worst scores across the board tend to accrue to Model 1 in its full sample and ensemble versions. On the other hand, Model 2 (full sample) concentrates most of the best scores and second best scores of all the models, but also some of the worst scores for Fatality. Model 4, in contrast, has most of the second best scores and a few top scores, but not a single worst score.

Of all the models, Model 2 (full sample) performs best in terms of Percent Correct, followed by Model 4 (full sample). The worst performer from this perspective is Model 1 (full sample), with a PC score several percentage points below the top models. The second score is Percent Correct by Class (PC_c). This score is calculated individually for each outcome class. Model 2 (full sample), has the best performance for outcomes No Injury and Injury, and the second best score for Fatality. Model 4 (full sample) has the best score for Fatality, and is second best for No Injury and Injury. Model 1 (full sample) has worst scores for No Injury and Injury whereas its ensemble version has the worst score for Fatality. It is important to note that PC and PC_c are heavily influenced by the most common category, something that can be particularly appreciated in the scores for Fatality. The scores for this class of outcome are generally high, despite the fact that the number of hits are relatively low; the high values of PC_c in this case are due to the high occurrence of correct non-events elsewhere in the table.

The models with the best performance in terms of B are Model 1 (full sample) for No Injury and Injury, and Model 3 (ensemble) for Fatality. Model 4 (full sample) is the second best performer for No Injury and Injury, and Model 4 (ensemble) is second best performer for Fatality. Model 1 (ensemble) has the worst bias for No Injury and Injury, whereas Model 2 (full sample) has the worst bias for Fatality.

No model performs uniformly best from the perspective of Critical Success Index (CSI). Model 2 (full sample) has the best CSI for No Injury, Model 2 (ensemble) has the best score for Injury, and Model 4 (ensemble) the best score for Fatality. On the other hand, Model 1 (ensemble) has the worst score for No Injury, Model 1 (full sample) the worst score for Injury, and Model 2 the worst score for Fatality. These scores indicates that the models are not particularly skilled at predicting the corresponding classes correctly, given the frequency with which they give false alarms or miss the class.

The lowest probability of false detection F in the case of No Injury is 19.17% for Model 2 (ensemble), with every other model having values lower than 21%, with the exception of Model 1 (full sample) that has a

score of 26.46%. With respect to Injury, the lowest probabilities range 35.8% and 35.29% and 35.35% for Models 4 (full sample) and 2 (full sample). In contrast, the highest probability of false detection for Injury is 45.12% for Model 1 (ensemble). The scores for F for Fatality are all extremely low as a consequence of the very low frequency of this class of outcome in the sample.

As with some other verification statistics, no model is consistently a best performer in terms of POD. Model 4 (full sample) has the highest probability of detection for No Injury (65.38%), followed by Model 2 (65.32%), whereas the worst probability of detection is by Model 1 (ensemble) with a score of 55.54%. In terms of Injury, all models have POD higher than 79%, and the highest score is 80.67% for Model 2 (ensemble). The exception is Model 1 (full sample), which has a considerably lower POD of Injury with a score of 73.36%. Lastly, in terms of Fatalities, all models have very low probabilities of detection, ranging from a high of 4.94% in the case of Model 4 (ensemble) to a worst score of 0.11% in the case of Model 2 (full sample).

Model 2 (full sample) has the best FAR statistic for No Injury, as only 25.05% of predictions for this class are false alarms. The next best score is by Model 4 (full sample), with only 25.23% of No Injury predictions being false alarms. The worst performance in this class is by Model 1 (ensemble), which produces almost a third of false alarms in its predictions of No Injury. In the case of Injury, the False Alarm Ratio ranges from a low of 29.15% by Model (ensemble), with every other model having scores lower than 30% except Model 1 (full sample), that gives almost 32% of false alarms. In terms of Fatality, the lowest FAR is also for Model 4 (ensemble) with only 17.86% of false alarms, whereas the worst performance is by Model 2 (full sample), which produces over 95% of false alarms.

The skill scores help to remove some of the ambiguity regarding the overall performance of a model. In this way, we know that Model 2 (full sample) does not do particularly well with the class Fatality - however, of all models, it tends to have the best overall performance. Its HSS, for example, suggests that it achieves 44.74% of correct predictions after removing correct predictions attributable to chance. In contrast, the lowest score is for Model 1 (ensemble), which only achieves 36.26% correct predictions after removing those attributable to chance. Model 2 (full sample) also has the highest PSS and the second highest GS. Model 4 (full sample) has the highest GS and the second highest HSS and PSS.

