Citation
McSwiggan, G. and Baddeley, A. and Nair, G. 2020. Estimation of relative risk for events on a linear network. Statistics and Computing. 30 (2): pp. 469-484. http://doi.org/10.1007/s11222-019-09889-7

Statistics and Computing manuscript No.
(will be inserted by the editor)

Estimation of relative risk for events on a linear network

Greg McSwiggan[1] · **Adrian Baddeley**[2,3] · **Gopalan Nair** [1,3]

Received: date / Accepted: date

Abstract Motivated by the study of traffic accidents on a road network, we discuss the estimation of the relative risk, the ratio of rates of occurrence of different types of events occurring on a network of lines. Methods developed for two-dimensional spatial point patterns can be adapted to a linear network, but their requirements and performance are very different on a network. Computation is slow and we introduce new techniques to accelerate it. Intensities (occurrence rates) are estimated by kernel smoothing using the heat kernel on the network. The main methodological problem is bandwidth selection. Binary regression methods, such as likelihood cross-validation and least squares cross-validation, perform tolerably well in our simulation experiments, but the Kelsall-Diggle density-ratio cross-validation method does not. We find a theoretical explanation, and propose a modification of the Kelsall-Diggle method which has better performance. The methods are applied to traffic accidents in a regional city, and to protrusions on the dendritic tree of a neuron.

Keywords Bandwidth selection · Cross-validation · Dendritic spines · Density ratio · Heat kernel · Kelsall-Diggle cross-validation · Road traffic accidents

Research supported by the Australian Research Council, Discovery Grants DP130102322 and DP130104470.

✉Adrian Baddeley
adrian.baddeley@curtin.edu.au
· [1] Department of Mathematics & Statistics, University of Western Australia, 35 Stirling Hwy, Nedlands WA 6009, Australia
· [2] Department of Mathematics & Statistics, Curtin University, GPO Box U1987, Perth WA 6845, Australia
· [3] Data61, CSIRO Leeuwin Centre, 65 Brockway Rd, Floreat WA 6014, Australia

1 Introduction

Statistical methodology for analysing a spatial pattern of events on a network of lines, such as traffic accidents on a road network, has recently become the focus of intensive research (Okabe and Sugihara, 2012, Baddeley et al, 2015, Chap. 17). Figure 1 shows one of our motivating datasets, which records the spatial locations of serious accidents on state-declared roads in the Australian regional city of Geelong over a three-year period. The analysis of such data presents many methodological and technical challenges.

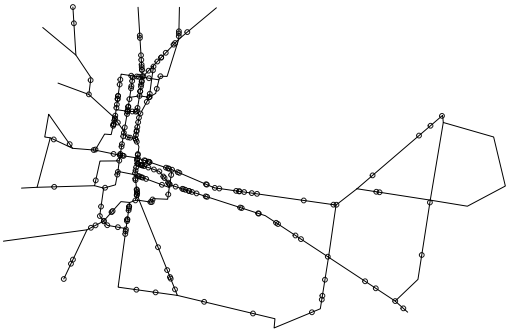

Fig. 1 High severity traffic accidents (circles) on state-declared roads (lines) in Geelong, Australia, 2009–2011. Downloaded from CrashStats interactive statistics database at `www.vicroads.vic.gov.au`. Figure is approximately 46 by 29 km across.

Previous research has focused on kernel estimation of the spatially-varying accident rate, expressed as the expected count of accidents per unit length of network in a fixed period, without adjusting for traffic volume or other explanatory variables. Kernel estimation on a network was developed by Borruso (2003, 2005, 2008); Downs and Horner (2007a,b, 2008); Xie and Yan (2008); Okabe et al (2009); Sugihara et al (2010) and summarised in Okabe and Sugi-

hara (2012, Chap. 9). We have proposed new kernel techniques in McSwiggan et al (2016); Rakshit et al (2019) and applied kernel smoothing to Figure 1 in McSwiggan et al (2016).

This paper aims to estimate the *relative risk*, the spatially-varying ratio of the intensities of two different types of events. Examples include the relative proportions of single-vehicle and multi-vehicle accidents; daytime and night-time accidents; high-speed and low-speed accidents; and those involving private and commercial vehicles. Figure 2 shows the Geelong crash data classified into "day" and "night" accidents (numbering 144 and 98 accidents, respectively). Another example, from neuroscience, is shown in Section 7.2.

Estimation of relative risk is different from estimation of the absolute accident rate, particularly with regard to the choice of smoothing bandwidth. For example, if the bandwidth is chosen to be *infinite*, the resulting smoothed function is constant with respect to spatial location; a constant accident rate is implausible, but a constant relative risk between two types of accidents is a reasonable null hypothesis in many applications.

Relative risk estimates are also less susceptible to Simpson's Paradox (Yule, 1903). For example, the traffic accident rate is influenced by the weather, but if we assume that weather has the same multiplicative effect on day and night accident rates, then the relative risk of day and night accidents can be estimated without needing to adjust for weather.

Relative risk estimation for spatial point patterns in two-dimensional space is well developed (Kelsall and Diggle, 1995a,b, 1998; Duong and Hazelton, 2003, 2005; Clark and Lawson, 2004; Diggle et al, 2005; Hazelton and Davies, 2009; Davies et al, 2016). In this paper we adapt and extend the two-dimensional techniques to point patterns on a linear network. Kernel estimates of relative risk can be obtained simply by taking the ratio of kernel estimates of the intensity functions (accident rates) of the two types of event. The main problem is to choose the smoothing bandwidth for kernel estimation, and to decide whether the numerator and denominator should be estimated using the same bandwidth (a "symmetric regimen", Davies et al, 2016) or whether different bandwidths may be permitted (Kelsall and Diggle, 1995a,b). Estimation on a linear network presents new challenges and exigencies: computation is much slower than in Euclidean space because the Fast Fourier Transform cannot be used; the leave-one-out kernel estimate is very costly.

In this paper, several standard methods for bandwidth selection for relative risk in two dimensions are adapted and extended to linear networks. These include the normal reference rule (Scott, 1992, p. 152), density-ratio cross-validation (Kelsall and Diggle, 1995a,b), and binary likelihood and binary least squares cross-validation (Kelsall and Diggle, 1998). Asymptotic performance and optimal bandwidths are derived; small-sample performance is evaluated in simulation experiments.

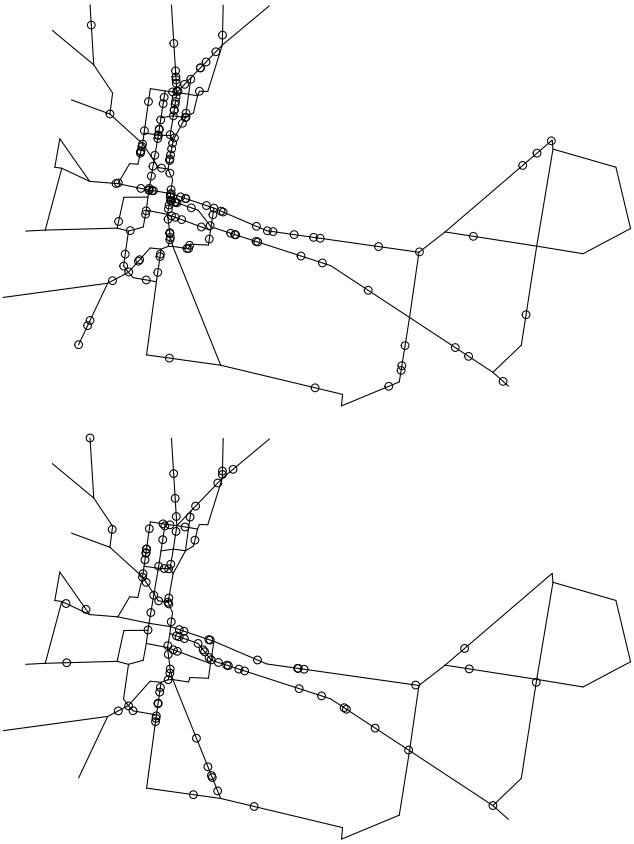

Fig. 2 Geelong data split into daytime (*Top*) and night-time (*Bottom*) accidents.

Kelsall and Diggle (1995a,b) reported that their density-ratio cross-validation method suffered sporadic "breakdowns" in which the selected bandwidths and resulting risk estimates were very unsatisfactory. On linear networks, these breakdowns occur even more frequently. We have found a theoretical explanation for breakdown, in any spatial domain, and propose a modification of the Kelsall-Diggle method to improve its performance. Our simulation experiments demonstrate this improvement.

Throughout this paper the smoothing kernel is chosen to be the classical heat kernel, the analogue for linear networks of the Gaussian kernel (McSwiggan et al, 2016). This is the Gaussian extension of the popular "equal-split continuous" rule (Okabe et al, 2009, Okabe and Sugihara, 2012, Chap. 9). However, the methods of this paper could be applied to any of the competing kernel estimators discussed in the literature.

Section 2 states necessary background. Section 3 describes the estimation of relative risk using direct adaptations of standard methods for two-dimensional data. Section 4 analyses the weaknesses of these methods and proposes our modified version. Section 5 proposes a fast ap-

proximation to the leave-one-out estimate which is needed for practical applications. Section 6 reports the results of a simulation experiment to measure performance of the methods. Section 7 reports our analysis of two datasets: the Geelong road accidents, and the spatial pattern of protrusions on the dendritic tree of a neuron. We end with a discussion in Section 8.

2 Background

2.1 Linear networks and point patterns

Following Ang et al (2012, Section 2) we define a linear network as the union $L = \cup_{i=1}^{N} l_i$ of finitely many line segments $l_i = [u_i, v_i] = \{w : w = tu_i + (1-t)v_i, 0 \le t \le 1\}$ in the two-dimensional plane, where $u_i, v_i \in \mathbb{R}^2$ are the endpoints of l_i. We assume that the intersection of any two segments l_i and l_j, with $i \neq j$, is either empty or is an endpoint of both segments.

Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ denote an observed point pattern on a linear network L, where each point x_i represents a location on L, and the number of points n is not fixed in advance. We assume that $\mathbf{x}$ is a realisation of a finite, simple point process $\mathbf{X}$ on L such that the total number of points has finite second moment.

2.2 Intensity

In this paper we work with the intensity or rate function $\lambda(u)$ of the point process of accidents, rather than the probability density $f(u)$ of the location of a typical accident. The intensity and probability density are very closely related, and our main results can easily be rephrased in terms of probability densities (Diggle and Marron, 1988).

