**Citation**
Baddeley, A. and Davies, T.M. and Rakshit, S. and Nair, G. and McSwiggan, G. 2022. Diffusion Smoothing for Spatial Point Patterns. Statistical Science. 37 (1): pp. 123-142.
http://doi.org/10.1214/21-STS825

Submitted to Statistical Science

# Diffusion Smoothing for Spatial Point Patterns

**Adrian Baddeley, Tilman M. Davies, Suman Rakshit, Gopalan Nair and Greg McSwiggan**

*Abstract.* Traditional kernel methods for estimating the spatially-varying density of points in a spatial point pattern may exhibit unrealistic artefacts, in addition to the familiar problems of bias and over- or under-smoothing. Performance can be improved by using diffusion smoothing, in which the smoothing kernel is the heat kernel on the spatial domain. This paper develops diffusion smoothing into a practical statistical methodology for two-dimensional spatial point pattern data. We clarify the advantages and disadvantages of diffusion smoothing over Gaussian kernel smoothing. Adaptive smoothing, where the smoothing bandwidth is spatially-varying, can be performed by adopting a spatially-varying diffusion rate: this avoids technical problems with adaptive Gaussian smoothing and has substantially better performance. We introduce a new form of adaptive smoothing using lagged arrival times, which has good performance and improved robustness. Applications in archaeology and epidemiology are demonstrated. The methods are implemented in open-source R code.

*AMS 2000 subject classifications:* Primary 62G07; secondary 62M30.
*Key words and phrases:* adaptive smoothing, bandwidth, heat kernel, kernel estimation, lagged arrival method, Richardson extrapolation.

*School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, GPO Box U1987, Perth WA 6845, Australia. (e-mail: Adrian.Baddeley@curtin.edu.au; Suman.Rakshit@curtin.edu.au). Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin 9054, New Zealand. (e-mail: tilman.davies@otago.ac.nz). Department of Mathematics and Statistics, University of Western Australia, 35 Stirling Hwy, Nedlands WA 6009, Australia. (e-mail: gopalan.nair@uwa.edu.au; qfengineers@gmail.com).*

## 1. INTRODUCTION

In the statistical analysis of spatial point pattern data [31, 43, 3], an important task is to estimate the spatially-varying density or occurrence rate of points. The standard nonparametric method is kernel estimation [30, 10, 63], typically using a Gaussian kernel. This paper develops an alternative methodology based on *diffusion smoothing* [18, 12, 6] in which the observed distribution of data points is smoothed by imitating the physical process of diffusion.

Diffusion smoothing resolves several important problems encountered in Gaussian kernel smoothing. These include failure to conserve mass; bias near the boundary of the spatial domain [30, 10, 44, 50]; the need for adaptive smoothing to avoid simultaneous over- and under-smoothing artefacts [1, 41, 70]; and physically impossible or implausible results [6].

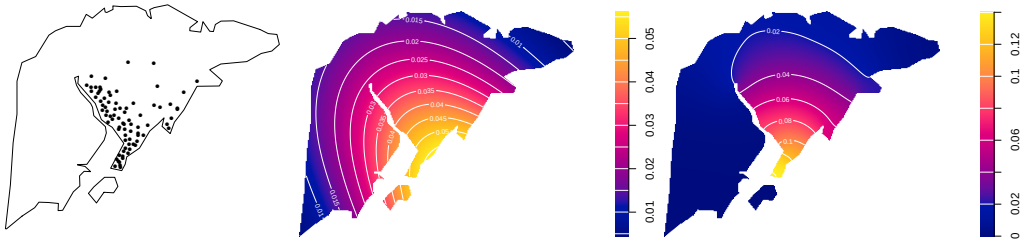

FIG 1. Left: *a synthetic pattern of points in a polygonal region.* Middle: *Gaussian kernel estimate of intensity (with uniform edge correction), showing tunnelling of mass.* Right: *diffusion estimate of intensity with equivalent bandwidth.*

The latter problem is illustrated in Figure 1. In the left panel is a synthetic point pattern dataset in an irregular domain $W$, extracted from real coastline data. The middle panel is a Gaussian kernel estimate of intensity. Although all the data points lie in the upper right half of $W$, the Gaussian kernel estimate has quite large values on the lower left half as well: about 36% of the mass has been transferred to the lower left half. Kernel mass has effectively "leapt" or "tunnelled" across the gulf between the two halves. Indeed about 2.5% of mass has been transferred to the small island where there were no data points. This behaviour would be unrealistic if the points represent terrestrial animals [6].

