Citation
Rakshit, S. and McSwiggan, G. and Nair, G. and Baddeley, A. 2021. Variable selection using penalised likelihoods for point patterns on a linear network. Australian & New Zealand Journal of Statistics. 63 (3): pp. 417-454. http://doi.org/10.1111/anzs.12341

Variable selection using penalised likelihoods for point patterns on a linear network

Suman Rakshit[1,3]*, Greg McSwiggan[2], Gopalan Nair[2] and Adrian Baddeley[3]

Curtin University and The University of Western Australia

Summary

Motivated by the analysis of a comprehensive database of road traffic accidents, we investigate methods of variable selection for spatial point process models on a linear network. The original data may include explanatory spatial covariates, such as road curvature, and 'mark' variables attributed to individual accidents, such as accident severity. The treatment of mark variables is new. Variable selection is applied to the canonical covariates, which may include spatial covariate effects, mark effects, and mark-covariate interactions. We approximate the likelihood of the point process model by that of a generalised linear model, in such a way that spatial covariates and marks are both associated with canonical covariates. We impose a convex penalty on the log likelihood, principally the elastic-net penalty, and maximise the penalised loglikelihood by cyclic coordinate ascent. A simulation study compares the performances of the lasso, ridge regression and elastic-net methods of variable selection on their ability to select variables correctly, and on their bias and standard error. Standard techniques for selecting the regularisation parameter γ often yielded unsatisfactory results. We propose two new rules for selecting γ which are designed to have better performance. The methods are tested on a small dataset on crimes in a Chicago neighbourhood, and applied to a large dataset of road traffic accidents in Western Australia.

Key words: elastic net; lasso; Poisson process; discretised models; generalised linear model

1. Introduction

The analysis of patterns of points on a network of lines is becoming widespread in applications. The lines could represent roads, rivers, railways, wires, cracks or nerve fibres, and the points give the locations of events or objects observed along these lines. For recent

*Author to whom correspondence should be addressed.
[1] SAGI-West, School of Molecular and Life Sciences, Curtin University, Bentley, WA, Australia
Email: suman.rakshit@curtin.edu.au
[2] Department of Mathematics and Statistics, The University of Western Australia, Crawley, WA, Australia
[3] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, WA, Australia

Acknowledgment. The support and collaboration of Main Roads Western Australia, especially Dr. Thandar Lim and Dr. Sannath Jayamanna, was essential to this project. Research was funded by Australian Research Council grant DP130102322. We also acknowledge support from the Grains Research and Development Corporation.

surveys, see Okabe & Sugihara (2012), Baddeley et al. (2020), and Baddeley, Rubak & Turner (2015, Chap. 17).

Figure 1 shows the locations of 14,562 traffic accidents recorded in Western Australia in 2011. The data, curated by the state government agency Main Roads WA, include spatial coordinates of each road segment; the spatial location of each accident; properties of the road, such as speed limit and curvature; and attributes of each accident, such as severity and time of day.

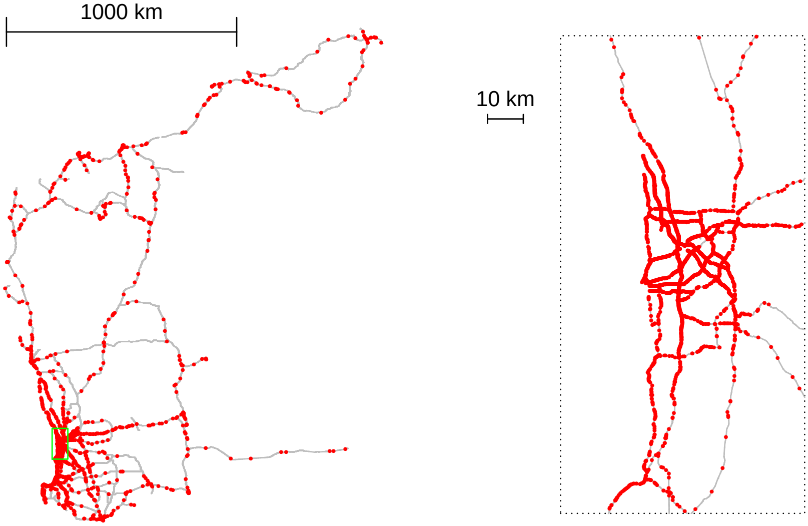

Figure 1. Traffic accidents (dots), recorded in 2011, on state roads (lines) in Western Australia. *Left:* entire state of Western Australia; *Inset right:* Perth metropolitan area.

For agencies that build and manage road networks, the main question of interest is the relationship between accident risk and explanatory covariates such as road geometry (curvature, width, number of lanes, distance to nearest intersection) and traffic management (speed limit, traffic lights, type of intersection). On the other hand, agencies that provide road safety advice tend to focus on attributes of the individual accident, such as the time of day, vehicle type, number of vehicles involved, and attributes of the driver such as fatigue and alcohol use. There is a methodological distinction between explanatory covariates and accident attributes; the techniques described here apply to both kinds of data.

A point pattern on a linear network can be modelled as a realisation of a spatial *point process* on the network (McSwiggan 2019, Baddeley, Rubak & Turner 2015, Chap. 17, Baddeley et al. 2020). The fundamental definition of a point process on a network is a simple special case of the general theory of point processes (Daley & Vere-Jones 2003). Some statistical methodology for point patterns on a network has been developed (Okabe &

Sugihara 2012; Baddeley, Rubak & Turner 2015, Chap. 17). However, the inhomogeneous geometry of the network defeats many of the standard modelling techniques of spatial statistics (Baddeley et al. 2017) and causes substantial computational challenges (Okabe & Satoh 2009; Rakshit, Baddeley & Nair 2019).

As a first step in this paper, we consider an inhomogeneous *Poisson point process* model for the spatial locations of the accidents. The inhomogeneous Poisson process is specified by its spatially-varying intensity or rate $\lambda(u)$, as a function of location u on the network. When the intensity is constant, the process is called a homogeneous Poisson point process and λ denotes the expected number of events per unit length. Explicit models will express $\lambda(u)$ as a function of explanatory spatial covariates $\boldsymbol{Z}(u)$, typically in the loglinear form $\lambda(u) = \exp(\boldsymbol{\beta}^\top \boldsymbol{Z}(u))$, after transformation of covariates. Even when the Poisson process is not an appropriate model, experience with two-dimensional point pattern data suggests that the Poisson likelihood is still an appropriate tool for estimating the intensity (Guan, Jalilian & Waagepetersen 2015; Waagepetersen & Guan 2009).

Variable selection becomes important when the number of explanatory variables is large, and especially when the models of interest include polynomial terms, factors with many levels, and interactions between variables, which all increase the number of canonical variables in the model. This applies to the WA road accident data, which include numerous variables listed in Tables 4 and 5. Some methods of variable selection are available for point process models in two dimensional space, including sufficient dimension reduction (Guan & Wang 2010) and penalised maximum likelihood (Yue & Loh 2015) as well as classical hypothesis tests and Akaike information criteria (Baddeley, Rubak & Turner 2015, pp. 335–338, 371–378, 512–513). In this paper we adapt penalised maximum likelihood methods, including the lasso, ridge regression and elastic net, to point process models on a linear network, and also extend them to the selection of mark variables.

The likelihood of the Poisson point process involves an integral over the network. Adapting an approach often used for one-dimensional and two-dimensional point processes, we shall approximate the integral in such a way that the approximate point process likelihood is formally equivalent to the likelihood of a generalised linear model (GLM), which may then be fitted using standard software (Brillinger & Preisler 1986; Berman & Turner 1992; Lindsey 1992, 1995; Baddeley & Turner 2000; Baddeley et al. 2010; Aarts, Fieberg & Matthiopoulos 2012; Fithian & Hastie 2013; Renner & Warton 2013).

A simple case of this approach arises when the accident locations are aggregated into accident counts for each road segment, and a count regression model is applied to the accident counts. This is an instance of the 'crash-frequency' approach to accident risk analysis (Lord & Mannering 2010) and is equivalent to assuming the accident rate is constant along each road segment. Such analysis is common because the road variables in most road accident databases are stored as constant values per road segment. However, as McSwiggan (2019, Chap. 2) pointed out, there are covariates, possibly influencing road accidents, that vary along a road segment, such as, sighting distance, curvature change rate, and distance to the nearest intersection. Aggregating these covariates at the level of road segments can cause substantial bias and loss of information (Koorey 2009; Baddeley et al. 2010; McSwiggan 2019). Our general approach avoids these inefficiencies by allowing fine spatial discretisation and accurate approximation of the likelihood.

Regularisation can improve the performance of techniques for fitting spatial point process models. For two-dimensional point processes, Baddeley & Turner (2000, pp. 296, 301, 307) used generalised additive models to estimate covariate effects as smooth functions. Renner (2013), Renner et al. (2015) and Yue & Loh (2015) used regularisation methods for variable selection, adapting the lasso (Tibshirani 1996) and elastic net (Zou & Hastie 2005) to two-dimensional point process models. The lasso produces sparse solutions (i.e. estimates of the parameter vector in which relatively few entries are non-zero) and thus supports parameter estimation and variable selection simultaneously. In the case of correlated predictors, the lasso tends to select one predictor from each set of highly-correlated predictors. The elastic net, which provides a compromise between the lasso and the ridge regression penalty (Hastie, Tibshirani & Friedman 2009, Sec. 3.4, pp. 61–79), tends to average the effects of the highly-correlated predictors and selects an averaged predictor for the model (Friedman, Hastie & Tibshirani 2010). Ridge regression typically shrinks the estimated coefficient values of the correlated predictors close to each other.

Algorithms for maximising these penalised likelihoods are well-developed in the case of the general linear model (Hastie, Tibshirani & Friedman 2009, Sec. 3.4.4, pp. 71–79). Recently, Friedman, Hastie & Tibshirani (2010) extended these methods for fitting generalised linear models using an efficient cyclic coordinate descent algorithm, which was shown to be faster than the least angle regression algorithm (Efron, Hastie & Tibshirani 2004). The cyclic coordinate descent algorithm is used in this paper.

Road traffic accident data may also include attributes of each individual accident, such as accident severity, type of collision, and number of vehicles involved. The accident records then constitute a 'marked point pattern' which can be modelled by a 'marked point process' (Stoyan, Kendall & Mecke 1995, Sec. 4.2, pp. 105-109, Baddeley 2010b). The analysis of marked point patterns confers valuable advantages (Baddeley, Rubak & Turner 2015, Chap. 14, Illian et al. 2008), including the ability to estimate the spatially-varying relative risk of different types of events, and to avoid confounding due to the effect of latent variables. In this paper we also develop variable selection for mark variables, which appears to be new.

The paper is organised as follows. Section 2 gives basic definitions including the Poisson point process on a linear network and its likelihood. Section 3 explains how Poisson point process models can be approximated by generalised linear models. The penalised likelihood and cyclical coordinate descent algorithm are described in Section 4. A simulation study comparing the performance of the variable selection methods is presented in Section 5. analysis of the Western Australia traffic accident data is described in Section 6. Variable selection for a *marked* point process model is explained in Section 7. The paper ends with a discussion in Section 8.

2. Point process models on a linear network

2.1. Data structure

The data consist of a linear network L, a spatial pattern of points $\mathbf{x}$ on L, and spatial covariate functions V on L, defined below.

A linear network is defined (Ang, Baddeley & Nair 2012) as the union $L = \bigcup_{i=1}^{m} s_i$ of a finite number m of line segments $s_1, \dots, s_m$ in the plane, where $s_i = [u_i, v_i] = \{w : w = tu_i + (1-t)v_i, 0 \le t \le 1\}$ is the line segment with endpoints u_i, v_i, belonging to the two-dimensional space.

A (finite, simple) *point pattern* $\mathbf{x}$ on L is a finite set $\mathbf{x} = \{x_1, \dots, x_n\}$ of distinct points $x_i \in L$, where $n \ge 0$. For any set $B \subset L$, let $N_{\mathbf{x}}(B) = N(\mathbf{x} \cap B)$ be the number of points of $\mathbf{x}$ lying in B.

A *spatial covariate* V on L is a real- or vector-valued function $V(u)$, $u \in L$. It is assumed that the values $V(u)$ are fixed and known (in principle) for all locations $u \in L$. In practice, the values may only be given at a set of sample locations. In any case, $V(u)$ must be known for some locations u other than the points of the point pattern.

2.2. Point processes

Following the general theory of point processes (Daley & Vere-Jones 2003) we can formally define a finite *point process* $\mathbf{X}$ on L as a mapping from a probability space $(\Omega, \mathcal{F}, P)$ to $(N_f, \mathcal{N}_f)$, where N_f denotes the class of all point patterns in L, and $\mathcal{N}_f$ is the smallest σ-field on N_f with respect to which $N_{\mathbf{X}}(B)$ is measurable, for all compact subsets $B \subseteq L$.

All point processes under consideration here are assumed to be finite and simple (that is, almost surely there are finitely many points and the point locations are distinct) and to possess an *intensity function* $\lambda(u)$, $u \in L$, defined to satisfy

$$\mathrm{E}[N_{\mathbf{X}}(B)] = \Lambda(B) = \int_B \lambda(u)\, \mathrm{d}_1 u, \tag{1}$$

for all measurable B in L, where $\mathrm{d}_1 u$ denotes integration with respect to arc length (Ang, Baddeley & Nair 2012; Jammalamadaka et al. 2013; Rakshit, Nair & Baddeley 2017; Baddeley et al. 2017). Heuristically, for an infinitesimal interval of length $\mathrm{d}_1 u$ centred at $u \in L$, the probability that the interval will contain a random point of $\mathbf{X}$ is equal to $\lambda(u)\,\mathrm{d}_1 u$.

2.3. Poisson process models

The Poisson point process on L is determined by its intensity function $\lambda(u)$, $u \in L$, and is well-defined whenever λ is non-negative and integrable over L. It is characterised by the properties that:

(PP1) the random variable $N(\mathbf{X} \cap B)$ has a Poisson distribution with mean $\mu_B = \int_B \lambda(u)\,\mathrm{d}_1 u$, for all measurable $B \subset L$;

(PP2) for disjoint subsets $B_1, B_2, \ldots B_m$ of L, the random variables $N(\mathbf{X} \cap B_1), N(\mathbf{X} \cap B_2), \ldots, N(\mathbf{X} \cap B_m)$ are independent;

(PP3) for any measurable $B \subseteq L$, conditional on $N(\mathbf{X} \cap B) = n$, the points $x_1, \ldots, x_n$ in $\boldsymbol{x} \cap B$ are independent and identically distributed with probability density $f(u) = \lambda(u)/\Lambda(B)$, for $u \in B$, and zero otherwise,

where (PP3) is a consequence of (PP1) and (PP2).

Explicit models for the intensity function $\lambda(u)$ could take any functional form. We shall consider *loglinear* intensity models

$$\lambda_{\boldsymbol{\beta}}(u) = \exp(\boldsymbol{\beta}^\top \mathbf{Z}(u)), \quad u \in L, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the p-dimensional parameter vector and $\mathbf{Z}(u) = (Z_1(u), \ldots, Z_p(u))^\top$ is the p-dimensional vector of canonical covariate functions. We assume $\int_L \lambda_{\boldsymbol{\beta}}(u)\,\mathrm{d}_1 u < \infty$, for all $\boldsymbol{\beta}$. Note that the canonical covariates $Z_j(u)$ could be transformations of the originally observed covariate functions $V(u)$, including dummy variables associated with different levels of a factor-valued covariate, and interaction terms involving several of the original covariates.