Formally the point process $\mathbf{X}$ on L is defined to have *intensity function* $\lambda(u)$, $u \in L$, if

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

for all intervals $B \subset L$, where $\mathrm{d}u$ denotes integration with respect to arc length along the network.

The intensity can be interpreted as the spatially-varying expected number of random points per unit length of network. If $\mathbf{X}$ is a *Poisson* process with intensity function $\lambda(u)$, then conditional on the number of points $n(\mathbf{X}) = n$, the locations of the points are independent and identically distributed with probability density $f(u) = \lambda(u)/\Lambda$, where $\Lambda = \int_L \lambda(u)\,\mathrm{d}u$. Using intensity rather than probability density simplifies some technical statements and makes it easier to allow dependence between points.

Measure-theoretic details will mostly be omitted, but we note that if (1) holds for intervals B in L, then it holds whenever B is a Borel subset of L, i.e. the intersection of L with a Borel subset of $\mathbb{R}^2$. *Campbell's theorem* on a network states that

$$E\left[\sum_{x_i \in \mathbf{X}} h(x_i)\right] = \int_L h(u)\lambda(u)\,\mathrm{d}u, \tag{2}$$

where h is any Borel-measurable function (on L) such that $\int_L |h(u)|\lambda(u)\,\mathrm{d}u < \infty$.

2.3 Kernel estimation of intensity

A kernel estimator of intensity takes the general form

$$\widehat{\lambda}(u) = \sum_{i=1}^{n} K(u \mid x_i), \qquad u \in L, \tag{3}$$

where $K(u \mid v)$ is the kernel function. Numerous kernel estimators have been proposed (Borruso, 2003, 2005, 2008; Downs and Horner, 2007a,b, 2008; Xie and Yan, 2008; Okabe et al, 2009; Sugihara et al, 2010; Okabe and Sugihara, 2012, Chap. 9; McSwiggan et al, 2016; Rakshit et al, 2019), but there is no consensus on the choice of the kernel K, and indeed some of the proposals do not satisfy basic requirements. Recently we showed (McSwiggan et al, 2016) that the popular, but computationally very expensive, "equal-split continuous" estimator (Okabe and Sugihara, 2012, Chap. 9) is formally equivalent, in a special case, to a diffusion estimator (Chaudhuri and Marron, 2000; Botev et al, 2010) obtained by taking K to be the classical heat kernel on the network. Consequently, this estimator enjoys many desirable statistical properties including unbiasedness and conservation of mass, and it can be calculated quickly by solving the classical time-dependent heat equation on the network. Accordingly we use the diffusion estimator in this study. However, the methods of this paper apply to any kernel estimator.

The diffusion estimator (McSwiggan et al, 2016) can be expressed (but is not computed) as

$$\widehat{\lambda}_h(u) = \widehat{\lambda}(u \mid \mathbf{x}, h) = \sum_{x_i \in \mathbf{x}} \kappa_t(u \mid x_i), \quad u \in L, \tag{4}$$

where $h > 0$ is the smoothing bandwidth and $t = h^2$. Here $\kappa_t(u \mid v)$ denotes the heat kernel, the analogue of the Gaussian density. In brief, $\kappa_t(u \mid v)$ is the probability density at u of the location, at time t, of a particle which executes Brownian diffusion on the network and which started at time $t = 0$ at location v. Elapsed time equals variance, so that the bandwidth is $h = \sqrt{t}$. The estimator $f_t(u) = \widehat{\lambda}_{\sqrt{t}}(u)$ satisfies the classical time-dependent *heat equation*

$$\frac{\partial f}{\partial t} = \frac{1}{2}\frac{\partial^2 f}{\partial u^2}. \tag{5}$$

The estimator (4) can be computed by solving the partial differential equation (5) numerically as a finite difference equation on a grid of locations along the network and a grid of time steps $0, t_1, t_2, \ldots, t_m$, where $t_m = h^2$. Details were presented in McSwiggan et al (2016).

2.4 Asymptotics for intensity estimation

Standard asymptotic results for kernel density estimation on the real line also hold for linear networks, because the connectivity of the network can be ignored for very small bandwidths (McSwiggan et al, 2016, Section 7.2). Suppose there are N i.i.d. observations from the probability density $f(u)$, assumed to be twice continuously differentiable. Let $N \to \infty$ and consider the heat kernel density estimator $\widehat{f}$ with bandwidth $h = h_N \to 0$ such that $Nh_N \to \infty$. Adapting Botev et al (2010, Theorem 1), for any location u that is not a vertex, the behavior of $\widehat{\lambda}(u)$ is asymptotically equivalent to that of the Gaussian kernel density estimator on the infinite real line, so that $\widehat{\lambda}(u)$ is asymptotically normal with asymptotic bias and variance

$$\mathbb{E}[\widehat{f}(u) - f(u)] = \frac{h^2}{2}\frac{\partial^2 f(u)}{\partial u^2} + O(h^4), \tag{6}$$

$$\mathsf{var}[\widehat{f}(u)] = \frac{f(u)}{2\sqrt{\pi}Nh} + o(1). \tag{7}$$

If $h = O(N^{-1/5})$, the mean square error is of order $O(N^{-4/5})$ and the estimator is pointwise consistent.

2.5 Bandwidth selection for intensity estimation

Techniques for selecting the smoothing bandwidth h for real-valued data (Silverman, 1986; Wand and Jones, 1995; Jones et al, 1996; Loader, 1999b) can also be adapted to linear networks (McSwiggan et al, 2016, Section 9). Asymptotic performance for large samples and small bandwidths is the same on a linear network as on the real line. However, computational complexity and cost will generally be much greater on a linear network, and could be prohibitive for some techniques.

2.5.1 Theoretically optimal bandwidth

Using (6)–(7), the asymptotically optimal bandwidth (minimizing asymptotic integrated MSE) is $h_A^* = \left(2\sqrt{\pi}N\,I(f)\right)^{-1/5}$, where $I(f) = \int_L (\partial^2 f(u)/\partial u^2)^2\,\mathrm{d}u$. Given a pilot estimate of f, it could be feasible to estimate $I(f)$ and calculate h_A^*, but in our experience this method is highly sensitive to the choice of pilot estimate.

2.5.2 Cross-validation for intensity estimation

Data-based bandwidth selection is computationally prohibitive for the popular "equal-split continuous" and "equal-split discontinuous" methods using standard algorithms (Okabe and Sugihara, 2012, Chapter 9), but is feasible using the finite difference method described above. Not only is the finite difference method much faster for computing the kernel estimate at a given bandwidth, but it also computes the kernel estimates at intermediate bandwidths $h_j = \sqrt{t_j}$ at no additional cost (McSwiggan et al, 2016).

Leave-one-out cross-validation (Silverman, 1986; Loader, 1999b, Sec. 5.3, pp. 87–95) selects the bandwidth h^* which maximises

$$A(h) = \sum_i \log \widehat{\lambda}_h^{-i}(x_i), \tag{8}$$

where $\widehat{\lambda}_h^{-i}(x_i)$ is the estimate of $\lambda(x_i)$ based on all the data except x_i:

$$\widehat{\lambda}_h^{-i}(x_i) = \widehat{\lambda}(x_i \mid \mathbf{x}\setminus\{x_i\}, h) = \sum_{j\neq i} \kappa_t(x_i \mid x_j) = \widehat{\lambda}_h(x_i) - \kappa_t(x_i \mid x_i). \tag{9}$$

While evaluation of $A(h)$ for a sequence of values h is feasible on a linear network, it is much more computationally intensive than on the real line, where for a fixed-bandwith kernel estimator with kernel k, the leave-one-out estimate $\widehat{\lambda}_{-i}(x_i) = \widehat{\lambda}(x_i) - k(0)$ can be calculated easily. Calculation of (8) on a network effectively requires us to run the finite difference algorithm separately for each data point.

For the pooled Geelong data in Figure 1, cross-validation using (8) selected a bandwidth of 2.55 km (McSwiggan et al, 2016, Section 9). Computation time was 200 seconds when the maximum bandwidth is 5 km, but only 75 seconds when the maximum bandwidth is 3 km. A fast version using an approximation described in Section 5 takes less than 1 second and yields an almost identical bandwidth of 2.50 km.

Alternative cross-validation methods are discussed in Bowman (1984); Cao et al (1994); Hu et al (2012); Zhang et al (2006). Weaknesses of data-based cross-validation methods are well known. Terrell (1990) argues that they "have often failed to be useful" because they choose an under-estimate of the best bandwidth and consequently produce too many artefacts. In the context of linear networks, the computational cost of cross-validation is especially high.

2.5.3 Rules of thumb

An alternative to cross-validation would be to adapt one of the popular rules of thumb for bandwidth selection. For d-dimensional data, Scott's rule of thumb (Scott, 1992, p. 152) is that the smoothing bandwidth for the ith Cartesian coordinate should be $h_i = n^{-1/(d+4)}\,s_i$, where s_i is the sample standard deviation of the ith coordinate values. Silverman's rule of thumb (Silverman, 1986, eq. (3.31), p. 48) is $h_i = (4/(d+2))^{1/(d+4)}\,n^{-1/(d+4)}s_i$. These are equivalent for $d = 2$. For the Geelong data, treated as a two-dimensional point pattern, this rule of thumb gives bandwidths of 2.56 and 2.24 km in the east-west and north-south directions, respectively.

These are quite close to the cross-validated choice of 2.55 km.

An alternative would be to apply Scott's or Silverman's rule for *one-dimensional* coordinate data to an orthogonal projection of the spatial points onto a one-dimensional axis chosen to maximise the sample standard deviation of the projected coordinates. The maximised standard deviation is $s = \sqrt{a}$, where a is the largest eigenvalue of the sample variance-covariance matrix of the spatial coordinates. Scott's rule would give

$$h = n^{-1/5}\sqrt{a}. \tag{10}$$

For the Geelong data this yields $h = 2.36$ km. Silverman's rule is inflated by the factor $(4/3)^{1/5} = 1.059$ and in the case of the Geelong data gives $h = 2.50$ km.

Terrell (1990) proposed slightly *oversmoothed* kernel estimates. On a linear network, while the general principle of oversmoothing is surely applicable, we have found it difficult to adapt Terrell's theoretical results to this setting.

