Australasian Transport Research Forum 2022 Proceedings
28-30 September, Adelaide, Australia
Publication website: http://www.atrf.info

Simulated Annealing for Modeling Crash-Counts

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: zeke.ahern@hdr.qut.edu.au

Abstract

Efficient estimation of modeling crash counts is complex and requires advanced knowledge, experience, ad hoc, and time-consuming processes. Challenges involved in developing the model include; identification of significant road factors for model specification, an adequate functional form and the type of models such as Poisson or Negative Binomial with the inclusion of random parameters. To overcome these barriers, the model specification is selected using a simulated annealing algorithm to identify the likely contributing factors to crashes on roads in Queensland, AU. The applicability of the approach is demonstrated by comparing it with the results found in Behara et al. (2021). As a result of the solution algorithm, human biases are minimized since decisions are updated based on goodness of fit rather than context-specific information. Additionally, time is reduced as more models can be tested more efficiently, and new solutions that are worse than the incumbent solution are considered, which enables the escape of local optimal regions that an analyst may not be able to identify.

1. Introduction

Statistical models such as Poisson regression (Arndt, 2004), Negative binomial (NB) or its variants (Yu et al., 2019), are the most common approach for estimating crash counts (Mannering and Bhat, 2014). This is because they are easy to estimate and can be derived to handle overdispersion in the model. These models only handle linear relationships between features and do not allow parameter estimates to vary between observations; recent studies have emphasized the need to account for unobserved heterogeneity in the data (Malyshkina and Mannering, 2010; Zeng et al., 2017). In these approaches, the distribution of the random parameters and the selection of potential explanatory variables must be assumed before estimation. This model specification process is highly dependent on human judgement to include context-specific information in the model. Furthermore, it is time-consuming, subject to ad hoc trial-and-error approaches, of which local-optima solutions are likely to be proposed. Therefore, a method that reduces the time and effort required to find an efficient solution while avoiding human biases with the available information is desirable.

The simulated annealing (SA) algorithm was inspired by the metal annealing process (Kirkpatrick et al., 1983). High temperatures cause metals to become liquidized and mold into a new structure during the annealing process. The temperature is gradually reduced until the new structure solidifies, allowing the metal to retain its newly acquired qualities while reducing structural flaws. SA was selected for this problem due to its success in similar problems (Khadka et al., 2020), convergence capability, and few hyperparameters that can be easily tuned (Lokupitiya et al., 2005).

This study proposes the SA solution algorithm to find the best model specification by optimizing the Bayesian information criterion (BIC). BIC is a criterion for model selection that prevents overfitting by penalizing the number of parameters (Markon and Krueger, 2004). The specification includes the most significant explanatory variables, the functional form, type of model, and selection of distributions for the random parameters. The specification is therefore a set of decision variables and thus can be formulated as a mathematical program.

2. Methodology

2.1. Notation

The following notation is used to formulate the problem into a mathematical program:

Table 1: Notation and definitions.

Notation	Definitions
N	number of road observations
X	vector of potential explanatory variables (road factors)
K	number of explanatory variables
x_k	The explanatory variable, $\forall k \in K$
$\hat{x_k}$	The transformed explanatory variable, $\forall k \in K$
α_k	Indicator variable taking value 1 if potential explanatory variable x_k is included; 0 otherwise
r_k	Indicator variable taking value 1 if potential explanatory variable x_k has an associated random parameter; 0 otherwise
D	set of distributions
d_k	variable to select a distribution from D applied to the associated random parameter.
M	Set of possible model types (e.g Poisson).
T	set of transformations
τ_k	variable to select the transformation from T applied to α_k
β	vector of coefficients for potential explanatory variables

The problem is formulated as a bilevel optimization problem. The lower-level objective function is maximum simulated log likelihood (MSL) to find β, constrained by the following decisions:

- α_k: the explanatory variables that specify the model
- τ_k: the transformations on α_k to create the functional form
- r_k: the indication that an explanatory variable has an associated random parameter
- d_k: the distribution drawn for the random parameter

The objective function at the upper level is minimizing BIC.

2.2. Mathematical Program

$$Min \quad BIC = -2ln(L) + \zeta ln(|I|)) \tag{1}$$

subject to:

$$\alpha_k = \begin{cases} 1, & \text{if } x_k \text{ is included;} \\ 0, & \text{otherwise} \end{cases} \qquad \forall n \in N \tag{2}$$

$$r_k = \begin{cases} 1, & \text{if } x_k \text{ is signified having a random parameter;} \\ 0, & \text{otherwise} \end{cases} \qquad \forall n \in N \tag{3}$$

$$\hat{x_k} = f(\tau_k, x_k) \qquad \forall k \in K, n \in N \tag{4}$$

$$\alpha_k \geq r_k \qquad \forall n \in N \tag{5}$$

$$\rho_k \leq U \qquad \forall n \in N \tag{6}$$

$$\rho_k \geq 0 \qquad \forall n \in N \tag{7}$$

$$\zeta = \sum_{k \in K} (\alpha_k + r_k) + q \tag{8}$$

$$f(m) = M_m \tag{9}$$

$$m \leq |M| \tag{10}$$

$$f(d_k) = D_{d_k} \qquad \forall k \in K \tag{11}$$

$$d_k = |D| \qquad \forall k \in K \tag{12}$$

$$q = \begin{cases} 0, & \text{if } f(m) \text{ = Poisson;} \\ 1, & \text{otherwise} \end{cases} \tag{13}$$