The right panel of Figure 1 shows the *diffusion* estimate of intensity for the same point pattern using an equivalent bandwidth. Very little mass (only about 5%) has migrated to the lower left half of the domain, and no mass has migrated to the island. This estimate would be much more realistic for terrestrial wildlife monitoring.

Use of the physical process of diffusion as a paradigm for smoothing is already well-developed in computer vision [71, 20] where it is inherent in the "scale space" approach [46, 47]. The benefits of this approach to statistical curve estimation were set out by Chaudhuri and Marron [18]. Otherwise the statistical literature on kernel smoothing does not frequently mention diffusion kernels; but the "reflected-kernel correction" for density estimation on the positive half-line ([42, 59, 32, 40], [63, p. 26]) is a special case of a diffusion kernel.

Botev *et al.* [12] developed formal statistical theory for diffusion kernel es-

timation of a probability density, mainly in one dimension. They argued that diffusion smoothing is an intrinsic solution to the problems of density estimation in a bounded domain. Edge corrections are unnecessary because the heat kernel is inherently tailored to the domain. They also showed that *adaptive* estimation, in which the amount of smoothing is spatially-varying, can be performed intrinsically by using a diffusion with spatially-varying speed and drift. They demonstrated that adaptive diffusion smoothing has substantially better performance than adaptive Gaussian smoothing.

Independently, Barry and McIntyre [6] drew attention to the "tunnelling" artefact illustrated in the middle panel of Figure 1, and proposed an estimator of point process intensity using random walks on a lattice. Their procedure can be regarded as a discrete approximation to a fixed-bandwidth diffusion smoother.

In this paper we reconcile and extend these approaches, with the goal of developing diffusion smoothing into a practical methodology for spatial point pattern data. We elucidate the correspondence between the "lattice-based" smoother of [6] and the continuous diffusion smoother of [12] — which is required, for example, to ensure that the two density estimates in Figure 1 are obtained using equivalent bandwidths. These results also show how to generalise the computational algorithm of [6] to non-square rectangular and hexagonal grids. Detailed algorithms are given, and implemented in open-source software. We study the sample behaviour of the diffusion estimate, including its behaviour at the boundary of the study region, and statistical properties of the diffusion estimator, and compare them with the Gaussian kernel estimator.

Corrections for boundary effects become more important in higher dimensions. The diffusion estimator does not require edge correction because it is intrinsically tailored to the study region. Indeed it is *both* unbiased for the uniform density *and* mass-conserving, whereas the Gaussian kernel estimator requires edge correction to satisfy either of these properties, and cannot satisfy them both simultaneously.

Spatial point pattern data sometimes exhibit very strong heterogeneity, for which fixed-bandwidth smoothing is profoundly unsatisfactory [1, 63, 41, 69, 26]. We develop *spatially adaptive* versions of diffusion smoothing. An elegant approach, advocated by Botev *et al.* [12], is to allow the diffusion parameters to vary over the spatial domain; we develop the special case of "variable-rate diffusion", where the diffusion has spatially-varying speed and zero drift. We also propose a new form of adaptive diffusion smoothing, "lagged-arrival diffusion", in which the diffusion process parameters are constant across the domain, but the data points enter the diffusion at different starting times, according to the smoothing bandwidths assigned to them. Lagged-arrival diffusion smoothing corresponds closely to sample-point-adaptive Gaussian smoothing. In both the variable-rate and lagged-arrival methods, the spatially-varying smoothing bandwidth must be determined by a suitable rule, such as the square-root rule of Abramson [1] based on a reference density such as a pilot estimate of the density. We find that misspecification of the reference density has far less effect on the lagged-arrival estimate than on the variable-rate estimate. The lagged-arrival method may be useful even for one-dimensional density estimation.

We complete our investigation by implementing and applying diffusion smoothing in practical contexts. Performance is evaluated using simulations, real-data examples, and special cases. Bandwidth selection is discussed. We measure the

discretisation error in the "lattice" approximation and show how it can be vastly reduced using numerical techniques. All the algorithms developed here are implemented in open-source R software in the package `spatstat` [4, 5].