5.3.2. Backcasting: verification statistics

Table 10 presents the results of the verification exercise for the case of our out-of-sample predictions (i.e., backasting). Qualitatively, the results are similar to those of the nowcasting experiments, but with a somewhat weaker performance of the ensemble models. This, again, supports the idea that these models might be overfitting the process, as discussed in reference to the aggregate forecasts (see Section 5.2). Models 2 (full

Table 9: Assessment of in-sample outcomes (nowcasting using 2017 data set, i.e., estimation data set)

Predicted	Observed Outcome			Verification Statistics									
Outcome	No Injury	Injury	Fatality	Percent Correct	Percent Correct by Class	Bias[1]	Critical Success Index[2]	Probability of False Detection[3]	Probability of Detection[4]	False Alarm Ratio[5]	Heidke Skill Score[6]	Peirce Skill Score[7]	Gerrity Score[8]
Model 1. Single-level/No opponent													
No Injury	50652	22504	150		69.07	**0.9293**	0.4988	0.2646	0.6421	0.309			
Injury	28232	62120	797	68.55	68.64	**1.0765**	0.5463	0.3636	0.7336	0.3185	0.3725	0.3696	0.1902
Fatality	2	51	3		99.39	0.0589	0.003	3e-04	0.0032	0.9464			
Model 2. Single-level/Opponent attributes													
No Injury	51531	17137	85		**72.9**	0.8715	**0.5362**	0.2011	0.6532	**0.2505**			
Injury	27355	67514	864	**72.36**	**72.42**	1.1306	0.598	0.3535	0.7973	0.2948	**0.4474**	**0.4429**	0.2265
Fatality	0	24	1		99.41	0.0263	0.001	1e-04	0.0011	0.96			
Model 3. Hierarchical: Traffic unit													
No Injury	51103	17296	79		72.55	0.8681	0.5309	0.2029	0.6478	0.2537			
Injury	27783	67338	868	72	72.05	1.1336	0.5942	0.3589	0.7953	0.2985	0.44	0.4357	0.2239
Fatality	0	41	3		99.4	0.0463	0.003	3e-04	0.0032	0.9318			
Model 4. Hierarchical: Opponent													
No Injury	51574	17316	84		72.82	0.8744	0.5356	0.2032	**0.6538**	0.2523			
Injury	27312	67336	863	72.28	72.33	1.128	0.5967	**0.3529**	0.7952	0.295	0.4458	0.4414	**0.2268**
Fatality	0	23	3		**99.41**	0.0274	0.0031	1e-04	0.0032	0.8846			
Model 1 Ensemble. Single-level/No opponent													
No Injury	34664	16434	63		69.88	0.8197	0.4393	0.1952	0.5554	0.3225			
Injury	27749	67120	829	69.31	69.35	1.1452	0.5985	0.4512	0.8032	0.2986	0.3626	0.3521	0.201
Fatality	0	10	39		99.39	0.0526	0.0414	1e-04	0.0419	0.2041			
Model 2 Ensemble. Single-level/Opponent attributes													
No Injury	35443	16145	60		70.62	0.8275	0.4508	**0.1917**	0.5678	0.3138			
Injury	26974	67437	829	70.05	70.08	1.1393	**0.6054**	0.4389	**0.8067**	0.2919	0.3784	0.3678	0.2106
Fatality	0	13	42		99.39	0.0591	0.0445	1e-04	0.0451	0.2364			
Model 3 Ensemble. Hierarchical: Traffic unit													
No Injury	35498	16204	60		70.62	0.8294	0.4512	0.1924	0.5688	0.3142			
Injury	26913	67379	828	70.04	70.08	1.1379	0.6052	0.4379	0.806	0.2916	0.3786	0.3681	0.2123
Fatality	0	13	45		99.39	**0.0622**	0.0476	1e-04	0.0482	0.2241			
Model 4 Ensemble. Hierarchical: Opponent													
No Injury	35553	16297	59		70.59	0.8318	0.4514	0.1935	0.5697	0.3151			
Injury	26852	67271	827	70.02	70.06	1.1361	0.6046	0.437	0.8049	**0.2915**	0.3783	0.3679	0.2127
Fatality	0	10	46		99.39	0.0601	**0.0488**	**1e-04**	**0.0494**	**0.1786**			

Note:
Bold numbers: best scores; underlined numbers: second best scores; red numbers: worst scores
[1] $B > 1$: class is overpredicted; $B < 1$: class is underpredicted;
[2] $CSI = 1$: perfect score; $CSI = 0$: no skill;
[3] $F = 0$: perfect score;
[4] $POD = 1$: perfect score;
[5] $FAR = 0$: perfect score;
[6] $HSS = 1$: perfect score; $HSS = 0$: no skill; $HSS < 0$: random is better;
[7] $PSS = 1$: perfect score; $PSS = 0$: no skill;
[8] $GS = 1$: perfect score; $GS = 0$: no skill.

