Australasian Transport Research Forum 2023 Proceedings
29 November – 1 December, Perth, Australia
Publication website: http://www.atrf.info

Improving crash frequency model estimation through multi-objective extensive hypothesis testing

Zeke Ahern[1], Paul Corry[2], Alexander Paz[1]

[1]School of Civil & Environmental Engineering, Queensland University of Technology, Brisbane, Australia

[2]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia

Email for correspondence (presenting author): zeke.ahern@hdr.qut.edu.au

1. Introduction

Poisson regression is commonly used to model crashes as road accidents typically follow a Poisson process in homogeneous conditions (Lord et al., 2005). However, this approach becomes complicated when the data contains too many zero values and is not evenly distributed. To handle over-dispersed data, the Negative Binomial (NB) model is a better option. Recent research has extended the traditional NB approach to account for possible unobserved heterogeneity by using random parameters and considering possible heterogeneity in the means and variances of these parameters (Behara et al., 2021). One of the main advantages of random parameter models is their ability to provide a more flexible and accurate representation of the data. Several studies have suggested advanced variations of random parameter models, such as the correlated random parameters approach, which is useful for accounting for the correlation between different sources of unobserved heterogeneity (Huo et al., 2020).

Despite extensive research and development to incorporate all these model variations, the time, knowledge, and complexity required for an analyst to perform these tasks can be limiting. Consider the systematic approach an analyst must take just to develop one model, which involves data collection, model selection, model fitting, model evaluation, and model implementation. Each stage requires an extensive number of decisions, including hypotheses about likely contributing factors, unobserved heterogeneity for those factors, possible distributional assumptions, data transformations, statistical model assumptions, and techniques for handling excessive zeros. An iterative process has been recommended, but the number and characteristics of the hypotheses tested are subject to available resources, human bias, and lack of exploration due to local optima (Paz et al., 2019). Furthermore, there is a restricted selection of hypotheses for model specification with limited or no out-of-sample validation (Chai and Draxler, 2014).

To build precise and efficient crash prediction models without sacrificing interpretability, an optimization-based framework is needed. Veeramisti et al. (2020) proposed a promising approach that uses metaheuristic search to estimate clusterwise safety performance functions, enabling the simultaneous estimation of the optimal number of clusters and associated safety performance functions. However, their framework cannot estimate generalized crash prediction models considering multiple likely contributing factors, non-linearities, or random parameters. Nevertheless, a metaheuristic search-based approach is effective in generating complex heterogeneous models in practical time frames. Therefore, a metaheuristic solution algorithm is proposed to effectively test many hypotheses and capture the strengths of various modelling approaches while minimizing the sensitivity to human, time, bias, and analysis intervention.

2. Methodology

Consider the equation for the expected number of crashes λ_i for observation (road segment) i modeled by the CRP approach as specified in below:

$$\lambda_{i|\breve{v}_\iota} = \exp\left[(\grave{X}_\iota{}^T\beta + \tilde{X}_\iota{}^T\Gamma\breve{v}_\iota + \breve{\epsilon}_i\right]$$

where:

- $\grave{X}_i$, is the vector of selected explanatory terms. Note that $\grave{X}_\iota \subset X_i$, where X_i is all potential explanatory variables in observation i.
- $\tilde{X}_\iota{}^T \subset \grave{X}_i$ Such that it only includes the random paramater explanatory variables.
- $\Gamma\breve{v}_\iota$ denotes a stochastic term following an analyst-specified distribution.
- Γ is a parameter matrix that provides the variance-covariance and possible correlation matrix of random parameters in the distribution of β_i
- $\breve{v}_\iota$ represents an unobservable random variable with a size of $K \times 1$ (where K denotes the number of random parameters). It has an average value of zero and a variance-covariance matrix that equals an identity matrix. This implies that the mean and variance-covariance matrix Σ for random parameters to be $E(\beta_i|\nu_i) = \beta_i$ and $Var(\beta_i|\nu_i) = \Gamma\Gamma^T$, respectively. $\breve{v}_\iota$can also vary with an appropriate distribution of the analyst.
- $\breve{\epsilon}_\iota$ potential error term following an analyst's distribution; For example, Gamma if Negative Binomial.