2.4. Maximum likelihood for Poisson point process model

Likelihood theory for the Poisson process on a network can be derived from the case of a Poisson process on the real line (Cox & Lewis 1966; Kutoyants 1998; Rathbun & Cressie 1994). Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ denote the observed point pattern on the network L.

The loglikelihood of the Poisson process with intensity function $\lambda(u)$ is

$$\ell = \log L = \sum_{i=1}^{n} \log \lambda(x_i) - \int_L \lambda(u)\, \mathrm{d}_1 u. \tag{3}$$

In particular, for the loglinear intensity function (2), the loglikelihood takes the form

$$\ell(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \sum_{i=1}^{n} \mathbf{Z}(x_i) - \int_L \exp\{\boldsymbol{\beta}^\top \mathbf{Z}(u)\}\, \mathrm{d}_1 u. \tag{4}$$

The score function is

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{Z}(x_i) - \int_L \mathbf{Z}(u) \exp\{\boldsymbol{\beta}^\top \mathbf{Z}(u)\}\, \mathrm{d}_1 u. \tag{5}$$

The Fisher information is

$$\boldsymbol{I}(\boldsymbol{\beta}) = \int_L \mathbf{Z}(u)\mathbf{Z}(u)^\top \exp\{\boldsymbol{\beta}^\top \mathbf{Z}(u)\}\, \mathrm{d}_1 u. \tag{6}$$

The loglikelihood (4) is concave as a function of $\boldsymbol{\beta}$, and achieves its maximum at a zero of the score function, except in degenerate cases. The maximum likelihood estimate is asymptotically multivariate normal with mean $\boldsymbol{\beta}$ and variance $\boldsymbol{I}(\boldsymbol{\beta})^{-1}$, under several asymptotic regimes, including ‘infill asymptotics’, in which $\lambda(u) = N\lambda_1(u)$ at stage $N = 1, 2, \ldots$.

3. Reduction to generalised linear models

The loglikelihood (3) of the Poisson point process involves an integral over the linear network L. For two-dimensional point processes, the counterpart of the integral in (3) is taken over a two-dimensional spatial domain and is approximated by a finite sum, in such a way that the approximate loglikelihood is equivalent to the loglikelihood of a GLM, which can be fitted using standard statistical software. In the most common implementation, the spatial domain is partitioned into subsets, the point process and covariate values are aggregated over these subsets, and the aggregated data follow a Poisson count regression or logistic regression (Lewis 1972; Brillinger 1978; Lindsey 1992, 1995; Baddeley et al. 2010; Renner & Warton 2013). A slightly different approach developed by Berman & Turner (1992) uses numerical quadrature, and the approximating model is a Poisson loglinear regression (Berman & Turner 1992; Baddeley & Turner 2000; Baddeley, Rubak & Turner 2015, Section 9.8). These techniques can be adapted to linear networks as we describe below.

3.1. Spatially discretised models

In this subsection we assume the linear network is partitioned into disjoint subsets $l_1, \dots, l_J$ with lengths $a_1, \dots, a_J$, respectively. The spatial covariate functions are assumed to be constant (or are approximated by a constant) on each subset,

$$\mathbf{Z}(u) = \mathbf{z}_j, \text{ for } u \in l_j, \tag{7}$$

where $\mathbf{z}_j = (z_{j1}, \dots, z_{jp})^\top$. The point process is aggregated over these subsets, so that the observable random variables are either the counts $N_j = N(\mathbf{X} \cap l_j)$ of points falling in each subset, or the indicators $Y_j = \mathbf{1}(N_j > 0)$ of the event, that at least one point of the process falls inside the subset, for $j = 1, \dots, J$. Correspondingly, let $\mu_j = \mathrm{E}[N_j]$ denote the expected number count in l_j, and $p_j = \mathrm{E}[Y_j] = \Pr(N_j > 0)$, the probability that at least one point falls inside l_j.

These assumptions are useful in two situations. In *Scenario A*, common in road accident research, the subsets l_j are the same as the original segments $s_1, \dots, s_m$, which defined the network, and the available covariates are constant along each segment s_i. In *Scenario B*, the subsets $l_1, \dots, l_J$ constitute a much finer subdivision of the network into short segments called 'lixels' (line picture elements), and the covariates are treated as approximately constant on each lixel.

Under the assumption (7) that covariates are constant on each subset, the inhomogeneous Poisson process with loglinear intensity (2) has constant intensity on each subset, and properties (PP1)–(PP2) imply that the counts N_j satisfy a Poisson loglinear regression, that is, $N_j \sim \text{Pois}(\mu_j)$ are independent random variables with means $\mu_j = \int_{l_j} \lambda(u) \, \mathrm{d}_1 u = a_j \exp(\boldsymbol{\beta}^\top \mathbf{z}_j)$. The loglikelihood (3) collapses to the loglikelihood of Poisson count regression

$$\ell_{\text{Pois}}(\boldsymbol{\beta}) = \sum_{j=1}^{J} [N_j \log \mu_j - \mu_j], \tag{8}$$

with linear predictor

$$\log \mu_j = \log a_j + \boldsymbol{\beta}^\top \mathbf{z}_j, \tag{9}$$

the counts N_j are sufficient for $\boldsymbol{\beta}$, and the loglikelihood can be maximised using standard software, treating the term $\log a_j$ in (9) as an 'offset'. If the covariates are only approximately constant on each subset, there is a loss of efficiency in approximating the Poisson process by a Poisson loglinear regression, and this has been explored for two-dimensional point processes by Baddeley et al. (2010).

Again, under the assumption (7) of constant intensity on each subset, the presence-absence indicators Y_j satisfy a complementary log-log regression, that is, $Y_1, \dots, Y_J$ are independent Bernoulli random variables with success probabilities

$$p_j = p_j(\boldsymbol{\beta}) = 1 - \exp(-\mu_j) = 1 - \exp\left\{-a_j \exp(\mathbf{z}_j^\top \boldsymbol{\beta})\right\}, \tag{10}$$

so that

$$\log(-\log(1 - p_j)) = \log a_j + \mathbf{z}_j^\top \boldsymbol{\beta}. \tag{11}$$

The loglikelihood based on the indicator variables Y_j is

$$\ell_{\text{cloglog}}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \left[Y_j \log \frac{p_j}{1 - p_j} + \log(1 - p_j) \right], \tag{12}$$

which again can be fitted using standard software, treating $\log a_j$ as an offset. Efficiency is lost when the counts N_j are replaced by the indicators Y_j; the Fisher information for both models (8) and (12) is derived in Baddeley et al. (2010), which shows that the relative efficiency is approximately $\bar{\mu}/(\exp(\bar{\mu}) - 1)$ in the case of a single covariate, where $\bar{\mu} = J^{-1} \sum_{j=1}^{J} \mu_j$ is the average expected number of points per subset.

The complementary log-log regression (11) can in turn be approximated by the logistic regression in which

$$\log \frac{p_j}{1 - p_j} = \log a_j + \mathbf{z}_j^\top \boldsymbol{\beta}. \tag{13}$$

This approximation is tolerably accurate provided $\mu_j < 0.4$, for all j, as shown in Baddeley et al. (2010). Logistic regression is the most popular approximation for model-fitting for two-dimensional spatial point process models (Renner & Warton 2013). It has some advantages in numerical performance over the other discretised likelihoods, because the logistic link is canonical (Baddeley et al. 2010, p. 1172).

In Scenario B, these generalised linear models are approximations of the original Poisson point process model, obtained by partitioning the spatial domain and aggregating the data. An important caveat about spatial aggregation is that models fitted using different discretisations or partitions of space are not equivalent in general. This is an instance of the 'ecological fallacy', 'modifiable unit area problem', or 'change-of-support problem' (Robinson 1950; Openshaw 1984; Cressie 1996; Banerjee & Gelfand 2002; Gotway & Young 2002). For spatial Poisson processes, Baddeley et al. (2010) showed that models obtained using different discretisations can even be logically incompatible. They also calculated the bias due to spatial discretisation, and showed that it depends crucially on the spatial regularity of the covariate function $\mathbf{Z}(u)$. These findings also apply to point processes on a linear network.

3.2. Berman–Turner device on a network

The device of Berman & Turner (1992) was adapted to point process models on linear networks by McSwiggan (2019). Integrals $\int_L f(u)\,\mathrm{d}_1 u$ are approximated by finite quadrature sums $\sum_{j=1}^{J} w_j f(u_j)$, where $u_1, \dots, u_J$ are sample points on L and $w_1, \dots, w_J$ are nonnegative weights summing to $|L|$, the length of the network. Methods for choosing the sample point locations u_j and the quadrature weights w_j are discussed by Berman & Turner (1992) and Baddeley & Turner (2000) for two dimensional domains, and by McSwiggan (2019) for linear networks. The sample points must include all the data points; we assume without loss of generality that $u_j = x_j$, for $j = 1, \dots, n$. Thus $J > n$, and we describe the sample points u_j, for $j > n$, as 'dummy' points. The Poisson process loglikelihood (4) is approximated by (Berman & Turner 1992; Baddeley & Turner 2000)

$$\ell(\boldsymbol{\beta}) \approx \sum_{j=1}^{J} \left[y_j \log \lambda_{\boldsymbol{\beta}}(u_j) - \lambda_{\boldsymbol{\beta}}(u_j) \right] w_j, \tag{14}$$

where y_j is the pseudo-response

$$y_j = \begin{cases} 1/w_j, & \text{if } j \leq n, \\ 0, & \text{if } j > n. \end{cases}$$

Following Berman & Turner (1992), we recognise that the approximate loglikelihood (14) is formally equivalent to the weighted loglikelihood of Poisson regression with responses y_j and weights w_j so that it can be maximized using standard software for fitting generalised linear models (Aitkin et al. 1989; Becker, Chambers & Wilks 1988; Chambers & Hastie 1992; Venables & Ripley 2002; Hastie & Tibshirani 1990; Faraway 2005). This can also be treated as the unweighted loglikelihood of Poisson loglinear regression with offset $\log w_j$, i.e., $\log(\lambda(u_j)) = \mathbf{z}_j^\top \boldsymbol{\beta} + \log w_j$, as shown in McSwiggan (2019).

4. Variable selection using regularisation methods

In this section, we describe the regularisation method of variable selection using penalised loglikelihoods (Tibshirani 1996). We demonstrate how to apply the method to the GLM approximations, given in Section 3, of the point process model. In Section 4.1, we define the general form of the penalised loglikelihood for GLMs, and in Section 4.2, we introduce the coordinate descent algorithm for solving the optimisation problem associated with penalised loglikelihoods.

4.1. Penalised loglikelihood

Let $\ell(\boldsymbol{\beta})$ denote any of the GLM loglikelihoods in (8), (12), and (14) with regression coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. We consider the penalised loglikelihood

$$\ell^{\mathrm{Pen}}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \gamma P_\alpha(\boldsymbol{\beta}), \tag{15}$$

where $\gamma(> 0)$ is the regularisation parameter and $P_\alpha(\boldsymbol{\beta})$ is the elastic-net penalty defined by

$$P_\alpha(\boldsymbol{\beta}) = \sum_{l=1}^{p} \left\{ \frac{(1-\alpha)}{2} \beta_l^2 + \alpha |\beta_l| \right\}, \tag{16}$$

for $0 \leq \alpha \leq 1$. Note that $P_\alpha(\boldsymbol{\beta})$ is a convex function of $\boldsymbol{\beta}$ for $0 \leq \alpha \leq 1$, and includes the lasso $(\alpha = 1)$ and the ridge penalty $(\alpha = 0)$ as special cases, while for any $0 < \alpha < 1$, it provides a compromise between the two penalties. In the rest of this section, we shall assume that α is known. Thus, the penalty (16), as a function of $\boldsymbol{\beta}$, is fully specified in the objective function in (15).

For a fixed γ (> 0), the aim is to maximise $\ell^{\mathrm{Pen}}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The choice of γ determines the amount of regularisation imposed on the regression coefficients – the larger the value of γ, the greater the amount of regularisation. When γ is close to zero, most variables under study are included in the regression model. In contrast, for large values of γ, we obtain a highly regularised model, in which the regression coefficients are greatly shrunk toward zero (see Hastie, Tibshirani & Friedman 2009, Chap. 3).

Since the penalties in (16) depend on the measurement scale of the covariates, it is typical to standardize the covariates before applying the penalised likelihood approach (Hastie, Tibshirani & Wainwright 2015, Chap. 2). Exceptions include covariates which are already measured in the same units or are transformed to an equivalent scale (e.g., between zero and unity).

To maximise the penalised loglikelihood in (15), some ideas of iteratively reweighted least-squares (IRLS) are adapted, particularly developed for the maximisation of the loglikelihood $\ell(\boldsymbol{\beta})$ (Nelder & Wedderburn 1972; Hillis & Davis 1994). At every iteration, a quadratic approximation of $\ell(\boldsymbol{\beta})$ is formed about the current estimates of the coefficients. These estimates are updated after every iteration, as explained below.

Let R_j, for $j = 1, \ldots, J$, denote the responses (either the binary responses Y_j or the count outcomes N_j) under consideration. Let $\mu_j = \mathrm{E}[R_j]$ be the mean responses, g be the link function that connects the linear predictor $\mathbf{z}_j^\top \boldsymbol{\beta}$ to μ_j by the relation $g(\mu_j) = \mathbf{z}_j^\top \boldsymbol{\beta}$, and g' be the derivative of g. For a given estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, we now define the 'working response'

(Friedman, Hastie & Tibshirani 2010) as

$$U_j(\tilde{\boldsymbol{\beta}}) = g(\tilde{\mu}_j) + g'(\tilde{\mu}_j)(R_j - \tilde{\mu}_j), \quad j = 1, \ldots, J, \tag{17}$$

where $\tilde{\mu}_j = g^{-1}(\mathbf{z}_j^\top \tilde{\boldsymbol{\beta}})$.

For a fixed γ, we now describe the iterative procedure used in maximising the penalised loglikelihood. Let $\hat{\boldsymbol{\beta}}^{(k)}$ denote the estimate of the parameter vector available at the kth step of the iterative process. Define

$$\omega_j^{(k)} = [\{g'(\mu_j)\}^2 \mathrm{var}(R_j)]^{-1}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k)}}, \quad j = 1, \ldots, J.$$

Then the quadratic approximation to the loglikelihood $\ell(\boldsymbol{\beta})$ about the current estimate $\hat{\boldsymbol{\beta}}^{(k)}$ is the second-order Taylor approximation

$$\tilde{\ell}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(k)}) = -\frac{1}{2J} \sum_{j=1}^{J} \omega_j^{(k)} (U_j(\hat{\boldsymbol{\beta}}^{(k)}) - \mathbf{z}_j^\top \boldsymbol{\beta})^2 + C(\hat{\boldsymbol{\beta}}^{(k)}), \tag{18}$$

where $C(\hat{\boldsymbol{\beta}}^{(k)})$, defined so that $\tilde{\ell}(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)}) = \ell(\hat{\boldsymbol{\beta}}^{(k)})$, is the constant term that does not depend on $\boldsymbol{\beta}$.