3 Estimation of relative risk

3.1 Relative risk function

Turning to the main goal of this paper, we now suppose there are two point patterns $\mathbf{x} = \{x_1, \ldots, x_m\}$ and $\mathbf{y} = \{y_1, \ldots, y_n\}$ observed on the same linear network L. Treating $\mathbf{x}$ and $\mathbf{y}$ as realisations of point processes $\mathbf{X}$ and $\mathbf{Y}$, respectively, our goal is to estimate the *logarithmic relative risk*

$$\rho(u) = \log \frac{\lambda_{\mathbf{X}}(u)}{\lambda_{\mathbf{Y}}(u)}, \quad u \in L, \tag{11}$$

where $\lambda_{\mathbf{X}}(u), \lambda_{\mathbf{Y}}(u)$ are the intensities of $\mathbf{X}, \mathbf{Y}$, respectively. The *plug-in estimator* is

$$\widehat{\rho}(u) = \widehat{\rho}_{h_1,h_2}(u) = \log \frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}, \quad u \in L, \tag{12}$$

where $\widehat{\lambda}_{\mathbf{X}}(u) = \widehat{\lambda}(u \mid \mathbf{x}, h_1)$ and $\widehat{\lambda}_{\mathbf{Y}}(u) = \widehat{\lambda}(u \mid \mathbf{y}, h_2)$ are diffusion kernel estimates (4) of $\lambda_{\mathbf{X}}(u)$ and $\lambda_{\mathbf{Y}}(u)$, computed from $\mathbf{x}$ and $\mathbf{y}$, using bandwidths h_1 and h_2, respectively. We note the warning by Loader (1999a) that plug-in estimators may perform poorly.

3.2 Bandwidth selection for relative risk

Three possible methods for selecting the bandwidths h_1, h_2 in (12), discussed in Kelsall and Diggle (1995a), are (**M1**) *separate* selection of h_1 based only on $\mathbf{x}$ and of h_2 based only on $\mathbf{y}$; (**M2**) *joint* selection of the pair (h_1, h_2) based on $\mathbf{x}$ and $\mathbf{y}$; (**M3**) *symmetric* selection of a common bandwidth $h = h_1 = h_2$ based on $\mathbf{x}$ and $\mathbf{y}$. Method M1 could use any of the criteria described in Section 2.5, while methods M2 and M3 require the introduction of new techniques.

As mentioned in the Introduction, bandwidth selection for relative risk is different in principle from bandwidth selection for intensity. For example, if the true intensities are proportional, say $\lambda_{\mathbf{X}}(u) = c\lambda_{\mathbf{Y}}(u)$ for some constant c, then $\rho(u) = \log c$ is constant, and an infinite bandwidth $h = \infty$ is typically optimal for estimating ρ, while estimation of $\lambda_{\mathbf{X}}(u)$ requires smaller bandwidths. Consequently, method M1 is unlikely to perform well when the relative risk is almost constant.

Assuming $\mathbf{X}$ and $\mathbf{Y}$ are independent Poisson processes, the estimator (12) is asymptotically normal with asymptotic pointwise bias and variance of the same form as given in Kelsall and Diggle (1995a, Sec. 2) for the two-dimensional case:

$$\mathbb{E}[\widehat{\rho}_{h_1,h_2}(u) - \rho(u)] \sim \frac{1}{2}h_1^2 \frac{\lambda''_{\mathbf{X}}(u)}{\lambda_{\mathbf{X}}(u)} - \frac{1}{2}h_2^2 \frac{\lambda''_{\mathbf{Y}}(u)}{\lambda_{\mathbf{Y}}(u)} \tag{13}$$

$$\mathsf{var}[\widehat{\rho}_{h_1,h_2}(u)] \sim \frac{1}{2\sqrt{\pi}} \left(\frac{1}{h_1 \lambda_{\mathbf{X}}(u)} + \frac{1}{h_2 \lambda_{\mathbf{Y}}(u)} \right) \tag{14}$$

so that the asymptotic mean integrated squared error is

$$\mathsf{MISE}[\widehat{\rho}_{h_1,h_2}(u)] \sim \frac{1}{2\sqrt{\pi}} \left(\frac{A_1}{h_1} + \frac{A_2}{h_2} \right) + \frac{1}{4}\left(h_1^4 B_{11} - 2h_1^2 h_2^2 B_{12} + h_2^4 B_{22} \right) \tag{15}$$

where $A_1 = A(\lambda_{\mathbf{X}})$ and $A_2 = A(\lambda_{\mathbf{Y}})$ are defined by $A(f) = \int_L f(u)^{-1}\,\mathrm{d}u$, and $B_{11} = B(\lambda_{\mathbf{X}}, \lambda_{\mathbf{X}})$, $B_{12} = B(\lambda_{\mathbf{X}}, \lambda_{\mathbf{Y}})$, $B_{22} = B(\lambda_{\mathbf{Y}}, \lambda_{\mathbf{Y}})$ are defined by $B(f,g) = \int_L (f''(u)/f(u)(g''(u)/g(u))\,\mathrm{d}u$. These approximations determine the asymptotically optimal bandwidths for methods M1, M2 and M3, and these have the same form as those given by Kelsall and Diggle (1995a, Sec. 2) for the two-dimensional case.

The literature is not unanimous on the relative merits of the three methods, but generally favours method M3, which constrains the two bandwidths to be equal. Kelsall and Diggle (1995b, p. 10) conclude that method M3 has some theoretical justification when $\lambda_{\mathbf{X}} \propto \lambda_{\mathbf{Y}}$, and report simulation experiments in which method M3 achieved the best performance. Davies et al (2016) demonstrate that halo-like artefacts can occur when relative risk is estimated using different smoothing bandwidths for the numerator and denominator. Support for method M3 is also given in Davies and Lawson (2018) and Diggle (2014, §9.3, p. 179ff.) with examples and technique described in Davies et al (2018, 2011). In this paper we explore using either M2 or M3.

3.3 Kelsall-Diggle cross-validation on linear networks

For two-dimensional point pattern data, Kelsall and Diggle (1995b) proposed a method for selecting the bandwidths (h_1, h_2) in (12), with or without the constraint $h_1 = h_2$, by

cross-validation based on integrated squared error. Adapted to a linear network, and re-expressed in terms of the intensity rather than the probability density, the criterion becomes

$$\widetilde{C}_{\mathrm{KD}}(h_1,h_2) = -\int_L \left[\log \frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u - 2\sum_{i=1}^{m} \frac{1}{\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i)} \log \frac{\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i)}{\widehat{\lambda}_{\mathbf{Y}}(x_i)} - 2\sum_{j=1}^{n} \frac{1}{\widehat{\lambda}_{\mathbf{Y}}^{-j}(y_j)} \log \frac{\widehat{\lambda}_{\mathbf{Y}}^{-j}(y_j)}{\widehat{\lambda}_{\mathbf{X}}(y_j)}, \tag{16}$$

where $\widehat{\lambda}_{\mathbf{X}}(u) = \widehat{\lambda}(u \mid \mathbf{x}, h_1)$ is the estimate (4) of $\lambda_{\mathbf{X}}(u)$ using bandwidth h_1, while $\widehat{\lambda}_{\mathbf{Y}}(u) = \widehat{\lambda}(u \mid \mathbf{y}, h_2)$ is the estimate of $\lambda_{\mathbf{Y}}(u)$ using bandwidth h_2, and $\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i), \widehat{\lambda}_{\mathbf{Y}}^{-j}(y_j)$ denote the corresponding leave-one-out estimates. The bandwidth pair (h_1,h_2) or (h,h) should be chosen to minimise the criterion (16).

3.4 Likelihood and least-squares cross-validation

A later paper by Kelsall and Diggle (1998) proposed a different approach to relative risk, using a connection with binary regression, which they argued is more flexible than the density-ratio approach. If $\mathbf{X}$ and $\mathbf{Y}$ are independent Poisson processes in $\mathbb{R}^2$ with intensity functions $\lambda_{\mathbf{X}}(u), \lambda_{\mathbf{Y}}(u)$, respectively, then the superimposition $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$ is Poisson with intensity $\lambda_{\mathbf{Z}}(u) = \lambda_{\mathbf{X}}(u) + \lambda_{\mathbf{Y}}(u)$, and a random point of $\mathbf{Z}$ at location u has probability $p(u) = \lambda_{\mathbf{X}}(u)/\lambda_{\mathbf{Z}}(u)$ of having originated from the process $\mathbf{X}$ rather than $\mathbf{Y}$.

Given data patterns $\mathbf{x}$, $\mathbf{y}$, define for each point $x_i \in \mathbf{x}$ for $i = 1,\ldots,m$

$$\hat{p}_i = \frac{\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i)}{\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i) + \widehat{\lambda}_{\mathbf{Y}}(x_i)}, \tag{17}$$

the estimated probability (estimated from all data other than x_i) that a point of $\mathbf{X} \cup \mathbf{Y}$ at location x_i would belong to $\mathbf{X}$ rather than $\mathbf{Y}$. Similarly, for all points $y_j \in \mathbf{y}$, $j = 1,\ldots,n$ define

$$\hat{q}_j = \frac{\widehat{\lambda}_{\mathbf{Y}}^{-j}(y_j)}{\widehat{\lambda}_{\mathbf{X}}(y_j) + \widehat{\lambda}_{\mathbf{Y}}^{-j}(y_j)},$$

the estimated probability that a point of $\mathbf{X} \cup \mathbf{Y}$ at y_j would belong to $\mathbf{Y}$ rather than $\mathbf{X}$.

Then the *likelihood cross-validation* criterion (Kelsall and Diggle, 1998) is the negative log-likelihood

$$\widetilde{C}_{\mathrm{LIK}}(h_1,h_2) = -\left[\sum_{i=1}^{m} \log(\hat{p}_i) + \sum_{j=1}^{n} \log(\hat{q}_j)\right]. \tag{18}$$

Minimisation of (18) has also been suggested by Azzalini et al (1989). The *least-squares* cross-validation criterion is

$$\widetilde{C}_{\mathrm{LSQ}}(h_1,h_2) = \sum_{i=1}^{m} (1-\hat{p}_i)^2 + \sum_{j=1}^{n} (1-\hat{q}_j)^2. \tag{19}$$

This is the loss criterion for least squares prediction of the status of each point given the locations of all points. It is known to work well in many contexts (Haerdle, 1990).

We shall use all of the cross-validation criteria listed above in our experiments.

4 Improved cross-validation method

Kelsall and Diggle (1995a,b) reported that, in two dimensions, their cross-validation criterion (16) suffered occasional "breakdowns" in which the selected bandwidth values were extreme and the resulting estimates very unsatisfactory. Similar breakdowns occurred in our experiments with the analogue of the Kelsall-Diggle criterion on a linear network (reported in Section 6 and the supplementary material).

