

Fast Kernel Smoothing of Point Patterns on a Large Network using Two-dimensional Convolution

Suman Rakshit¹, Tilman Davies², M. Mehdi Moradi³, Greg McSwiggan⁴, Gopalan Nair⁴, Jorge Mateu³ and Adrian Baddeley¹ 

¹*Curtin University, Perth, Australia*

²*University of Otago, Dunedin, New Zealand*

³*University Jaume I, Castellón, Spain*

⁴*University of Western Australia, Perth, Australia*

E-mail: adrian.baddeley@curtin.edu.au

Summary

We propose a computationally efficient and statistically principled method for kernel smoothing of point pattern data on a linear network. The point locations, and the network itself, are convolved with a two-dimensional kernel and then combined into an intensity function on the network. This can be computed rapidly using the fast Fourier transform, even on large networks and for large bandwidths, and is robust against errors in network geometry. The estimator is consistent, and its statistical efficiency is only slightly suboptimal. We discuss bias, variance, asymptotics, bandwidth selection, variance estimation, relative risk estimation and adaptive smoothing. The methods are used to analyse spatially varying frequency of traffic accidents in Western Australia and the relative risk of different types of traffic accidents in Medellín, Colombia.

Key words: bandwidth; intensity; linear network; spatial point pattern.

1 Introduction

This paper develops a fast non-parametric method for estimating the spatially varying density or intensity of points observed on a network of lines. In applications, the lines could represent roads, rivers, railways, electrical wires or nerve fibres, while the points could be the locations of traffic accidents, street crimes, roadside trees, invasive species, kiosks, animal nests or neurological features; see Okabe and Sugihara (2012) and Baddeley *et al.* (2015, chapter 17) for surveys and recent work by O'Donnell *et al.* (2014) and Isaak *et al.* (2014).

The analysis of network data is challenging because of geometrical complexities, unique methodological problems and the huge volume of data. Here, we analyse three datasets of increasing size with progressively more complicated features. The first of these, shown in Figure 1, gives the street address locations of 116 crimes in an area close to the University of Chicago, on a street network of 503 individual segments. The most challenging dataset in the paper is shown in Figure 2; it gives the locations of 14 562 traffic accidents on a road network



Figure 1. Chicago crime location data. Street network and street address locations of crimes near the University of Chicago. Chicago Weekly News. Map width 1 280 ft. Arrow points North.



Figure 2. Traffic accidents (black dots) recorded in the year 2011 on the state road network (grey lines) of the southern half of Western Australia. Map width 1 460 km.

of 115 169 segments in the state of Western Australia. This was the motivating dataset for our study. The accident locations and the road network itself are highly concentrated in the urban area along the west coast of the state. Major highways are clearly visible from the accident pattern.

Estimates of the spatially varying frequency of events on a network are needed for emergency response planning, accident research, urban modelling and other purposes. The technique that immediately comes to mind is kernel density estimation (Silverman, 1986). For spatial point pattern data in two dimensions, kernel estimates are conceptually simple (Diggle, 1985; Bithell, 1990) and very fast to compute using the fast Fourier transform (FFT) (Silverman, 1982). However, for points on a network, kernel estimation is mathematically intricate and computationally expensive (Okabe *et al.*, 2009; Sugihara *et al.*, 2010; Okabe & Sugihara, 2012, chapter 9; McSwiggan *et al.*, 2016; Moradi *et al.*, 2017). The fastest algorithm available hitherto (McSwiggan *et al.*, 2016) has computation times that increase quadratically with bandwidth; current implementations can easily deal with datasets of the size of Figure 1 but have difficulty with Figure 2, with computation times ranging from seconds to hours depending on the spatial resolution and error tolerance.

In this paper, we describe a kernel estimator of probability density or point process intensity on a network based on two-dimensional (2D) kernels. Our proposed estimator can be seen as a rigorously justified modification and extension of an idea of Borruso (2008). The estimator can be computed very rapidly using the FFT. It can be viewed as formally equivalent to the standard fixed-bandwidth edge-corrected kernel estimator for 2D spatial point patterns (Diggle, 1985; Bithell, 1990) and enjoys similar statistical properties. When compared with the diffusion method of McSwiggan *et al.* (2016), the main disadvantage of our estimator is a slight decrease in statistical performance. Existing methods of bandwidth selection can easily be adapted to our setting. The technique easily lends itself to theoretical analysis and to extensions such as estimation of the spatially varying relative risk of two different types of accidents. We also extend our proposed technique to adaptive (variable bandwidth) estimation, in order to accommodate extreme variation in the density of events and of the network itself. Unlike other techniques, our method is robust against errors or changes in the connectivity of the network.

Section 2 provides definitions, background and literature. Section 3 introduces our proposed method, and Section 4 describes a fast computational implementation. The method is demonstrated on the Chicago crime data of Figure 1 in Section 5. Theoretical properties are described in Section 6, including exact formulae for bias and variance, variance estimators and asymptotic behaviour. Section 7 reports the results of simulation experiments comparing the statistical performance of our proposed estimator with that of the diffusion estimator. Section 8 sketches strategies for bandwidth selection. Section 9 briefly describes the estimation of spatially varying relative risk and spatial smoothing of data observed at sample points on a network. Section 10 applies these techniques to the relative risk of different types of traffic accident in Medellín, Colombia. Section 11 describes the extension to adaptive (variable bandwidth) estimation, and this is applied in Section 12 to analyse data on road accidents in the state of Western Australia (Figure 2). We end with a discussion. The Supporting Information provides code, additional results and details of the simulation experiments.

2 Background and Definitions

2.1 Kernel Estimation of Density and Intensity in Two-dimensional

Readers will be familiar with the classical kernel estimator of probability density $f(x)$ for real-valued observations x_1, \dots, x_n given by $\hat{f}(x) = n^{-1} \sum_{i=1}^n \kappa(x - x_i)$ for $x \in \mathbb{R}$,

where the kernel κ is a probability density function on the real line. Without the normalising factor $1/n$,

$$\hat{\lambda}(x) = \sum_{i=1}^n \kappa(x - x_i), \quad x \in \mathbb{R}, \quad (1)$$

is the kernel estimator of the intensity function $\lambda(x)$ of the point process assumed to have generated the data. This paper focuses on the estimation of intensities but can easily be rephrased in terms of probability densities (Diggle & Marron, 1988).

In 2D Euclidean space, suppose we observe a spatial pattern of points $\mathbf{x} = \{x_1, \dots, x_n\}$ inside a restricted spatial domain ('window') $W \subset \mathbb{R}^2$. Diggle (1985) proposed the edge-corrected estimator of intensity

$$\hat{\lambda}^U(u) = \frac{1}{c_W(u)} \sum_{i=1}^n \kappa(u - x_i), \quad u \in W, \quad (2)$$

where κ is now a probability density on \mathbb{R}^2 and $c_W(u) = \int_W \kappa(u - v)dv$ is the mass of the kernel centred at u that falls inside the window; see also Bithell (1990). The factor $1/c_W(u)$ in (2) is often called the 'uniform' edge correction, because $\hat{\lambda}^U$ is a pointwise unbiased estimator when the true intensity is uniform: if $\lambda(u) \equiv \lambda > 0$, then $\mathbb{E}[\hat{\lambda}^U(u)] \equiv \lambda$. Jones (1993) proposed the alternative estimator

$$\hat{\lambda}^{JD}(u) = \sum_{i=1}^n \frac{\kappa(u - x_i)}{c_W(x_i)}, \quad u \in W, \quad (3)$$

often confusingly called the 'Diggle correction' in spatial statistics literature: we shall call it the 'Jones–Diggle' correction. This estimator conserves total mass, that is, $\int_W \hat{\lambda}^{JD}(u)du = n$.

2.2 Linear Networks

We define a linear network L as the union $L = \cup_{i=1}^N l_i$ of finitely many line segments $l_i = [u_i, v_i] = \{w : w = tu_i + (1-t)v_i, 0 \leq t \leq 1\}$ in the 2D 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 for $i \neq j$ is either empty or an endpoint of both segments. The total length of a subset $B \subseteq L$ is denoted by $|B|$.

Let $\mathbf{x} = \{x_1, \dots, x_n\}$ be 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 (Daley & Vere-Jones, 2003, chapter 5).

The intensity function $\lambda(u)$ of \mathbf{X} gives the expected number of points per unit length of network in the vicinity of location u . Formally, \mathbf{X} has intensity function $\lambda(u)$, $u \in L$, if

$$E[N(\mathbf{X} \cap B)] = \int_B \lambda(u) d_1 u, \quad (4)$$

for all measurable $B \subset L$, where $N(\mathbf{X} \cap B)$ is the number of points of \mathbf{X} falling in B and $d_1 u$ denotes integration with respect to arc length on the network. Thus, an intensity function on the

network has values with dimension length^{-1} , points per unit length of network. Campbell's formula on a network states that

$$E \left[\sum_{x_i \in \mathbf{X}} h(x_i) \right] = \int_L h(u) \lambda(u) d_1 u, \quad (5)$$

where h is any Borel measurable real function on L such that $\int_L |h(u)| \lambda(u) d_1 u < \infty$.

2.3 Literature on Intensity Estimation on a Network

Given an observed point pattern \mathbf{x} on a network L , the task is to estimate the intensity function $\lambda(u)$ defined in (4).

There is currently no general agreement on how to perform kernel smoothing on a network (Borruso, 2003, 2005, 2008; Downs & Horner, 2007a, 2007b, 2008; Xie & Yan, 2008; Okabe *et al.*, 2009; Sugihara *et al.*, 2010; Okabe & Sugihara, 2012, chapter 9; McSwiggan *et al.*, 2016; Moradi *et al.*, 2017).

Xie and Yan (2008) proposed that intensity on a network should be estimated by a modification of the one-dimensional kernel estimator (1) in which the vector distance $x - x_i$ is formally replaced by the shortest-path distance in the network, $d_L(x, x_i)$, giving

$$\hat{\lambda}(u) = \sum_{i=1}^n \kappa(d_L(u, x_i)), \quad u \in L, \quad (6)$$

where κ is a one-dimensional kernel, a probability density on the real line. However, this is fallacious, as pointed out by Okabe and Sugihara (2012) and McSwiggan *et al.* (2016). Kernel estimates are sums of probability densities, but in (6), the 'kernel' $k_i(u) = \kappa(d_L(u, x_i))$ for data point x_i is not a probability density on L : it typically has mass far greater than 1, resulting in an extremely biased estimator of the true intensity $\lambda(u)$.