We substitute $\tilde{\ell}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(k)})$ for $\ell(\boldsymbol{\beta})$ in the penalised loglikelihood (15) at the kth iteration. The updated $\hat{\boldsymbol{\beta}}^{(k+1)}$ is then obtained by computing

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{-\tilde{\ell}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(k)}) + \gamma P_\alpha(\boldsymbol{\beta})\}. \tag{19}$$

The objective function in (19) is a quadratic approximation of the negative penalised loglikelihood. Instead of maximising the penalised loglikelihood, a standard practice is to solve the equivalent minimisation problem in (19) (Friedman, Hastie & Tibshirani 2010). Because the objective function in (19) is a convex function of $\boldsymbol{\beta}$, there exist algorithms that converge to a global optimum by iterative application of (19) (Hastie, Tibshirani & Wainwright 2015, Chap. 5). One such algorithm is given below.

4.2. Cyclical coordinate descent algorithm

The cyclical coordinate descent algorithm minimises each coefficient β_j one-at-a-time and converges to a global minimiser for the convex objective function in (19) under mild conditions (Hastie, Tibshirani & Wainwright 2015, Chap. 2). Friedman et al. (2007) first explored this algorithm in the linear regression setting with multiple predictors. It was subsequently extended by Friedman, Hastie & Tibshirani (2010) for maximising penalised loglikelihoods of GLMs, and implemented in open source software (Friedman et al. 2019).

A coordinate descent step to partially optimise (19) with respect to β_l involves holding the coefficients other than β_l fixed at their current estimates, and then solving the resulting optimisation problem using the results in Friedman et al. (2007), developed for the univariate case. This step is repeated for each β_l, $l = 1 \ldots, p$, one after another, to complete a single cycle.

The lth step, for any $l \in \{1, \ldots, p\}$, in a cycle can be described as follows. Suppose we have finished the kth iteration for some $k \geq 1$. Then, after substituting (18) in (19), the lth step of the $(k+1)$th iteration amounts to solving

$$\underset{\beta_l}{\operatorname{argmin}} \frac{1}{2J} \sum_{j=1}^{J} \omega_j^{(k)} \left[U_j(\hat{\boldsymbol{\beta}}^{(k)}) - \sum_{i=0, i \neq l}^{p} \hat{\beta}_i^{(k)} z_{ji} - \beta_l z_{jl} \right]^2 + \gamma \left[\frac{(1-\alpha)}{2} \beta_l^2 + \alpha |\beta_l| \right]. \tag{20}$$

A solution to the above minimisation problem is the coordinate-wise update, as described in Donoho & Johnstone (1995), for coordinate l, which is given by

$$\hat{\beta}_l^{(k+1)} = \frac{S\left(\sum\limits_{j=1}^{J} \omega_j^{(k)} z_{jl} \left[U_j(\hat{\boldsymbol{\beta}}^{(k)}) - \tilde{U}_j^{(l)}(\hat{\boldsymbol{\beta}}^{(k)}) \right], \gamma\alpha \right)}{\sum\limits_{j=1}^{J} \omega_j^{(k)} z_{jl}^2 + \gamma(1-\alpha)}, \tag{21}$$

where $\tilde{U}_j^{(l)}(\hat{\boldsymbol{\beta}}^{(k)}) = \sum_{i=0, i\neq l}^{p} \hat{\beta}_i^{(k)} z_{ji}$ and $S(u, v)$ is the soft-thresholding function

$$S(u,v) = \operatorname{sign}(u)(|u| - v)_+ = \begin{cases} u - v, & \text{if } u > 0 \text{ and } v < |u|, \\ u + v, & \text{if } u < 0 \text{ and } v < |u|, \\ 0, & \text{if } v \geq |u|. \end{cases} \tag{22}$$

After each cycle, the coordinate-wise updates (21) are used to update the quadratic term $\tilde{\ell}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(k)})$ in the objective function in (19). Then the next cycle of the algorithm reestimates the model parameters using (21). These two steps of updating $\tilde{\ell}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(k)})$ and computing the parameter estimates continue until convergence. See Tseng (2001) for an overview of the convergence properties of coordinate descent in convex problems.

It follows from (21) that the estimated coefficients are non-zero when the ridge penalty is imposed using $\alpha = 0$ in (16). For any non-zero α, it follows from (22) that the shrinkage in estimated coefficients depends on the value of γ. Therefore, a crucial question is how to obtain an appropriate value of γ for the regularised model in (15). A simple approach is to use some form of cross-validation, for example 10-fold cross-validation, for determining an optimum value of γ (Friedman, Hastie & Tibshirani 2010). As described in Hastie, Tibshirani

& Friedman (2009, Chap. 7), two values commonly used in practice are γ_{min} and γ_{1se}, where γ_{min} is the value of γ that minimises the cross-validation error and γ_{1se} is the value that provides the most regularised model with an error no more than one standard error above the minimum cross-validation error. There may be practical limitations with these two particular choices. In some applications, the selection rule γ_{min} could produce a model with almost the full set of variables, while the rule γ_{1se} produces a highly regularised model with hardly any variables selected (see Sections 6 and 7 for examples). We propose two additional rules for selecting γ, given by

$$\gamma_{\text{avg}} = (\gamma_{\text{min}} + \gamma_{\text{1se}})/2 \qquad \text{and} \qquad \gamma_{\text{gmean}} = \sqrt{\gamma_{\text{min}} \cdot \gamma_{\text{1se}}}\,, \tag{23}$$

the arithmetic mean and geometric mean of γ_{min} and γ_{1se}, respectively. These additional choices provide some means of compromise between the two extreme models chosen using γ_{min} and γ_{1se}. The performance of γ_{avg} for variable selection is assessed in a simulation study alongside γ_{min} and γ_{1se} in Section 5. We have used γ_{gmean} to select variables under the Berman–Turner approximation for the WA accident data, analysed in Sections 6 and 7.

5. Simulation study

We performed a simulation study to evaluate the performance of the lasso, ridge regression and elastic net methods in selecting spatial covariates for a point process model on a linear network. R code for performing the simulations is provided in an online supplement.

For practical reasons, the simulations were performed on a relatively simple network L, shown in Figure 2, which represents the street network surrounding the University of Chicago, and has featured in many recent research papers (Ang, Baddeley & Nair 2012; Jammalamadaka et al. 2013; Rakshit, Nair & Baddeley 2017; Baddeley et al. 2017).

Ten spatial covariate functions $Z_1(u), \ldots, Z_{10}(u)$ were generated (once only) as independent realisations of a stationary Gaussian random field in two dimensions, with mean 1 and exponential variogram with sill 2.4 and range 100 feet, restricted to the network. This is similar to the simulation studies in Thurman & Zhu (2014) and Yue & Loh (2015), where the covariates are generated using independent realizations of the same Gaussian random field.

The true point process model was a Poisson process with intensity depending only on the first five covariates

$$\lambda(u) = 0.00015 \exp\left\{10Z_1(u) + 9Z_2(u) + 8Z_3(u) + 7Z_4(u) + 6Z_5(u)\right\}, \quad u \in L. \tag{24}$$

Figure 3 shows the synthetically generated covariates $Z_1, \ldots, Z_5$ and one realization of the Poisson process with intensity (24).

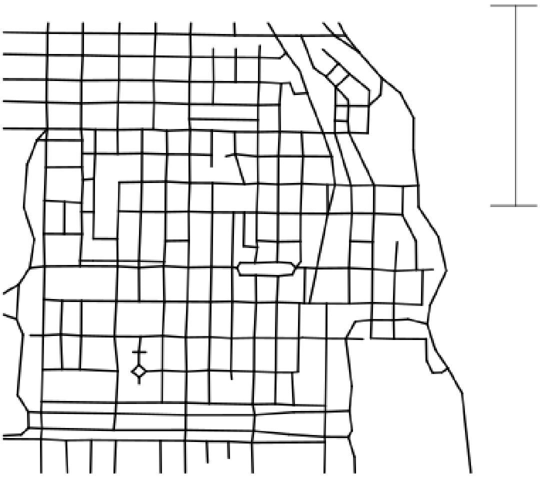

Figure 2. Street network around the University of Chicago. Scale bar (top right) is 500 feet.

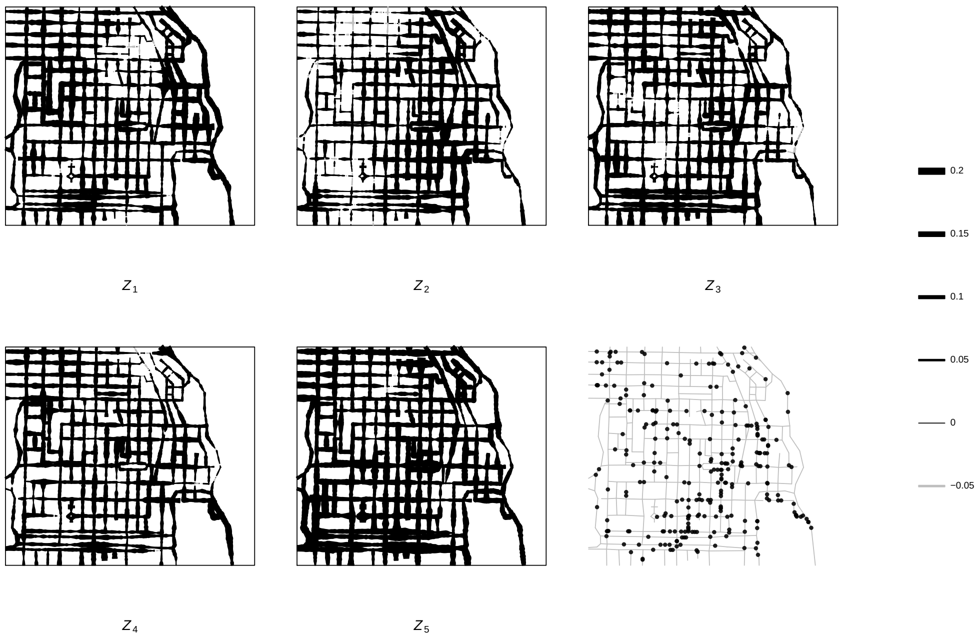

Figure 3. The covariates $Z_1, \dots, Z_5$ used in the simulation study. Covariate value at each location on the network is represented by line thickness, with scale shown at the right. Bottom right panel shows one of the simulated point patterns, a realisation of the inhomogeneous Poisson process whose log-intensity is a linear combination (24) of these covariates.

The experiment generated 1000 simulated realizations of the Poisson process with intensity (24) using a version of the thinning algorithm of Lewis & Shedler (1979). To each realization we fitted the Poisson point process model with intensity

$$\lambda(u) = 0.00015 \exp\left\{ \sum_{l=1}^{10} \beta_j Z_j(u) \right\}, \quad u \in L, \tag{25}$$

approximating the likelihood using the loglinear Poisson regression model for discretised counts (8), logistic regression for discretised presence-absence indicators (12) with (13), and the Berman–Turner approximation (14).

Results of the experiment are reported in Tables 1, 2 and 3. Table 1 reports estimated bias and standard error of the estimates of the regression coefficients $\beta_1, \ldots, \beta_5$, which are actually present in the model (24). Table 2 reports estimated bias and standard error for the remaining coefficients $\beta_6, \ldots, \beta_{10}$, which are zero in the true model. Table 3 reports, for each covariate Z_l, the observed fraction of outcomes in which the coefficient β_l is nonzero, so that the covariate is included in the fitted model. This helps to assess how well a variable selection method can perform in selecting the variables that are originally present in the model and rejecting the ones that are absent from the model.

Tables 1 and 2 report the empirical bias and standard error in estimating the regression coefficients using the lasso, ridge regression, and elastic net (with $\alpha = 0.5$), based on the regularisation parameter selection rule γ_{min}. In these tables, the first three columns show, respectively, the method of variable selection, the number of lixels or dummy points used in fitting, and the type of approximating generalised linear model. The labels logistic and Poisson in the third column correspond to the approximations using logistic regression and Poisson count regression, respectively, fitted by minimising (15) for the likelihoods (12) and (8), respectively. The label B–T represents the Berman–Turner approximation (14).

Table 1 concerns the first five coefficients, which were nonzero in the true model. The biases are all negative, reflecting the well-known fact that regularised methods result in estimates biased towards zero. For the largest two coefficients β_1 and β_2, the lasso produced smaller biases than ridge regression, for all choices of approximate likelihood. On the other hand, for the smallest two coefficients β_4 and β_5, ridge regression produced smaller estimated biases than the lasso, for all approximate likelihoods. The elastic net biases lay typically between the lasso and ridge biases. The estimated biases for β_3 are very close to each other for all three variable selection methods. For the approximate likelihoods based on discretisation, logistic regression produced smaller estimated bias than loglinear Poisson regression, for all scales of discretisation and all variable selection methods. Amongst the logistic regression approximations, the discretisation using 3370 lixels produced the minimum biases. Thus, increasing the fineness of the discretisation does not always increase the estimation accuracy – a phenomenon of numerical integration which is well recognised in this context (Baddeley et al. 2010). For the Berman–Turner approximation, the estimated biases are close to each other for all three choices of the number of dummy points on the network. For all likelihood approximations and variable selection methods, the coefficients representing larger effect sizes were estimated with less bias than the ones representing relatively smaller effect sizes.

Table 2 concerns the coefficients which were zero in the true model. Coefficients β_6 and β_8 were estimated with large bias. This reflects the inability of the regularisation parameter selection rule γ_{min} to exclude some of the variables that are absent in the model; we address this issue below. Of the three penalties, ridge regression produced the greatest bias, while the lasso and elastic-net results are close to each other. Overall, the Berman–Turner approximation produced smaller biases than the discretised GLM approximations, for all the coefficients. The Berman–Turner fits with relatively large numbers of dummy points (3063 and 5516) produced smaller biases than the fits using smaller numbers (1684) of dummy points.

The large estimated biases in Table 2, especially for the coefficients β_6 and β_8, indicate that the corresponding variables Z_6 and Z_8 will be selected in the final model for a high percentage of the total simulations. This is undesirable because one of the main objectives is to exclude variables that are absent in the true model. Although the selection rule γ_{min} provides accurate estimates of the non-zero coefficients, it often fails to discard unrelated variables. To overcome this problem, Friedman, Hastie & Tibshirani (2010) proposed the selection rule γ_{1se}, described in Section 4.

Table 3 reports the proportion of simulated outcomes in which each coefficient was estimated to be non-zero, and was therefore included in the fitted model. We considered the selectors γ_{min} and γ_{1se}, and also a proposed compromise, γ_{avg}, defined in (23). The results in Table 3 clearly show that the use of γ_{min} provides the highest proportion of non-zero estimates for the coefficients β_6,...,β_{10}. In contrast, the use of γ_{1se} provided the most strongly regularised models, and even failed to select the variables with large effects for some models. For example, the Berman–Turner fit with 5516 dummy points for the selector γ_{1se} produced coefficient estimates equal to zero (in almost all the simulations) for all the 10 variables. This behaviour of γ_{1se} indicates that it may fail to select any variables, even the variables with large effect sizes.