2.3. Data

Transport Main Roads provided the data, which detail head-on collisions in the state of Queensland. This contains 1875 unique road segments with 67 explanatory variables that may explain a head-on collision. The data was used to compare with an analyst to find the best model in a previous study (Behara et al., 2021). In the analysts study, it was found that a random parameter model following the Lindley distribution provides better estimation ability as opposed to a traditional Normal distribution. Therefore, we aim for SA to reveal a similar result in order to save time and effort in future models with different datasets. For a complete description of the data, see Behara et al. (2021), we have provided a summary of the number of head-on collisions in Figure 1.

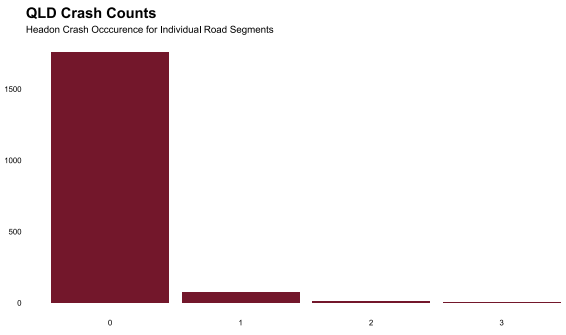

Figure 1. Excessive zeros for QLD crash data.

2.4. Optimisation Algorithm

In this study, we use the SA algorithm with a heuristic that calculates the initial temperature, as seen in Ahern et al. (2022). Calculating the initial temperature using a heuristic reduces the number of sensitive hyperparameters in the algorithm. Subsequently, our termination criterion

is time-based with 50000 seconds of CPU time dedicated. Other hyperparameters, such as the temperature reduction factor and the number of iterations before reducing the temperature, require tuning. Due to the sensitivity of these hyperparameters, a covering array was constructed and five seeded experiments were carried out for each value of the covering array. The experiment that contained the lowest mean BIC value was reported and compared with (Behara et al., 2021), which appears in Table 2[1].

Table 2. Behara (2021) NB-Lindley model calculated with MSL; BIC is 747.24

Effect	Coefficient	Std. Error	z-values	Pr $\|z\| > Z$
Constant	-12.03409	1.18888	-5.00	0.0000***
US	1.93452	0.82065	2.36	0.0184*
RSMS	2.26579	1.03969	2.18	0.0293*
LNMCV	9.03728	1.33912	6.75	0.0000***
RSHS	3.89542	0.76918	5.00	0.0000***
LNAADT	2.55426	1.17746	2.17	0.0301*
Curve50	-0.75197	0.52946	-1.42	0.1555
RSHS ($Std.Dev.$) Lindley	-0.60642	0.20195	-3.00	0.0027**
LNAADT ($Std.Dev.$) Lindley	0.01078	0.14085	0.08	0.9390
Curve50 ($Std.Dev.$) Lindley	0.33032	0.31032	1.06	0.2871
NB ScalParm	0.21355	0.30613	0.70	0.4854

In order for SA to work, we must define the sets in Table 1, which can be observed as follows:

- $D = \{uniform, normal, triangular, lindley, gamma\}$
- $T = \{fixed, log, sqr, factorial, exp, squared, cubed\}$
- $M = \{Poisson, NB, GP, COMP\}$

The best hyperparameters for the SA algorithm were found using the covering arrays, which deduce that the temperature reduction factor α and the steps taken before the temperature is reduced T_{step} should be 0.99 and 10 respectively. The convergence plot for these hyperparameters can be seen in Figure 2. This result confirms that SA was able to use exploration that deviated from the current best solution before discovering a new best.

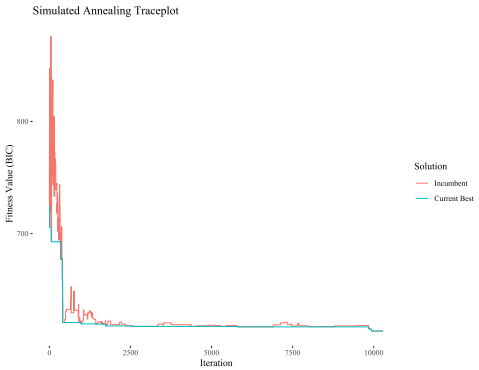

Figure 2. BIC optimised through SA

[1]We converted the model in the paper to estimate MSL in order to accurately benchmark against.