Borruso (2003, 2005, 2008) offered several *ad hoc* proposals for kernel smoothing of network data including the 'Euclidean, divide-by-length' smoother (Borruso, 2008)

$$\hat{\lambda}^B(u) = \frac{N(\mathbf{x} \cap b(u, r))}{|L \cap b(u, r)|}, \quad u \in \mathbb{R}^2, \quad (7)$$

where $b(u, r) = \{v \in L : \|v - u\| \leq r\}$ is the 2D disc of fixed radius $r > 0$ centred at the query location u . That is, (7) is the number of data points, divided by the total network length, within a disc of radius r . This was described as a 'pure' density estimate (Borruso, 2008, p. 382) but no justification or statistical properties were given. Borruso (2008) allows u to be any location in 2D space.

Kernel estimators on a network were investigated systematically by Okabe *et al.* (2009) and Sugihara *et al.* (2010) and summarised in Okabe and Sugihara (2012, chapter 9). A family of kernel estimators based on weighted sums of path lengths, generalising and correcting (6), was proposed. Two kernel estimators were found to satisfy many of the desired properties: the 'equal-split discontinuous' and 'equal-split continuous' rules (Okabe & Sugihara, 2012, sections 9.2.2 and 9.2.3). The 'continuous' rule has good properties but is extremely slow to compute by the original algorithm of Sugihara *et al.* (2010; see Okabe & Sugihara, 2012, algorithm 9.3, section 9.3.3), while the 'discontinuous' rule is faster but has less desirable properties (Okabe & Sugihara, 2012, section 9.3.2). Both methods require a kernel on the real line with bounded support, which excludes the Gaussian kernel.

McSwiggan *et al.* (2016) developed a statistically principled kernel estimator on a linear network by exploiting the connection between kernel smoothing and diffusion (Chaudhuri & Marron, 2000; Botev *et al.*, 2010). Their diffusion estimator $\widehat{\lambda}^H(u)$ can be expressed as a kernel sum

$$\widehat{\lambda}^H(u) = \sum_{i=1}^n \kappa_\tau(x_i, u), \quad u \in L, \quad (8)$$

where $\kappa_\tau(x_i, v)$ is the heat kernel, the probability density at time $t = \tau^2$ of a particle executing a Brownian motion on the network, starting from location x_i at time 0. McSwiggan *et al.* (2016, lemma 2, p. 12) showed that the diffusion estimator is formally equivalent to the ‘equal-split continuous’ rule (Sugihara *et al.*, 2010; Okabe & Sugihara, 2012, chapter 9) applied to the Gaussian kernel. The diffusion estimator conserves mass and is unbiased when the true density is uniform (McSwiggan *et al.*, 2016, section 7). These properties strongly suggest that (8) should be regarded as the ‘canonical’ kernel estimator on a network. The diffusion estimator also satisfies the classical time-dependent heat equation so that (8) can be computed rapidly by solving the heat equation up to time $t = \tau^2$. Table 1 in Section 5 gives computation times in an example. However, computation on a large network can be costly, depending on the bandwidth and spatial resolution (Section 12.4).

Moradi *et al.* (2017) introduced a network counterpart of (3) using the shortest-path distance,

$$\widehat{\lambda}^M(u) = \sum_{i=1}^n \frac{\kappa(d_L(u, x_i))}{A_L(x_i)}, \quad u \in L, \quad (9)$$

where κ is a one-dimensional kernel and $A_L(u) = \int_L \kappa(d_L(u, v))d_1v$. The estimator (9) conserves mass and is a continuous function on the network if κ is continuous. Computation of (9) is costly in large networks.

3 Proposed Method

In this paper, we propose a kernel estimator on a linear network based on a 2D smoothing kernel. The original motivation was speed: this estimator can be expressed in terms of 2D convolutions of the kernel, so it can be computed very rapidly using the FFT. However, it also has theoretical and methodological advantages.

Definition 1. Let $\mathbf{x} = \{x_1, \dots, x_n\}$ be a point pattern on a linear network L . Let κ denote a bivariate kernel function, that is, a probability density on \mathbb{R}^2 . The convolution kernel estimator of intensity is, with the uniform correction,

$$\widehat{\lambda}^U(u) = \frac{1}{c_L(u)} \sum_{i=1}^n \kappa(u - x_i), \quad u \in L, \quad (10)$$

Table 1. Computation time (in seconds) for the diffusion algorithm and the convolution algorithm (U = uniform correction; J = Jones–Diggle correction) applied to the Chicago data with different bandwidths σ or τ (in feet).

σ or τ	50	100	150	200	250	300	350	400	450	500
Diffusion	0.053	0.077	0.116	0.176	0.246	0.327	0.429	0.546	0.676	0.824
Convolution (U)	0.044	0.041	0.041	0.041	0.041	0.042	0.044	0.042	0.042	0.041
Convolution (J)	0.041	0.041	0.041	0.042	0.043	0.041	0.041	0.041	0.041	0.038

and with the Jones–Diggle correction

$$\widehat{\lambda}^{JD}(u) = \sum_{i=1}^n \frac{\kappa(u - x_i)}{c_L(x_i)}, \quad u \in L, \quad (11)$$

where

$$c_L(u) = \int_L \kappa(v - u) d_1 v, \quad u \in L. \quad (12)$$

Notice that (10) and (11) are similar to (2) and (3), respectively, and could have been formally ‘derived’ by replacing the 2D integral over the window W by the one-dimensional integral over the network L . However, a rigorous justification needs to be found, and statistical properties analysed, in this new setting.

The sum in (10) is the usual kernel estimator of intensity in two dimensions without edge correction. The denominator $c_L(u)$ defined in (12) is the convolution of the kernel κ with arc-length measure on the network. Both functions are now evaluated only at locations on the network.

Intuitively, it may seem wrong to use a 2D kernel to estimate a one-dimensional density. However, the normalisation in (10)–(11) ensures that the result has the correct dimension. Values of κ have dimension $length^{-2}$ (points per unit area), while values of $c_L(\cdot)$ are integrals of κ against length so have dimension $length^{-1}$. Consequently, (10) and (11) have the correct dimension $length^{-1}$ (points per unit length).

In the special case where κ is the uniform density on a disc of fixed radius $r > 0$, that is, $\kappa(u) = a 1\{\|u\| \leq r\}$, where $a = 1/(\pi r^2)$, the uniform correction estimator (10) reduces to

$$\widehat{\lambda}^U(u) = \frac{N(\mathbf{x} \cap b(u, r))}{|L \cap b(u, r)|}, \quad u \in L, \quad (13)$$

the number of points divided by the total length of network in a disc of radius r centred on u . This is Borruso’s estimator (7) but now restricted to query locations u lying on the network. The Jones–Diggle correction estimator (11) reduces to

$$\widehat{\lambda}^{JD}(u) = \sum_{x_i \in b(u, r)} \frac{1}{|L \cap b(x_i, r)|}, \quad u \in L. \quad (14)$$

This corresponds to associating, with each data point x_i , a unit mass that is then uniformly spread over the part of the network lying within Euclidean distance r of x_i .

Returning to the general case, the normalising factor $c_L(u)$ in (12) is also very closely related to the ‘network density’ proposed by Borruso (2003), $B(u) = \frac{1}{m} \sum_{j=1}^m \kappa(u - v_j)$, $u \in \mathbb{R}^d$,

where v_1, \dots, v_m are random sample points on the network. The function $B(u)$ for $u \in \mathbb{R}^2$ was used by Borruso (2003) in order to investigate the spatial density of the lines themselves. By Campbell’s formula (5), if the sample points are uniformly distributed, $\mathbb{E}[B(u)] = c_L(u)/|L|$ for $u \in L$. Thus, $c_L(u)$ or $B(u)$ has the interpretation of a reference or baseline value that would be obtained if the data points were uniformly randomly distributed on the network.

The uniform correction estimator (10) could also be motivated by techniques for estimating spatially varying relative risk in a case–control study (Bithell, 1991; Kelsall & Diggle, 1995a). The data points x_i are viewed as the locations of cases, and a random sample of points uniformly distributed on the network will serve as the controls. The standard estimator of relative risk is

the ratio of kernel estimates of probability density for the cases and for the controls. This is effectively the same as (10).

It is theoretically permissible to choose a kernel κ that is not isotropic, that is, not invariant under rotation. This seems undesirable in practice, except in situations where the coordinate system is not isometric, such as the latitude–longitude coordinates on a globe. We shall assume that κ is isotropic.

Unlike estimators of intensity based on path distances in the network, the convolution estimators are robust against errors in the geometry of the network. If κ is uniformly continuous, the quantities (10), (12) and (11) are continuous functions of the point pattern \mathbf{x} and the linear network L under the Wasserstein metric (Rüschendorf, 1994).

4 Fast Computation

The sums in (10)–(11) and the integral (12) can be recognised as convolutions of the kernel κ with different measures M on \mathbb{R}^2 . These may be computed rapidly using the FFT, using the rule (Pinsky, 2002, pp. 13, 92; Wand & Jones, 1995, appendix D)

$$\kappa * M = \mathcal{F}^{-1}(\mathcal{F}(\kappa) \cdot \mathcal{F}(M)), \quad (15)$$

where $*$ denotes convolution, \mathcal{F} is the Fourier transform and ‘ \cdot ’ is pointwise multiplication of functions.

The sum in (10) is the convolution of the kernel κ with the counting measure on data points (Silverman, 1982). It would usually be computed by discretising the point pattern \mathbf{x} onto a fine grid of pixels (where each pixel value equals the number of data points falling in the pixel) and then applying (15). Similarly, the estimator (11) is the convolution of κ with the measure that puts mass $1/c_L(x_i)$ on data point x_i for $i = 1, \dots, n$. A similar computational strategy can be used.

The denominator $c_L(u)$ in (12) is the convolution of κ with arc-length measure on the network. Here, one can discretise the network onto a pixel grid, where each pixel value equals the length of network segments intersecting the pixel, and apply (15). To suppress artefacts such as aliasing and numerical underflow, we recommend that pixel size should be smaller than one-tenth of the smoothing bandwidth.