Table 10: Assessment of out-of-sample outcomes (backcasting using 2016 data set)

Predicted	Observed Outcome			Verification Statistics									
Outcome	No Injury	Injury	Fatality	Percent Correct	Percent Correct by Class	Bias[1]	Critical Success Index[2]	Probability of False Detection[3]	Probability of Detection[4]	False Alarm Ratio[5]	Heidke Skill Score[6]	Peirce Skill Score[7]	Gerrity Score[8]
Model 1. Single-level/No opponent													
No Injury	61683	27447	184		68.53	**0.9221**	0.4955	0.269	0.6368	0.3094			
Injury	35172	74073	915	68.03	68.12	**1.0842**	0.538	0.3684	0.729	0.3276	0.3628	0.3604	0.1882
Fatality	5	85	10		99.4	0.0902	0.0083	5e-04	0.009	0.9			
Model 2. Single-level/Opponent attributes													
No Injury	62736	21013	106		**72.32**	0.8657	**0.5318**	0.2056	0.6477	**0.2519**			
Injury	34124	80569	996	**71.81**	**71.86**	1.1386	0.5893	0.3585	0.793	**0.3036**	**0.4373**	**0.4336**	**0.2241**
Fatality	0	23	7		**99.44**	0.0271	0.0062	**1e-04**	0.0063	**0.7667**			
Model 3. Hierarchical: Traffic unit													
No Injury	62248	21133	107		72.01	0.8619	0.5271	0.2068	0.6427	0.2544			
Injury	34610	80433	996	71.5	71.55	1.1421	0.5862	0.3634	0.7916	0.3068	0.431	0.4274	0.2205
Fatality	2	39	6		99.43	0.0424	0.0052	2e-04	0.0054	0.8723			
Model 4. Hierarchical: Opponent													
No Injury	62788	21247	102		72.23	0.8686	0.5312	0.2078	**0.6482**	0.2537			
Injury	34071	80331	1000	71.72	71.77	1.1358	0.5877	**0.358**	0.7906	0.3039	0.4354	0.4318	0.2233
Fatality	1	27	7		99.43	0.0316	0.0062	1e-04	0.0063	0.8			
Model 1 Ensemble. Single-level/No opponent													
No Injury	42896	20230	95		69.26	0.8162	0.4387	0.2011	0.5538	0.3215			
Injury	34546	79692	962	68.67	68.73	1.1519	0.588	0.4522	0.7968	0.3082	0.3539	0.3447	0.1831
Fatality	15	91	15		99.35	0.1129	0.0127	6e-04	0.014	0.876			
Model 2 Ensemble. Single-level/Opponent attributes													
No Injury	43485	19937	95		69.76	0.82	0.446	**0.1981**	0.5614	0.3154			
Injury	33954	80009	961	69.16	69.23	1.1487	**0.5928**	0.4446	**0.7997**	0.3038	0.3644	0.355	0.1883
Fatality	20	103	15		99.34	0.1289	0.0126	7e-04	0.014	0.8913			
Model 3 Ensemble. Hierarchical: Traffic unit													
No Injury	43526	20033	92		69.73	0.8217	0.446	0.199	0.5619	0.3162			
Injury	33915	79915	964	69.13	69.19	1.1474	0.5923	0.4441	0.7988	0.3038	0.3639	0.3546	0.1885
Fatality	18	102	16		99.34	0.1269	0.0134	7e-04	0.0149	0.8824			
Model 4 Ensemble. Hierarchical: Opponent													
No Injury	43561	20160	94		69.67	0.8238	0.4458	0.2003	0.5624	0.3174			
Injury	33875	79762	959	69.07	69.14	1.1456	0.5914	0.4436	0.7974	0.304	0.3629	0.3538	0.1886
Fatality	25	107	17		99.34	**0.1393**	**0.0141**	7e-04	**0.0159**	0.8859			

Note:
Bold numbers: best scores; underlined numbers: second best scores; red numbers: worst scores
[1] $B > 1$: class is overpredicted; $B < 1$: class is underpredicted;
[2] $CSI = 1$: perfect score; $CSI = 0$: no skill;
[3] $F = 0$: perfect score;
[4] $POD = 1$: perfect score;
[5] $FAR = 0$: perfect score;
[6] $HSS = 1$: perfect score; $HSS = 0$: no skill; $HSS < 0$: random is better;
[7] $PSS = 1$: perfect score; $PSS = 0$: no skill;
[8] $GS = 1$: perfect score; $GS = 0$: no skill.

sample) and 4 (full sample) are again identified as the best overall performers, and particularly Model 2 (full sample) performs somewhat more adroitly with respect to Fatality in backcasting than it did in nowcasting.