Equation 1 is presented to demonstrate the potential flexibility of the proposed approach. Specifically, when $\breve{v}_\iota = \varnothing$, the equation reduces to a fixed effects model subject to the error term selected by an analyst $\breve{\epsilon}_i$. Moreover, if $\breve{\Gamma}$ is a parameter matrix with all zeros except for coefficients along the diagonals, the model is reduced to random parameters count model.
Therefore, a mathematical programming problem was formulated and solved with the assistance of a Harmony Search solution algorithm to construct the most appropriate values for $\breve{\Gamma}$, $\breve{v}_\iota$, $\breve{X}_\iota$, and $\breve{\epsilon}$. This aids in the effective use of the CRP model, optimized in accordance with the objective function. This strategy entails assessing various modeling factors, including the type of probabilistic model, pertinent factors and their transformations, the possibility of correlated or uncorrelated random parameters, and their corresponding parametric distributions. The objective is to determine β, $\breve{\Gamma}$, $\breve{v}_\iota$, $\breve{X}_\iota$, and $\breve{\epsilon}$, with the use of decision variables, and therefore, it conducts searches independently of the analyst, time bias, and knowledge. Consequently, these variables will adapt based on the available data, ensuring the best fit.

3. Results

A dataset, as analyzed in Behara (2021), was utilized, provided by the Queensland Department of Transport and Main Roads (TMR). This dataset encompasses 1,875 unique roadway segments in Queensland and outlines potential factors that might explain head-on collisions. Included in the dataset are 67 potentially associated factors. The emphasis of this experiment was to investigate the application of multiple objectives, a concept not explored in the original study. A Harmony Search (HS) solution algorithm was employed to solve the mathematical program.

Figure 1: Pareto Frontier found by the Harmony Search Algorithm

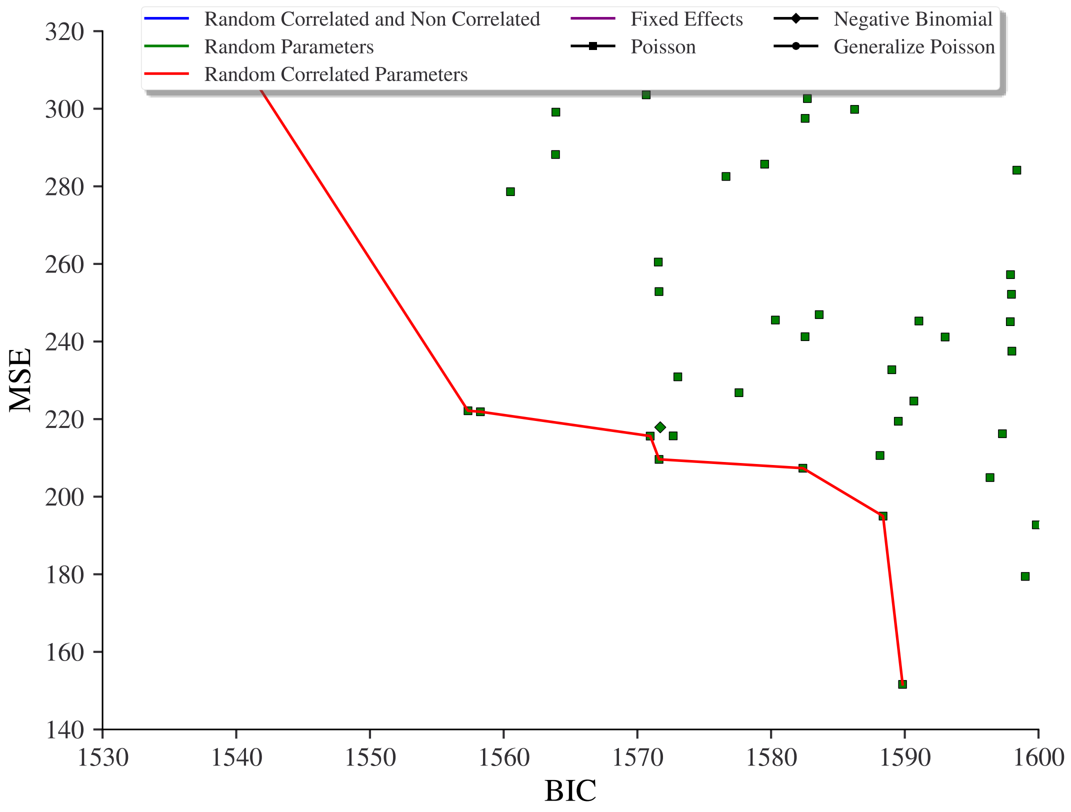

Table 1: Poisson model found through the HS algorithm. BIC: 574.25 Loglikihood: -241.98