The selector γ_{avg} tends to balance the two extreme behaviours of γ_{min} and γ_{1se}. The results in Table 3 show that the proportion of the non-zero coefficients are similar for the selectors γ_{avg} and γ_{min}. On the other hand, the selectors γ_{avg} and γ_{1se} produced similar proportions for the coefficients $\beta_6, \ldots, \beta_{10}$. This shows that the selector γ_{avg} performs reasonably well in both selecting the variables with non-zero effect sizes and excluding the ones with zero effect sizes.

6. Western Australia traffic accident data

6.1. Data description

The Western Australian road network and accident data, plotted in Figure 1, were provided by the state government agency Main Roads WA. These data have now been made publicly available on the agency's website as part of the Western Australian Whole of Government Open Data Policy. There are 5386 accidents recorded. For practical purposes we have reduced the data complexity by simplifying the road geometry, so that the original dataset of over 600000 individual road segments has been reduced to 281740 segments with a total length of approximately 19263 km.

Table 4 describes the covariates for the WA state road network used in the analysis. For each of these, the covariate value is constant along each road segment. Figure 4 illustrates the three covariates `SPD_LIM`, `KERB_L`, and `SHLDR`. Because our analysis is motivated by the road characteristics that are of primary interest to Main Roads, we chose not to include covariates based on network distance, i.e., distance between two points on a network as the shortest distance between the points along the network (see Ang, Baddeley & Nair (2012) for a formal definition).

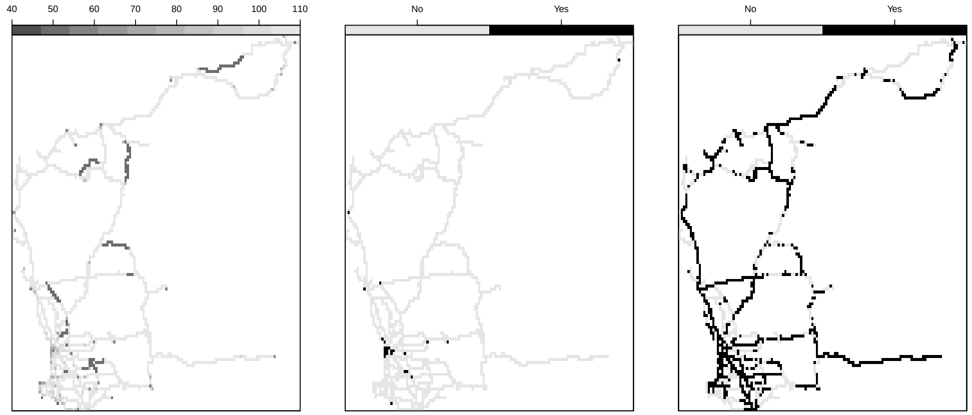

Figure 4. WA road characteristics (left to right): (i) `SPD_LIM`, (ii) `KERB_L` and (iii) `SHLDR`

6.2. Variable selection

In this section, we select variables for modelling accidents on the Western Australian road network (shown in Figure 1) using the lasso, ridge regression, and elastic net (with $\alpha = 0.5$) methods described in Section 4. From Main Roads WA, we obtained data on 11 road characteristics, listed in Table 4. All the numeric variables were linearly rescaled to the range

$[0, 1]$ by first subtracting the minimum value and then dividing by the range of that variable. We further created additional 33 canonical variables from these initial 11 variables. Thus, the total number of variables entered into the selection process is 44. The additional variables consisted of the scaled quadratic terms of the numeric variables, computed by first taking squares and scaling afterwards, and two-way interaction terms associated with the factors `SHLDR`, `KERB_L` and `KERB_R`. For each of these three factors, the interaction terms were created between the given factor and the remaining factors and scaled numeric variables. The coefficient estimates for regularised models are computed using the three approximations, described in Section 2, and are reported below in Table 7.

We fitted the WA accident data using the Berman–Turner approximation with 563764 dummy points. We also fitted two discretised models: Poisson count regression and logistic regression, using the 281740 road segments as the aggregation units. The logarithm of road segment lengths appear in the offset terms in (9) and (13), respectively.

Figure 5 shows how the cross-validation mse (left panel) and fraction of deviance explained (right panel) vary with the number of variables in the lasso method when analysing the WA accident pattern based on the Berman–Turner (top panel), logistic regression (middle panel) and Poisson count regression (bottom panel) approximations. The cross-validation curve shows the out-of-sample performance of the model, while the fraction of deviance explained is computed solely based on the training data. The two vertical dotted lines in the cross-validation error plots correspond to the selectors γ_{min} and γ_{1se}. For brevity, we did not include the plots analogous to Figure 5 corresponding to the ridge regression and elastic-net, but these can be produced using the R-scripts and datasets provided in the online supplement document, along with other details.

Nonetheless, the estimated coefficients corresponding to all the three variable selection methods are presented in Table 7. Furthermore, for each of these variable selection methods, the main results needed to select an appropriate regularised model are reported in Table 6. It presents the four selectors γ_{min}, γ_{1se}, γ_{avg}, and γ_{gmean}, along with the percentage deviation explained and the number of variables selected by each of them. In the case of the lasso, when γ_{min} is used, a large number of variables is selected for all three approximations. In contrast, a parsimonious model is obtained using either of the selectors γ_{1se} and γ_{avg}. For the two discretised models, the selectors γ_{1se} and γ_{avg} picked models of similar sizes, but the cross-validation error for γ_{avg} was lower than that for γ_{1se}. Consequently, we used γ_{avg} for the discretised models. Using γ_{avg} for the logistic and Poisson approximations, models with 22 and 27 variables, respectively, were obtained. Also, the percentage deviance explained adopting γ_{avg} is very close to the percentage deviance explained by the full model.

For the Berman–Turner approximation, the use of γ_{1se} and γ_{avg} produced very parsimonious models with two and three variables, respectively. In this case, we chose γ_{gmean},

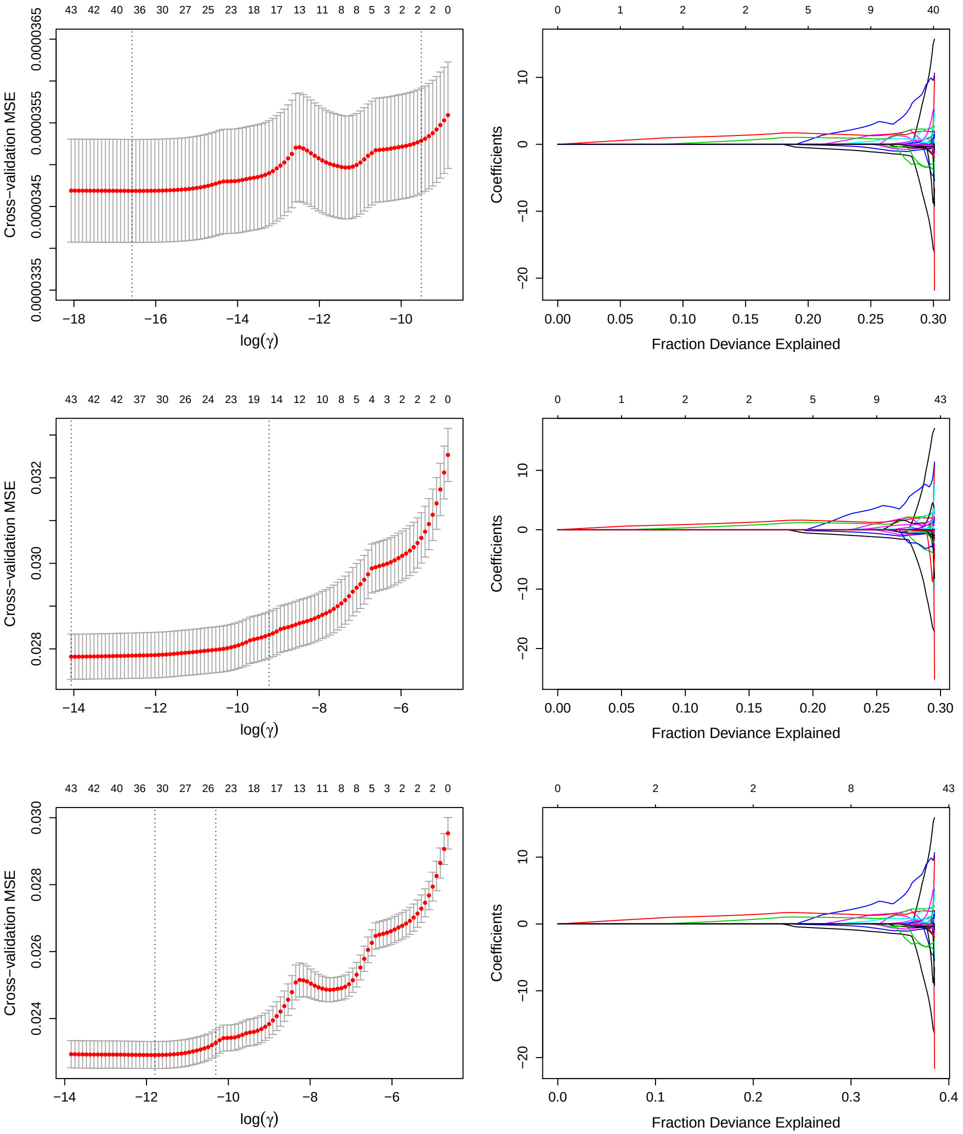

Figure 5. Cross-validation diagnostics for selecting the regularisation parameter γ for the Western Australian road accident data. Lasso method applied to the Berman–Turner (*Top row*), logistic regression (*Middle row*) and Poisson count regression (*Bottom row*) approximations. *Left column*: cross-validation error curve (horizontal dotted line) and corresponding upper and lower standard deviation curves (error bars) are plotted against $\log(\gamma)$. Number of variables in the lasso is shown along the top margin. The two vertical dotted lines correspond to γ_{min} and $\gamma_{1\text{se}}$. *Right column*: coefficient estimates against fraction of deviance explained.

the logarithm of which corresponds to the middle value between the two vertical dotted lines in Figure 5. This choice yielded a model with 17 variables, and explained 27.67% deviance out of 30.09% deviance, explained by the full model.

In the case of ridge regression and elastic net, we again selected γ_{avg} for the two discretised models, and the justification is similar to the one given for the lasso method. For the Berman–Turner approximation with the elastic-net penalty, we used the selection rule γ_{gmean}, similar to the lasso case. However, when γ_{gmean} was used for the ridge regression, very low percentage deviance was explained, and consequently, the selector γ_{min} was used for the ridge regression. Important information about the selected models is presented in Table 6 with bold font.

Table 7 shows that the main effects of `TOT_S`, `N_LANE`, `KERB_R`, and `KERB_L` are selected by the lasso and elastic-net with relatively large positive estimates, whereas the main effects of $\texttt{SPD\_LIM}^2$ and `FLDWY` with relatively large negative estimates. The interaction effects `TOT_S`×`KERB_R`, `TOT_S`×`SHLDR`, and `TOT_S`×`KERB_L` are selected with relatively large absolute values as their estimates. The ridge regression, as expected, selects all these variables but shrinks their coefficients towards zero.

6.3. Interpretation of fitted models

Since the selected models in Table 7 are all log-linear models of accident intensity (2), the interpretation of estimated coefficients is similar for each of them. To illustrate how to interpret these model coefficients, we consider the lasso solution, given in the first numeric column of Table 7, obtained by applying the Berman–Turner approximation. The interpretation is straightforward when only main effects are present (Baddeley, Rubak & Turner 2015, Section 9.3). However, in the presence of interaction terms, the effect of a covariate on log-intensity depends on the values (for quantitative variables) or levels (for categorical factors) of other covariates. For example, the effect of `TOT_S` may vary across the levels of three factors `KERB_L`, `SHLDR`, and `KERB_R`. We divide all the effects associated with `TOT_S` by the scale value of 31.6, which can be computed using the maximum and minimum values provided in Table 4, to obtain the effect sizes in the original measurement unit (meter) of `TOT_S`.

After adjusting for the scaling, the coefficient of `TOT_S` main effect is 0.14, and the coefficients of the interaction effects `TOT_S` × `KERB_L`, `TOT_S` × `SHLDR`, and `TOT_S` × `KERB_R` are -0.0008, 0.038, and -0.067, respectively. Holding covariates, other than `TOT_S`, `KERB_L`, `SHLDR`, and `KERB_R`, constant, the intensity of road accident would increase by a factor of $\exp(0.14) = 1.15$ if the seal-width (`TOT_S`) increased by one meter in a road with no kerbs and no shoulder-padding. If right-kerb (`KERB_R`) is present in a road with no left-kerb and no shoulder-padding, the increase in the intensity would be by a factor of $\exp(0.14 - 0.067) = 1.08$, for a one meter increase in the seal-width. When all the three road characteristics are present, the increase would be by a factor of $\exp(0.14 - 0.0008 + 0.038 - 0.067) = 1.12$.

The fitted models appear to imply that an increase in road seal-width would lead to an increase in accident rate. We caution strongly against inferring such causal connections from a perfunctory inspection of the fitted models.

Firstly, the road properties should not be treated as fixed covariates. The road network is continually being modified in response to events on the network, including traffic accidents, traffic congestion and changes in usage. It would be more appropriate to regard both the accidents and the road properties as observational data. Our analysis is then a form of conditional regression of the accident pattern on the road properties, and the fitted effects represent correlations between accidents and road properties. For example, it would be appropriate to say that the fitted model indicates that accident rate is positively correlated with road seal-width (and vice versa); and a possible explanation is that, in those road segments with a history of accidents, the authorities will try to improve safety by widening the road.

Secondly, the accident 'rate' $\lambda(u)$ defined in our point process model (2) is the rate of accidents per unit length during a specified period. This may be the appropriate measure of accident rate for emergency planning purposes, but for road accident research, it would be more appropriate to estimate the accident risk relative to traffic volume, that is, the probability of an accident per vehicle per unit length during a specified period. If the traffic volume (number of cars per hour) along each road segment is known, then our modelling approach can be modified to estimate accident risk simply by including the logarithm of traffic volume as an offset in the linear predictor. Unfortunately, the available traffic volume data tend to be inadequate, because they are aggregated, or available only for the roads with very high volumes. This is one reason why estimation of relative risk is important.

Thus, the paradoxical positive correlation between accidents and road seal-width could simply be attributable to the fact that roads with more traffic tend to be widened to handle the traffic.

7. Variable selection for marked point processes

In a 'marked' point pattern, the spatial locations x_i are augmented by values m_i, called marks, containing information about the event at x_i.