6. Further considerations

As discussed in Section 3, there is a rich selection of modelling approaches that are applicable to crash severity analysis. Based on the literature, we limited our empirical assessment of modelling strategies to only one model, namely the ordinal logit. On the other hand, since the modelling strategies discussed here all relate to the specification of the latent function and data subsetting, it is a relatively simple matter to extend them to other modelling approaches. For example, take Expression 9 and add a random component μ_k as follows:

$$y^*_{itk} = \sum_{l=1}^{L} \alpha_l p_{itkl} + \sum_{m=1}^{M} \beta_m u_{tkm} + \sum_{q=1}^{Q} \kappa_q c_{kq} + \sum_{r=1}^{R} \delta_r o_{jvkr} + \mu_k + \epsilon_{itk} \tag{21}$$

The addition of the random component in this fashion would help to capture, when appropriate, unobserved heterogeneity at the level of the crash (this is similar to the random intercepts approach in multi-level modelling; also see Mannering et al., 2016). As a second example, take Expressions 6 to 11 and add a random component to a hierarchical coefficient, to obtain:

$$\begin{aligned} \beta_m u_{tkm} &= \big(\beta_{m1} + \beta_{m2} p_{itk2} + \cdots + \beta_{mL} p_{itkL} + \mu_{mk}\big) u_{tkm} \\ &= \beta_{m1} u_{tkm} + \beta_{m2} p_{itk2} u_{tkm} + \cdots + \beta_{mL} p_{itkL} u_{tkm} + \mu_{mk} u_{tkm} \end{aligned} \tag{22}$$

This is similar to the random slopes strategy in multi-level modelling.

We do not report results regarding other modelling strategies. On the one hand, more sophisticated modelling frameworks are generally capable of improving the performance of a model. On the other hand, there are well-known challenges in the estimation of more sophisticated models (see Lenguerrand et al., 2006, p. 47, for a discussion of convergence issues in models with mixed effects; Mannering et al., 2016, p. 13, for some considerations regarding the complexity and cost of estimating more complex models; and Bogue et al., 2017, p. 27, on computational demands of models with random components). The additional cost and complexity of more sophisticated modelling approaches would, in our view, have greatly complicated our empirical assessment, particularly considering the large size of the sample involved in this research (a data set with over 164,000 records in the case of the full sample models). That said, we experimented with a model with random components using monthly subsets of data to find that, indeed, estimation takes considerably longer, is more demanding in terms of fixing potential estimation quirks, and in the end resulted in variance components that could not be reliably estimated as different from zero (results can be consulted in the source R Notebook). For this reason, we choose to leave the application of more sophisticated models as a matter for future research.

7. Concluding remarks

The study of crash severity is an important component of accident research, as seen from a large and vibrant literature and numerous applications. Part of this literature covers different modelling strategies that can be used to model complex hierarchical, multievent outcomes such as the severity of injuries following a collision. In this paper, our objective has been to assess the performance of different strategies to model opponent effects in two-vehicle crashes. In broad terms, three strategies were considered: 1) incorporating opponent-level variables in the model; 2) single- versus multi-level model specifications; and 3) sample

subsetting and estimation of separate models for different types of individual-opponent interactions. The empirical evaluation was based on data from Canada's National Crash Database and the application of ordered probit models. A suite of models that implemented the various strategies considered was estimated using data from 2017. We then assessed the performance of the models using one information criterion (AIC). Furthermore, the predictive performance of the models was assessed in terms of both nowcasting (in-sample predictions) and backcasting (out-of-sample predictions), the latter using data from 2016.

The results of the empirical assessment strongly suggest that incorporating opponent effects can greatly improve the goodness-of-fit and predictive performance of a model. Two modelling strategies appear to outperform the rest: a relatively simple single-level modelling approach that incorporates opponent effects, and a hierarchical modelling approach with nested opponent effects. There was some evidence that subsetting the sample can improve the results in some cases (e.g., when modelling the severity of crashes involving active travelers or motorcyclists), but possibly at the risk of overfitting the process and leading to worse out-of-sample performance. In this paper we did not compare individual models in our ensemble approach, but we suggest that this is an avenue for future research.

The results of this research should be informative to analysts interested in crashes involving two parties, since it provides some useful guidelines regarding the specification of opponent effects. Not only do opponent effects improve the goodness of fit and performance of models, they also add rich insights into their effects. The focus on this paper was on performance, and for space reasons it is not possible to include an examination of the best model without failing to do it justice. We plan to report the results of the best-fitting model in a future paper.