Effect	Transformation	Coefficient	Standard Error	z-values	Prob |z|>Z
Constant	no	-7.80	3.22	-2.42	0.0156*
MCV (multi-combination vehicular traffic)	log	1.11	0.14	8.13	0.0000***
SP (highway speed)	sqrt	0.27	0.41	50.00	0.0000***
M_Curve (mountainous terrain curve)	no	0.32	0.00	50.00	0.0000***
RT (rolling terrain)	no	0.68	0.32	2.13	0.0330*
Nlanes (number of lanes)	sqrt	-3.36	1.36	-2.47	0.0137*
ATLM (audio tactile line markings)	no	0.57	0.00	50.00	0.0000***
UD (urban double)	no	-15.40	0.00	-50.00	0.0000***
UD random parameter (Std Dev. Triangular)		0.44	0.00	50.00	0.0000***

In the analysis, a solution from the Pareto front most consistent with the existing literature on road safety was chosen. This approach has its merits, as it ensures the selection of the most realistic solution. It's crucial to emphasize that this method doesn't supplant the role of an analyst, but rather offers a supplementary modelling approach to attain efficient estimation. The framework merely aids in efficient modelling with minimal prior input. The responsibility of selecting the most suitable modelling approach still lies with the analysts. The solution selected for this analysis, as presented in section 3, was based on this analysis, and an examination of the effects was conducted as follows.

The analysis probed the impacts of various explanatory variables on the likelihood of crashes. The results were in line with similar studies in the literature. Primarily, it was determined that rolling terrain (RT) augmented the probability of crashes. Studies such as Agbelie (2016); Azimi et al. (2020) corroborated these findings, attributing them to limited visibility and restricted alignment.

The present study's framework has identified that road curvature on mountainous terrain (M_Curve) significantly increases the likelihood of vehicular crashes. This finding is consistent with prior research conducted by Rusli et al. (2018) in Malaysia, which has demonstrated the effectiveness of passing lanes in reducing the probability of multi-vehicle crashes on rural mountainous highways. As such, the researchers proposed the inclusion of passing lanes as a crucial factor in road design for such areas. By utilizing this framework and validating the model through the previous study, potential solutions for road design have been identified. However, it is important to note that the applicability of these findings is currently limited to Malaysia, although they have provided novel insights that were previously unknown to us.

The framework hypothesised that formation width (FW) rural single carriageways (RS) the formation width of rural single roads increased the likelihood of crashes, although the increase was only slight. This finding is consistent with the literature, which suggests that wider roads promote risk-driving behaviour and reduced driver attention. Cost-saving measures could be considered as a result (Behara et al., 2021).

The hypothesis put forth by the framework is that the use of multi-combination vehicular traffic (MCV) increases the probability of vehicular accidents, which is supported by existing literature highlighting factors such as their size and weight, limited visibility, longer stopping distances, driver fatigue, and mechanical issues (AASHTO, 2010; Akgüngör and Doan, 2009). It is noteworthy that while this hypothesis aligns with the existing body of literature, the extensive hypothesis testing framework has established a logarithmic association between MCV usage and crash risk. Consequently, it can be inferred that a substantial increase in MCV occupancy has a diminishing impact on crash risk. This finding corroborates the notion that extant standards governing MCV utilization are sustainable. It is imperative to mention that the same phenomenon has been explicated in the original study by Behara et al. (2021).

In this study, our framework revealed that increasing number of lanes (NLANES) on a roadway has the potential to decrease the frequency of crashes. This is primarily due to the reduction of congestion, increased capacity, traffic separation, and more opportunities for drivers to regain control (Mecheri et al., 2017). However, the effect of MCV use on the probability of a crash may differ based on various factors, including traffic volumes, speed limits, driver behaviour, and roadway design. Our analysis considers unobserved heterogeneity that may arise from the number of lanes, which may increase the risk to roads, as captured by our random parameter term that followed a triangular distribution. Furthermore, it is essential to consider that adding lanes can be an expensive approach and may not always be the most effective solution for improving safety or reducing congestion, and a comprehensive approach should be used with this information (Mannering and Washburn, 2020).

Next, the framework found that urban double carriageways (UD) greatly improved road safety as opposed to the alternative urban single carriageways (US) and RS available within the data. This is likely due to the separate travel lines, limited access, design standards, and speed management (Rezapour et al., 2019).

Lastly, the framework found that presence of Audio Tactile Lane Marking (ATLM) increased crash likelihood. However, this could be a spurious correlation, and other factors related to road safety might be driving the observed increase in crash likelihood, rather than the ATLM themselves. This could be due to endogeneity, as they are often placed in areas that are prone to crashes as mentioned in Behara et al. (2021). Therefore, careful consideration must be evaluated to determine if ATLM markings are causing drivers to behave in riskier ways.