For example, if each m_i is a categorical value, then each point is effectively assigned to one of several discrete categories, and the point pattern is effectively divided into several different types of points. Figure 6 shows an example in which a spatial point pattern of crime locations is augmented by classifying the crimes into different types. In the original

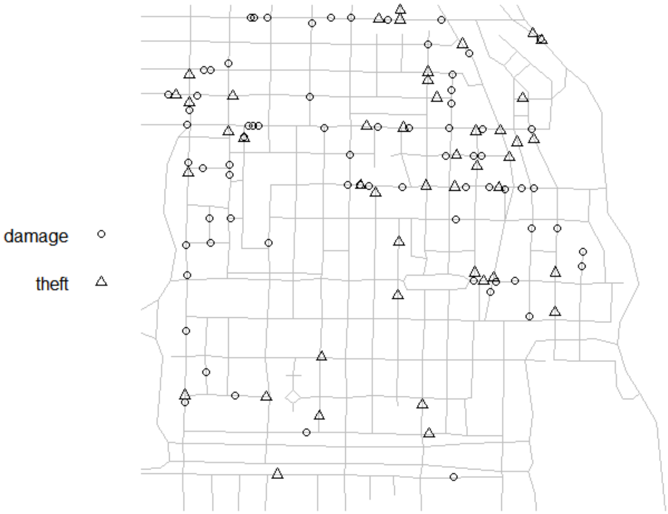

Figure 6. Crime events on the Chicago street network (lines) with the information on the type of crime (*open circle*: damage and *open triangle*: theft).

dataset, these crime events were categorised into seven categories, namely assault, burglary, theft, damage, robbery, trespass and car-theft. For simplicity, we have grouped these crime categories into two broad categories, namely theft and damage. The new theft category is created by grouping the initial theft and car-theft events together; the new damage category consists of the remaining five crime categories.

In many applications, the mark m_i is a multivariate observation consisting of several 'mark variables'. For the Western Australian road accident dataset, Table 5 shows some of the mark variables available.

There is an important methodological distinction between marks and spatial covariates. Covariates can be observed at any spatial location, and are usually treated as explanatory variables. Marks are attributes of the point events and are typically treated as part of the 'response' in a statistical model (Baddeley 2010a, Baddeley, Rubak & Turner 2015, p. 147, Baddeley 2010b). Despite this distinction, we shall show that our variable selection methods can also be applied to marked point patterns on a linear network. This appears to be new.

7.1. Theory

Theoretical foundations of marked point processes were established by Matthes (1963) and are sketched in Matthes, Kerstan & Mecke (1978, p 7), Stoyan, Kendall & Mecke (1995, Sec. 4.2, pp. 105-109) and Baddeley (2010b).

A point x on the network L, with an associated mark m belonging to some general set of possible marks $\mathcal{M}$, is regarded as an ordered pair $(x, m) \in L \times \mathcal{M}$. Impose the very general assumption that $\mathcal{M}$ is a separable metric space. Then a marked point process $\mathbf{X}$ on the network L, with marks belonging to $\mathcal{M}$, is defined as a point process on $L \times \mathcal{M}$ with the condition that $N(L \times \mathcal{M}) < \infty$, almost surely. This condition ensures that the process of points without marks is well-defined (Matthes, Kerstan & Mecke 1978, p. 7; Stoyan, Kendall & Mecke 1995, pp. 105-109). A realisation of $\mathbf{X}$ is a (finite) marked point pattern, that is, an unordered set $\{(x_1, m_1), \ldots, (x_n, m_n)\}$, where $x_1, \ldots, x_n$ are the point locations in L and $m_1, \ldots, m_n$ are the corresponding mark values.

A 'multitype' point process on L is a marked point process in which the marks are categorical values, say $\mathcal{M} = \{1, \ldots, c\}$, which effectively label the points into c different types.

A Poisson marked point process on L with marks in $\mathcal{M}$ is defined as a Poisson process on $L \times \mathcal{M}$ with the condition that $\mathrm{E}[N(B \times \mathcal{M})] < \infty$, for all bounded $B \subset L$, or equivalently, that the expected total number of marked points is finite.

In order to define the intensity function and likelihood of a marked point process, we must designate a measure μ on $\mathcal{M}$ which serves as the reference measure for integration over $\mathcal{M}$. Then a marked point process $\mathbf{X}$ on L with marks in $\mathcal{M}$ has intensity function $\lambda(u, m)$, $u \in L$, $m \in \mathcal{M}$, if

$$\mathrm{E}[N(B \times M)] = \int_B \int_{\mathcal{M}} \lambda(u, m)\, \mathrm{d}\mu(m)\, \mathrm{d}_1 u, \tag{26}$$

and the log-likelihood of the Poisson marked point process with intensity function $\lambda(u, m)$ is, from Baddeley (2010a) extended to linear networks,

$$\ell = \sum_{i=1}^{n} \log \lambda(x_i, m_i) - \int_L \int_{\mathcal{M}} \lambda(u, m)\, \mathrm{d}\mu(m)\, \mathrm{d}_1 u. \tag{27}$$

The only case considered here is the multitype point process with mark space $\mathcal{M} = \{1, 2, \ldots, c\}$. Taking the reference measure μ to be the counting measure on $\mathcal{M}$, equations (26) and (27) become

$$\mathrm{E}[N(B \times M)] = \int_B \sum_{m=1}^{c} \lambda(u, m)\, \mathrm{d}_1 u \tag{28}$$

and

$$\ell = \sum_{i=1}^{n} \log \lambda(x_i, m_i) - \int_L \sum_{m=1}^{c} \lambda(u, m)\, \mathrm{d}_1 u. \tag{29}$$

The Berman–Turner device was extended to multitype point patterns by Baddeley & Turner (2000) and we now adapt this to linear networks. Given a quadrature scheme $U =$

$\{u_1, \dots, u_{n+n_0}\}$ for the spatial locations on the network, take the product

$$V = U \times \mathcal{M} = \{(u_j, m) : j = 1, \dots, n + n_0; m = 1, \dots, c\}.$$

Define the pseudo-response y_{jm} corresponding to (u_j, m) so that $y_{jm} = 1$, if u_j is a data point (i.e. if $j \leq n$) and $m = m_j$ is the corresponding observed mark, and $y_{jm} = 0$ otherwise. The approximate likelihood is then

$$\ell(\beta) = \sum_{j=1}^{n+n_0} \sum_{m=1}^{c} (y_{jm} \log \lambda(u_j, m) - \lambda(u_j, m)) w_j. \tag{30}$$

7.2. Chicago data

For expository purposes, and in order to test the performance of the technique, we first consider the Chicago crime data in Fig. 6. The data do not include spatial covariates, but since there is clear evidence of spatial inhomogeneity, we use the Cartesian coordinates as surrogate covariates. We fitted a model in which the intensity at spatial location (x, y), for crimes of type m, is a log-quadratic function of the coordinates,

$$\begin{aligned} \log \lambda(x, y, m) &= \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2 \\ &\quad + \mathbf{1}(m = 1) \left(\beta_6 + \beta_7 x + \beta_8 y + \beta_9 x^2 + \beta_{10} xy + \beta_{11} y^2\right), \end{aligned} \tag{31}$$

where the mark values `theft` and `damage` are encoded as $m = 0, 1$, respectively.

We fitted the model (31) using the lasso applied to the Berman–Turner approximation, with 1800 dummy points on the Chicago network, for both theft and damage categories. The Cartesian coordinates x and y were rescaled to have mean zero and sample variance 1 over the quadrature points.

Table 8 reports the coefficient estimates, with a dot indicating that the coefficient estimate was zero, so that the associated covariate was not selected. The left side of the table gives the coefficients $\beta_0, \dots, \beta_5$ in (31), while the right side of the table gives $\beta_6, \dots, \beta_{11}$. The right side of the table is associated with the effect of the mark variable (crime type) and its interaction with spatial location. For this model the relative risk of damage to theft is

$$r(x, y) = \frac{\lambda(x, y, 1)}{\lambda(x, y, 0)} = \exp(\beta_6 + \beta_7 x + \beta_8 y + \beta_9 x^2 + \beta_{10} xy + \beta_{11} y^2) \tag{32}$$

and the conditional probability that a crime at location (x, y) is a damage is $p(x, y) = r(x, y)/(1 + r(x, y))$. Consequently, $r(x, y)$ and $p(x, y)$ depend only on coefficients in the right half of Table 8.

Note that the variable-selection algorithm (Friedman, Hastie & Tibshirani 2010) does not obey the usual rules of nested models, which require that, if an interaction term is selected, then the corresponding main effects must also be selected. For example, the last row of Table 8 indicates that $\hat{\beta}_{11} \neq 0$ but $\hat{\beta}_5 = 0$, violating the nesting rule. It would be more appropriate to constrain the algorithm to respect the nesting rules.

For comparison, backward stepwise model selection using AIC, started from a maximum likelihood fit of the full model (31), selected the covariates corresponding to coefficients $\beta_0, \beta_2, \beta_5$ and β_6, β_7. The weaknesses of stepwise model selection are well known, and we expect that the penalised maximum likelihood methods will have better performance in the presence of correlated covariates.

7.3. Western Australian road accident data

Here we analyse the Western Australian accident data classified into two types according to the accident severity (high or low). Figure 7 shows the spatial patterns of the two accident types, and Figure 8 shows a closer view.

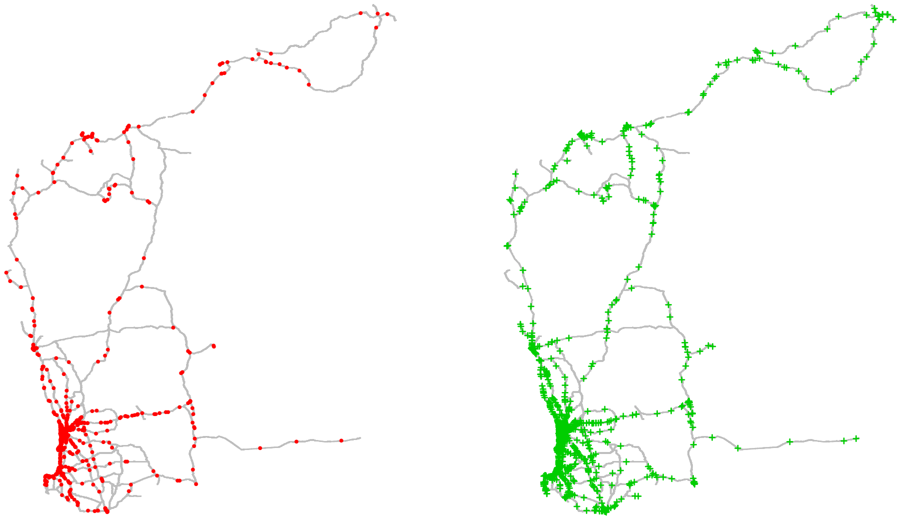

Figure 7. Traffic accidents on the state road network of Western Australia, separated into high severity (*Left*) and low severity (*Right*) accidents.

For the Western Australian road accident data with marks indicating accident severity, Figure 9 shows the cross-validation mse plotted against regularisation parameter γ (in the left column) and the coefficient estimates plotted against fraction deviance explained (right column), for the lasso (top row), ridge regression (middle row) and elastic-net (bottom row) methods applied to the Berman–Turner approximation (30).

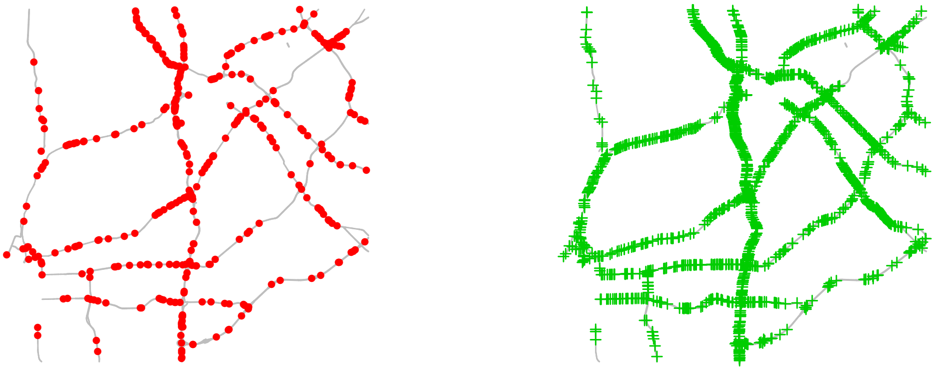

Figure 8. Closeup of previous figure in an area of northern metropolitan Perth.

For the lasso and elastic-net penalties, the rule γ_{min} selected 50 and 74 variables, respectively, yielding models with a large number of variables, while the rule γ_{1se} selected zero variables. Consequently, for these two penalties, we considered the respective γ_{gmean} values for fitting regularised models to the marked WA accident pattern based on the Berman–Turner approximation. The logarithms of γ_{gmean} values for the lasso and elastic-net penalties are -13.27 and -13.46, respectively. For both lasso and elastic-net, the corresponding γ_{gmean} values selected reasonably sparse models with 16 and 31 variables, respectively.

In the case of ridge regression, the selection rules γ_{1se}, γ_{avg}, and γ_{gmean} produced models with less than 3% deviance explained. In contrast, using γ_{min}, we obtained a model that explained 22.24% of deviance, which is very close to the maximum of 24.32% deviance explained by the model without any regularisation. Consequently, we selected $\gamma_{\text{min}} = \exp(-10.83)$ in the case of ridge regression.

Table 9 shows the coefficient estimates obtained using lasso, ridge regression and elastic-net penalties applied to the Berman–Turner approximation. The results for lasso and elastic-net methods are broadly in agreement, which is not unexpected given the similarity of these methods. It is interesting that the speed limit `SPD_LIM` is not selected. The right-hand half of the table is associated with the effect of the mark variable (accident severity), in a similar way to that explained for the Chicago data. In the right-hand half of the table, two and six variables are selected by the lasso and elastic-net, respectively; all the selected coefficients are negative, so that the relative risk of a severe accident is predicted to decrease as the values of these covariates increase.

For brevity, we have only included the results based on the Berman–Turner method in the table. However, the two discretised methods described in Section 3 can also be easily extended for the variable selection in marked point patterns.

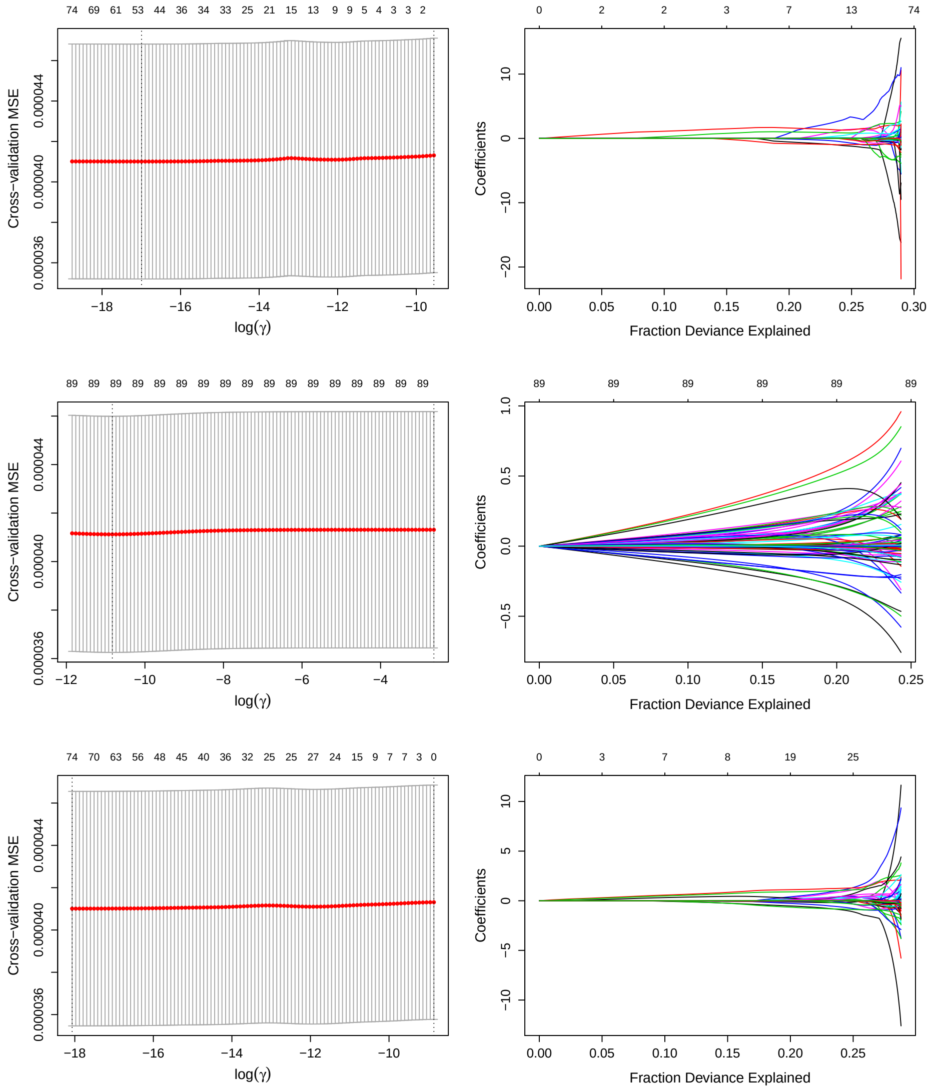

Figure 9. Cross-validation diagnostics for selecting the regularisation parameter γ for the Western Australian road accident data with Mark indicating accident severity. Lasso (*Top row*), ridge (*Middle row*), and elastic-net (*Bottom row*) penalties are applied to the Berman–Turner approximation. *Left column*: cross-validation error (horizontal dotted line) and corresponding upper and lower standard error bars are plotted against $\log(\gamma)$; two vertical dotted lines correspond to γ_{min} and $\gamma_{1\text{se}}$. *Right column*: coefficient estimates against fraction of deviance explained. Number of variables is shown along the top margin.