The analysis also opens up a few avenues for future research. First, for reasons discussed in Section 6, we did not consider more sophisticated modelling approaches, such as models with random components, partial proportional odds, ranked ordered models, or multinomial models, to mention just a few possibilities. Secondly, we only considered the performance of the models when making predictions for the full sample. That is, the submodels in the ensembles were not compared in detail, just their aggregate results when predicting the full sample. However, the goodness-of-fit was not uniformly better for any one modelling strategy when the data were subset, and it is possible that individual models perform better for a certain subset than competitors that are part of a better ensemble, overall. For this reason, we suggest that additional work with ensemble approaches is warranted. Finally, it is clear that the models do not generally do well when predicting the least frequent class of outcome, namely Fatality. It would be worthwhile to further investigate approaches for so-called imbalanced learning, a task where Torrao et al. (2014) have already made some headway.

Finally, as an aside, this paper is, to the best of our knowledge, the first example of reproducible research

in crash severity analysis. By providing the data *and* code for the analysis, it is our hope that this will allow other researchers to easily verify the results, and to extend them. A common practice in the machine learning community is to use canonical data set to demonstrate the performance of new techniques. Sharing code and data has remained relatively rare in transportation research, and we would like to suggest that the data sets used in this research could constitute one such canonical data set for future methodological developments.

Acknowledgments

This research was supported by a Research Excellence grant from McMaster University. The authors wish to express their gratitude to the Office of the Provost, the Faculty of Science, and the Faculty of Engineering for their generous support. The comments of four anonymous reviewers are also duly acknowledged: their feedback helped to improve the quality and clarity of presentation of this paper.

Appendix

Variable definitions in Canada's National Collision Database.

Table 11: Contents of National Collision Database: Collision-level variables

Variable	Description	Notes
C_CASE	Unique collision identifier	Unique identifier for collisions
C_YEAR	Year	Last two digits of year.
C_MNTH	Month	14 levels: January - December; unknown; not reported by jurisdiction.
C_WDAY	Day of week	9 levels: Monday - Sunday; unknown; not reported by jurisdiction.
C_HOUR	Collision hour	25 levels: hourly intervals; unknown; not reported by jurisdiction.
C_SEV	Collision severity	4 levels: collision producing at least one fatality; collision producing non-fatal injury; unknown; not reported by jurisdiction.
C_VEHS	Number of vehicles involved in collision	Number of vehicles: 1-98 vehicles involved; 99 or more vehicles involved; unknown; not reported by jurisdiction.

Table 11: Contents of National Collision Database: Collision-level variables *(continued)*

Variable	Description	Notes
C_CONF	Collision configuration	21 levels: SINGLE VEHICLE: Hit a moving object (e.g. a person or an animal); Hit a stationary object (e.g. a tree); Ran off left shoulder; Ran off right shoulder; Rollover on roadway; Any other single vehicle collision configuration; TWO-VEHICLES SAME DIRECTION OF TRAVEL: Rear-end collision; Side swipe; One vehicle passing to the left of the other, or left turn conflict; One vehicle passing to the right of the other, or right turn conflict; Any other two vehicle - same direction of travel configuration; TWO-VEHICLES DIFFERENT DIRECTION OF TRAVEL: Head-on collision; Approaching side-swipe; Left turn across opposing traffic; Right turn, including turning conflicts; Right angle collision; Any other two-vehicle - different direction of travel configuration; TWO-VEHICLES, HIT A PARKED VEHICLE: Hit a parked motor vehicle; Choice is other than the preceding values; unknown;not reported by jurisdiction.
C_RCFG	Roadway configuration	15 levels: Non-intersection; At an intersection of at least two public roadways; Intersection with parking lot entrance/exit, private driveway or laneway; Railroad level crossing; Bridge, overpass, viaduct; Tunnel or underpass; Passing or climbing lane; Ramp; Traffic circle; Express lane of a freeway system; Collector lane of a freeway system; Transfer lane of a freeway system; Choice is other than the preceding values; unknown;not reported by jurisdiction.
C_WTHR	Weather condition	10 levels: Clear and sunny; Overcast, cloudy but no precipitation; Raining; Snowing, not including drifting snow; Freezing rain, sleet, hail; Visibility limitation; Strong wind; Choice is other than the preceding values; unknown;not reported by jurisdiction.
C_RSUR	Road surface	12 levels: Dry, normal; Wet; Snow (fresh, loose snow); Slush, wet snow; Icy, packed snow; Debris on road (e.g., sand/gravel/dirt); Muddy; Oil; Flooded; Choice is other than the preceding values; unknown;not reported by jurisdiction.
C_RALN	Road alignment	9 levels: Straight and level; Straight with gradient; Curved and level; Curved with gradient; Top of hill or gradient; Bottom of hill or gradient; Choice is other than the preceding values; unknown;not reported by jurisdiction.

