Response to reviewers

5/8/2020

We wish to thank four anonymous reviewers for their feedback on our paper. In this letter we respond to their comments and describe the changes made to the paper in response. Reviewer comments are in black and our responses in blue.

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 new, 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 tried 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'.

> 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 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.

Thank you for your thoughtful comments on our paper.

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

Thank you for this comment. We have changed "accidents" to "collisions" and "crashes".

"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.

Furthermore, the rewritten introduction now includes this:

> 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.

"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. As noted above, 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.

Thank you for this comment. To address it, the introduction was substantially rewritten to remove some background information 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. The strategies contribute to the empirical context by systematizing the different approaches to specify a model and prepare the data for the analysis of crashes involving two parties.

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

To bring together several items in the second summation in the formula. We changed the description to avoid confusion.

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. We might ask why the mechanisms of injury for the two groups are different, and our answer lies, in part, in the opposite party: it is not the same thing when to pedestrians collide than when a heavy vehicle collides with a pedestrian. By considering the atributes of the opponent explicitly, we provide more contextual information and account for the differences in mechanisms of injury severity. This is done implicitly when subsetting the data, since then we consider exclusively one type of collision, say, a heavy vehicle hitting a pedestrian.

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 seldom reported in detail since most everybody assumes that it is standard. As we were working on this research, we found this aspect of the project is as complex as data analysis and modelling, and as important. Since we could not find good, complete discussions of this topic, and data preprocessing is an essential component of the strategies discussed in the paper, we would rather maintain this as part of the paper. We feel that this is particularly important given our push for reproducible research.

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

The statement actually says:

> "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".

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.

Overfitting is a combination of sample size and the number of parameters. 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. It is challenging to look at the number of coefficients and conclude from that alone that there is overfitting. Instead, evidence of overfitting is 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 results of our assessment strongly suggest that there is some overfitting when we take an ensemble approach, since the performance of this approach deteriorates more when backcasting. But this is relative: even when "overfitting" the models still produce good results according to APE and WAPE. The set of verification statistics shows that the ensemble models do not perform as well, but also their performance is not abysmal.

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?

> This is a tempting idea. Unfortunately, Models 3 and 4 cannot be combined due to perfect multicollinearity.

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 hope that after taking into account your thoughtful suggestions you will find that the paper has improved.

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 strategies discussed in the paper. The process, as the paper hopefully illustrates, is not necessarily straightforward. Our preference is to keep this aspect of the process well documented and transparent, which is also in line with our push for reproducible research.

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.

   The models actually do not do a poor job when backcasting. See Table 6: the Weighted Average Prediction Errors (although higher than for the nowcasting case) are still less than one percent for every model/ensemble of 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?

   \textcolor{blue}{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.