8. Discussion

We have considered only Poisson point process models. Variable selection for non-Poisson point processes, specifically for log-Gaussian Cox processes, is feasible for two-dimensional point patterns (Thurman & Zhu 2014; Yue & Loh 2015; Thurman et al. 2015).

We expect that the same techniques could be applied on a linear network. We have not attempted this because there are some unresolved difficulties in constructing Cox models on a linear network (Baddeley et al. 2017; Anderes et al. 2020). Gibbs point process models on a network are in the early stages of development (van Lieshout 2018).

A mature methodology for model selection on a linear network should also deal with directed networks such as rivers and streams (Cressie et al. 2006; Ver Hoef & Peterson 2010; Ver Hoef, Peterson & Theobald 2006) and point events which occur exactly at a vertex of the network (such as traffic accidents at a road intersection). Alternative techniques include the Osborne descent algorithm (Osborne, Presnell & Turlach 2000b,a) and the adaptive lasso (Zou 2006).

All computations were performed in R (R Core Team 2019) using the packages spatstat (Baddeley & Turner 2005; Baddeley, Rubak & Turner 2015), glmnet (Friedman et al. 2019) and gstat (Grler, Pebesma & Heuvelink 2016). The gstat package was used only to generate the spatial covariates $Z_1, \ldots, Z_{10}$ in Section 5. The `lppm` function in spatstat fits inhomogeneous Poisson process models to point patterns on a linear network (Baddeley & Turner 2005; Baddeley, Rubak & Turner 2015, Chap. 17). Instead of using `lppm` directly, we used the underlying internal functions to extract the data for the approximating generalised linear models, and then used glmnet to perform the penalised model-fitting and variable selection. Future versions of spatstat will allow this procedure to be performed using `lppm` alone, with appropriate command arguments. All figures were produced using spatstat, except that Figures 5 and 9 were produced using glmnet.

The glmnet package does not respect the usual rules for variable selection which require that, if an interaction term is selected, then the corresponding main effect terms must also be selected. This issue is not specific to spatial point process models. See Lim & Hastie (2019) for a newer package that does do this correctly.

Supplementary materials

The supplementary materials provide the data (as `RDS` files) and R scripts required for reproducing the results and plots in the paper. Two R scripts `Main.R` and `Utils.R` have been provided. The `Main.R` script contains the codes for reproducing the figures and contents of the tables provided in the paper; `Utils.R` contains the utility functions used in the computation. Due to the large size of the datasets, we have put them in the author's `Github` repository, and it can be accessed using the link: `https://github.com/rakstats/VarSelectOnLinnet`.

References

Aarts, G., Fieberg, J. & Matthiopoulos, J. (2012). Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution* **3**, 177–187.

Aitkin, M., Anderson, D., Francis, B. & Hinde, J. (1989). *Statistical Modelling in GLIM.* Oxford: Oxford University Press.

Anderes, E., Møller, J., Rasmussen, J.G. et al. (2020). Isotropic covariance functions on graphs and their edges. *Annals of Statistics* **48**, 2478–2503.

Ang, Q., Baddeley, A. & Nair, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics* **39**, 591–617.

Baddeley, A. (2010a). Modelling strategies. In *Handbook of Spatial Statistics*, eds. A. Gelfand, P. Diggle, M. Fuentes & P. Guttorp, chap. 20. Boca Raton: CRC Press, pp. 339–369.

Baddeley, A. (2010b). Multivariate and marked point processes. In *Handbook of Spatial Statistics*, eds. A. Gelfand, P. Diggle, M. Fuentes & P. Guttorp, chap. 21. Boca Raton: CRC Press, pp. 371–402.

Baddeley, A., Berman, M., Fisher, N., Hardegen, A., Milne, R., Schuhmacher, D., Shah, R. & Turner, R. (2010). Spatial logistic regression and change-of-support for Poisson point processes. *Electronic Journal of Statistics* **4**, 1151–1201. doi:10.1214/10-EJS581.

Baddeley, A., Nair, G., Rakshit, S. & McSwiggan, G. (2017). 'Stationary' point processes are uncommon on linear networks. *STAT* **6**, 68–78.

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G. & Davies, T.M. (2020). Analysing point patterns on networks — A review. *Spatial Statistics* (In press).

Baddeley, A., Rubak, E. & Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R.* London: Chapman and Hall/CRC.

Baddeley, A. & Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Australian and New Zealand Journal of Statistics* **42**, 283–322.

Baddeley, A. & Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software* **12**, 1–42. URL `http://www.jstatsoft.org/v12/i06/`.

Banerjee, S. & Gelfand, A. (2002). Prediction, interpolation and regression for spatially misaligned data. *Sanhkya A* **64**, 227–245.

Becker, R., Chambers, J. & Wilks, A. (1988). *The NEW S Language.* London: Chapman and Hall.

Berman, M. & Turner, T. (1992). Approximating point process likelihoods with GLIM. *Applied Statistics* **41**, 31–38.

Brillinger, D. (1978). Comparative aspects of the study of ordinary time series and of point processes. In *Developments in Statistics*, ed. P. Krishnaiah. New York, London: Academic Press, pp. 33–133.

Brillinger, D. & Preisler, H. (1986). Two examples of quantal data analysis: a) multivariate point process, b) pure death process in an experimental design. In *Proceedings, XIII International Biometric Conference, Seattle.* International Biometric Society, pp. 94–113.

Chambers, R. & Hastie, T. (eds.) (1992). *Statistical models in S.* Monterey: Wadsworth and Brooks/Cole.

Cox, D. & Lewis, P. (1966). *The Statistical Analysis of Series of Events.* London: Methuen.

Cressie, N. (1996). Change of support and the modifiable area unit problem. *Geographical Systems* **3**, 159–180.

Cressie, N., Frey, J., Harch, B. & Smith, M. (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological and Environmental Statistics* **11**, 127–150.

Daley, D. & Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods.* New York: Springer-Verlag, 2nd edn.

Donoho, D. & Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

EFRON, B., HASTIE, T. JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–499.

FARAWAY, J. (2005). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Chapman and Hall/CRC.

FITHIAN, W. & HASTIE, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics* **7**, 1917–1939.

FRIEDMAN, J., HASTIE, T., HÖFLING, H. & TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302–332.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R., SIMON, N., B.NARASIMHAN & QIAN, J. (2019). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.* URL `http://CRAN.R-project.org/package=glmnet`. R package version 2.0-18.

GOTWAY, C. & YOUNG, L. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association* **97**, 632–648.

GRLER, B., PEBESMA, E. & HEUVELINK, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal* **8**, 204–218. URL `https://journal.r-project.org/archive/2016/RJ-2016-014/index.html`.

GUAN, Y., JALILIAN, A. & WAAGEPETERSEN, R. (2015). Quasi-likelihood for spatial point processes. *Journal of the Royal Statistical Society, Series B* **77**, 677–697.

GUAN, Y. & WANG, H. (2010). Sufficient dimension reduction for spatial point processes directed by Gaussian random fields. *Journal of the Royal Statistical Society, Series B* **72**, 367–387.

HASTIE, T. & TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall/CRC.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2009). *The Elements of Statistical Learning.* New York: Springer-Verlag, 2nd edn.

HASTIE, T., TIBSHIRANI, R. & WAINWRIGHT, M. (2015). *Statistical learning with sparsity: the lasso and generalizations.* CRC press.

HILLIS, S.L. & DAVIS, C.S. (1994). A simple justification of the iterative fitting procedure for generalized linear models. *The American Statistican* **48**, 288–289.

ILLIAN, J., PENTTINEN, A., STOYAN, H. & STOYAN, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns.* Chichester: John Wiley and Sons.

JAMMALAMADAKA, A., BANERJEE, S., MANJUNATH, B. & KOSIK, K. (2013). Statistical analysis of dendritic spine distributions in rat hippocampal cultures. *BMC Bioinformatics* **14**.

KOOREY, G. (2009). Road data aggregation and sectioning considerations for crash analysis. *Transportation Research Record: Journal of the Transportation Research Board* , 61–68.

KUTOYANTS, Y. (1998). *Statistical Inference for Spatial Poisson Processes.* No. 134 in Lecture Notes in Statistics, New York: Springer.

LEWIS, P. (1972). Recent results in the statistical analysis of univariate point processes. In *Stochastic Point Processes*, ed. P. Lewis. New York: John Wiley and Sons, pp. 1–54.

LEWIS, P. & SHEDLER, G. (1979). Simulation of non-homogeneous Poisson processes by thinning. *Naval Logistics Quarterly* **26**, 406–413.

LIM, M. & HASTIE, T. (2019). *glinternet: Learning Interactions via Hierarchical Group-Lasso Regularization.* URL `http://CRAN.R-project.org/package=glinternet`. R package version 1.0-9.

LINDSEY, J. (1992). *The Analysis of Stochastic Processes using GLIM.* Berlin: Springer.

LINDSEY, J. (1995). *Modelling Frequency and Count Data.* Oxford: Oxford University Press.

LORD, D. & MANNERING, F. (2010). The statistical analysis of crash frequency data: a review and assessment of methodological alternatives. *Transportation Research A* **44**, 291–305.

Matthes, K. (1963). Stationäre zufällige Punktfolgen. *Jahresbericht Deutsche Mathematische Vereinigung* **66**, 66–79.

Matthes, K., Kerstan, J. & Mecke, J. (1978). *Infinitely Divisible Point Processes.* Chichester: John Wiley and Sons.

McSwiggan, G. (2019). Spatial point process methods for linear networks with applications to road accident analysis. Ph.D. thesis, University of Western Australia.

Nelder, J. & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

Okabe, A. & Satoh, T. (2009). Spatial analysis on a network. In *The SAGE Handbook on Spatial Analysis*, eds. A. Fotheringham & P. Rogers, chap. 23. London: SAGE Publications, pp. 443–464.

Okabe, A. & Sugihara, K. (2012). *Spatial Analysis Along Networks.* New York: John Wiley and Sons.

Openshaw, S. (1984). *The Modifiable Area Unit Problem.* Norwich: Geo Books.

Osborne, M.R., Presnell, B. & Turlach, B.A. (2000a). A new approach to variable selection in least squares problems. *IMA journal of numerical analysis* **20**, 389–403.

Osborne, M.R., Presnell, B. & Turlach, B.A. (2000b). On the lasso and its dual. *Journal of Computational and Graphical statistics* **9**, 319–337.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rakshit, S., Baddeley, A. & Nair, G. (2019). Efficient code for second-order analysis of events on a linear network. *Journal of Statistical Software* **90**, 1–37. doi:10.18637/jss.v090.i01. URL `https://www.jstatsoft.org/v090/i01`.

Rakshit, S., Nair, G. & Baddeley, A. (2017). Second-order analysis of point patterns on a network using any distance metric. *Spatial Statistics* **22**, 129–154.

Rathbun, S. & Cressie, N. (1994). Asymptotic properties of estimators of the parameters of spatial inhomogeneous Poisson point processes. *Advances in Applied Probability* **26**, 122–154.

Renner, I. (2013). Advances in presence-only methods in ecology. Phd thesis, University of New South Wales.

Renner, I., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S., Popovic, G. & Warton, D. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution* **6**, 366–379. doi:10.1111/2041-210X.12352.

Renner, I. & Warton, D. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* **69**, 274–281.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* **15**, 351–357.

Stoyan, D., Kendall, W. & Mecke, J. (1995). *Stochastic Geometry and its Applications.* Chichester: John Wiley and Sons, 2nd edn.

Thurman, A., Fu, R., Guan, Y. & Zhu, J. (2015). Regularized estimating equations for model selection of clustered spatial point processes. *Statistica Sinica* **25**, 173–188.

Thurman, A.L. & Zhu, J. (2014). Variable selection for spatial Poisson point processes via a regularization method. *Statistical Methodology* **17**, 113–125.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications* **109**, 475–494.

van Lieshout, M.N.M. (2018). Nearest-neighbour Markov point processes on graphs with Euclidean edges. *Advances in Applied Probability* **50**, 12751293. doi:10.1017/apr.2018.60.

Venables, W. & Ripley, B. (2002). *Modern Applied Statistics with S-Plus.* New York: Springer, 4th edn.

VER HOEF, J. & PETERSON, E. (2010). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association* **105**, 6–18.

VER HOEF, J., PETERSON, E. & THEOBALD, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics* **13**, 449–464.

WAAGEPETERSEN, R. & GUAN, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society, Series B* **71**, 685–702.

YUE, Y. & LOH, J. (2015). Variable selection for inhomogeneous spatial point process models. *Canadian Journal of Statistics* **43**, 288–305.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**, 1418–1429.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 301–320.

Table 1. Bias (standard error) in estimating the non-zero regression coefficients for the spatial covariates, shown in Fig. 3, on the Chicago network, using the lasso, ridge regression and elastic-net. The selection rule γ_{min} was used to choose the regularisation parameter.