Table 11: Contents of National Collision Database: Collision-level variables *(continued)*

Variable	Description	Notes
C_TRAF	Traffic control	21 levels: Traffic signals fully operational; Traffic signals in flashing mode; Stop sign; Yield sign; Warning sign; Pedestrian crosswalk; Police officer; School guard, flagman; School crossing; Reduced speed zone; No passing zone sign; Markings on the road; School bus stopped with school bus signal lights flashing; School bus stopped with school bus signal lights not flashing; Railway crossing with signals, or signals and gates; Railway crossing with signs only; Control device not specified; No control present; Choice is other than the preceding values; unknown; not reported by jurisdiction.

Note:

Source NCDB available from https://open.canada.ca/data/en/data set/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a

Source data files for analysis also available from https://drive.google.com/open?id=12aJtVBaQ4Zj0xa7mtfqxh0E48hKCb_XV

Table 12: Contents of National Collision Database: Traffic unit-level variables

Variable	Description	Notes
V_ID	Vehicle sequence number	Number of vehicles: 1-98; Pedestrian sequence number: 99; unknown.
V_TYPE	Vehicle type	21 levels: Light Duty Vehicle (Passenger car, Passenger van, Light utility vehicles and light duty pick up trucks); Panel/cargo van (<= 4536 KG GVWR Panel or window type of van designed primarily for carrying goods); Other trucks and vans (<= 4536 KG GVWR); Unit trucks (> 4536 KG GVWR); Road tractor; School bus; Smaller school bus (< 25 passengers); Urban and Intercity Bus; Motorcycle and moped; Off road vehicles; Bicycle; Purpose-built motorhome; Farm equipment; Construction equipment; Fire engine; Snowmobile; Street car; Data element is not applicable (e.g. dummy vehicle record created for pedestrian); Choice is other than the preceding values; unknown; not reported by jurisdiction.
V_YEAR	Vehicle model year	Model year; dummy for pedestrians; unknown; not reported by jurisdiction.

Note:

Source NCDB available from https://open.canada.ca/data/en/data set/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a

Source data files for analysis also available from https://drive.google.com/open?id=12aJtVBaQ4Zj0xa7mtfqxh0E48hKCb_XV

Table 13: Contents of National Collision Database: Personal-level variables

Variable	Description	Notes
P_ID	Person sequence number	Sequence number: 1-99; Not applicable (dummy for parked vehicles); not reported by jurisdiction.
P_SEX	Person sex	5 levels: Male; Female; Not applicable (dummy for parked vehicles); unknown (runaway vehicle); not reported by jurisdiction.
P_AGE	Person age	Age: less than 1 year; 1-98 years old; 99 years or older; Not applicable (dummy for parked vehicles); unknown (runaway vehicle); not reported by jurisdiction.
P_PSN	Person position	Person position: Driver; Passenger front row, center; Passenger front row, right outboard (including motorcycle passenger in sidecar); Passenger second row, left outboard, including motorcycle passenger; Passenger second row, center; Passenger second row, right outboard; Passenger third row, left outboard;...; Position unknown, but the person was definitely an occupant; Sitting on someone's lap; Outside passenger compartment; Pedestrian; Not applicable (dummy for parked vehicles); Choice is other than the preceding values; unknown (runaway vehicle); not reported by jurisdiction.
P_ISEV	Medical treatment required	6 levels: No Injury; Injury; Fatality; Not applicable (dummy for parked vehicles); Choice is other than the preceding values; unknown (runaway vehicle); not reported by jurisdiction.
P_SAFE	Safety device used	11 levels: No safety device used; Safety device used; Helmet worn; Reflective clothing worn; Both helmet and reflective clothing used; Other safety device used; No safety device equipped (e.g. buses); Not applicable (dummy for parked vehicles); Choice is other than the preceding values; unknown (runaway vehicle); not reported by jurisdiction.
P_USER	Road user class	6 levels: Motor Vehicle Driver; Motor Vehicle Passenger; Pedestrian; Bicyclist; Motorcyclist; Not stated/Other/Unknown.

Note:

Source NCDB available from https://open.canada.ca/data/en/data set/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a

Source data files for analysis also available from https://drive.google.com/open?id=12aJtVBaQ4Zj0xa7mtfqxh0E48hKCb_XV

References

Amoh-Gyimah, R., Aidoo, E.N., Akaateba, M.A., Appiah, S.K., 2017. The effect of natural and built environmental characteristics on pedestrian-vehicle crash severity in ghana. International Journal of Injury Control and Safety Promotion 24, 459–468. doi:10.1080/17457300.2016.1232274

Aziz, H.M.A., Ukkusuri, S.V., Hasan, S., 2013. Exploring the determinants of pedestrian-vehicle crash severity in new york city. Accident Analysis and Prevention 50, 1298–1309. doi:10.1016/j.aap.2012.09.034