Alternatively, for appropriate choices of the kernel, $c_L(u)$ can be computed analytically. Decompose $c_L(u)$ into contributions from each network segment:

$$c_L(u) = \int_L \kappa(u - v) d_1 v = \sum_{j=1}^N \int_{l_j} \kappa(u - v) d_1 v = \sum_{j=1}^N c_{l_j}(u). \quad (16)$$

When κ is the 2D isotropic Gaussian kernel with mean 0 and standard deviation σ , it can easily be calculated that

$$c_{l_j}(u) = \phi_\sigma(u_2) |\Phi_\sigma(s_j - u_1) - \Phi_\sigma(u_1)|, \quad (17)$$

where ϕ_σ and Φ_σ are, respectively, the probability density function and cumulative distribution function of the Gaussian $N(0, \sigma^2)$ distribution, $s_j = |l_j|$ is the length of the segment l_j and (u_1, u_2) are the coordinates of u in a coordinate system in which the origin is an endpoint of l_j and the first coordinate axis is parallel to l_j . The function (17) can be evaluated rapidly at a fine grid of pixels u .

We implemented these procedures in the R language using the package `spatstat` (Baddeley & Turner, 2005; Baddeley *et al.*, 2015). Timings are reported for the implementation of (10)–(12) using byte-compiled R code (Tierney, 2001) and the `fftw` library for FFT computation (Frigo & Johnson, 2005; Rahim, 2017).

5 Chicago Example

Figure 1 shows the nearest street address locations of crimes reported between 25 April 2002 and 8 May 2002 in the neighbourhood of the University of Chicago (Illinois, USA). The original crime map was published in the *Chicago Weekly News* in 2002. The data were presented and analysed by Ang *et al.* (2012) and in Baddeley *et al.* (2015, chapter 17), Baddeley *et al.* (2017), Moradi *et al.* (2017) and Rakshit *et al.* (2017).

Table 1 shows computation times for our proposed convolution algorithm (Section 3) and the diffusion algorithm (Section 2.3) applied to the Chicago data for various values of bandwidth. Computation time increases steeply with bandwidth for the diffusion algorithm but is roughly constant for the convolution algorithm. However, for a given bandwidth τ , the diffusion algorithm also provides kernel estimates for a finely spaced sequence of bandwidths η with $0 < \eta \leq \tau$.

Comparison of the methods is complicated by the fact that the diffusion bandwidth τ and convolution bandwidth σ have different meanings. The diffusion bandwidth τ can be interpreted as a path distance on the network, while the convolution bandwidth σ is a Euclidean distance in 2D space. A rough conversion rule is needed. To investigate this, one may inspect a scatterplot of shortest-path distance against Euclidean distance for a random sample of points on the network. We found that an approximately proportional relationship between shortest-path distance $\delta_{\text{SP}}(u, v)$ and Euclidean distance $\|u - v\|_2$ holds for some kinds of networks, in particular for street networks. That is,

$$\delta_{\text{SP}}(u, v) \approx \beta \|u - v\|_2 \quad (18)$$

for independent uniform random locations u and v on L , where β is the conversion factor. For the Chicago network, we estimate $\beta \approx 1.25$. In any network, we have $\|u - v\|_2 \leq \delta_{\text{SP}}(u, v)$ for all u, v . In a perfect grid composed of $a \times b$ rectangles, it is easy to show that $\delta_{\text{SP}}(u, v) = \|u - v\|_1 + q(u, v)$, where $\|\cdot\|_1$ denotes the L^1 norm and $q(u, v) \leq 2(a + b)$. Because $\|u - v\|_1 \leq \sqrt{2}\|u - v\|_2$, the approximation (18) with $\beta = 1.25$ should be reasonable in any rectangular grid.

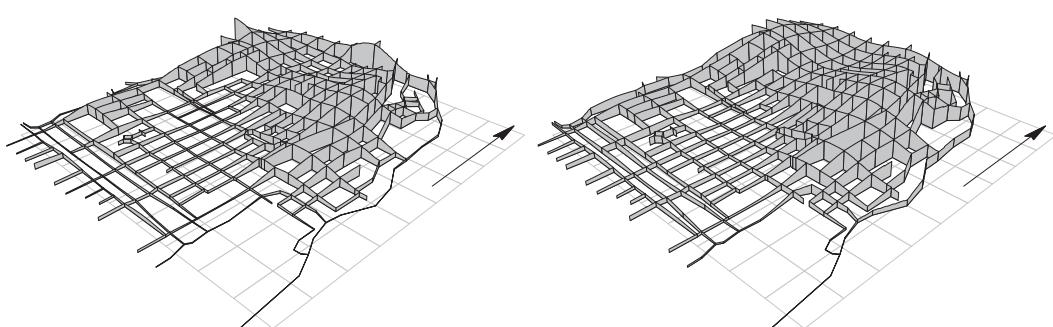


Figure 3. Kernel estimates of intensity for Chicago data. Perspective views with height representing function value. Left: diffusion estimate with bandwidth 125 ft. Right: convolution method with bandwidth 100 ft and uniform correction. Arrow points North.

For general networks, we shall estimate the conversion factor β by generating a large sample of independent uniform random points on the network, evaluating the pairwise Euclidean distances and pairwise shortest-path distances and fitting a proportional regression (18) with standard deviation proportional to the mean.

Figure 3 shows intensity estimates for the Chicago crimes using the diffusion and convolution methods, with comparable bandwidths 125 and 100 ft, respectively. These are three-dimensional perspective views in the style of Okabe and Sugihara (2012) in which the function values are represented by the heights of vertical walls erected above each segment of the network. While their overall appearance is very similar, the two estimates show some clear differences near the boundary, for example, on the eastern-most road and the northwest terminal nodes. The eastern-most road is geographically close to crime locations, but the corresponding network path distance is much longer.

6 Theoretical Properties

6.1 Bias

If \mathbf{X} is a point process on L with intensity function $\lambda(u)$, $u \in L$, then by Campbell's formula (5), the estimators (10) and (11) have mean values

$$\mathbb{E}[\widehat{\lambda}^U(u)] = \frac{1}{c_L(u)} \mathbb{E} \sum_i \kappa(u - x_i) = \frac{\int_L \kappa(u - v) \lambda(v) d_1 v}{\int_L \kappa(u - v) d_1 v}, \quad (19)$$

$$\mathbb{E}[\widehat{\lambda}^{JD}(u)] = \int_L \frac{\kappa(u - v)}{c_L(v)} \lambda(v) d_1 v, \quad (20)$$

for $u \in L$. The right-hand sides can be regarded as smoothed versions of the intensity function. In particular, if \mathbf{X} has uniform intensity, $\lambda(u) \equiv \lambda > 0$, then (19) gives $\mathbb{E}[\widehat{\lambda}^U(u)] \equiv \lambda$ so that the uniform correction estimator (10) is pointwise unbiased, while the Jones–Diggle correction estimator (11) is not:

$$\mathbb{E}[\widehat{\lambda}^{JD}(u)] = \lambda \int_L \frac{\kappa(u - v)}{c_L(v)} d_1 v, \quad u \in L. \quad (21)$$

On the other hand, the Jones–Diggle estimator conserves mass, $\int_L \widehat{\lambda}^{JD}(u) d_1 u = n$, while the uniform estimator does not. This echoes the behaviour of the 2D estimators (2)–(3). In contrast, the diffusion estimator (8) and equal-split continuous estimator (Okabe & Sugihara, 2012, chapter 9) are both mass conserving and unbiased for the uniform intensity.

6.2 Variance

If \mathbf{X} is a Poisson point process on L with intensity function $\lambda(u)$, then the general variance formula $\text{var}[\sum_{i=1}^n h(x_i)] = \int_L h(u)^2 \lambda(u) d_1 u$ holds for any function $h : L \rightarrow \mathbb{R}$ for which the right-hand side is integrable. Hence, the pointwise variances of the uniform correction estimator (10) and Jones–Diggle correction estimator (11) are respectively

$$\text{var}[\widehat{\lambda}^U(u)] = \frac{1}{c_L(u)^2} \int_L \kappa(u - v)^2 \lambda(v) d_1 v, \quad (22)$$

$$\text{var}[\widehat{\lambda}^{\text{JD}}(u)] = \int_L \frac{\kappa(u-v)^2}{c_L(v)^2} \lambda(v) d_1 v, \quad (23)$$

for $u \in L$. In particular, if \mathbf{X} is a uniform Poisson process with intensity λ , then

$$\text{var}[\widehat{\lambda}^{\text{U}}(u)] = \frac{\lambda}{c_L(u)^2} \int_L \kappa(u-v)^2 d_1 v, \quad u \in L, \quad (24)$$

$$\text{var}[\widehat{\lambda}^{\text{JD}}(u)] = \lambda \int_L \frac{\kappa(u-v)^2}{c_L(v)^2} d_1 v, \quad u \in L. \quad (25)$$

The variance is not constant over the network, even if the true intensity is uniform. The variances (22) and (23) can be estimated unbiasedly from data $\{x_1, \dots, x_n\}$ by (respectively)

$$\widehat{V}_{\text{U}}(u) = \frac{1}{c_L(u)^2} \sum_{i=1}^n \kappa(u-x_i)^2, \quad (26)$$

$$\widehat{V}_{\text{JD}}(u) = \sum_{i=1}^n \frac{\kappa(u-x_i)^2}{c_L(x_i)^2}, \quad (27)$$

for $u \in L$. These expressions can again be evaluated using FFT methods, in roughly the same time as computing the intensity estimates (10) and (11) themselves. This is another advantage of the convolution method: although analogous formulae exist for the variances of other intensity estimators, their computational cost is usually prohibitive.

6.3 Asymptotics

Here, we briefly sketch some asymptotic properties of our proposed intensity estimator and its asymptotic equivalence with other estimators.

Consider the kernel with bandwidth $\sigma > 0$ defined by $\kappa_\sigma(u) = \sigma^{-2} \kappa_1(u/\sigma)$ for $u \in \mathbb{R}^2$, where κ_1 is a given, fixed, probability density on \mathbb{R}^2 . We assume $\kappa_1(x)$ is continuous at $x = 0$. The estimators (10) and (11) computed using $\kappa = \kappa_\sigma$ will be denoted by $\widehat{\lambda}_\sigma^{\text{U}}$ and $\widehat{\lambda}_\sigma^{\text{J}}$, respectively.

6.3.1 Large bandwidth, fixed data

Suppose the point pattern dataset \mathbf{x} is fixed and the bandwidth σ increases, then the convolution estimate converges to a constant intensity on the entire network L .