The article is structured as follows. In Section 2 we provide brief background on kernel smoothing estimators of spatial intensity functions. We define the fixed-bandwidth diffusion estimator in Section 3, stating its key properties. Section 4 explains the discrete approximation to the diffusion: this is useful for exposition, necessary for practical implementation (including the ability to accommodate different grid geometries), and reconciles the estimators developed independently in the literature. Section 5 applies the diffusion estimator to a novel dataset of ancient Māori sites in New Zealand. Adaptive diffusion is tackled in Section 6, beginning with the theory of [12] for the variable-rate estimator, and subsequently introducing the new lagged-arrival estimator. Section 7 applies the adaptive diffusion estimators to an epidemiological dataset. Section 8 reports on a simulation study evaluating the diffusion estimators and bandwidth selection methods. We end with a discussion and commentary on future research avenues in Section 9.

## 2. BACKGROUND

The methods described here can be applied in $d$-dimensional Euclidean space $\mathbb{R}^d$ for any $d \geq 1$, but for simplicity we consider only the two-dimensional plane $\mathbb{R}^2$. A typical spatial location in $\mathbb{R}^2$ will be denoted by a single letter $x$. We will usually avoid explicit mention of spatial coordinates, but when necessary, the Cartesian coordinates will be denoted $u$ and $v$, so that $x = (u, v)$.

The observed data consist of a spatial point pattern $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ in a spatial domain $W \subset \mathbb{R}^2$, where $n \geq 0$ is not fixed in advance, and $x_i \in W$ for $i = 1, \ldots, n$. We regard $\boldsymbol{x}$ as a realisation of a spatial point process $\boldsymbol{X}$ in $W$ assumed to have an *intensity function* $\lambda(x)$, $x \in W$, defined so that the number $N(B) = n(\boldsymbol{X} \cap B)$ of points falling in any given Borel set $B \subseteq W$ has expectation $\mathbb{E}[N(B)] = \int_B \lambda(x)\,\mathrm{d}x$. See [24, 25] or [31, 43, 3]. An important and often-used model for $\boldsymbol{X}$ is a Poisson process, implying that $N(B)$ has a Poisson distribution and that given $N(W) = n$, the locations of the $n$ points in $W$ are independent and identically distributed with probability density $f(x) = \lambda(x)/\Lambda$, where $\Lambda = \int_W \lambda(x)\,\mathrm{d}x$. Estimation of the intensity function $\lambda(x)$ is effectively equivalent to density estimation.

Kernel estimators of $f(x)$ or $\lambda(x)$ for two-dimensional spatial point patterns were described in [30, 10, 63]. The "fixed-bandwidth" kernel estimator of intensity is

$$\widehat{\lambda}_\sigma(x) = \sum_{i=1}^{n} k_\sigma(x - x_i), \quad x \in W, \tag{1}$$

where $\sigma > 0$ is the smoothing bandwidth, $k_\sigma(x) = \sigma^{-2}k(x/\sigma)$ is the kernel with bandwidth $\sigma$, and $k(x)$ is the template kernel, a probability density on $\mathbb{R}^2$, often taken to be the bivariate standard normal density. This estimator is biased because of edge effects. Bias-corrected estimators include the "uniform" correction [30]

$$\widehat{\lambda}^{(U)}_\sigma(x) = \frac{1}{c_W(\sigma, x)} \sum_{i=1}^{n} k_\sigma(x - x_i), \quad x \in W, \tag{2}$$

and the Jones or "Jones-Diggle" correction [44]

$$\widehat{\lambda}^{(J)}_\sigma(x) = \sum_{i=1}^{n} \frac{k_\sigma(x - x_i)}{c_W(\sigma, x_i)}, \quad x \in W, \tag{3}$$

where in both cases $c_W(\sigma, x) = \int_W k_\sigma(y - x)\,\mathrm{d}y$ is the mass of the kernel centred at $x$ which lies within $W$. The Jones-Diggle correction preserves total mass, $\int_W \widehat{\lambda}^{(J)}_\sigma(x)\,\mathrm{d}x = n$, while the uniform correction is unbiased for the uniform density, $\mathbb{E}[\widehat{\lambda}^{(U)}_\sigma(x)] = \lambda$ if the true intensity is constant $\lambda(x) \equiv \lambda$. The corrections (2) and (3) are workable solutions to edge effect bias, but are not entirely satisfactory, because both properties cannot be satisfied at the same time, and because such corrections also inflate the estimator variance.

Variances of these estimators can be calculated using point process methods. For a Poisson process with intensity function $\lambda(x)$, the variance of $\sum_{i=1}^n g(x_i)$ is $\int g(x)^2\lambda(x)\,\mathrm{d}x$ (by first principles; see [24, p. 188] or [3, p. 173]). Calculations for a special case are given in Appendix B. For a general, non-Poisson point process there is an explicit formula for the estimator variance in terms of the first and second moment intensities (cf. Lemma 6 below).