Method	lixel/	Approx	$\beta_1 = 10.0$	$\beta_2 = 9.0$	$\beta_3 = 8.0$	$\beta_4 = 7.0$	$\beta_5 = 6.0$
	dummy	model					
Lasso	1810	logistic	-0.521 (1.64)	-0.782 (1.67)	-0.725 (1.79)	-1.39 (2.21)	-1.62 (2.22)
	1810	Poisson	-1.39 (1.91)	-1.52 (1.97)	-1.44 (2.01)	-1.92 (2.39)	-2.04 (2.42)
	3370	logistic	-0.252 (1.49)	-0.601 (1.48)	-0.773 (1.71)	-1.25 (2.05)	-1.47 (2.06)
	3370	Poisson	-0.86 (1.59)	-1.12 (1.65)	-1.23 (1.87)	-1.65 (2.2)	-1.82 (2.25)
	6481	logistic	-0.516 (1.5)	-0.659 (1.45)	-0.932 (1.75)	-1.25 (1.98)	-1.46 (2.01)
	6481	Poisson	-0.869 (1.58)	-0.95 (1.55)	-1.19 (1.86)	-1.46 (2.06)	-1.67 (2.14)
	1684	B–T	-0.69 (1.47)	-0.682 (1.4)	-1.03 (1.72)	-1.48 (2.03)	-1.6 (2.04)
	3063	B–T	-0.642 (1.47)	-0.731 (1.42)	-1.11 (1.8)	-1.31 (1.93)	-1.42 (1.95)
	5516	B–T	-0.595 (1.45)	-0.854 (1.48)	-1.09 (1.8)	-1.43 (2.03)	-1.55 (2.03)
Ridge	1810	logistic	-0.703 (1.634)	-0.937 (1.674)	-0.737 (1.674)	-1.157 (1.945)	-1.388 (1.967)
	1810	Poisson	-1.652 (2.077)	-1.759 (2.120)	-1.500 (1.970)	-1.735 (2.165)	-1.866 (2.205)
	3370	logistic	-0.434 (1.454)	-0.762 (1.469)	-0.769 (1.591)	-1.030 (1.805)	-1.254 (1.814)
	3370	Poisson	-1.131 (1.714)	-1.369 (1.786)	-1.280 (1.813)	-1.457 (1.961)	-1.632 (2.008)
	6481	logistic	-0.729 (1.515)	-0.854 (1.494)	-0.941 (1.647)	-1.042 (1.752)	-1.268 (1.790)
	6481	Poisson	-1.134 (1.690)	-1.196 (1.670)	-1.227 (1.783)	-1.272 (1.829)	-1.491 (1.905)
	1684	B–T	-0.925 (1.543)	-0.875 (1.439)	-1.049 (1.638)	-1.279 (1.803)	-1.401 (1.818)
	3063	B–T	-0.902 (1.542)	-0.979 (1.513)	-1.147 (1.720)	-1.142 (1.730)	-1.271 (1.739)
	5516	B–T	-0.858 (1.537)	-1.105 (1.599)	-1.132 (1.719)	-1.238 (1.793)	-1.368 (1.809)
E-net	1810	logistic	-0.599 (1.661)	-0.844 (1.709)	-0.747 (1.765)	-1.345 (2.154)	-1.574 (2.173)
	1810	Poisson	-1.450 (1.958)	-1.578 (2.007)	-1.450 (2.002)	-1.871 (2.337)	-1.998 (2.376)
	3370	logistic	-0.310 (1.505)	-0.653 (1.483)	-0.776 (1.693)	-1.197 (1.993)	-1.416 (2.001)
	3370	Poisson	-0.938 (1.634)	-1.193 (1.692)	-1.252 (1.870)	-1.611 (2.151)	-1.784 (2.201)
	6481	logistic	-0.575 (1.517)	-0.713 (1.478)	-0.936 (1.730)	-1.194 (1.932)	-1.415 (1.955)
	6481	Poisson	-0.940 (1.617)	-1.015 (1.588)	-1.200 (1.848)	-1.415 (2.005)	-1.632 (2.082)
	1684	B–T	-0.748 (1.501)	-0.729 (1.420)	-1.031 (1.707)	-1.424 (1.967)	-1.543 (1.988)
	3063	B–T	-0.704 (1.502)	-0.789 (1.447)	-1.114 (1.779)	-1.263 (1.891)	-1.382 (1.895)
	5516	B–T	-0.653 (1.485)	-0.910 (1.513)	-1.096 (1.779)	-1.376 (1.967)	-1.499 (1.981)

Table 2. Bias (standard error) in estimating the misspecified regression coefficients. The selection rule γ_{min} was used to choose regularisation parameters.

Method	lixel/ dummy	Approx model	$\beta_6 = 0$	$\beta_7 = 0$	$\beta_8 = 0$	$\beta_9 = 0$	$\beta_{10} = 0$
Lasso	1810	logistic	-0.491 (1.235)	0.019 (1.267)	-0.265 (1.237)	-0.155 (1.407)	0.165 (1.164)
	1810	Poisson	-0.408 (1.006)	0.034 (1.061)	-0.249 (1.020)	-0.101 (1.194)	0.157 (0.937)
	3370	logistic	-0.198 (1.050)	0.005 (1.158)	-0.183 (1.127)	-0.187 (1.303)	0.053 (1.050)
	3370	Poisson	-0.181 (0.914)	-0.003 (1.008)	-0.167 (0.983)	-0.168 (1.126)	0.040 (0.900)
	6481	logistic	-0.132 (0.961)	-0.009 (1.068)	-0.085 (1.030)	-0.125 (1.227)	0.091 (0.961)
	6481	Poisson	-0.119 (0.900)	-0.013 (1.003)	-0.071 (0.956)	-0.112 (1.140)	0.089 (0.902)
	1684	B–T	-0.291 (0.938)	0.071 (0.977)	-0.038 (0.932)	0.177 (1.124)	0.001 (0.906)
	3063	B–T	-0.112 (0.909)	0.005 (1.046)	-0.067 (0.963)	-0.176 (1.153)	0.012 (0.928)
	5516	B–T	-0.118 (0.894)	0.004 (1.006)	-0.137 (0.965)	-0.198 (1.148)	0.034 (0.900)
Ridge	1810	logistic	-0.596 (1.448)	-0.015 (1.520)	-0.411 (1.487)	-0.190 (1.676)	0.160 (1.365)
	1810	Poisson	-0.507 (1.202)	-0.005 (1.278)	-0.428 (1.259)	-0.164 (1.413)	0.192 (1.128)
	3370	logistic	-0.235 (1.267)	-0.032 (1.403)	-0.306 (1.381)	-0.232 (1.568)	0.062 (1.265)
	3370	Poisson	-0.226 (1.129)	-0.044 (1.255)	-0.334 (1.243)	-0.238 (1.385)	0.071 (1.110)
	6481	logistic	-0.174 (1.189)	-0.042 (1.324)	-0.190 (1.293)	-0.174 (1.492)	0.128 (1.179)
	6481	Poisson	-0.162 (1.115)	-0.053 (1.255)	-0.208 (1.215)	-0.177 (1.392)	0.137 (1.110)
	1684	B–T	-0.365 (1.161)	0.068 (1.235)	-0.142 (1.183)	0.202 (1.366)	0.003 (1.112)
	3063	B–T	-0.151 (1.114)	-0.041 (1.281)	-0.190 (1.200)	-0.237 (1.400)	0.057 (1.125)
	5516	B–T	-0.170 (1.116)	-0.033 (1.264)	-0.284 (1.234)	-0.291 (1.415)	0.058 (1.112)
E-net	1810	logistic	-0.512 (1.278)	0.013 (1.323)	-0.295 (1.289)	-0.160 (1.462)	0.171 (1.201)
	1810	Poisson	-0.426 (1.045)	0.013 (1.098)	-0.284 (1.062)	-0.111 (1.228)	0.164 (0.975)
	3370	logistic	-0.207 (1.095)	-0.001 (1.206)	-0.209 (1.182)	-0.208 (1.358)	0.060 (1.098)
	3370	Poisson	-0.184 (0.950)	-0.016 (1.044)	-0.191 (1.027)	-0.183 (1.169)	0.052 (0.934)
	6481	logistic	-0.144 (1.011)	-0.017 (1.118)	-0.099 (1.090)	-0.140 (1.279)	0.099 (1.005)
	6481	Poisson	-0.124 (0.937)	-0.019 (1.055)	-0.092 (1.008)	-0.126 (1.184)	0.100 (0.939)
	1684	B–T	-0.307 (0.987)	0.072 (1.035)	-0.061 (0.982)	0.180 (1.159)	0.003 (0.945)
	3063	B–T	-0.114 (0.953)	-0.002 (1.092)	-0.090 (1.008)	-0.188 (1.201)	0.023 (0.963)
	5516	B–T	-0.129 (0.945)	0.000 (1.060)	-0.158 (1.018)	-0.220 (1.200)	0.031 (0.941)

Table 3. Proportion of simulated outcomes in which each coefficient is included in the model, when the regularisation parameter γ is selected by mse cross-validation (min), one standard error stronger than MSE (1se) or the average of these two rules (avg).

Method	lixel/	Model	γ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	dummy		rule	(10)	(9)	(8)	(7)	(6)	(0)	(0)	(0)	(0)	(0)
Lasso	1810	logistic	min	1.00	1.00	1.00	1.00	1.00	0.68	0.68	0.67	0.63	0.64
			avg	1.00	1.00	1.00	0.90	0.81	0.14	0.06	0.09	0.05	0.09
			1se	0.89	0.91	0.88	0.51	0.29	0.01	0.00	0.00	0.00	0.01
	1810	Poisson	min	1.00	1.00	1.00	1.00	1.00	0.67	0.66	0.65	0.60	0.64
			avg	1.00	1.00	1.00	0.86	0.73	0.11	0.03	0.04	0.03	0.06
			1se	0.73	0.73	0.70	0.35	0.21	0.00	0.00	0.00	0.00	0.00
	3370	logistic	min	1.00	1.00	1.00	1.00	1.00	0.66	0.67	0.66	0.62	0.66
			avg	1.00	1.00	0.99	0.78	0.65	0.02	0.01	0.02	0.01	0.02
			1se	0.48	0.48	0.43	0.14	0.07	0.00	0.00	0.00	0.00	0.00
	3370	Poisson	min	1.00	1.00	1.00	1.00	1.00	0.64	0.65	0.63	0.60	0.64
			avg	1.00	1.00	0.99	0.75	0.59	0.02	0.00	0.01	0.01	0.01
			1se	0.33	0.34	0.30	0.10	0.06	0.00	0.00	0.00	0.00	0.00
	6481	logistic	min	1.00	1.00	1.00	1.00	1.00	0.64	0.66	0.65	0.59	0.64
			avg	1.00	1.00	0.98	0.72	0.50	0.00	0.00	0.00	0.00	0.00
			1se	0.04	0.04	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	6481	Poisson	min	1.00	1.00	1.00	1.00	1.00	0.63	0.64	0.63	0.59	0.64
			avg	1.00	1.00	0.99	0.72	0.51	0.00	0.00	0.00	0.00	0.00
			1se	0.04	0.04	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	1684	B–T	min	1.00	1.00	1.00	1.00	1.00	0.64	0.66	0.64	0.59	0.63
			avg	1.00	1.00	0.99	0.79	0.62	0.01	0.00	0.00	0.00	0.01
			1se	0.29	0.30	0.26	0.10	0.06	0.00	0.00	0.00	0.00	0.00
	3063	B–T	min	1.00	1.00	1.00	1.00	1.00	0.66	0.66	0.64	0.60	0.65
			avg	1.00	1.00	0.98	0.73	0.54	0.00	0.00	0.00	0.00	0.00
			1se	0.05	0.05	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	5516	B–T	min	1.00	1.00	1.00	1.00	1.00	0.64	0.64	0.62	0.58	0.63
			avg	1.00	1.00	0.98	0.71	0.51	0.00	0.00	0.00	0.00	0.00
			1se	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E-net	1810	logistic	min	1.00	1.00	1.00	1.00	1.00	0.74	0.75	0.74	0.70	0.73
			avg	1.00	1.00	1.00	0.94	0.85	0.18	0.10	0.13	0.07	0.12
			1se	0.91	0.91	0.89	0.58	0.35	0.02	0.01	0.01	0.00	0.01
	1810	Poisson	min	1.00	1.00	1.00	1.00	1.00	0.74	0.74	0.73	0.68	0.69
			avg	1.00	1.00	1.00	0.89	0.77	0.16	0.06	0.08	0.05	0.08
			1se	0.75	0.74	0.71	0.43	0.28	0.02	0.00	0.01	0.00	0.00
	3370	logistic	min	1.00	1.00	1.00	1.00	1.00	0.74	0.76	0.75	0.71	0.75
			avg	1.00	1.00	0.99	0.82	0.66	0.04	0.02	0.03	0.02	0.04
			1se	0.49	0.52	0.48	0.21	0.12	0.00	0.00	0.00	0.00	0.00
	3370	Poisson	min	1.00	1.00	1.00	1.00	1.00	0.70	0.72	0.70	0.68	0.70
			avg	1.00	1.00	0.99	0.77	0.61	0.02	0.00	0.02	0.01	0.02
			1se	0.32	0.34	0.30	0.12	0.08	0.00	0.00	0.00	0.00	0.00
	6481	logistic	min	1.00	1.00	1.00	1.00	1.00	0.72	0.76	0.72	0.69	0.73
			avg	1.00	1.00	0.99	0.74	0.52	0.01	0.00	0.00	0.00	0.00
			1se	0.04	0.04	0.03	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	6481	Poisson	min	1.00	1.00	1.00	1.00	1.00	0.70	0.71	0.69	0.66	0.70
			avg	1.00	1.00	0.99	0.74	0.52	0.00	0.00	0.00	0.00	0.00
			1se	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	1684	B–T	min	1.00	1.00	1.00	1.00	1.00	0.73	0.74	0.72	0.70	0.72
			avg	1.00	1.00	0.99	0.82	0.64	0.02	0.01	0.01	0.01	0.02
			1se	0.31	0.32	0.28	0.14	0.08	0.00	0.00	0.00	0.00	0.00
	3063	B–T	min	1.00	1.00	1.00	1.00	1.00	0.72	0.73	0.73	0.69	0.73
			avg	1.00	1.00	0.99	0.77	0.55	0.01	0.00	0.00	0.00	0.00
			1se	0.05	0.05	0.04	0.02	0.01	0.00	0.00	0.00	0.00	0.00
	5516	B–T	min	1.00	1.00	1.00	1.00	1.00	0.70	0.73	0.70	0.68	0.71
			avg	1.00	1.00	0.99	0.74	0.53	0.00	0.00	0.00	0.00	0.00
			1se	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4. Spatial covariates associated with the WA state road network, alongside the maximum (Max) and minimum (Min) values used to produce scaled measurements.

Covariate name	Type	Description	Max	Min
SPD_LIM	numeric	Legal speed limit	40	110
H_CURVE	numeric	Radius of horizontal curve in metres	0	99999
TOT_P	numeric	Width of pavement in meters	0	33.9
TOT_S	numeric	Width of road-side seal in metres	0	31.6
TRFABL	numeric	Width of trafficable road surface in metres	0	96.8
N_LANE	integer	Number of lanes	1	7
SHLDR	binary	Presence of shoulder-padding	0	1
KERB_L	binary	Presence of kerb on left side of road	0	1
KERB_R	binary	Presence of kerb on right side of road	0	1
FLDWY	binary	Presence of floodway	0	1
BRDG	binary	Presence of bridge over road	0	1

Table 5. Some mark variables attributed to accidents in the WA data.