To prove this, because κ_1 is continuous at 0, as $\sigma \rightarrow \infty$, we have $\kappa_\sigma(x-y) \sim \sigma^{-2} \kappa_1(0)$ for any fixed $x, y \in L$. Hence, $\widehat{\lambda}_\sigma^{\text{U}}(u) \rightarrow \bar{\lambda}$ and $\widehat{\lambda}_\sigma^{\text{J}}(u) \rightarrow \bar{\lambda}$, uniformly in u , where $\bar{\lambda} = n/|L|$ is the average intensity over the network.

By comparison, the diffusion estimate converges as $\sigma \rightarrow \infty$ to a function which is constant on each connected component of the network (McSwiggan *et al.*, 2016, section 7.2).

6.3.2 Increasing intensity, decreasing bandwidth

Standard asymptotic results for kernel density estimation on the infinite real line can be generalised to a linear network, adapting the approach of Botev *et al.* (2010). Let $N \rightarrow \infty$ and suppose the true intensity is $\lambda_N(u) = N \lambda_1(u)$ on L , where $\lambda_1(u)$ is twice continuously differentiable. Assume the bandwidth σ_N satisfies $\sigma_N \rightarrow 0$ and $N\sigma_N \rightarrow \infty$. Then adapting Botev

et al. (2010, theorem 1), for any location u that is not a vertex, the behaviour of $\widehat{\lambda}_\sigma^U(u)/N$ and $\widehat{\lambda}_\sigma^J(u)/N$ is asymptotically equivalent to that would occur if the network were an infinite straight line. Both estimators are asymptotically equivalent to the kernel smoother on the real line with (one-dimensional) kernel $\zeta_\sigma(x) = \kappa_\sigma((0, x))/Z_1(\sigma)$, where $Z_1(\sigma)$ is the normalising constant $Z_1(\sigma) = \int_{-\infty}^{\infty} \kappa_\sigma((0, x))dx$. By change of variables $Z_1(\sigma) = (1/\sigma)Z_1(1)$ so that $\zeta_\sigma(x) = \sigma^{-1}\zeta_1(x/\sigma)$ in accordance with the usual scaling behaviour of a kernel in one dimension.

In the special case where the 2D kernel κ_σ is the isotropic Gaussian density with standard deviation σ in each coordinate, the corresponding one-dimensional density ζ_σ is the Gaussian density with standard deviation σ , and the normalised bias and variance of the convolution estimator $\widehat{\lambda}^U(u)/N$ satisfy, for any fixed $u \in L$,

$$\mathbb{E}[N^{-1}\widehat{\lambda}^U(u) - \lambda_1(u)] = \frac{\sigma^2}{2} \frac{\partial^2 \lambda_1}{\partial u^2} + O(\sigma^4), \quad (28)$$

$$\text{var}[N^{-1}\widehat{\lambda}^U(u)] = \frac{\lambda_1(u)}{2\sqrt{\pi}N\sigma} + o(1), \quad (29)$$

and similarly for $N^{-1}\widehat{\lambda}^{JD}(u)$. At a vertex v of degree m , the estimate $\widehat{\lambda}^U(v)$ is equal to $1/m$ times the sum of m asymptotically independent contributions with asymptotic bias and variance of the form (28)–(29); hence, these asymptotics also hold at a vertex.

7 Simulation Experiments

Simulation experiments are necessary in order to compare the performance of our convolution method (10)–(11) with that of the diffusion method (8).

7.1 Description of Experiments

We simulated eight different scenarios, illustrated in Figure 4. In the top row, the network L is taken to be the street network from the Chicago data (Figure 1). In the bottom row, L is a subset of the Western Australian road network (Figure 2) representing the southern part of the city of Perth, comprising 18 870 road segments.

In the first column of Figure 4, the intensity is constant on the network. In the second column, a mixture of four anisotropic Gaussian probability densities with mean vectors lying inside the convex hull of the network was chosen arbitrarily and evaluated only at locations on the network. In the third column, a realisation of a stationary log-Gaussian random field was generated and evaluated only at locations on the network. In the fourth column, an intensity function was computed by applying the diffusion method to the original point pattern data for the relevant network. Full details are given in the Supporting Information. Each intensity was scaled to integrate to an expected sample size of 500. These eight functions were then held fixed for the rest of the experiment.

For each given scenario with true intensity function λ on L , we generated 100 simulated realisations of a Poisson process with intensity λ . For each realisation, we computed the intensity estimates $\widehat{\lambda}^U$, $\widehat{\lambda}^{JD}$ for several bandwidths σ and computed the diffusion estimate $\widehat{\lambda}^H$ over the corresponding bandwidths $\tau = \beta\sigma$, where β is the conversion factor in (18). We used $\beta \approx 1.25$ for the Chicago network and $\beta \approx 1.42$ for the southern Perth network. Pixel images were computed with resolution 128×128 for the Chicago network and $2\,048 \times 2\,048$ for the southern

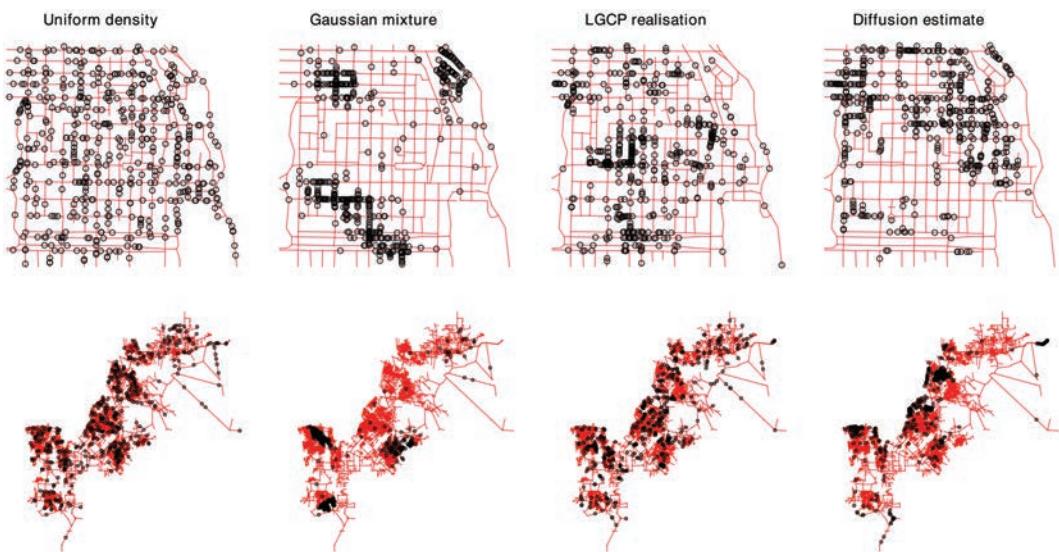


Figure 4. Typical simulated realisations for each of the eight scenarios. Top row: Chicago street network. Bottom row: southern part of the city of Perth, extracted from the Western Australian road network. The Gaussian mixture and log-Gaussian Cox process (LGCP) realisation scenarios are based on an initial two-dimensional surface defined on W . The diffusion estimate scenario is based on the original data observed on the relevant network. Simulated realisations all have size $n = 500$. [Colour figure can be viewed at wileyonlinelibrary.com]

Perth network to ensure sufficient accuracy. Performance was measured by computing, for each estimate $\hat{\lambda}$, the integrated squared error $\text{ISE} = \int (\hat{\lambda} - \lambda)^2$, and by noting the computation time.

7.2 Results

Figure 5 shows statistical performance of the estimators. Boxplots show the spread of ISE values for the diffusion estimates computed from 100 simulated realisations and are plotted against the diffusion kernel bandwidth τ indicated at the top of each panel. Curves show the mean integrated squared error (MISE) for the convolution method, computed by evaluating the analytic formulae for bias and variance in Section 6 and plotted against the convolution kernel bandwidth σ indicated at the bottom of each panel. We verified that these analytic formulae accurately predicted the sample means of ISE obtained in the simulations.

For the scenarios with uniform intensity, shown in the top row of Figure 5, MISE decreases with increasing bandwidth. This is expected, because the optimal estimate is constant, which is obtained as the bandwidth approaches infinity. The uniform correction estimator $\hat{\lambda}^U$ has the best empirical performance, as expected.

For the other three scenarios, performance is optimised at an intermediate value of the bandwidth. This is the familiar trade-off between bias and variance. For the convolution method, this trade-off can be quantified using the analytic formulae for bias and variance in Section 6.

In the Gaussian mixture and log-Gaussian Cox process scenarios, the competing methods have roughly equal performance overall. For large bandwidths, the diffusion method has better performance than the convolution method. Interestingly, this advantage disappears near the optimal bandwidth, where the two methods have roughly equal MISE. This suggests that, provided we do not greatly oversmooth or undersmooth, the convolution estimators are capable of performing as well as the diffusion method. The uniform correction has a slight advantage over the Jones–Diggle correction.