Adaptive estimation, where the amount of smoothing is spatially-varying, is discussed in Section 6.1.

## 3. DIFFUSION ESTIMATORS (FIXED-BANDWIDTH)

This section provides a formal definition of the diffusion estimator of point process intensity, and states its main properties.

The key fact motivating the definition is that the Gaussian kernel satisfies the Fourier heat equation on the infinite plane $\mathbb{R}^2$. This suggests that, when data are observed in a bounded window $W$, we should replace the Gaussian kernel by a solution of the heat equation in $W$ [18, 12].

### 3.1 The heat kernel

Unless otherwise stated, the domain $W \subset \mathbb{R}^2$ is assumed to be a regular compact set, that is, $W$ is bounded and is the closure of its interior. For simplicity, we also assume $W$ has piecewise-differentiable boundary $\partial W$, although more irregular boundaries can be permitted.

The heat kernel can be defined as the transition probability density of a Brownian motion on $W$ with reflecting boundary [39, 67]. That is, for each $y \in W$, the function $\kappa_t(\cdot \mid y)$ is the probability density of the location $B_t$ at time $t$ of a standard Brownian motion in $W$, with reflecting boundary at $\partial W$, started at position $B_0 = y$. Here it is useful to give an equivalent definition in the language of differential equations.

**Definition 1 (Heat Conduction Problem)** *Suppose $g$ is a real-valued function on $W$. The* heat conduction problem *on $W$ with initial condition $g$ is the problem of finding a solution $f_t(x), t \geq 0, x \in W$ to*

1. *the classical Fourier time-dependent* heat equation

$$\frac{\partial}{\partial t} f_t(x) = \frac{1}{2}\,\nabla^2 f_t(x), \quad t > 0, \tag{4}$$

*at every location $x$ in the interior of $W$, where $\nabla^2 = \partial^2/\partial u^2 + \partial^2/\partial v^2$ is the Laplacian operator with respect to coordinates $x = (u, v)$;*

2. *the* Neumann boundary condition

$$(\nabla f_t(x)) \cdot \nu(x) = 0, \qquad x \in \partial W, \quad t > 0 \tag{5}$$

*at each boundary location $x \in \partial W$, where $\nabla = (\partial/\partial u, \partial/\partial v)$ is the gradient operator and $\nu(x)$ is the normal vector to the boundary of $W$ at $x$; and*

3. *the initial condition $f_0(x) = g(x)$ for all $x \in W$.*

*Here it is required that $f_t(x)$ be differentiable with respect to $t$ and twice-differentiable with respect to $x$, for $t > 0$.*

The heat kernel can now be defined as the Green's function for the heat conduction problem. That is,

**Theorem 1** *The solution $f_t(x)$ of the heat conduction problem on $W$ with any initial condition $g$ can be expressed as*

$$f_t(x) = \int_W \kappa_t(x \mid y) g(y) \, \mathrm{d}y \tag{6}$$

*where $\kappa_t(x \mid y)$, $t \geq 0$, $x, y \in W$, is a unique function called the* heat kernel *on $W$, which is differentiable with respect to $t$ and twice-differentiable with respect to $x$ and $y$ for $t > 0$.*

See [13, Chap. 10]. By the principle of superposition we can regard the heat kernel $\kappa_t(x \mid y)$ as the impulse response, that is, the unique solution of the heat conduction problem with initial condition $f_0(x) = \delta(x - y)$, where $\delta$ is the Dirac delta function.

A simple analytic expression for the heat kernel on $W$ is not available in general. Infinite series expansions are available [13, Chap. 10] and they are computable for simple shapes such as rectangles, studied in Appendix A. For spatial domains of general shape, the heat kernel can be evaluated by numerically solving the heat equation, and we shall follow this approach in Section 4.

### 3.2 The diffusion estimator

The diffusion estimator of intensity can now be defined, in the case of a fixed smoothing bandwidth.

**Definition 2** *Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a point pattern in $W$. For any location $x \in W$, the (fixed-bandwidth)* diffusion estimate *of the intensity function $\lambda(x)$, with bandwidth $\sigma$, is*

$$\widehat{\lambda}_t(x) = \sum_{i=1}^{n} \kappa_t(x \mid x_i), \tag{7}$$

*where $t = \sigma^2$, and $\kappa_t$ is the heat kernel defined in Theorem 1.*

The diffusion estimator (7) is the solution at time $t$ of the heat conduction problem with initial condition $f_0(x) = \sum_{i=1}^{n} \delta(x - x_i)$, corresponding to unit masses at the data points.