Beguería, S., 2006. Validation and evaluation of predictive models in hazard assessment and risk management. Natural Hazards 37, 315–329. doi:10.1007/s11069-005-5182-6

Bogue, S., Paleti, R., Balan, L., 2017. A modified rank ordered logit model to analyze injury severity of occupants in multivehicle crashes. Analytic Methods in Accident Research 14, 22–40. doi:10.1016/j.amar.2017.03.001

Casetti, E., 1972. Generating models by the expansion method: Applications to geographic research. Geographical Analysis 4, 81–91.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027. doi:10.1016/j.aap.2006.04.009

Chen, F., Song, M.T., Ma, X.X., 2019. Investigation on the injury severity of drivers in rear-end collisions between cars using a random parameters bivariate ordered probit model. International Journal of Environmental Research and Public Health 16. doi:10.3390/ijerph16142632

Chiou, Y.C., Hwang, C.C., Chang, C.C., Fu, C., 2013. Modeling two-vehicle crash severity by a bivariate generalized ordered probit approach. Accident Analysis and Prevention 51, 175–184. doi:10.1016/j.aap.2012.11.008

Dissanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes. Accident Analysis and Prevention 34, 609–618. doi:10.1016/s0001-4575(01)00060-4

Duddu, V.R., Penmetsa, P., Pulugurtha, S.S., 2018. Modeling and comparing injury severity of at-fault and not at-fault drivers in crashes. Accident Analysis and Prevention 120, 55–63. doi:10.1016/j.aap.2018.07.036

Effati, M., Thill, J.C., Shabani, S., 2015. Geospatial and machine learning techniques for wicked social science problems: Analysis of crash severity on a regional highway corridor. Journal of Geographical Systems 17, 107–135. doi:10.1007/s10109-015-0210-x

Gong, L.F., Fan, W.D., 2017. Modeling single-vehicle run-off-road crash severity in rural areas: Accounting for unobserved heterogeneity and age difference. Accident Analysis and Prevention 101, 124–134. doi:10.1016/j.aap.2017.02.014

Haleem, K., Gan, A., 2013. Effect of driver's age and side of impact on crash severity along urban freeways: A mixed logit approach. Journal of Safety Research 46, 67–76. doi:10.1016/j.jsr.2013.04.002

Hedeker, D., Gibbons, R.D., 1994. A random-effects ordinal regression-model for multilevel analysis. Biometrics 50, 933–944.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis and Prevention 108, 27–36. doi:10.1016/j.aap.2017.08.008

Islam, S., Jones, S.L., Dye, D., 2014. Comprehensive analysis of single- and multi-vehicle large truck at-fault crashes on rural and urban roadways in alabama. Accident Analysis & Prevention 67, 148–158. doi:https://doi.org/10.1016/j.aap.2014.02.014

Khan, G., Bill, A.R., Noyce, D.A., 2015. Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity. Transportation Research Part C-Emerging Technologies 50, 86–96. doi:10.1016/j.trc.2014.10.003

Kim, J.K., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in california: A mixed logit analysis of heterogeneity due to age and gender. Accident Analysis and Prevention 50, 1073–1081. doi:10.1016/j.aap.2012.08.011

Kim, K., Nitz, L., Richardson, J., Li, L., 1995. PERSONAL and behavioral predictors of automobile crash and injury severity. Accident Analysis and Prevention 27, 469–481. doi:10.1016/0001-4575(95)00001-g

Lee, C., Li, X.C., 2014. Analysis of injury severity of drivers involved in single- and two-vehicle crashes on highways in ontario. Accident Analysis and Prevention 71, 286–295. doi:10.1016/j.aap.2014.06.008

Lenguerrand, E., Martin, J.L., Laumon, B., 2006. Modelling the hierarchical structure of road crash data - application to severity analysis. Accident Analysis and Prevention 38, 43–53. doi:10.1016/j.aap.2005.06.021

Li, L., Hasnine, M.S., Habib, K.M.N., Persaud, B., Shalaby, A., 2017. Investigating the interplay between the attributes of at-fault and not-at-fault drivers and the associated impacts on crash injury occurrence and severity level. Journal of Transportation Safety & Security 9, 439–456. doi:10.1080/19439962.2016.1237602

Ma, J.M., Kockelman, K.M., Damien, P., 2008. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. Accident Analysis and Prevention 40, 964–975. doi:10.1016/j.aap.2007.11.002

Maddala, G.S., 1986. Limited-dependent and qualitative variables in econometrics. Cambridge university press.