This motivates us to re-visit the original derivation of the Kelsall-Diggle method. Define the integrated squared error of estimation of ρ

$$\mathrm{ISE}(\widehat{\rho}) = \int_L (\widehat{\rho}(u) - \rho(u))^2 \, \mathrm{d}u. \tag{20}$$

We now discuss approximation of the ISE from data, along the same lines as Kelsall and Diggle (1995b), but expressed in terms of point process intensity rather than probability density. We use a slightly more general formulation so that we may revisit it.

4.1 General derivation

If the plug-in estimator (12) is used, expanding the square in (20) gives

$$\mathrm{ISE}(\widehat{\rho}) = \int_L \left[\log \frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u - 2\int_L \log \frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)} \log \frac{\lambda_{\mathbf{X}}(u)}{\lambda_{\mathbf{Y}}(u)} \, \mathrm{d}u + \int_L \left[\log \frac{\lambda_{\mathbf{X}}(u)}{\lambda_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u. \tag{21}$$

The last term on the right hand side of (21) is constant in any given application, and may be omitted for optimisation purposes. The middle term on the right hand side involves the unknown true intensities. Following the approach of Kelsall and Diggle (1995a,b) we would replace the true intensities by approximations, based on a Taylor expansion of the logarithm:

$$\log \lambda_{\mathbf{X}}(u) \approx \log \lambda_{\mathbf{X}}^0(u) + \frac{1}{\lambda_{\mathbf{X}}^0(u)} \left[\lambda_{\mathbf{X}}(u) - \lambda_{\mathbf{X}}^0(u)\right] = \log \lambda_{\mathbf{X}}^0(u) + \frac{\lambda_{\mathbf{X}}(u)}{\lambda_{\mathbf{X}}^0(u)} - 1 \tag{22}$$

$$\log \lambda_{\mathbf{Y}}(u) \approx \log \lambda_{\mathbf{Y}}^0(u) + \frac{\lambda_{\mathbf{Y}}(u)}{\lambda_{\mathbf{Y}}^0(u)} - 1, \tag{23}$$

where $\lambda^0_{\mathbf{X}}(u), \lambda^0_{\mathbf{Y}}(u)$ are some chosen "reference estimates" to be discussed below. Note that the Taylor expansions are performed about the reference estimates. The approximations (22)–(23) give

$$\log\frac{\lambda_{\mathbf{X}}(u)}{\lambda_{\mathbf{Y}}(u)} \approx \log\frac{\lambda^0_{\mathbf{X}}(u)}{\lambda^0_{\mathbf{Y}}(u)} + \frac{\lambda_{\mathbf{X}}(u)}{\lambda^0_{\mathbf{X}}(u)} - \frac{\lambda_{\mathbf{Y}}(u)}{\lambda^0_{\mathbf{Y}}(u)} \tag{24}$$

and hence

$$\begin{aligned}\log\frac{\lambda_{\mathbf{X}}(u)}{\lambda_{\mathbf{Y}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)} \approx{}& \log\frac{\lambda^0_{\mathbf{X}}(u)}{\lambda^0_{\mathbf{Y}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)} \\ &+ \frac{\lambda_{\mathbf{X}}(u)}{\lambda^0_{\mathbf{X}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)} \\ &+ \frac{\lambda_{\mathbf{Y}}(u)}{\lambda^0_{\mathbf{Y}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}.\end{aligned} \tag{25}$$

Collecting terms we obtain the cross-validation criterion

$$\begin{aligned}C(h_1,h_2) = {}& \int_L \left[\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u - 2\int_L \log\frac{\lambda^0_{\mathbf{X}}(u)}{\lambda^0_{\mathbf{Y}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u \\ &- 2\int_L \frac{\lambda_{\mathbf{X}}(u)}{\lambda^0_{\mathbf{X}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u + 2\int_L \frac{\lambda_{\mathbf{Y}}(u)}{\lambda^0_{\mathbf{Y}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u.\end{aligned} \tag{26}$$

The last two terms on the right hand side of (26) still contain the unknown true intensity functions $\lambda_{\mathbf{X}}, \lambda_{\mathbf{Y}}$. Following Kelsall and Diggle (1995a,b) these terms can be estimated from data by "leave-one-out averaging" (Hall and Marron, 1991), or equivalently the Campbell-Mecke formula (Mecke, 1967):

$$\int_L \frac{\lambda_{\mathbf{X}}(u)}{\lambda^0_{\mathbf{X}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u \approx \mathbb{E}\left[\sum_{i=1}^m \frac{1}{\widehat{\lambda}^0_{\mathbf{X}}(x_i)}\log\frac{\widehat{\lambda}^{-i}_{\mathbf{X}}(x_i)}{\widehat{\lambda}_{\mathbf{Y}}(x_i)}\right], \tag{27}$$

where

$$\begin{aligned}\widehat{\lambda}^{-i}_{\mathbf{X}}(x_i) &= \widehat{\lambda}(x_i \mid \mathbf{x}\setminus\{x_i\}, h) = \sum_{j\neq i}\kappa_{h_1}(x_j \mid x_i) \\ &= \widehat{\lambda}_{\mathbf{X}}(x_i) - \kappa_{h_1}(x_i \mid x_i)\end{aligned} \tag{28}$$

is the leave-one-out estimate of intensity of $\mathbf{X}$ based on all points of $\mathbf{x}$ except the query point x_i. The approximation (27) would be exact, by the Campbell-Mecke formula, if the leave-one-out estimator was non-random, so heuristically we expect the right-hand side of (27) to be a consistent estimator of the left-hand side. Similarly we approximate

$$\int_L \frac{\lambda_{\mathbf{Y}}(u)}{\lambda^0_{\mathbf{Y}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u \approx \mathbb{E}\left[\sum_{j=1}^n \frac{1}{\lambda^0_{\mathbf{Y}}(y_j)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(y_j)}{\widehat{\lambda}^{-j}_{\mathbf{Y}}(y_j)}\right], \tag{29}$$

yielding the empirical cross-validation criterion

$$\begin{aligned}\widetilde{C}(h_1,h_2) = {}& \int_L \left[\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u - 2\int_L \log\frac{\lambda^0_{\mathbf{X}}(u)}{\lambda^0_{\mathbf{Y}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u \\ &- 2\sum_{i=1}^m \frac{1}{\lambda^0_{\mathbf{X}}(x_i)}\log\frac{\widehat{\lambda}^{-i}_{\mathbf{X}}(x_i)}{\widehat{\lambda}_{\mathbf{Y}}(x_i)} - 2\sum_{j=1}^n \frac{1}{\lambda^0_{\mathbf{Y}}(y_j)}\log\frac{\widehat{\lambda}^{-j}_{\mathbf{Y}}(y_j)}{\widehat{\lambda}_{\mathbf{X}}(y_j)}.\end{aligned} \tag{30}$$

4.2 Derivation of Kelsall–Diggle criterion

Kelsall and Diggle (1995b) choose the reference estimates $\lambda^0_{\mathbf{X}}(u)$ and $\lambda^0_{\mathbf{Y}}(u)$ in (26) and (30) to be the *current* estimates $\widehat{\lambda}_{\mathbf{X}}(u) = \widehat{\lambda}(u \mid \mathbf{x}, h_1)$ and $\widehat{\lambda}_{\mathbf{Y}}(u) = \widehat{\lambda}(u \mid \mathbf{y}, h_2)$, respectively. This allows some algebraic simplification of (26) yielding

$$\begin{aligned}C_{\mathrm{KD}}(h_1,h_2) = {}& -\int_L \left[\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u - 2\int_L \frac{\lambda_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{X}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u \\ &+ 2\int_L \frac{\lambda_{\mathbf{Y}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u.\end{aligned} \tag{31}$$

Similarly taking the reference intensities at the data points to be the current empirical estimates, $\lambda^0_{\mathbf{X}}(x_i) = \widehat{\lambda}^{-i}_{\mathbf{X}}(x_i)$ and $\lambda^0_{\mathbf{Y}}(y_j) = \widehat{\lambda}^{-j}_{\mathbf{Y}}(y_j)$, the empirical cross-validation criterion (30) becomes

$$\begin{aligned}\widetilde{C}_{\mathrm{KD}}(h_1,h_2) = {}& -\int_L \left[\log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u - 2\sum_{i=1}^m \frac{1}{\widehat{\lambda}^{-i}_{\mathbf{X}}(x_i)}\log\frac{\widehat{\lambda}^{-i}_{\mathbf{X}}(x_i)}{\widehat{\lambda}_{\mathbf{Y}}(x_i)} \\ &- 2\sum_{j=1}^n \frac{1}{\widehat{\lambda}^{-j}_{\mathbf{Y}}(y_j)}\log\frac{\widehat{\lambda}^{-j}_{\mathbf{Y}}(y_j)}{\widehat{\lambda}_{\mathbf{X}}(y_j)}.\end{aligned} \tag{32}$$

A possible explanation for the breakdown of $\widetilde{C}_{\mathrm{KD}}$ is now clear. The general form of the cross-validation criterion (26) is derived by replacing the true intensities $\lambda_{\mathbf{X}}(u)$ and $\lambda_{\mathbf{Y}}(u)$ by Taylor approximations (22) and (23) about the "reference" estimates $\lambda^0_{\mathbf{X}}(u)$ and $\lambda^0_{\mathbf{Y}}(u)$, respectively. In the case of the Kelsall-Diggle cross-validation criterion (32) the reference estimates are taken to be the current kernel estimates $\widehat{\lambda}_{\mathbf{X},h_1}(u)$ and $\widehat{\lambda}_{\mathbf{Y},h_2}(u)$. For small bandwidths, these estimates could be highly biased because of undersmoothing, and Taylor expansions about these estimates could yield poor approximations to the true intensities.