4. Conclusion

In summary, this journal paper proposes an optimization framework to assist with the development of crash data count models. Traditional methods are prone to bias and may overlook unique specifications and insights present in the data. The proposed framework incorporates a mathematical programming formulation that minimizes two objectives, the Bayesian Information Criterion (BIC) and mean-square prediction error (MSE) and uses metaheuristic solution algorithms to efficiently search for an efficient solution. The effectiveness of the framework is evaluated using two real-world datasets and a synthetic dataset, demonstrating its ability to estimate crash data count models accurately and efficiently, reducing computational complexity, bias, and sub-standard processes for analysts to retrieve efficient estimation. While there were limitations to the search process and the multi-objective approach, the proposed framework provides an efficient and effective approach to developing crash data count models and has the potential to provide valuable insights for researchers and practitioners in the field. Overall, the framework has performed well given the complexity and challenges of the problem, indicating its potential to improve future analysis in this field.

In the analysis of the Queensland data, likely contributing factors to predict crash likelihood were evaluated, and a solution from the Pareto front, most consistent with existing literature on road safety, was selected. An investigation was then conducted on the selected solution to examine the impact of explanatory variables on crash likelihood. It was found that road curvature on mountainous terrain significantly escalates the likelihood of vehicular crashes. The analysis also unveiled that broader roads spur risk-driving behavior and decrease driver attention, and the employment of MCV enhances the probability of vehicular accidents. Moreover, adding more lanes to a roadway harbors the potential to lessen the frequency of crashes. However, this can be a costly method and may not always be the most effective solution for enhancing safety or easing congestion. Finally, it was discovered that ATLM markings amplified crash likelihood. Therefore, it is essential to assess if they incite drivers to adopt riskier behaviors.

5. References

Lord, D., S.P. Washington, and J.N. Ivan, Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, Vol. 37, No. 1, 2005, pp. 35–46.

Behara, K. N., A. Paz, O. Arndt, and D. Baker, A random parameters with heterogeneity in means and Lindley approach to analyze crash data with excessive zeros: A case study of head-on heavy vehicle crashes in Queensland. *Accident Analysis & Prevention*, Vol. 160, 2021, p. 106308.

Huo, X., J. Leng, Q. Hou, and H. Yang, A Correlated Random Parameters Model with Heterogeneity in Means to Account for Unobserved Heterogeneity in Crash Frequency Analysis. *Transportation Research Record*, Vol. 2674, No. 7, 2020, pp. 312–322.

Paz, A., C. Arteaga, and C. Cobos, Specification of mixed logit models assisted by an optimization framework. *Journal of Choice Modelling*, Vol. 30, 2019, pp. 50–60.

Chai, T. and R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE) - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, Vol. 7, No. 3, 2014, pp. 1247–1250.

Veeramisti, N., A. Paz, and J. Baker, A framework for corridor-level traffic safety network screening and its implementation using Business Intelligence. *Safety Science*, Vol. 121, 2020, pp. 100–110.

Agbelie, B. R. D. K., Random-parameters analysis of highway characteristics on crash frequency and injury severity, 2016.

Azimi, G., A. Rahimi, H. Asgari, and X. Jin, Severity analysis for large truck rollover crashes using a random parameter ordered logit model. *Accident Analysis & Prevention*, Vol. 135, 2020, p. 105355.

Rusli, R., M. M. Haque, A. P. Afghari, and M. King, Applying a random parameters Negative Binomial Lindley model to examine multi-vehicle crashes along rural mountainous highways in Malaysia. *Accident Analysis & Prevention*, Vol. 119, 2018, pp. 80–90.

AASHTO, *Highway Safety Manual*, Vol. 2. Washington, D.C., 2010.

Akgüngör, A. P. and E. Doan, An application of modified Smeed, adapted Andreassen and artificial neural network accident models to three metropolitan cities of Turkey. *Scientific Research and Essay*, Vol. 4, No. 9, 2009, pp. 906–913.

Mecheri, S., F. Rosey, and R. Lobjois, The effects of lane width, shoulder width, and road cross-sectional reallocation on drivers' behavioral adaptations. *Accident Analysis and Prevention*, Vol. 104, 2017, pp. 65–73.

Mannering, F. and S. Washburn, *Principles of highway engineering and traffic analysis*, Vol. 28. Wiley, 2020.

Rezapour, M., M. Moomen, and K. Ksaibati, Ordered logistic models of influencing factors on crash injury severity of single and multiple-vehicle downgrade crashes: A case study in Wyoming. *Journal of Safety Research*, Vol. 68, 2019, pp. 107–118.