Mark variable	Type	Description
Severity	binary	Severity of accident (Low/High)
Day_No	factor	Day number in the week
Time	factor	Time of the accident
Spd_Fact	binary	True if speeding was a cause
Police	binary	True if police case was filed
Inattention	binary	True if driver inattention was a cause
Fatigue	binary	True if fatigue was a cause
Cond	factor	Weather-dependent road condition

Table 6. Percentage deviance explained (Deviance), as compared to the deviance explained by a full model (Max-deviance), and the number of variables (N) selected in regularised models corresponding to the parameters γ_{min}, $\gamma_{1\text{se}}$, γ_{avg}, and γ_{gmean} for the WA accident pattern analysis; rows written in bold font correspond to the selected models.

Method	Approximation	γ	$\log\gamma$	Deviance (%)	Max-deviance (%)	N
lasso	B-T	γ_{min}	-16.58	29.95	30.09	36
lasso	B-T	$\gamma_{1\text{se}}$	-9.51	15.46	30.09	2
lasso	B-T	γ_{avg}	-10.20	18.26	30.09	3
lasso	**B-T**	$\boldsymbol{\gamma_{\text{gmean}}}$	**-13.04**	**27.67**	**30.09**	**17**
lasso	Logistic	γ_{min}	-14.06	29.53	29.53	43
lasso	Logistic	$\gamma_{1\text{se}}$	-9.23	27.10	29.53	15
lasso	**Logistic**	$\boldsymbol{\gamma_{\text{avg}}}$	**-9.91**	**28.04**	**29.53**	**22**
lasso	Logistic	γ_{gmean}	-11.64	29.20	29.53	28
lasso	Poisson	γ_{min}	-11.80	38.27	38.54	33
lasso	Poisson	$\gamma_{1\text{se}}$	-10.31	37.42	38.54	24
lasso	**Poisson**	$\boldsymbol{\gamma_{\text{avg}}}$	**-10.80**	**37.92**	**38.54**	**27**
lasso	Poisson	γ_{gmean}	-11.05	38.03	38.54	27
ridge	**B-T**	$\boldsymbol{\gamma_{\text{min}}}$	**-10.14**	**22.69**	**24.93**	**44**
ridge	B-T	$\gamma_{1\text{se}}$	-1.95	0.00	24.93	44
ridge	B-T	γ_{avg}	-2.64	0.09	24.93	44
ridge	B-T	γ_{gmean}	-6.04	2.59	24.93	44
ridge	Logistic	γ_{min}	-7.16	24.49	24.57	44
ridge	Logistic	$\gamma_{1\text{se}}$	-6.13	22.30	24.57	44
ridge	**Logistic**	$\boldsymbol{\gamma_{\text{avg}}}$	**-6.52**	**23.38**	**24.57**	**44**
ridge	Logistic	γ_{gmean}	-6.64	23.58	24.57	44
ridge	Poisson	γ_{min}	-6.38	30.42	31.93	44
ridge	Poisson	$\gamma_{1\text{se}}$	-5.17	26.30	31.93	44
ridge	**Poisson**	$\boldsymbol{\gamma_{\text{avg}}}$	**-5.60**	**28.13**	**31.93**	**44**
ridge	Poisson	γ_{gmean}	-5.77	28.76	31.93	44
e-net	B-T	γ_{min}	-17.37	29.92	29.93	43
e-net	B-T	$\gamma_{1\text{se}}$	-8.44	3.49	29.93	3
e-net	B-T	γ_{avg}	-9.14	14.04	29.93	4
e-net	**B-T**	$\boldsymbol{\gamma_{\text{gmean}}}$	**-12.91**	**27.40**	**29.93**	**25**
e-net	Logistic	γ_{min}	-13.37	29.34	29.35	44
e-net	Logistic	$\gamma_{1\text{se}}$	-9.84	27.48	29.35	29
e-net	**Logistic**	$\boldsymbol{\gamma_{\text{avg}}}$	**-10.50**	**28.05**	**29.35**	**33**
e-net	Logistic	γ_{gmean}	-11.60	28.69	29.35	36
e-net	Poisson	γ_{min}	-13.15	38.33	38.34	43
e-net	Poisson	$\gamma_{1\text{se}}$	-10.64	36.93	38.34	33
e-net	**Poisson**	$\boldsymbol{\gamma_{\text{avg}}}$	**-11.25**	**37.46**	**38.34**	**35**
e-net	Poisson	γ_{gmean}	-11.89	37.88	38.34	41

Table 7. Estimates of the regression coefficients corresponding to the scaled WA road variables.

Variable	B-T (lasso)	Logistic (lasso)	Poisson (lasso)	B-T (ridge)	Logistic (ridge)	Poisson (ridge)	B-T (e-net)	Logistic (e-net)	Poisson (e-net)
(Intercept)	-9.408	-10.182	-11.184	-8.320	-8.446	-8.273	-9.090	-9.997	-10.699
SPD_LIM	·	2.807	9.975	-0.398	-0.446	-0.362	·	2.025	4.887
H_CURVE	·	·	·	0.007	0.012	0.005	·	0.005	0.678
TOT_P	·	·	0.385	0.243	0.291	0.205	0.559	1.267	2.490
TOT_S	4.422	6.349	9.317	0.401	0.499	0.334	2.217	4.328	6.272
TRFABL	·	·	·	0.102	0.123	0.087	·	·	·
N_LANE	1.884	1.422	·	0.368	0.420	0.306	1.706	1.311	0.871
SPD_LIM^2	-1.593	-4.242	-10.992	-0.543	-0.641	-0.481	-1.604	-3.606	-6.338
H_CURVE^2	·	·	·	-0.000	-0.000	-0.000	·	·	·
TOT_P^2	·	·	·	0.190	0.232	0.164	·	·	-1.148
TOT_S^2	·	·	-2.792	0.278	0.343	0.235	0.234	0.371	-1.189
TRFABL^2	·	·	·	0.022	0.027	0.019	·	·	·
N_LANE^2	·	·	·	0.257	0.295	0.216	0.282	0.257	-0.125
SHLDR	·	-0.030	-0.944	0.028	0.053	0.016	·	-0.285	-0.658
KERB_L	1.445	1.909	1.929	0.700	0.764	0.649	1.536	2.032	2.013
KERB_R	1.901	2.145	2.323	0.614	0.706	0.575	1.601	2.285	2.389
FLDWY	-1.065	-0.834	-0.675	-0.381	-0.468	-0.326	-0.952	-0.873	-0.824
BRDG	0.137	0.123	0.275	0.248	0.251	0.214	0.377	0.308	0.362
SPD_LIM×KERB_L	0.300	0.267	·	0.278	0.292	0.259	0.474	0.465	0.355
H_CURVE×KERB_L	·	·	·	-0.000	-0.000	0.000	·	·	·
TOT_P×KERB_L	·	·	·	0.260	0.271	0.253	-0.044	-0.773	-0.666
TOT_S×KERB_L	-0.025	-2.070	-3.441	0.280	0.298	0.270	-0.117	-1.503	-2.519
TRFABL×KERB_L	·	·	·	0.093	0.101	0.088	·	·	·
N_LANE×KERB_L	·	1.159	2.254	0.319	0.352	0.288	·	1.117	1.740
SHLDR×KERB_L	-0.763	-0.701	-0.324	-0.075	-0.173	-0.018	-0.785	-0.791	-0.561
KERB_R×KERB_L	-0.418	-0.597	-0.557	0.365	0.341	0.393	-0.444	-0.708	-0.622
FLDWY×KERB_L	·	·	·	-0.004	-0.006	-0.003	·	·	·
BRDG×KERB_L	·	·	-0.035	0.066	0.044	0.076	·	-0.092	-0.076
SPD_LIM×SHLDR	0.068	0.119	0.939	-0.035	-0.019	-0.051	0.077	0.235	0.595
H_CURVE×SHLDR	·	·	·	0.007	0.012	0.006	·	0.003	0.598
TOT_P×SHLDR	·	·	·	0.188	0.238	0.153	0.417	0.315	·
TOT_S×SHLDR	1.187	1.054	2.022	0.246	0.309	0.200	1.324	1.955	2.809
TRFABL×SHLDR	·	·	-1.728	0.045	0.054	0.038	·	-0.334	-1.901
N_LANE×SHLDR	·	·	-0.000	0.188	0.217	0.156	·	-0.475	-0.671
KERB_R×SHLDR	-0.246	-0.315	-0.146	-0.122	-0.173	-0.059	-0.441	-0.335	-0.194
FLDWY×SHLDR	·	-0.362	-0.543	-0.161	-0.199	-0.136	-0.262	-0.479	-0.548
BRDG×SHLDR	0.820	0.839	0.723	0.272	0.282	0.225	0.686	0.769	0.704
SPD_LIM×KERB_R	·	·	-0.140	0.184	0.207	0.183	0.062	·	-0.001
H_CURVE×KERB_R	·	·	·	-0.001	-0.002	-0.000	·	·	·
TOT_P×KERB_R	·	-0.150	-0.713	0.204	0.225	0.209	-0.431	-1.253	-1.516
TOT_S×KERB_R	-2.128	-2.244	-3.250	0.221	0.250	0.224	-0.481	-1.572	-2.393
TRFABL×KERB_R	·	·	·	0.079	0.090	0.077	·	·	·
N_LANE×KERB_R	·	·	1.019	0.270	0.308	0.250	·	0.569	0.886
FLDWY×KERB_R	·	·	·	-0.004	-0.006	-0.003	·	·	-0.019
BRDG×KERB_R	-0.118	-0.110	-0.350	-0.006	-0.003	0.021	-0.346	-0.249	-0.389

Table 8. Variable selection results for the log-quadratic model of the Chicago data including crime type (Fig. 6) using the lasso penalty applied to the Berman–Turner approximation.

Variable	Coefficient	Estimate	Variable	Coefficient	Estimate
Intercept	β_0	-6.61	Damage	β_6	0.34
x	β_1	·	Damage $\times x$	β_7	-0.27
y	β_2	0.38	Damage $\times y$	β_8	0.24
x^2	β_3	·	Damage $\times x^2$	β_9	·
xy	β_4	0.16	Damage $\times xy$	β_{10}	·
y^2	β_5	·	Damage $\times y^2$	β_{11}	-0.09

Table 9. Estimates of regression coefficients for scaled road variables, computed using the lasso, ridge and elastic-net penalties applied to the Berman–Turner approximation, for the Western Australia accident data with mark indicating high (Severity = 1) and low (Severity = 0) severity.

Variable	B-T	B-T	B-T	Variable	B-T	B-T	B-T
	(lasso)	(ridge)	(e-net)		(lasso)	(ridge)	(e-net)
(Intercept)	-9.351	-8.751	-9.346	Severity	-0.945	-0.375	-0.913
SPD_LIM	·	-0.369	·	Severity×SPD_LIM	·	-0.219	·
H_CURVE	·	0.007	·	Severity×H_CURVE	·	0.001	·
TOT_P	·	0.241	0.466	Severity×TOT_P	·	-0.062	·
TOT_S	2.956	0.402	2.039	Severity×TOT_S	·	-0.020	·
TRFABL	·	0.102	·	Severity×TRFABL	·	-0.012	·
N_LANE	1.523	0.365	1.604	Severity×N_LANE	·	-0.003	·
SPD_LIM^2	-1.410	-0.512	-1.559	Severity×SPD_LIM^2	·	-0.219	·
H_CURVE^2	·	-0.000	·	Severity×H_CURVE^2	·	-0.000	·
TOT_P^2	·	0.188	·	Severity×TOT_P^2	·	0.003	·
TOT_S^2	·	0.276	0.221	Severity×TOT_S^2	·	0.022	·
TRFABL^2	·	0.022	·	Severity×TRFABL^2	·	-0.000	·
N_LANE^2	·	0.254	0.271	Severity×N_LANE^2	·	0.015	·
SHLDR	·	0.042	·	Severity×SHLDR	·	-0.076	·
KERB_L	1.550	0.715	1.502	Severity×KERB_L	·	-0.091	-0.086
KERB_R	1.158	0.632	1.527	Severity×KERB_R	·	-0.099	·
FLDWY	-1.007	-0.374	-0.947	Severity×FLDWY	·	-0.103	·
BRDG	0.081	0.250	0.374	Severity×BRDG	·	0.013	·
SPD_LIM×KERB_L	0.105	0.290	0.460	Severity×SPD_LIM×KERB_L	·	-0.026	·
H_CURVE×KERB_L	·	-0.000	·	Severity×H_CURVE×KERB_L	·	-0.000	·
TOT_P×KERB_L	·	0.262	-0.004	Severity×TOT_P×KERB_L	·	-0.029	·
TOT_S×KERB_L	·	0.282	-0.023	Severity×TOT_S×KERB_L	·	-0.031	·
TRFABL×KERB_L	·	0.093	·	Severity×TRFABL×KERB_L	·	-0.009	·
N_LANE×KERB_L	·	0.318	·	Severity×N_LANE×KERB_L	·	-0.002	·
SHLDR×KERB_L	-0.614	-0.062	-0.732	Severity×SHLDR×KERB_L	·	-0.075	-0.099
KERB_R×KERB_L	-0.200	0.383	-0.330	Severity×KERB_R×KERB_L	-0.430	-0.147	-0.395
FLDWY×KERB_L	·	-0.004	·	Severity×FLDWY×KERB_L	·	-0.001	·
BRDG×KERB_L	·	0.071	·	Severity×BRDG×KERB_L	·	-0.018	·
SPD_LIM×SHLDR	·	-0.017	0.069	Severity×SPD_LIM×SHLDR	·	-0.028	·
H_CURVE×SHLDR	·	0.008	·	Severity×H_CURVE×SHLDR	·	0.001	·
TOT_P×SHLDR	·	0.191	0.448	Severity×TOT_P×SHLDR	·	0.016	·
TOT_S×SHLDR	1.375	0.249	1.277	Severity×TOT_S×SHLDR	·	0.026	·
TRFABL×SHLDR	·	0.045	·	Severity×TRFABL×SHLDR	·	0.003	·
N_LANE×SHLDR	·	0.189	·	Severity×N_LANE×SHLDR	·	0.022	·
KERB_R×SHLDR	-0.417	-0.107	-0.442	Severity×KERB_R×SHLDR	·	-0.078	-0.016
FLDWY×SHLDR	·	-0.156	-0.235	Severity×FLDWY×SHLDR	·	-0.038	·
BRDG×SHLDR	0.796	0.270	0.672	Severity×BRDG×SHLDR	·	0.042	·
SPD_LIM×KERB_R	·	0.198	0.041	Severity×SPD_LIM×KERB_R	·	-0.042	·
H_CURVE×KERB_R	·	-0.001	·	Severity×H_CURVE×KERB_R	·	-0.000	·
TOT_P×KERB_R	·	0.208	-0.327	Severity×TOT_P×KERB_R	·	-0.039	·
TOT_S×KERB_R	-0.090	0.225	-0.360	Severity×TOT_S×KERB_R	·	-0.041	-0.003
TRFABL×KERB_R	·	0.080	·	Severity×TRFABL×KERB_R	·	-0.012	·
N_LANE×KERB_R	·	0.271	·	Severity×N_LANE×KERB_R	·	-0.011	·
FLDWY×KERB_R	·	-0.004	·	Severity×FLDWY×KERB_R	·	-0.001	·
BRDG×KERB_R	·	-0.000	-0.306	Severity×BRDG×KERB_R	·	-0.040	·