4.3 Our proposed alternative

We propose taking the "reference" estimates $\lambda^0_{\mathbf{X}}(u), \lambda^0_{\mathbf{Y}}(u)$ in (26) and (30) to be *oversmoothed* kernel estimates obtained by setting h_1, h_2 to the maximum values under consideration, say H_1, H_2. The Kelsall-Diggle argument then leads to our proposed "modified Kelsall-Diggle" cross-validation

criterion,

$$\begin{aligned}\widetilde{C}_{\mathrm{OVER}}(h_1,h_2) = &\int_L \left[\log \frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u \\ &- 2\int_L \log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)} \log\frac{\widehat{\lambda}(u \mid \mathbf{x}, H_1)}{\widehat{\lambda}(u \mid \mathbf{y}, H_2)}\,\mathrm{d}u \\ &- 2\sum_{i=1}^{m} \frac{1}{\widehat{\lambda}^{-i}(x_i \mid \mathbf{x}, H_1)} \log\frac{\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i)}{\widehat{\lambda}_{\mathbf{Y}}(x_i)} \\ &- 2\sum_{j=1}^{n} \frac{1}{\widehat{\lambda}^{-j}(y_j \mid \mathbf{y}, H_2)} \log\frac{\widehat{\lambda}_{\mathbf{Y}}^{-j}(y_j)}{\widehat{\lambda}_{\mathbf{X}}(y_j)}.\end{aligned} \tag{33}$$

This has marginally greater computational cost than the Kelsall-Diggle criterion (32) due to the addition of the second term on the right hand side of (33).

Our proposal, to use an over-smoothed estimate as the reference for the Taylor expansion, could be compared to the use of "pre-smoothed" estimates by Hall et al (1992).

An even simpler alternative could be to take the reference intensities to be constant, $\lambda_{\mathbf{X}}^0(u) = m/|L|$ and $\lambda_{\mathbf{Y}}^0(u) = n/|L|$, where $m = n(\mathbf{x})$ and $n = n(\mathbf{y})$ are the observed numbers of points in the patterns $\mathbf{x}$ and $\mathbf{y}$. This would yield the cross-validation criterion

$$\begin{aligned}\widetilde{C}_{\mathrm{UNIF}}(h_1,h_2) = &\int_L \left[\log \frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\right]^2 \mathrm{d}u \\ &- 2(\log\frac{m}{n}) \int_L \log\frac{\widehat{\lambda}_{\mathbf{X}}(u)}{\widehat{\lambda}_{\mathbf{Y}}(u)}\,\mathrm{d}u \\ &- 2\frac{|L|}{m}\sum_{i=1}^{m} \log\frac{\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i)}{\widehat{\lambda}_{\mathbf{Y}}(x_i)} \\ &- 2\frac{|L|}{n}\sum_{j=1}^{n} \log\frac{\widehat{\lambda}_{\mathbf{Y}}^{-j}(y_j)}{\widehat{\lambda}_{\mathbf{X}}(y_j)}.\end{aligned} \tag{34}$$

This criterion is computationally cheaper than the other cross-validation criteria, but may lead to suboptimal choices.

Other strategies include numerically stabilising the cross-validation by adding a small constant value to the reference intensities (Hazelton and Davies, 2009; Bowman and Azzalini, 1997). Instead of constraining $h_1 = h_2$ it would be possible to use the constraint $h_1/h_2 = (n_1/n_2)^{-1/5}$, or to allow $h_1 \neq h_2$ and introduce a penalty for discrepancy between them e.g. $(h_1 - h_2)^2$, or simply to constrain the bandwidths to be greater than a certain realistic minimum value. The latter option is discussed in Section 7.1.

5 Approximation to leave-one-out estimate

As noted in Section 2.5.2, computation of the leave-one-out estimates of intensity $\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i), \widehat{\lambda}_{\mathbf{Y}}^{-j}(y_j)$ is more complicated on a linear network than in two-dimensional space. Exact calculation of $\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i)$ (say) would require us to run the heat equation solver for the point pattern $\mathbf{x}^{-i} = \mathbf{x} \setminus \{x_i\}$. The solver would have to be executed n times to obtain all the values $\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i)$ for $i = 1, \ldots, n$.

An alternative is to use the relation $\widehat{\lambda}_{\mathbf{X}}^{-i}(x_i) = \widehat{\lambda}(x_i) - \kappa(x_i \mid x_i)$ from (9), and to find an approximation for $\kappa(x_i \mid x_i)$. Invoking McSwiggan et al (2016, equ. (23)) or Kostrykin et al (2007, Corollary 3.4) we can write $\kappa(u \mid u)$ as an infinite sum

$$\kappa_t(u \mid u) = \sum_{\Pi} a(\Pi)\phi_{\sqrt{t}}(\ell(\Pi)) \tag{35}$$

over all possible cycles $\Pi = (v_0, \ldots, v_{m+1})$ in the network, with $m \geq 0$, where $v_0 = u$, $v_{m+1} = u$ and $v_1, \ldots, v_m$ are vertices. Here $\ell(\Pi)$ is the total length of the path Π, and $a(\Pi)$ is a combinatorial coefficient, while ϕ_σ is the Gaussian density with mean 0 and standard deviation σ.

A simple approximation is obtained by truncating the sum in (35), retaining only the terms with $m = 0$ or $m = 1$ steps. This could be portrayed as the analogue of a first order Taylor approximation. If u lies on a segment of length $s = s(u)$ and is a distance $x = x(u)$ from the left endpoint, and if the left and right endpoints have degree d and d', respectively, the proposed approximation to $\kappa(u \mid u)$ is

$$\begin{aligned}\kappa_t(u \mid u) \approx \kappa_\sigma^*(u) = &\phi_\sigma(0) + \left(\frac{2}{d} - 1\right)\phi_\sigma(2x(u)) \\ &+ \left(\frac{2}{d'} - 1\right)\phi_\sigma(2(s(u) - x(u))),\end{aligned} \tag{36}$$

where $\sigma = \sqrt{t}$. This approximation has the advantage that it can be computed rapidly from the spatial coordinates and network geometry.

The approximation (36) is likely to be very accurate when $\sigma \ll s(u)$, and is likely to become progressively less accurate as σ increases. To improve the performance for large t, we constrain the approximation (36) to be greater than or equal to $1/|L|$, which is the limiting value of $\kappa_t(u \mid u)$ as $t \to \infty$.

Results in the online supplement demonstrate that the approximation (36) is highly satisfactory for this purpose.

6 Simulation Experiments

Kelsall and Diggle (1995b) compared the performance of their proposed bandwidth selection method with that of other methods, using a suite of simulations on the one-dimensional unit interval. We have run analogous experiments on a linear network, using all of the bandwidth selection criteria mentioned above.

6.1 Description of experiments

Kelsall and Diggle (1995b) considered nine different scenarios by combining three possible choices for the relative risk

function $r(u) = \lambda_{\mathbf{X}}(u)/\lambda_{\mathbf{Y}}(u)$ with three possible choices for the denominator intensity $d(u) = \lambda_{\mathbf{Y}}(u)$. The numerator intensity is then $\lambda_{\mathbf{X}}(u) = r(u)d(u)$. The three possible risk functions $r(u)$ were a constant function and two Gaussian densities. The three possible denominator intensities $d(u)$ were a constant function and two linear transformations of the sine function.

Figure 3 shows the linear network used in our experiments. It has a total length of 4.1 units and a diameter of 1.25 units (sc. the maximum path distance between any two points) and is inscribed in the unit square.

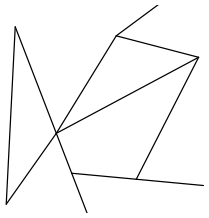

Fig. 3 Linear network used in the experiments.

Three risk functions $r(u)$ were used in our experiments. Risk function 1 is constant; risk functions 2 and 3 have a single peak, obtained by evaluating the heat kernel (bandwidths 0.4 and 0.12, respectively) for a single data point placed at the centre of the network. Similarly we used three functions for the denominator intensity $d(u)$, namely $d_1(x,y) \equiv 1$, $d_2(x,y) = 1+(1/2)\sin(2\pi x)$ and $d_3(x,y) = 1+(3/4)\sin(4\pi x)$, where (x,y) are the Cartesian coordinates. These six functions are plotted in the Supplementary Material.

6.2 Representative results

Here we present detailed results for one case, with risk function 2 and denominator function 3, shown in Figure 4. Simulated realisations were generated with fixed numbers of points, $n(\mathbf{x}) = 50$ and $n(\mathbf{y}) = 200$.

Figure 5 shows boxplots of the ISE values attained by each of the bandwidth selection methods. Here `sco` indicates Scott's rule of thumb as adapted in (10); `KD` is Kelsall–Diggle cross-validation (16); `mod` is our modification (33); `lik` is likelihood cross-validation (18); and `lsq` is least squares cross-validation (19). Each bandwidth selection method was applied to the same set of 100 simulated realisations. For each simulated dataset the bandwidth, or pair of bandwidths, selected by each method was used to smooth the data, yielding an estimate of ρ, and the ISE for this estimate was computed from (20) using the true value of $\rho(u)$ which is known exactly in the simulation experiment. The upper panel of Figure 5 shows boxplots of the ISE values obtained when the bandwidths are constrained to be equal according to method M3, and the lower panel when they are not constrained (method M2).

Table 1 reports the fraction of outcomes in which the bandwidth selected by each method is equal to the minimum or maximum bandwidth value considered. For the Kelsall-Diggle method in the symmetric case, 41% of the selected bandwidths equal the maximum permitted bandwidth. While large bandwidths may be quite satisfactory in some cases, selecting the minimum available bandwidth will almost always produce a poor estimate of relative risk, and this happens frequently in the asymmetric case ($h_1 \neq h_2$).

Figure 6 shows a scatterplot matrix for the values of bandwidth h obtained by each of the methods in the constrained case $h_1 = h_2 = h$, method M3. Interestingly, our modified version of the Kelsall-Diggle method yields bandwidths which are highly correlated with the likelihood cross-validation and least squares cross-validation methods, and

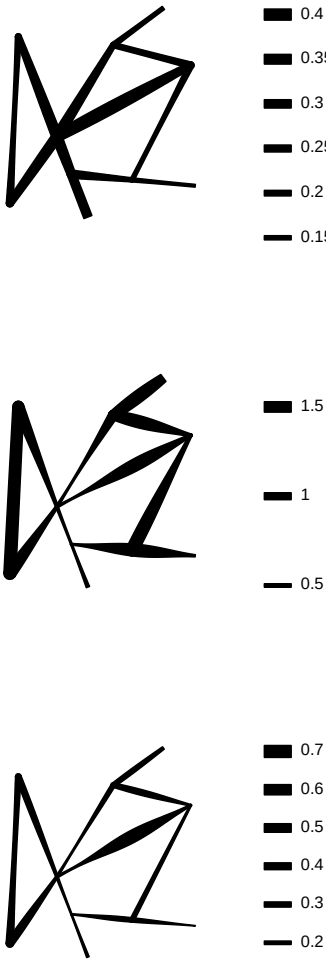

Fig. 4 Simulation experiment reported in this section. *Top:* relative risk $r(u)$. *Middle:* denominator intensity $\lambda_{\mathbf{Y}}(u) = d(u)$. *Bottom:* numerator intensity $\lambda_{\mathbf{X}}(u) = r(u)d(u)$. Plots are in the style of Xie and Yan (2008), with line thickness proportional to function value.

are only weakly correlated with the original Kelsall-Diggle method. Scott's rule of thumb has low correlation with all other methods, suggesting that it would be unwise to use the rule-of-thumb bandwidth estimate as an initial guess at the cross-validated bandwidth estimate.