Table 3. Poisson model found through the simulated annealing algorithm; BIC is 613.97

Effect	Transformation	Coefficient	Std.Error	z-values	Pr $\|z\| > Z$
LNLYR	sqrt	10.82676	0.00000	5.00	0.0000***
LNKMYR	log	-18.63380	0.00000	-5.00	0.0000***
LNADT	sqrt	5.02481	0.00000	5.00	0.0000***
Nlanes2	no	3.19753	0.00000	5.00	0.0000***
L50	no	0.31270	0.00000	5.00	0.0000***
SW_{PS}	cubed	-0.08866	0.00000	-5.00	0.0000***
L50 ($Std.Dev.$) Uniform		0.03772	0.00000	5.00	0.0000***
SW_{PS} ($Std.Dev.$) Lindley		0.04300	0.00000	5.00	0.0000***

Table 3 shows the final model. SA was able to significantly improve the fit of the model from the BIC value of 747.24 to 613.97. The explanatory variable LNAADT[2] was identified in the analyst specification to be included in the final model. This variable is the most frequently included in the literature, so it was reassuring that SA confirmed this (Mannering et al., 2020). However, almost all other effects are unique. SA was able to apply a wide range of transformations on the data and fit a mixture of random parameter distributions; for example, the SW_{PS}[3] variable was associated with the Lindley distribution, but the $L50$ [4] random parameter distribution was uniformly associated. Although the Lindley distribution has unique properties, Behara et al. (2021) may have been biased to use this distribution in the reported specification to validate the claim. SA therefore was able to consider a multiple of unique, and time-consuming hypothesis tests to create a specification that is non-biased to existing knowledge from an analyst.

3. Conclusion

In this experiment, we have found that the simulated annealing algorithm is capable of identifying efficient solutions in advanced statistical models related to crash data. Decisions are optimized in the iteration-based framework, which reduces the biases an analyst may introduce in model development. The algorithm is data-dependent, and hence other efficient solutions could be found if other datasets were used. In general, we found that wrapping a metaheuristic around a statistical model saves significant time and effort, even outperforming the results of Behara et al. (2021).

[2]Logarithmic of annual average daily transit.
[3]Shoulder width.
[4]Level terrain, longer than 50% of the corresponding road length.

4. References

Z. Ahern, A. Paz, and P. Corry. Approximate multi-objective optimization for integrated bus route design and service frequency setting. *Transportation Research Part B: Methodological*, 155:1–25, 1 2022. ISSN 0191-2615. doi: 10.1016/J.TRB.2021.10.007.

O. K. Arndt. *Relationship between unsignalised intersection geometry and accident rates*. PhD thesis, Queensland University of Tecnology, Brisbane, 2004.

K. N. Behara, 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*, 160:106308, 9 2021. ISSN 0001-4575. doi: 10.1016/J.AAP.2021.106308.

M. Khadka, A. Paz, and A. Singh. Generalised clusterwise regression for simultaneous estimation of optimal pavement clusters and performance models. *International Journal of Pavement Engineering*, 21(9):1122–1134, 7 2020. ISSN 1477268X. doi: 10.1080/10298436.2018.1521970. URL https://www.tandfonline.com/action/journalInformation?journalCode=gpav20.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 5 1983. ISSN 00368075. doi: 10.1126/science.220.4598.671. URL http://science.sciencemag.org/.

R. S. Lokupitiya, L. E. Borgman, and R. Anderson-Sprecher. Simulation of storm occurrences using simulated annealing. *Journal of Climate*, 18(21):4394–4403, 11 2005. ISSN 08948755. doi: 10.1175/JCLI3546.1.

N. V. Malyshkina and F. L. Mannering. Zero-state Markov switching count-data models: An empirical assessment. *Accident Analysis & Prevention*, 42(1):122–130, 2010. doi: https://doi.org/10.1016/j.aap.2009.07.012. URL http://www.sciencedirect.com/science/article/pii/S0001457509001778.

F. Mannering, C. R. Bhat, V. Shankar, and M. Abdel-Aty. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic methods in accident research*, 25:100113, 2020. ISSN 2213-6657.

F. L. Mannering and C. R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1:1–22, 2014. ISSN 22136657. doi: 10.1016/j.amar.2013.09.001.

K. E. Markon and R. F. Krueger. An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics*, 34(6):593–610, 11 2004. ISSN 00018244. doi: 10.1007/s10519-004-5587-0. URL https://link.springer.com/article/10.1007/s10519-004-5587-0.

R. Yu, Y. Wang, M. Quddus, and J. Li. A marginalized random effects hurdle negative binomial model for analyzing refined-scale crash frequency data. *Analytic methods in accident research*, 22:100092, 2019.

Z. Zeng, W. Zhu, R. Ke, J. Ash, Y. Wang, J. Xu, and X. Xu. A generalized nonlinear model-based mixed multinomial logit approach for crash data analysis. *Accident Analysis & Prevention*, 99:51–65, 2017.