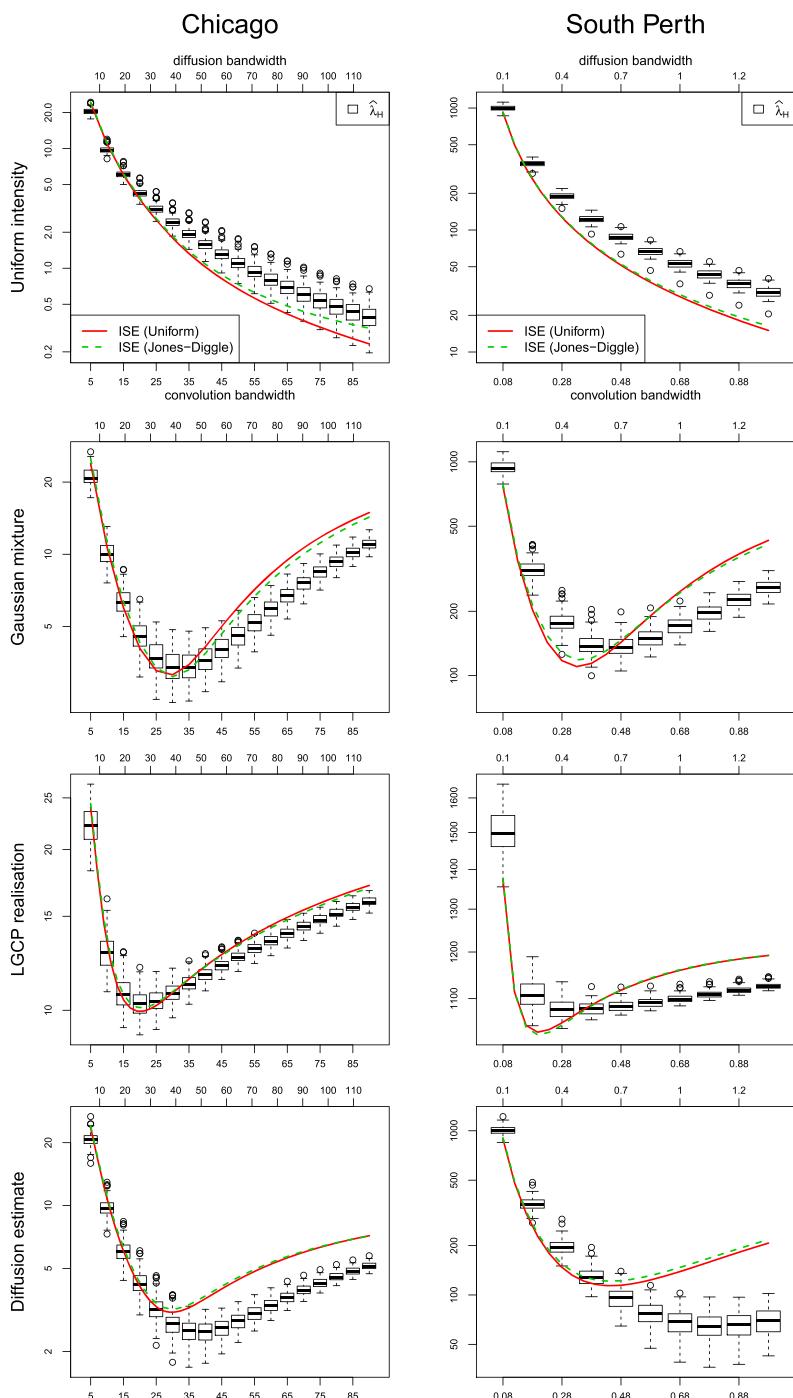


Figure 5. Integrated squared error (ISE) of the convolution and diffusion estimators applied to simulated data. Left column: Chicago network; right column: southern Perth network. Rows represent the four scenarios in Figure 4. The bottom horizontal axis in each panel shows the bandwidth σ of the convolution estimator; the top horizontal axis shows the bandwidth τ of the diffusion estimator. Boxplots show the numerically computed ISEs for the diffusion estimator. Lines show the theoretically calculated ISEs for $\hat{\lambda}^U$ (red, solid) and $\hat{\lambda}^{JD}$ (green, dashed). Bandwidths in feet for Chicago and in kilometres for southern Perth. LGCP, log-Gaussian Cox process. [Colour figure can be viewed at wileyonlinelibrary.com]

The diffusion estimate scenario shows results that clearly favour the diffusion estimator $\hat{\lambda}^H$ for large bandwidths. This is unsurprising because the behaviour of the diffusion estimate near a terminal vertex precisely matches the behaviour of the ‘true’ intensity surface in this synthetic problem. For nearly optimal bandwidths, there is no great difference in performance.

7.3 Computation Times

For the convolution method, execution times were roughly constant as a function of bandwidth and were roughly independent of the spatial pattern of points. Execution times were 0.06 s to produce a 128×128 pixel image for the Chicago network and 20 s for a $2\,048 \times 2\,048$ image for the southern Perth network.

In both networks, execution time for $\hat{\lambda}^H$ is comparable with that for $\hat{\lambda}^U$ and $\hat{\lambda}^{JD}$ for small to moderate bandwidths but rises steadily with increasing bandwidth, reaching 0.2 and 80 s in Chicago and southern Perth respectively at the largest bandwidth considered. Details are given in the Supporting Information.

8 Bandwidth Selection

The extensive literature on bandwidth selection in kernel density estimation, surveyed by Silverman (1986), Wand & Jones (1995) and Jones *et al.* (1996), suggests that we are unlikely to find a data-based method of bandwidth selection on a network that performs uniformly well. Rather, a variety of methods should be developed and deployed.

Classical univariate bandwidth selectors have been extended to the multivariate/spatial case (Sain *et al.*, 1994; Duong & Hazelton, 2003, 2005; Davies *et al.*, 2018, section 3). Many of these methods can be adapted to the setting of a linear network. They include likelihood cross-validation based on the Poisson point process likelihood

$$\text{cv}(\sigma) = \sum_{i=1}^n \log(\hat{\lambda}_\sigma^{-i}(x_i)) - \int_L \hat{\lambda}_\sigma(u) d_1 u, \quad (30)$$

where $\hat{\lambda}_\sigma(u)$ is either (10) or (11) and $\hat{\lambda}_\sigma^{(-i)}(x_i)$ is the corresponding ‘leave-one-out’ kernel estimate at x_i , defined by omitting the contribution from x_i from the sum in (10) or (11) accordingly. We could omit the integral term in (30) when considering the Jones–Diggle estimator (11) because the integral is constant and equal to the number of data points, ignoring errors of approximation. The leave-one-out estimates can easily be computed from the full estimates by subtracting the contribution associated with x_i , which in both cases is $\kappa_\sigma(0)/c_L(x_i)$. Here lies another advantage of the convolution method, because leave-one-out calculations with the diffusion estimate are much harder (McSwiggan *et al.*, 2016, section 7.1).

Simple rules of thumb for bandwidth selection are useful for many purposes, such as initial approximation and choosing the range of trial values σ for assessment by cross-validation. In d dimensions, there is Scott’s rule of thumb (Scott, 1992, p. 152) that the j -th Cartesian coordinate ($j = 1, \dots, d$) should be smoothed with bandwidth $\sigma_j = c_{n,d} s_j$, where s_j is the sample standard deviation of the j -th Cartesian coordinate values for the data locations, and $c_{n,d} = ((d+2)n)^{-1/(d+4)}$, where n is the number of data points. On a linear network, we propose a modification of Scott’s rule for $d = 1$,

$$\sigma = c_{n,1} \bar{s} = (3n)^{-1/5} \bar{s}, \quad (31)$$

where $\bar{s} = (s_1^2 + s_2^2)^{1/2}$ is the root sum of squares of the sample standard deviations. When the network is a single infinite line, the rule (31) yields the same value as Scott's rule in one dimension.

Terrell (1990) argues that automatic bandwidth selection should be avoided by inexperienced users in favour of an optimal 'oversmoothing' rule. A similar recommendation surely applies to network data.

9 Relative Risk and Smoothing on a Network

Here, we sketch how the convolution method can be generalised for the purpose of estimating spatially varying relative risk and for smoothing numerical responses observed at sample points on a network.

9.1 Weighted Kernel Estimators

Numerical weights w_i associated with the data points x_i can be incorporated in the kernel estimators (10)–(11) giving $\widehat{\lambda}^U(u) = c_L(u)^{-1} \sum_{i=1}^n w_i \kappa(u - x_i)$ and $\widehat{\lambda}^{JD}(u) = \sum_{i=1}^n w_i \kappa(u - x_i) / c_L(x_i)$ for $u \in L$. Weighted kernel estimators serve many purposes. The weight may represent multiplicity: for example, if w_i is the number of vehicles involved in the traffic accident that occurred at location x_i , then the weighted intensity estimate is the spatially varying number of vehicles involved in accidents per unit length. Weights may also be used to adjust for a denominator: if $w_i = 1/v_i$, where v_i is the traffic volume (average number of cars passing this location per hour), then the weighted intensity estimate is a spatially varying accident risk (accidents per car-km).

9.2 Relative Risk

In spatial relative risk estimation (Bithell, 1991; Kelsall & Diggle, 1995a, 1995b), data points are classified into different types, and we seek to estimate the spatially varying relative frequency of each type of point.

For simplicity, assume there are only two types of points, and we observe two point patterns \mathbf{x}, \mathbf{y} containing the points of the first and second types, respectively, on the same network L . The goal is to estimate the logarithmic relative risk $\rho(u) = \log(\lambda_X(u)/\lambda_Y(u))$, where $\lambda_X(u), \lambda_Y(u)$ are the intensity functions of the underlying point processes \mathbf{X}, \mathbf{Y} of points of each type.

A standard approach is to estimate ρ using a plug-in estimator based on kernel estimators of λ_X and λ_Y . It would be possible to use different smoothing bandwidths for the two point patterns, although there are theoretical arguments for using equal bandwidths (Kelsall & Diggle, 1995a). Recent research suggests that unequal bandwidths can cause undesirable 'halo' artefacts in the estimate of ρ (Davies *et al.*, 2016), at least in adaptive estimation. Hence, we recommend using the same bandwidth σ to estimate both λ_X and λ_Y . If the uniform correction (10) is used, then the plug-in estimate of ρ is

$$\widehat{\rho}^U(u) = \log \frac{\sum_i \kappa_\sigma(u - x_i)}{\sum_j \kappa_\sigma(u - y_j)}, \quad (32)$$

in which the edge correction factor $c_L(u)$ has cancelled out (Hazelton, 2008). This is equivalent to the usual uniform correction estimate of relative risk in 2D space, restricted to locations on the network. For the Jones–Diggle correction (11), the plug-in estimate of ρ is

$$\hat{\rho}^{\text{JD}}(u) = \log \frac{\sum_i \kappa_\sigma(u - x_i)/c_L(x_i)}{\sum_j \kappa_\sigma(u - y_j)/c_L(y_j)}. \quad (33)$$

Bandwidth selection for the optimal estimation of relative risk is different from bandwidth selection for intensity estimation. For spatial relative risk in two dimensions, several techniques for bandwidth selection are canvassed by Diggle (2003), Kelsall and Diggle (1995a), Diggle *et al.* (2005), Lawson and Williams (1993), Hazelton and Davies (2009) and Davies *et al.* (2018) and summarised in Davies *et al.* (2018, section 4.2). These can be adapted immediately to a linear network: we do not discuss them here.

For statistical inference, Hazelton and Davies (2009) introduced asymptotic p -value surfaces for relative risk. These could also be adapted to the network setting.