Mannering, F., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1–16. doi:10.1016/j.amar.2016.04.001

Montella, A., Andreassen, D., Tarko, A.P., Turner, S., Mauriello, F., Imbriani, L.L., Romero, M.A., 2013. Crash databases in australasia, the european union, and the united states review and prospects for improvement. Transportation Research Record 128–136. doi:10.3141/2386-15

Mooradian, J., Ivan, J.N., Ravishanker, N., Hu, S., 2013. Analysis of driver and passenger crash injury severity using partial proportional odds models. Accident Analysis and Prevention 58, 53–58. doi:10.1016/j.aap.2013.04.022

Osman, M., Mishra, S., Paleti, R., 2018. Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. Accident Analysis and Prevention 118, 289–300. doi:10.1016/j.aap.2018.05.004

Penmetsa, P., Pulugurtha, S.S., Duddu, V.R., 2017. Examining injury severity of not-at-fault drivers in two-vehicle crashes. Transportation Research Record 164–173. doi:10.3141/2659-18

Provost, F., Kohavi, R., 1998. Glossary of terms. Journal of Machine Learning 30, 271–274.

Rana, T.A., Sikder, S., Pinjari, A.R., 2010. Copula-based method for addressing endogeneity in models of severity of traffic crash injuries application to two-vehicle crashes. Transportation Research Record 75–87. doi:10.3141/2147-10

Rifaat, S.M., Chin, H.C., 2007. Accident severity analysis using ordered probit model. Journal of Advanced Transportation 41, 91–114. doi:10.1002/atr.5670410107

Roorda, M.J., Paez, A., Morency, C., Mercado, R., Farber, S., 2010. Trip generation of vulnerable populations in three canadian cities: A spatial ordered probit approach. Transportation 37, 525–548. doi:10.1007/s11116-010-9263-3

Sasidharan, L., Menendez, M., 2014. Partial proportional odds model-an alternate choice for analyzing pedestrian crash injury severities. Accident Analysis and Prevention 72, 330–340. doi:10.1016/j.aap.2014.07.025

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39, 955–963. doi:10.1016/j.aap.2006.12.016

Savolainen, P.T., Mannering, F., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention 43, 1666–1676. doi:10.1016/j.aap.2011.03.025

Shamsunnahar, Y., Eluru, N., Pinjari, A.R., Tay, R., 2014. Examining driver injury severity in two vehicle crashes - a copula based approach. Accident Analysis and Prevention 66, 120–135. doi:10.1016/j.aap.2014.01.018

Tang, J.J., Liang, J., Han, C.Y., Li, Z.B., Huang, H.L., 2019. Crash injury severity analysis using a two-layer stacking framework. Accident Analysis and Prevention 122, 226–238. doi:10.1016/j.aap.2018.10.016

Tay, R., Choi, J., Kattan, L., Khan, A., 2011. A multinomial logit model of pedestrian-vehicle crash severity. International Journal of Sustainable Transportation 5, 233–249. doi:10.1080/15568318.2010.497547

Torrao, G.A., Coelho, M.C., Rouphail, N.M., 2014. Modeling the impact of subject and opponent vehicles on crash severity in two-vehicle collisions. Transportation Research Record 53–64. doi:10.3141/2432-07

Train, K., 2009. Discrete choice methods with simulation, 2nd Edition. ed. Cambridge University Press, Cambridge.

Wang, K., Yasmin, S., Konduri, K.C., Eluru, N., Ivan, J.N., 2015. Copula-based joint model of injury severity and vehicle damage in two-vehicle crashes. Transportation Research Record 158–166. doi:10.3141/2514-17

Wang, X.K., Kockelman, K.M., 2005. Use of heteroscedastic ordered logit model to study severity of occupant injury - distinguishing effects of vehicle weight and type, in: Statistical Methods; Highway Safety Data, Analysis, and Evaluation; Occupant Protection; Systematic Reviews and Meta-Analysis, Transportation Research Record. pp. 195–204.

White, S., Clayton, S., 1972. Some effects of alcohol, age of driver, and estimated speed on the likelihood

668 of driver injury. Accident Analysis & Prevention 4.

669 Wu, Q., Chen, F., Zhang, G.H., Liu, X.Y.C., Wang, H., Bogus, S.M., 2014. Mixed logit model-based
670 driver injury severity investigations in single- and multi-vehicle crashes on rural two-lane highways. Accident
671 Analysis and Prevention 72, 105–115. doi:10.1016/j.aap.2014.06.014

672 Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury
673 severity. Accident Analysis & Prevention 59, 506–521. doi:https://doi.org/10.1016/j.aap.2013.06.040

CRediT Author Statement

5/8/2020

Antonio Paez: Conceptualization, Methodology, Validation, Formal analysis, Data Curation, Writing - Original Draft, Project Administration **Hanny Hassan:** Conceptualization, Writing - Review & Editing, Funding acquisition **Mark Ferguson:** Writing - Review & Editing **Saiedeh Razavi:** Conceptualization, Supervision, Funding acquisition