Analogous figures for the bandwidths h_1 and h_2 respectively, in the case where the bandwidths are permitted to be different, are given in the online supplement. They show that the bandwidth h_2, which serves to smooth the denominator, is frequently chosen to be an extremely small or extremely large value. This appears to be the main cause of breakdown

	Minimal			Maximal		
METHOD	h_1	h_2	h	h_1	h_2	h
sco	0	0.00	0	0.00	0.00	0.00
KD	0.01	0.10	0	0.25	0.83	0.41
mod	0	0.80	0	0.02	0.08	0.24
lik	0	0.27	0	0.02	0.48	0.23
lsq	0	0.13	0	0.02	0.58	0.23

Table 1 Fraction of outcomes of each bandwidth-selection method in which the selected bandwidth is equal to the minimum permitted bandwidth (*Minimal*) or the maximum permitted bandwidth (*Maximal*). Here h_1, h_2 are the bandwidths selected jointly without any constraint (method M2), and h is the symmetric bandwidth (method M3). Simulation experiment case $i = 2, j = 3$.

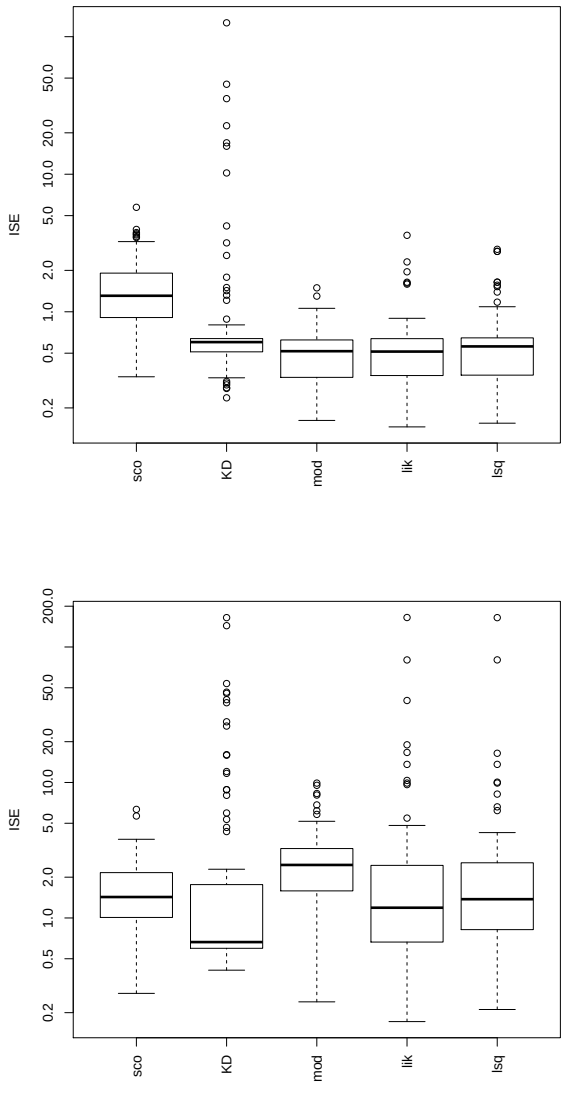

Fig. 5 Boxplots of ISE values achieved by different methods for bandwidth selection in a simulation experiment (case $i = 2, j = 3$). *Top*: bandwidths h_1, h_2 are constrained to be equal (method M3). *Bottom:* bandwidths unconstrained (method M2). Note logarithmic scale for ISE.

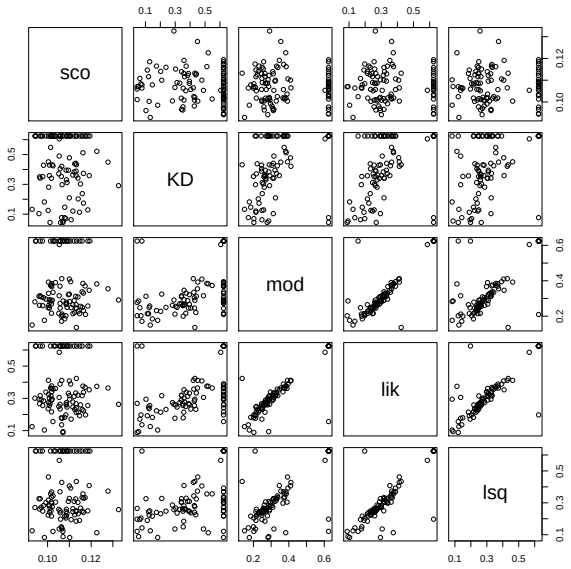

Fig. 6 Scatterplot matrix for the bandwidth values selected by each method under the constraint $h_1 = h_2 = h$, method M3. Simulation experiment case $i = 2, j = 3$.

in bandwidth selection when the bandwidths are not constrained to be equal.

6.3 Summary of performance

The online supplement to this paper gives detailed results from the suite of nine simulation experiments. Following is a summary of the main findings.

Bandwidth values selected using the fast approximation (36) to the leave-one-out estimate agreed very closely with those selected using the exact leave-one-out estimate, giving us confidence that the approximation is reliable.

The most significant finding is that estimates of relative risk were often much less accurate if we allowed $h_1 \neq h_2$ than when we constrained $h_1 = h_2$. This applied to all of the cross-validation methods. This may appear paradoxical unless we remember that the cross-validation criterion is

only a data-based estimate of true performance, so that unconstrained minimisation of the cross-validation criterion is not guaranteed to produce better true performance than constrained minimisation. In our experiments, method M3 consistently outperformed method M2.

Investigation showed that when $h_1 \neq h_2$, the selected value of h_1 was usually appropriate, but that the selected value of h_2 was frequently much too small. Plots of the cross-validation criteria in individual examples often showed a steep decline in $C(h_1,h_2)$ as $h_2 \to 0$. Since the expressions for the cross-validation criteria are symmetric in $\mathbf{x}$ and $\mathbf{y}$, this one-sided behaviour is probably attributable to the different numbers of points in the two patterns.

Each method exhibits occasional "breakdown" in which the estimate is quite poor. Cross-validation methods have better performance than Scott's rule-of-thumb overall. However, the Scott rule of thumb is computationally cheaper, and is less susceptible to breakdown – it is "consistently mediocre".

The Kelsall-Diggle method often has higher ISE and higher frequency of breakdown than other cross-validation methods. In some cases the K-D method was unusable, with an infinite median ISE. The K-D method and our modified method often gave quite different results, lending support to the argument about the Taylor expansion.

Somewhat surprisingly, our modified method, the likelihood cross-validation and the least square cross validation method often selected quite similar bandwidths and gave similar results. Our modified method typically has the lowest frequency of breakdowns and the lowest median ISE, although its performance is mediocre in some cases.

6.4 General comments on experiments

Statistical performance will depend on the maximum bandwidth specified when running the bandwidth selection algorithm, because several of the methods have a high probability of selecting the maximum bandwidth.

In their experiments, Kelsall and Diggle (1995a,b) measured the performance of estimators by the median ISE. Our figures suggest that summaries such as the median and mean of ISE could be hard to interpret because of the very different shapes of the distributions of ISE values obtained from each method.

The theoretical analysis presented by Kelsall and Diggle (1995a,b) assumed that $\lambda_{\mathbf{X}}, \lambda_{\mathbf{Y}}$ are bounded away from zero. Their (and our) experiments include scenarios where the minimum density is very small, so this could explain the poor performance.

7 Examples

Two real data examples are studied here. Evidence for spatially-varying relative risk is weak in the first example, and very strong in the second example.

7.1 Geelong road accidents

An important question for road safety management is whether some specific locations have high accident risk at night, after allowing for the inherently greater baseline risk of nighttime driving. For the Geelong data classified into day and night accidents in Figure 2, we considered estimation of the relative rate of night versus day accidents. The modified Scott rule of thumb gave bandwidths of about 2.7 km for each pattern. We computed the Kelsall-Diggle (32), modified Kelsall-Diggle (33), likelihood (18) and least squares (19) cross-validation criteria for relative risk. We nominated a maximum bandwidth of $h_{\max} = 5$ km, and searched over a grid of $N = 400$ candidate values of bandwidth $h = (k/N)^{1/2} h_{\max}$ for $k = 1, \ldots, N$. The maximum bandwidth was also used to compute the reference intensities for our modified criterion. The selected bandwidths are shown in Table 2. Total time to compute all four criteria was about 3 minutes if leave-one-out estimates were calculated exactly, and about 5 seconds if the fast approximation (36) was used.

	SYMMETRIC	ASYMMETRIC	
METHOD	h	h_1	h_2
`Scott`	2.76	2.76	2.68
`KD`	5.00 (∞)	5.00	5.00
`mod`	5.00 (∞)	5.00	3.42
`lik`	5.00 (∞)	2.92	0.25
`lsq`	5.00 (∞)	2.60	0.25

Table 2 Automatically-selected bandwidths for the Geelong accidents separated into night and day accidents. Symmetric bandwidths selected by method M3; asymmetric bandwidth pairs by method M2; exact calculation. The symbol ∞ indicates that infinite bandwidth achieved a better cross-validation score than the selected bandwidth, $C(\infty,\infty) < C(h,h)$.

Optimal bandwidths selected using the Kelsall-Diggle criterion (32) and our modification (33) are acceptable values. The likelihood and least squares criteria produce acceptable bandwidths in the symmetric case, but in the asymmetric case the bandwidth h_2 for the denominator is too small: the corresponding estimates of relative risk have values as high as 10^{15}. Figure 7 shows the estimate of the ratio of night to day intensities using symmetric bandwidth 5 km. For reference, the overall ratio of night to day accidents is $98/144 = 0.68$. The figure suggests that the relative risk of night time to day time accidents is up to 4 times higher on some of the more remote roads. This is plausible for reasons

including the higher speed limits and the absence of street lighting along the remote roads. However, the calculation does not include adjustment for diurnal differences in traffic volume.