9.3 Spatial Smoothing

Suppose we observe numerical responses z_1, \dots, z_n at a set of sample locations x_1, \dots, x_n and seek to estimate the spatially varying mean response. In the Nadaraya–Watson approach to smooth regression (Nadaraya, 1964; Watson, 1964; Nadaraya, 1989), the estimate of the mean response $Z(u)$ at a location u is a weighted average of the observed responses z_i at nearby points x_i , weighted by the smoothing kernel. For our convolution method, the uniform and Jones–Diggle corrections give respectively

$$\hat{Z}^U(u) = \frac{\sum_i z_i \kappa(u - x_i)}{\sum_i \kappa(u - x_i)}, \quad (34)$$

$$\hat{Z}^{\text{JD}}(u) = \frac{\sum_i z_i \kappa(u - x_i)/c_L(x_i)}{\sum_i \kappa(u - x_i)/c_L(x_i)} \quad (35)$$

for locations $u \in L$. These estimates can be calculated rapidly using the techniques described in Section 4. Bandwidth selection for spatial smoothing can be based on minimising the sum of squared errors of interpolation $S = \sum_i [z_i - \hat{Z}^{-i}(x_i)]^2$, where $\hat{Z}^{-i}(x_i)$ is the leave-one-out estimate of $Z(x_i)$ obtained by omitting terms associated with x_i from both numerator and denominator of (34)–(35).

10 Traffic Accidents on Urban Roads of Medellín

Figure 6 shows the locations of traffic accidents in the urban area of Medellín, Colombia, in 2016, which were published in the OpenData portal of Medellín Town Hall. There are 10 764 points, classified by accident severity as ‘property damage’, ‘personal injury’ or ‘fatal’ accidents (comprising 4 627, 6 004 and 133 points, respectively). The three types of accident are displayed in separate panels in Figure 6. The road network has 54 164 road segments and total length 1 244 km.

We first estimated the intensity of each type of accident using the convolution estimate with the uniform correction (10). Results are shown in the Supporting Information. For these individual intensity estimates, bandwidths were selected using our extension (31) of Scott’s rule of thumb, yielding bandwidths of 0.36, 0.36 and 0.67 km for property damage, personal injury and fatal accidents, respectively. For comparison, likelihood cross-validation yielded bandwidths of 0.24, 0.24 and 1.07 km, respectively.

Figure 7 shows plug-in estimates of relative risk for different types of accidents, relative to accidents that caused only property damage. Because of the complexity of the network, the



Figure 6. Traffic accidents on the urban road network of Medellín in 2016. Left: property damage; middle: personal injury; and right: fatal. Map width 9.9 km.

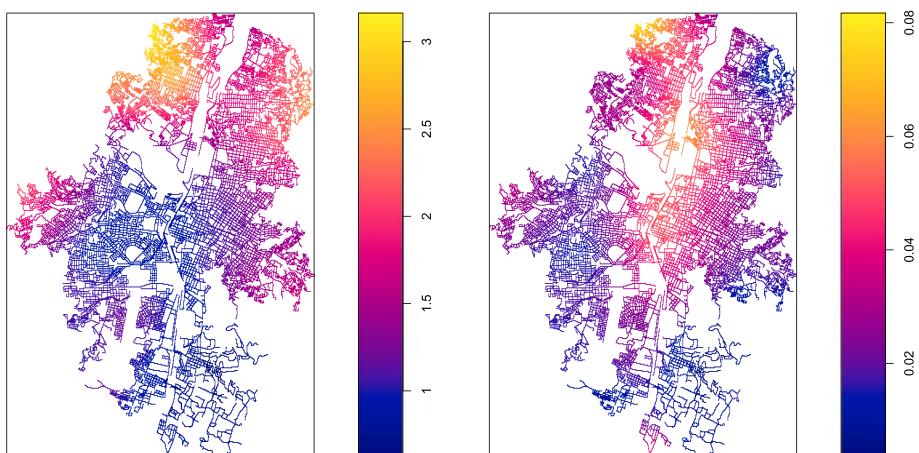


Figure 7. Estimated relative risk of different types of accidents, relative to accidents that caused only property damage. Left: personal injury; right: fatality. [Colour figure can be viewed at wileyonlinelibrary.com]

relative risk values are now rendered as colours rather than vertical heights. The estimates were computed as ratios of intensity estimates (10), using the same bandwidth σ for numerator and denominator. For estimation of relative risk, σ was selected using the weighted mean integrated squared error criterion of Hazelton (2008), available in the R package `sparr` (Davies *et al.*, 2018). This yielded bandwidths of 0.715 km for the relative risk of personal injury to property damage and 0.88 km for the relative risk of fatality to property damage. Each panel in Figure 7 is a 500×500 pixel image and required 20 s of computation, while 128×128 images need only 3 s.

The left panel of Figure 7 shows the relative risk of accidents involving personal injury relative to accidents involving property damage. The average relative risk is $6\ 004/4\ 627 \approx 1.3$. Estimated relative risks that are three times higher than average occur in the northern part of the urban area. The right panel shows the relative risk of fatal accidents to those involving property

damage (average relative risk $133/4\ 627 \approx 0.029$). Estimated relative risks almost three times higher than average occur in central and northern areas.

11 Adaptive Smoothing

The Western Australian road accident data shown in Figure 2 exhibit huge spatial variation in the concentration of data points. In such situations, it is well known that fixed-bandwidth kernel estimation can perform poorly. Dense concentrations of accidents in the urban areas will be oversmoothed, and sparse accidents in the remote desert will be undersmoothed. Adaptive (variable bandwidth) kernel estimation can perform substantially better in this context (Abramson, 1982; Hall & Marron, 1988).

A novel complication is that the network itself has huge spatial variation. This makes it more difficult to judge by eye the density of accidents per unit length of road. It also causes computational difficulties, because a much finer spatial resolution is required in urban areas than in the remote desert.

A pragmatic approach might be to divide the state of Western Australia into different subregions that are then analysed separately. We show one example in Section 12.3. To avoid artefacts at the boundaries between subregions, one could instead construct a spatially varying bandwidth function $\sigma(u)$, $u \in L$ and estimate the intensity by

$$\widehat{\lambda}^U(u) = \frac{1}{c_L(u, \sigma(u))} \sum_{i=1}^n \kappa_{\sigma(x_i)}(u - x_i), \quad u \in L, \quad (36)$$

analogous to the uniform correction (10), although it does not retain the same unbiasedness property. Estimators of the form (36) in 2D were proposed by Marshall and Hazelton (2010). An alternative is to allow each data point x_i to have its own smoothing bandwidth σ_i and estimate the intensity by

$$\widehat{\lambda}^{JD}(u) = \sum_{i=1}^n \frac{\kappa_{\sigma_i}(u - x_i)}{c_L(x_i, \sigma_i)}, \quad u \in L, \quad (37)$$

analogous to the Jones–Diggle correction (11), where $c_L(u, \sigma) = \int_L \kappa_\sigma(u - v) d_1 v$ for $u \in L$. This estimator conserves mass, that is, $\int_L \widehat{\lambda}^{JD}(u) d_1 u = n$.

Direct computation of either (36) or (37) would be very costly. For computational efficiency, we can follow the partitioning strategy of Davies and Baddeley (2018, section 4) in which continuously varying bandwidths σ are mapped to discretised bandwidths $\sigma^* = h(\sigma)$, where the function h takes only the values h_1, \dots, h_m with $m \ll n$.

For the Jones–Diggle style estimate (37), let $\sigma_i^* = h(\sigma_i)$ be the discretised bandwidth associated with data point x_i . Then $\widehat{\lambda}^{JD}(u)$ is approximated by $\widehat{\lambda}^{JD*}(u) = \sum_{i=1}^n \kappa_{\sigma_i^*}(u - x_i)/c_L(x_i, \sigma_i^*)$. This can be computed by dividing \mathbf{x} into subpatterns $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}$ according to the value of the discretised bandwidth, computing the fixed-bandwidth kernel estimate for each subpattern and summing:

$$\widehat{\lambda}^{JD*}(u) = \sum_{j=1}^m \sum_{i:\sigma_i^*=h_j} \frac{\kappa_{h_j}(u - x_i)}{c_L(x_i, h_j)} = \sum_{j=1}^m \widehat{\lambda}_{h_j}^{JD}(u | \mathbf{x}^{(j)}), \quad u \in L, \quad (38)$$

where $\widehat{\lambda}_\sigma^{JD}(u | \mathbf{y})$ is the Jones–Diggle corrected estimate (11) with bandwidth σ computed for point pattern \mathbf{y} .

For the uniform-style estimate (36), the discretised approximation to $\widehat{\lambda}^U(u)$ is $\widehat{\lambda}^{U*}(u) = c_L(u, h(\sigma(u)))^{-1} \sum_{i=1}^n \kappa_{\sigma^*(x_i)}(u - x_i)$. To compute this, we would again partition \mathbf{x} into subpatterns according to the value of $\sigma_i^* = h(\sigma(x_i))$, but evaluate the uncorrected fixed-bandwidth estimate for each subpattern

$$\widehat{\lambda}_{h_j}(u \mid \mathbf{x}^{(j)}) = \sum_{i:\sigma_i^*=h_j} \kappa_{h_j}(u - x_i), \quad u \in L, \quad (39)$$

then sum these estimates and normalise:

$$\widehat{\lambda}^{U*}(u) = \frac{1}{c_L(u, \sigma^*(u))} \sum_{j=1}^m \widehat{\lambda}_{h_j}(u \mid \mathbf{x}^{(j)}), \quad u \in L. \quad (40)$$

Total computation time for either (40) or (38) is about m times the cost of the fixed-bandwidth Jones corrected estimate.

The approximations (38) and (40) can produce artefacts associated with abrupt transitions in the value of $\sigma^*(u)$; these could be avoided by using the three-dimensional FFT technique proposed by Davies and Baddeley (2018), instead of partitioning the bandwidths.

It remains to choose the bandwidths σ_i associated with each data point or the bandwidth function $\sigma(u)$ for locations $u \in L$. Assume we have a pilot estimate of intensity $\widetilde{\lambda}(x_i)$ for each data point x_i , which could be any type of intensity estimate that can be calculated rapidly. The prescription of Abramson (1982) is to compute initial bandwidths inversely proportional to the square root of the estimated density,

$$a_i = \left(\frac{\widetilde{\lambda}(x_i)}{n} \right)^{-1/2}, \quad (41)$$

then to derive the smoothing bandwidths

$$\sigma_i = \sigma^\# \frac{a_i}{a}, \quad (42)$$

where $a = (\prod_i a_i)^{1/n}$ is the geometric mean of the initial bandwidths and $\sigma^\#$ is the global bandwidth. Adaptive bandwidth selection reduces to choosing the value of $\sigma^\#$.