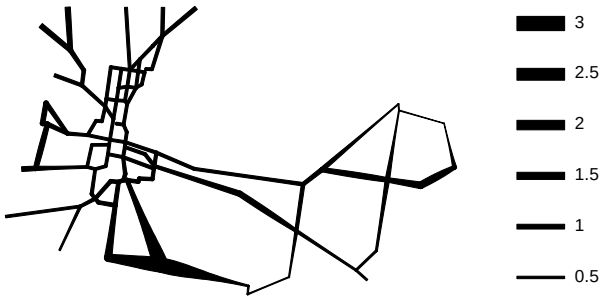

Fig. 7 Relative risk of night versus day accidents using bandwidth 5 km. Line width proportional to relative risk.

Figure 8 shows contours of the cross-validation criteria (33) and (18) as functions of (h_1, h_2), for the Geelong data split into day and night accidents. Our modified criterion (33) has convex contours and a clearly defined minimum in this case. The likelihood criterion (18) is quite regular for large bandwidth values, but has a steep slope when h_2 is small, which explains the incorrect choice of h_2 in the unconstrained case. Contour plots for the other cross-validation criteria are given in an online supplement.

Figure 8 suggests that a practical remedy for the selection of incorrect bandwidths might be simply to restrict attention to bandwidths larger than a data-dependent threshold. Consideration of (17) suggests using the mean nearest-neighbour distance between each type of accident. For the Geelong data, the mean distance from a daytime accident location to the nearest nighttime accident location is 0.98 km.

The Geelong data also include information on the number of vehicles involved in the accident. Single-vehicle accidents include accidents occuring when a driver loses control of the vehicle, and accidents involving a pedestrian. There were 100 single-vehicle accidents, 115 two-vehicle accidents, 21 three-vehicle and 6 four-vehicle accidents. Figure 9 shows the estimated ratio of accident rates of single- and multiple-vehicle accidents, again using the bandwidth 5 km selected by the symmetric method M3. For reference, the ratio of numbers of single-vehicle to multiple-vehicle accidents is $100/142 = 0.70$.

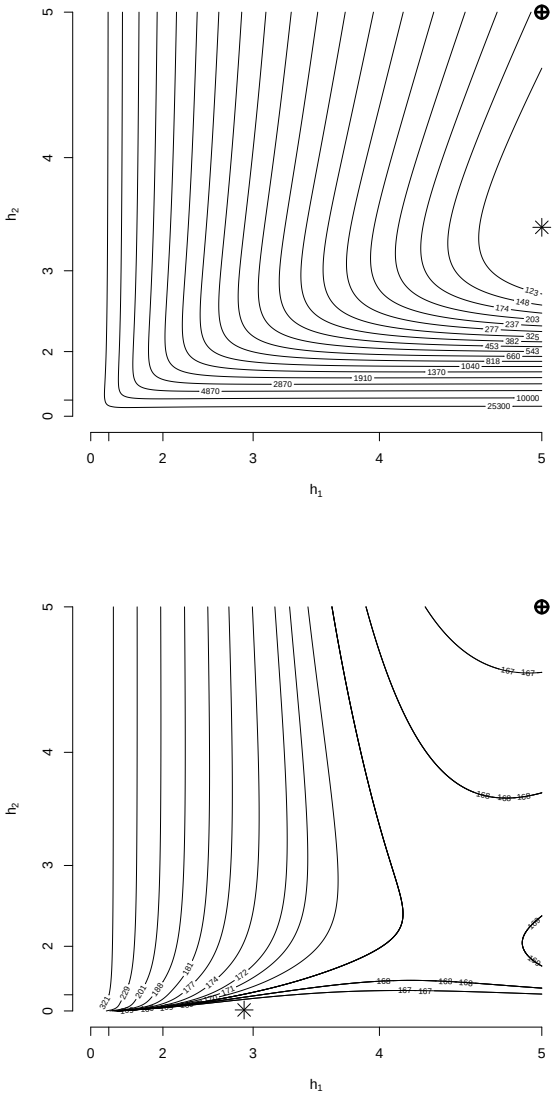

Fig. 8 Contours of cross-validation criterion as a function of the smoothing bandwidths h_1, h_2, for the Geelong data separated into night and day accidents. *Top*: modified Kelsall-Diggle criterion (33). *Bottom:* negative likelihood cross-validation criterion (18). Geelong data, relative risk, night versus day. Symbol $\oplus$ indicates optimal symmetric bandwidth h; symbol $*$ indicates optimal joint bandwidths (h_1, h_2).

7.2 Dendritic spines data

Figure 10 shows the dendritic spines data studied in Jammalamadaka et al (2013); Baddeley et al (2014). The network represents one branch of the dendritic tree of a neuron. The points are the locations of small protrusions called spines, which are classified into three types: mushroom, stubby and thin. Key research questions concern the spatial distribution of spines on the network, and differences in spatial distribution between different types of spines (Jammalamadaka et al, 2013). Analysis in Baddeley et al (2014) suggested that the mushroom and stubby types are uniformly distributed while the thin types are found more frequently near the ends of the dendritic tree, at the left side of the Figure.

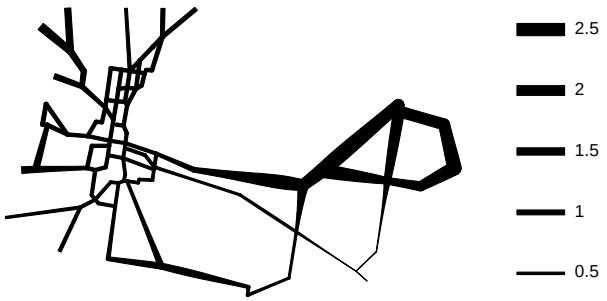

Fig. 9 Relative risk of single-vehicle versus multiple-vehicle accidents in the Geelong data, using bandwidth 5 km. Line width proportional to relative risk.

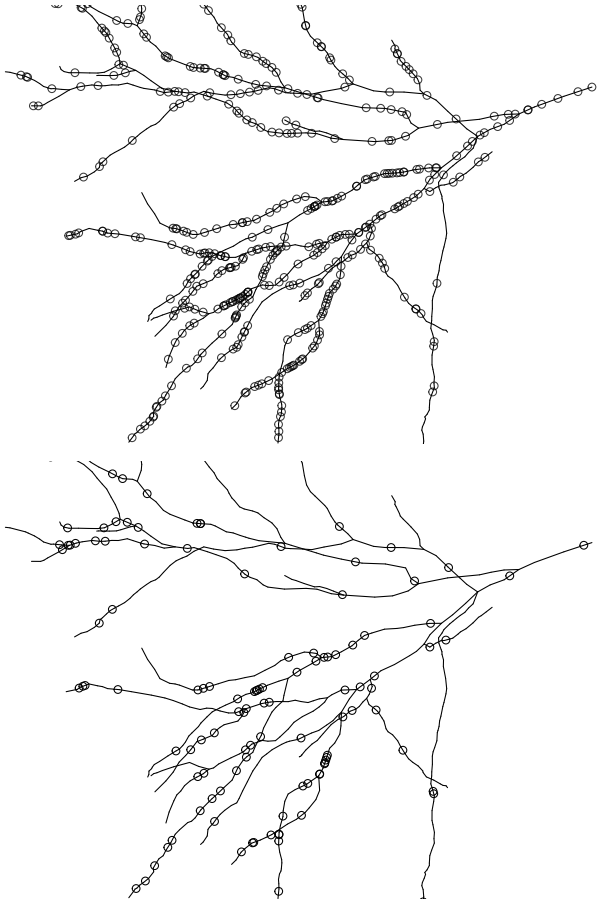

Fig. 10 Dendritic spine data. One branch of the dendritic tree of a neuron, showing the positions of dendritic spines, of "stubby" or "mushroom" type (*Top*) and "thin" type (*Bottom*).

Bandwidth selection was performed using the fast approximation (36) to the leave-one-out estimates. Mean and median nearest neighbour distances were less than 7 microns and the Scott rule gave bandwidths of 12 to 18 microns. Bandwidths from 15 to 300 microns were considered, incurring a total computation time of 81 seconds (whereas the exact computation would have taken 147 minutes). Contour plots for the cross-validation criteria are given in an online supplement. The bandwidths selected by each method are shown in Table 3.

	SYMMETRIC	ASYMMETRIC	
METHOD	h	h_1	h_2
`Scott`	17.6	17.6	12.5
`KD`	82.2	84.9	300
`mod`	83.5	93.7	15
`lik`	79.4	84.9	300
`lsq`	77.9	68.7	300

Table 3 Automatically-selected bandwidths (fast method) for the dendritic spines, relative risk of 'thin' type against 'other' types.

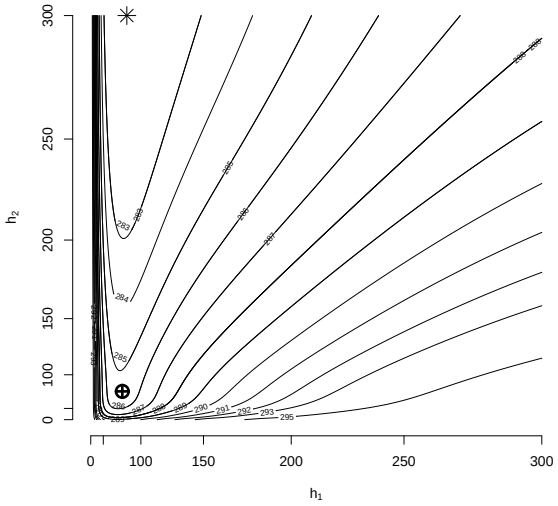

Fig. 11 Contours of likelihood cross-validation criterion (18) for dendrite data, relative risk of 'thin' type against 'other' types. Symbol $\oplus$ indicates optimal symmetric bandwidth h; symbol $*$ indicates optimal joint bandwidths (h_1, h_2).

Figure 11 shows the contours of the likelihood cross-validation criterion for the dendrite data, indicating strong support for a value of h_1 around 80 microns, but supporting a range of h_2 values. Other contour plots are given in the online supplements. Figure 12 shows the estimated relative risk using the bandwidths selected by our modified method. The ratio of counts of thin spines to other spines is $115/451 = 0.26$.

This dataset contains strong evidence for spatially-varying relative risk and, perhaps as a consequence, there is broad agreement between the different cross-validation methods for bandwidth selection.