Adaptive estimation of relative risk was developed by Davies and Hazelton (2010), including asymptotic tolerance contours. Davies *et al.* (2016) advocated the use of equal bandwidths in estimating the numerator and denominator of risk, extending Kelsall and Diggle (1995a) to the adaptive case. These results can be extended to linear networks.

12 Traffic Accidents in Western Australia

Figure 2 shows the spatial locations of 14 562 road traffic accidents recorded in 2011 in the state of Western Australia. The data were provided by the Western Australian state government Department of Main Roads and are made available for publication as part of the Western Australian Whole of Government Open Data Policy. The road network was simplified from 626 031 to 115 169 segments, with a total length of 97 165 km. Accident records with identical locations were removed because these typically arise from a single accident involving multiple vehicles.

12.1 Fixed-bandwidth Estimation

We first use the fixed-bandwidth estimators (10) and (11) to estimate the intensity of accidents on the network. Bandwidths were chosen using likelihood cross-validation (30), giving $\sigma = 9.1$ km for the uniform correction and 10.9 km for the Jones–Diggle correction. The modified Scott rule (31) gives $\sigma = 8.9$ km, which would have been acceptable.

Figure 8 plots the fixed-bandwidth intensity estimate using the Jones–Diggle correction (11) with bandwidth 10.9 km. A logarithmic colour scale is used in order to retain visual detail, because the intensity estimates vary by 20 orders of magnitude across the image. To avoid artefacts, we enforced a minimum intensity value of 10^{-5} accidents per km (corresponding to an average of one accident across the entire network). The result of the uniform correction (shown in the Supporting Information) was very similar. Both corrections show the effect of oversmoothing in the urban areas along the west coast and undersmoothing in the remote east. These are common features of fixed bandwidth estimates applied to spatially heterogeneous patterns (Loftsgaarden & Quesenberry, 1965; Breiman *et al.*, 1977; Davies & Baddeley, 2018).

12.2 Adaptive Bandwidth Estimation

It is evident from Figure 2 that accidents are highly concentrated in one area, the capital city, Perth. This contains more than 75% of the state's population, while the remote east has very low population density and traffic volume. Although the road network itself is also much denser in Perth than elsewhere, the estimated accident rate per kilometre shown in Figure 8 is still much higher in Perth. Fixed-bandwidth estimators are not suitable for analysing such spatially heterogeneous point patterns.

Figure 9 shows the adaptive estimate of intensity for the Western Australian road accidents using the Jones–Diggle style estimator (37). The pilot estimate of intensity was the fixed-bandwidth Jones–Diggle corrected convolution estimate (11) with bandwidth $\sigma = 10.9$ km shown in Figure 8. Adaptive bandwidths σ_i for each data point x_i were then derived using Abramson's rule (41)–(42) with global bandwidth $\sigma^\# = 10.9$ km. Values of σ_i ranged between

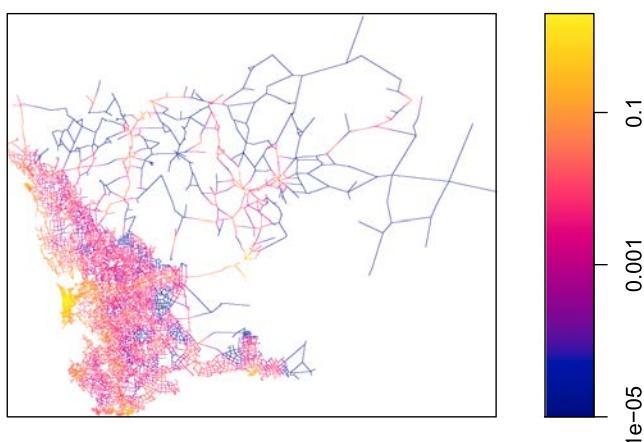


Figure 8. Fixed-bandwidth estimate of intensity for the accidents on the Western Australian road network using the Jones–Diggle correction with $\sigma = 10.9$ km. Logarithmic colour map. Intensity values are accidents per km. [Colour figure can be viewed at wileyonlinelibrary.com]

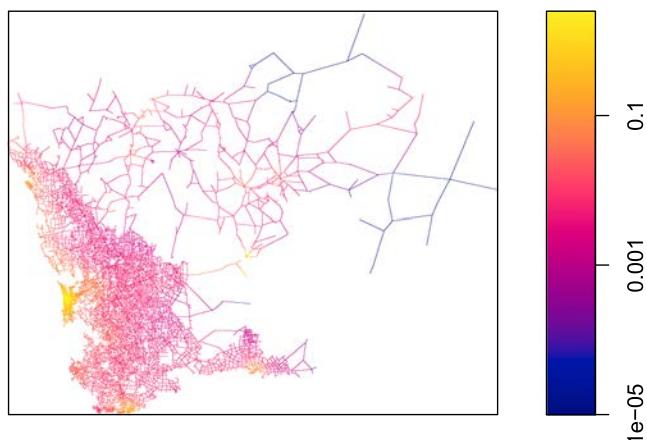


Figure 9. Adaptive bandwidth intensity estimate for the accidents on the Western Australian road network. Logarithmic colour map. Intensity values are accidents per km. [Colour figure can be viewed at wileyonlinelibrary.com]

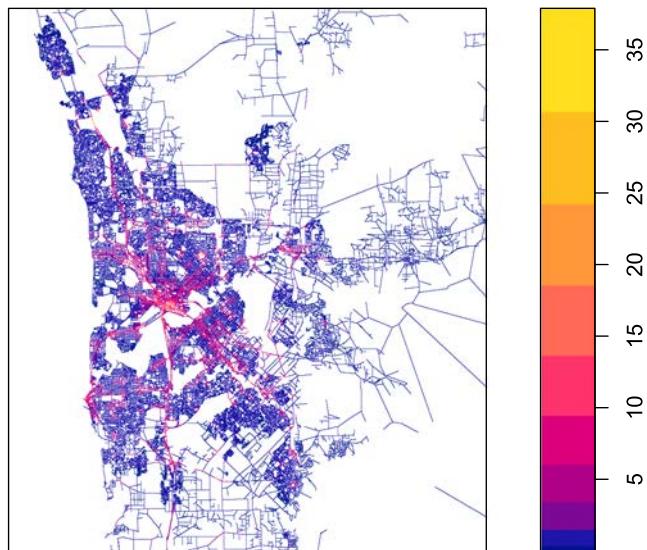


Figure 10. Adaptive bandwidth intensity estimate for the accidents in metropolitan Perth. Linear colour map with gamma-corrected colour sequence. Intensity values are accidents per km. Map is 60 km wide. [Colour figure can be viewed at wileyonlinelibrary.com]

7.5 and 171 km. The adaptive bandwidths σ_i were divided into 30 groups of equal size, and the adaptive intensity estimate was computed using (38).

The plot shows that the adaptive kernel-smoothed estimates have removed undersmoothing in the remote east, while retaining slightly more detail in the urban west.

12.3 Perth Metropolitan Area

Figure 10 shows an adaptive estimate of accident intensity for the metropolitan Perth area, a 60×67 -km subregion of Figure 2. The pilot estimate was a fixed-bandwidth Jones–Diggle estimate with bandwidth 0.35 km selected by cross-validation. Adaptive bandwidths ranged

from 0.1 to 1.1 km. The figure reveals the urban road system in exquisite detail, including the main freeways and their tributaries, the central business district and the coastal town of Fremantle. The Supporting Information shows even finer detail in the Central Business District (CBD).

12.4 Computation Times

The Supporting Information reports computation time for the Western Australian accident data. Using reasonable settings, fixed-bandwidth convolution estimates took approximately 12 s and did not depend on σ , as expected. For adaptive estimation, computation time ranged from 2 to 8 min, increasing with the number of discretised bandwidths m . As expected, computation time for the diffusion estimator increased dramatically with bandwidth τ and ranged from 12 s to 6 h.

Timings depended crucially on algorithm parameters, including the spatial resolution of the resulting pixel image (pixel width ϵ), the step-size δ of the sample points along the network for the diffusion algorithm and the number m of discretised bandwidths in the adaptive estimator. For the diffusion algorithm, computation time is roughly proportional to δ^{-3} for small δ . Our implementation of the diffusion algorithm chooses $\delta = \epsilon$ by default, and for this choice, the diffusion estimator is as fast as the convolution estimators. However, this causes substantial bias in the diffusion estimates, because the Western Australian network contains many short urban road segments and the diffusion algorithm effectively rounds these upward to a multiple of δ . A reasonably accurate estimate is obtained only when $\delta \leq 0.05$ km, leading to computation times between 1 and 6 h.

12.5 Findings

Intensity estimates for the Western Australian road accident data range from 0 to 40 accidents per km. The highest intensity values occur in the main conurbations and along major highways.

The intensity of accidents is a measure of accident frequency (number of accidents per km) rather than risk (probability of an accident occurring to a given vehicle at this location). The intensity estimates may be useful for planning emergency response but need further analysis if they are to be used for road safety research. Estimation of accident risk would require a denominator such as the traffic volume (number of vehicles passing the given location per hour) and explanatory variables such as speed limit.

The choice of graphical display method for the intensity estimate was quite influential. The intensity values could have been represented as vertical heights (Figure 3), colours (with different kinds of colour map shown in Figures 7–10) or line thicknesses (see the Supporting Information). Borruso (2008) uses perspective views of a surface obtained by extrapolating the intensity function onto the 2D plane. The 2D convolution approach is well suited to this kind of visualisation, and we showcase such an interactive three-dimensional graphic of fixed and adaptive intensity estimates of the WA road accident data focusing on observations in the Perth Central Business District at http://stats.otago.ac.nz/~tdavies/wacbd_hen.html (for further details on this graphic, see the Supporting Information). Naturally, the best choice regarding visualisation depends on the complexity of the network and the variability in the estimated intensity values.

Instead of removing duplicate points, an alternative would have been to count the number of vehicles involved in each accident and retain this as an attribute of the accident location. This would have enabled separate study of single-vehicle and multiple-vehicle accidents. The number of vehicles could also serve as a weight associated with each location: the weighted kernel estimator would provide the spatially varying number of vehicles involved in accidents per unit length, as discussed in Section 9.1.

13 Discussion

The convolution method introduced in this paper has many advantages. It is very fast to compute for any value of bandwidth. There is an exact variance formula and a variance estimator that can easily be computed. The calculation of leave-one-out estimates is easy. Estimation can easily be confined to a subregion of space.

Compared with other techniques that depend on path lengths in the network, the convolution method is much less sensitive to changes or errors in the connectivity of the network. This may be either advantage or disadvantage in different contexts. It may be considered inappropriate to allow mass to ‘tunnel’ between parts of the network that are not interconnected; a similar issue arises in 2D density estimation (Barry & McIntyre, 2011). However, this property also makes it much easier to estimate spatio-temporal variation when the network itself is changing over time: for example, urban road networks are continually being modified.

The main disadvantage of the convolution method is its slightly suboptimal statistical performance. There is also an inherent limit on spatial resolution determined by computer memory limits for the pixel grid. However, this can be overcome by dividing the spatial domain into tiles and computing the estimate on one tile at a time.

Simple rules of thumb for bandwidth selection, such as (31), performed well, and we recommend them for exploratory analysis.

Warnings for practitioners include sensitivity to the choice of spatial resolution, the possibility of bias due to coarse resolution and the influence of the choice of graphical display method (see Sections 4, 12.4 and 12.5).

An artefact of all available techniques for kernel smoothing on a network is that, for example, a high accident rate on a busy road will spread unrealistically onto the adjacent quiet roads. Remedies include segmenting the network, treating different classes of roads differently or accounting for traffic volume, all of which are forms of semi-parametric modelling, amenable to the methods of O’Donnell *et al.* (2014).

Acknowledgements

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 the Australian Research Council (Baddeley, Nair, and Rakshit: discovery grants DP 130102322 and DP 130104470); the European Union GEO-C project <http://www.geo-c.eu/> (Moradi: H2020-MSCA-ITN-2014, grant agreement 642332), the Royal Society of New Zealand, Marsden Fund (Davies: Fast Start grant 15-UOO-092); Grains Research and Development Corporation, Australia (Rakshit: SAGI-3 project); and the Spanish Ministry of Science and Education (Mateu: grant MTM2016-78917-R). The Medellín data were provided by Alcaldía de Medellín; we thank Fernando Benitez for fruitful discussion.

Calculations were performed in the R language using the packages `spatstat` 1.53-2 (Baddeley & Turner, 2005; Baddeley *et al.*, 2015) and `sparr` 2.1-13 (Davies *et al.*, 2018) and the C library `fftw` 3.3.7 (Frigo & Johnson, 2005; Rahim, 2017). Our code is provided as a Supporting Information to this article and will be released in `spatstat`.

References

- Abramson, I. (1982). On bandwidth estimation in kernel estimates—a square root law. *Annals. Stat.*, **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. *Scand. J. Stat.*, **39**, 591–617.

- Baddeley, A., Nair, G., Rakshit, S. & McSwiggan, G. (2017). ‘Stationary’ point processes are uncommon on linear networks. *STAT*, **6**(1), 68–78.
- Baddeley, A., Rubak, E. & Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman and Hall/CRC.
- Baddeley, A. & Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *J. Stat. Softw.*, **12**(6), 1–4. URL www.jstatsoft.org, ISSN: 1548–7660.
- Barry, R. & McIntyre, J. (2011). Estimating animal densities and home range in regions with irregular boundaries and holes: a lattice-based alternative to the kernel density estimator. *Ecol. Model.*, **222**, 1666–1672.
- Bitthell, J. (1990). An application of density estimation to geographical epidemiology. *Stat. Med.*, **9**, 691–701.
- Bitthell, J. (1991). Estimation of relative risk functions. *Stat. Med.*, **10**, 1745–1751.
- Borruso, G. (2003). Network density and the delimitation of urban areas. *Trans. GIS*, **7**, 177–191.
- Borruso, G. (2005). Network density estimation: analysis of point patterns over a network. In *Computational Science and its Applications — ICCSA 2005*, Eds. Gervasi, O., Gavrilova, M., Kumar, V., Laganà, A., Lee, H., Mun, Y., Taniar, D. & Tan, C. Berlin/Heidelberg: Springer, Number 3482 in Lecture Notes in Computer Science, pp. 126–132.
- Borruso, G. (2008). Network density estimation: a GIS approach for analysing point patterns in a network space. *Trans. GIS*, **12**, 377–402.
- Botev, Z., Grotowski, J. & Kroese, D. (2010). Kernel density estimation via diffusion. *Ann. Stat.*, **38**(5), 2916–2957.
- Breiman, L., Meisel, W. & Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, **19**, 135–144.
- Chaudhuri, P. & Marron, J. (2000). Scale space view of curve estimation. *Ann. Stat.*, **28**, 408–428.
- Daley, D. & Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods*, 2nd ed. New York: Springer-Verlag.
- Davies, T. & Baddeley, A. (2018). Fast computation of spatially adaptive kernel estimates. *Stat. Comput.*, **28**, 937–956.
- Davies, T. & Hazelton, M. (2010). Adaptive kernel estimation of spatial relative risk. *Stat. Med.*, **29**, 2423–2437.
- Davies, T., Jones, K. & Hazelton, M. (2016). Symmetric adaptive smoothing regimens for estimation of the spatial relative risk function. *Comput. Stat. Data Anal.*, **101**, 12–28.
- Davies, T., Marshall, J. & Hazelton, M. (2018). Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk. *Stat. Med.*, **37**, 1191–1221.
- Diggle, P. (1985). A kernel method for smoothing point process data. *J. R. Stat. Soc.: Ser. C: Appl. Stat.*, **34**, 138–147.
- Diggle, P. (2003). *Statistical Analysis of Spatial Point Patterns*, 2nd ed. London: Hodder Arnold.
- Diggle, P. & Marron, J. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. *J. Am. Stat. Assoc.*, **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. *Appl. Stat.*, **54**, 645–658.
- Downs, J. & Horner, M. (2007a). Characterising linear point patterns. In *Proceedings of the GIScience Research UK Conference (GISRUK)*, Ed. Winstanley, A. Maynooth, Ireland, County Kildare, Ireland: National University of Ireland Maynooth, 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, pp. 3–5.
- Downs, J. & Horner, M. (2008). Spatially modelling pathways of migratory birds for nature reserve site selection. *Int. J. Geogr. Inf. Sci.*, **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). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *J. Multivar. Anal.*, **93**, 417–433.
- Frigo, M. & Johnson, S. G. (2005). The design and implementation of FFTW3. *Proc. IEEE*, **93**(2), 216–231. Special issue on “Program Generation, Optimization, and Platform Adaptation”.
- Hall, P. & Marron, J. (1988). Variable window width kernel estimation of probability densities. *Probab. Theory Relat. Fields*, **80**, 37–49.
- Hazelton, M. (2008). Kernel estimation of risk surfaces without the need for edge correction. *Stat. Med.*, **27**(12), 2269–2272.
- Hazelton, M. & Davies, T. (2009). Inference based on kernel estimates of the relative risk function in geographical epidemiology. *Biom. J.*, **51**, 98–109.
- Isaak, D., Peterson, E., Ver Hoef, J. *et al.* (2014). Applications of spatial statistical network models to stream data. *Wiley Interdiscip. Rev. Water*, **1**(3), 277–294.
- Jones, M. (1993). Simple boundary corrections for kernel density estimation. *Stat. Comput.*, **3**, 135–146.

- Jones, M., Marron, J. & Sheather, S. (1996). A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc.*, **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. *Stat. Med.*, **14**, 2335–2342.
- Lawson, A. & Williams, F. (1993). Applications of extraction mapping in environmental epidemiology. *Stat. Med.*, **12**, 1249–1258.
- Loftsgaarden, D. O. & Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.*, **36**, 1049–1051.
- Marshall, J. & Hazelton, M. (2010). Boundary kernels for adaptive density estimators on regions with irregular boundaries. *J. Multivar. Anal.*, **101**, 949–963.
- McSwiggan, G., Baddeley, A. & Nair, G. (2016). Kernel density estimation on a linear network. *Scand. J. Stat.*, **44**(2), 324–345.
- Moradi, M. M., Rodríguez-Cortés, F. & Mateu, J. (2017). On kernel-based intensity estimation of spatial point patterns on linear networks. *J. Comput. Graph. Stat.*, **27**(2), 302–311. <https://doi.org/10.1080/10618600.2017.1360782>.
- Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**, 141–142.
- Nadaraya, E. (1989). *Nonparametric Estimation of Probability Densities and Regression Curves*. Dordrecht: Kluwer.
- O'Donnell, D., Rushworth, A., Bowman, A. & Scott, E. (2014). Flexible regression models over river networks. *Applied Statistics (Journal of the Royal Statistical Society, Series C)*, **63**(1), 47–63.
- Okabe, A., Satoh, T. & Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *Int. J. Geogr. Inf. Sci.*, **23**, 7–32.
- Okabe, A. & Sugihara, K. (2012). *Spatial Analysis along Networks*. New York: John Wiley and Sons.
- Pinsky, M. (2002). *Introduction to Fourier Series and Wavelets. Number 102 in Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society.
- Rahim, K. (2017). *fftwtools package. R package version 0.9-8*.
- Rakshit, S., Nair, G. & Baddeley, A. (2017). Second-order analysis of point patterns on a network using any distance metric. *Spatial Statistics*, **22**(1), 129–154.
- Rüschorf, L. (1994). Wasserstein metric. In *Encyclopedia of Mathematics*, Ed. Hazewinkel, M., pp. 487–488. Dordrecht, NL: Kluwer.
- Sain, S., Baggerly, K. & Scott, D. (1994). Cross-validation of multivariate densities. *J. Am. Stat. Assoc.*, **89**, 807–817.
- Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: John Wiley and Sons.
- Silverman, B. (1982). Kernel density estimation using the fast Fourier transform. *Appl. Stat.*, **31**, 93–99.
- Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- 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)*, pp. 827–832. Tokyo: IEEE.
- Terrell, G. (1990). The maximal smoothing principle in density estimation. *J. Am. Stat. Assoc.*, **85**, 470–476.
- Tierney, L. (2001). Compiling R: a preliminary report. In *DSC 2001: Proceedings of the Second International Workshop on Distributed Statistical Computing*, Eds. Hornik, K. & Leisch, F. Vienna, Austria: Technische Universität Wien.
- Wand, M. & Jones, M. (1995). *Kernel Smoothing*. Boca Raton, FL: Chapman and Hall.
- Watson, G. (1964). Smooth regression analysis. *Sankhya A*, **26**, 359–372.
- Xie, Z. & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Comput. Environ. Urban. Syst.*, **32**, 396–406.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

[Received May 2018, accepted April 2019]