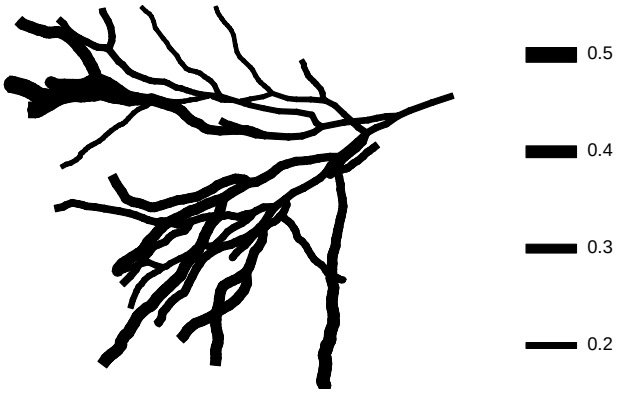

Fig. 12 Estimated relative risk of "thin" type against other types for the dendritic spine data. Line width proportional to relative risk. Bandwidth 83.5 microns, selected by our modified cross-validation method.

8 Discussion

This paper has demonstrated that existing methodology for estimating relative risk in spatial point patterns in two dimensions can be adapted to point patterns on a linear network, with broadly similar results. However, the known weaknesses of cross-validation methods seem to be amplified on a linear network. We identified at least one explanation for these problems and proposed a solution which may also be applicable to the two-dimensional case. Additionally the computational challenges of data analysis are much greater on a network; kernel estimates take much longer to compute, and leave-one-out estimates cannot easily be calculated. We proposed a workable approximation to the leave-one-out calculation.

Our proposed techniques could be applied to any of the kernel estimators that have been proposed in the literature on linear networks. However, the diffusion (heat) kernel estimator has many practical advantages. The finite-element algorithm for solving the heat equation is fast, and it automatically provides kernel estimates for a sequence of intermediate values of bandwidth as well as for the desired bandwidth. Total computation time increases quadratically with the bandwidth, so it becomes important to choose a sharp upper bound on the maximum bandwidth to be considered.

In a suite of experiments, we found that suboptimal (and sometimes unusable) estimates were obtained if the smoothing bandwidths of numerator and denominator were permitted to be different. Simple rules of thumb performed reasonably well, and were the least susceptible to "breakdown". Overall best performance was achieved by our modification of the Kelsall-Diggle density ratio cross-validation method.

We recommend using the diffusion (heat) kernel estimator, and to select the bandwidth using cross-validation with a symmetric bandwidth, using our one-step approximation to the leave-one-out estimator. An *infinite* bandwidth may be valid and can easily be included in the calculations. Two-dimensional convolution smoothing (Rakshit et al, 2019) could be used as a first approximation.

Adaptive smoothing in the style of Abramson (1982) can be implemented using the slicing algorithm of Davies and Baddeley (2018). Bandwidth selection can be performed using the same cross-validation criteria as above (applied to the global bandwidth parameter).

There are many avenues for future research. Extension to more than two types of points is straightforward. For faster computation in very large networks, convolution kernels should be considered (Rakshit et al, 2019). It would be useful to extend the methods of Hazelton and Davies (2009), for identifying regions of (statistically) significantly elevated relative risk, to linear networks. It remains a challenge to adapt the oversmoothing principle of Terrell (1990) to a linear network.

We believe our modification to the Kelsall-Diggle cross-validation criterion would also perform well for two-dimensional spatial and spatio-temporal point patterns.

Acknowledgements We thank Dr Tilman Davies and the referees for insightful comments.

References

Abramson I (1982) On bandwidth estimation in kernel estimates – a square root law. Annals of Statistics 10(4):1217–1223

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

Azzalini A, Bowman A, Haerdle W (1989) On the use of nonparametric regression for model checking. Biometrika 76:1–11

Baddeley A, Jammalamadaka A, Nair G (2014) Multitype point process analysis of spines on the dendrite network of a neuron. Applied Statistics (Journal of the Royal Statistical Society, Series C) 63(5):673–694, DOI 10.1111/rssc.12054

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

Borruso G (2003) Network density and the delimitation of urban areas. Transactions in GIS 7:177–191

Borruso G (2005) Network density estimation: Analysis of point patterns over a network. In: Gervasi O, Gavrilova M, Kumar V, Laganà A, Lee H, Mun Y, Taniar D, Tan C (eds) Computational Science and its Applications — ICCSA 2005, no. 3482 in Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, pp 126–132

Borruso G (2008) Network density estimation: A GIS approach for analysing point patterns in a network space. Transactions in GIS 12:377–402

Botev Z, Grotowski J, Kroese D (2010) Kernel density estimation via diffusion. Annals of Statistics 38(5):2916–2957

Bowman A (1984) An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71:353–360

Bowman AW, Azzalini A (1997) Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations. Oxford University Press, Oxford

Cao R, Cuevas A, Gonzáles-Manteiga W (1994) A comparative study of several smoothing methods in density estimation. Computational Statistics and Data Analysis 17:153–176

Chaudhuri P, Marron J (2000) Scale space view of curve estimation. Annals of Statistics 28:408–428

Clark AB, Lawson AB (2004) An evaluation of non-parametric relative risk estimators for disease maps. Computational Statistics and Data Analysis 47:63–78

Davies T, Baddeley A (2018) Fast computation of spatially adaptive kernel estimates. Statistics and Computing 28:937–956

Davies T, Lawson A (2018) An evaluation of likelihood-based bandwidth selectors for spatial and spatiotemporal kernel estimates, submitted for publication

Davies T, Jones K, Hazelton M (2016) Symmetric adaptive smoothing regimens for estimation of the spatial relative risk function. Computational Statistics and Data Analysis 101:12–28

Davies T, Marshall J, Hazelton M (2018) Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk. Statistics in Medicine 37:1191–1221

Davies TM, Hazelton ML, Marshall JC (2011) sparr: Analyzing spatial relative risk using fixed and adaptive kernel density estimation in R. Journal of Statistical Software 39(1):1–14, URL http://www.jstatsoft.org/v39/i01/

Diggle P (2014) Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, 3rd edn. Chapman and Hall/CRC, Boca Raton, FL

Diggle P, Marron J (1988) Equivalence of smoothing parameter selectors in density and intensity estimation. Journal of the American Statistical Association 83:793–800

Diggle P, Zheng P, Durr P (2005) Non-parametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. Applied Statistics 54:645–658

Downs J, Horner M (2007a) Characterising linear point patterns. In: Winstanley A (ed) Proceedings of the GIScience Research UK Conference (GISRUK), Maynooth, Ireland, National University of Ireland Maynooth, County Kildare, Ireland, pp 421–424

Downs J, Horner M (2007b) Network-based kernel density estimation for home range analysis. In: Proceedings of the 9th International Conference on Geocomputation, Maynooth, Ireland

Downs J, Horner M (2008) Spatially modelling pathways of migratory birds for nature reserve site selection. International Journal of Geographical Information Science 22(6):687–702

Duong T, Hazelton M (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics 15(1):17–30

Duong T, Hazelton M (2005) Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics 32:485–506

Haerdle W (1990) Applied Nonparametric Regression. Cambridge University Press, Cambridge

Hall P, Marron J (1991) Local minima in cross-validation functions. Journal of the Royal Statistical Society, Series B 53:245–252

Hall P, Marron J, Park B (1992) Smoothed cross-validation. Probability Theory and Related Fields 92:1–20

Hazelton M, Davies T (2009) Inference based on kernel estimates of the relative risk function in geographical epidemiology. Biometrical Journal 51:98–109

Hu S, Poskitt DS, Zhang X (2012) Bayesian adaptive bandwidth kernel density estimation of irregular multivariate distributions. Computational Statistics and Data Analysis 56:732–740

Jammalamadaka A, Banerjee S, Manjunath B, Kosik K (2013) Statistical analysis of dendritic spine distributions in rat hippocampal cultures. BMC Bioinformatics 14(287)

Jones M, Marron J, Sheather S (1996) A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association 91(433):401–407

Kelsall J, Diggle P (1995a) Kernel estimation of relative risk. Bernoulli 1:3–16

Kelsall J, Diggle P (1995b) Non-parametric estimation of spatial variation in relative risk. Statistics in Medicine 14:2335–2342

Kelsall J, Diggle P (1998) Spatial variation in risk of disease: a nonparametric binary regression approach. Applied Statistics 47:559–573

Kostrykin V, Potthoff J, Schrader R (2007) Heat kernels on metric graphs and a trace formula. In: Germinet F, Hislop P (eds) Adventures in Mathematical Physics, no. 447 in Contemporary Mathematics, American Mathematical Society, Providence, RI, pp 175–198

Loader C (1999a) Bandwidth selection: classical or plug-in? Annals of Statistics 27(2):415–438

Loader C (1999b) Local Regression and Likelihood. Springer, New York

McSwiggan G, Baddeley A, Nair G (2016) Kernel density estimation on a linear network. Scandinavian Journal of Statistics 44(2):324–345

Mecke J (1967) Stationäre zufällige maße auf lokalkompakten abelschen gruppen. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 9:36–58

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

Okabe A, Satoh T, Sugihara K (2009) A kernel density estimation method for networks, its computational method and a GIS-based tool. International Journal of Geographical Information Science 23:7–32

Rakshit S, Davies T, Moradi M, McSwiggan G, Nair G, Mateu J, Baddeley A (2019) Fast kernel smoothing of point patterns on a large network using 2D convolution. International Statistical Review DOI 10.1111/insr.12327, in press. Published online 06 June 2019

Scott D (1992) Multivariate Density Estimation. Theory, Practice and Visualization. John Wiley and Sons, New York

Silverman B (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London

Sugihara K, Satoh T, Okabe A (2010) Simple and unbiased kernel function for network analysis. In: ISCIT 2010 (International Symposium on Communication and Information Technologies), IEEE, pp 827–832, DOI 10.1109/ISCIT.2010.5665101

Terrell G (1990) The maximal smoothing principle in density estimation. Journal of the American Statistical Association 85:470–476

Wand M, Jones M (1995) Kernel Smoothing. Chapman and Hall

Xie Z, Yan J (2008) Kernel density estimation of traffic accidents in a network space. Computers, Environment and Urban Systems 32:396–406

Yule GH (1903) Notes on the theory of association of attributes in Statistics. Biometrika 2:121–134

Zhang X, King ML, Hyndman RJ (2006) A Bayesian approach to bandwidth selection for multivariate kernel estimation. Computational Statistics and Data Analysis 50:3009–3031