It is appropriate to define the bandwidth as $\sigma = \sqrt{t}$, because the heat kernel $\kappa_t$ is closely connected to the isotropic Gaussian density $\varphi_\sigma$ with standard deviation

$\sigma = \sqrt{t}$. If $W$ is the infinite plane, then $\kappa_t(x \mid y) = \varphi_\sigma(x-y)$. If $W$ is a rectangle, then $\kappa_t(x \mid y)$ is an infinite sum of terms of the form $\varphi_\sigma(x - y + z)$ given in Appendix A.

In this paper we focus on estimation of the intensity $\lambda(x)$, but the estimator (7) divided by $n$ is precisely the diffusion estimator of probability density defined by Botev *et al.* [12]. It is also closely related to the random walk estimate of intensity described by Barry and McIntyre [6], as we elucidate in Section 4.1.

Note especially that it is not necessary to introduce edge corrections for the diffusion estimate, in contrast to the situation for Gaussian kernel estimates. We show below that edge correction is "intrinsic" to the diffusion estimate.

If the data points $x_i$ have weights $w_i \in \mathbb{R}$, the weighted version of (7) is $\widehat{\lambda}_t(x) = \sum_{i=1}^n w_i \, \kappa_t(x \mid x_i)$, following the usual rationale [3, pp. 173–174].

### 3.3 Properties of the diffusion estimate

Sample properties of the diffusion estimate (7) can be deduced from analytic properties of the heat kernel. They would also be expected from the intuition that the heat kernel is the transition probability density of Brownian motion.

**Lemma 1** *The heat kernel $\kappa_t(x \mid y)$ in $W$ defined in Theorem 1 satisfies*

1. *symmetry, $\kappa_t(x \mid y) = \kappa_t(y \mid x)$ for all $x, y \in W$;*
2. *conservation of mass, $\int_W \kappa_t(x \mid y) \, \mathrm{d}x = 1$ for all $y \in W$;*
3. *the semigroup property*

$$\kappa_{t+s}(y \mid x) = \int_W \kappa_s(y \mid z) \kappa_t(z \mid x) \, \mathrm{d}z, \tag{8}$$

   *for any $s, t > 0$ and $x, y \in W$;*
4. *reliance on paths: if $x, y \in W$ are not connected by a path in $W$, then $\kappa_t(x \mid y) = 0$ for all $t$;*
5. *convergence to uniform: if the interior of $W$ is path-connected, then $\kappa_t(x \mid y) \to 1/|W|$ as $t \to \infty$, uniformly in $x$ and $y$, where $|W|$ is the area of $W$.*

Here a *path* in $W$ between two points $x, y \in W$ is a continuous curve (i.e. a continuous image of the unit interval), lying entirely in $W$, whose endpoints are $x$ and $y$. A set $W$ is *path-connected* if every pair of points $x, y \in W$ can be joined by a path in $W$.

Sample properties of the diffusion estimate follow directly:

**Lemma 2** *The diffusion estimate* (7) *satisfies:*

1. *conservation of mass, $\int_W \widehat{\lambda}_t(x) \, \mathrm{d}x = n$ for all $t > 0$;*
2. *the reproductive property*

$$\widehat{\lambda}_{t+s}(x) = \int_W \kappa_s(x \mid z) \widehat{\lambda}_t(z) \, \mathrm{d}z, \tag{9}$$

   *for any $s, t > 0$ and $x \in W$;*
3. *convergence to uniform: if the interior of $W$ is path-connected, then $\widehat{\lambda}_t(x) \to n/|W|$ as $t \to \infty$, uniformly in $x$.*

If the interior of $W$ is not path-connected, but consists of several path-connected components or "islands" $W_1, \ldots, W_m$, then the results above apply to each component $W_j$. The diffusion estimator preserves total mass on each component:

$$\int_{W_j} \widehat{\lambda}_t(x)\,\mathrm{d}x = n(\boldsymbol{x} \cap W_j) \qquad \text{for all } t. \tag{10}$$

As $t \to \infty$ the diffusion estimator converges to a uniform density on each path-connected component of $W$ with value $\lambda_j = n(\boldsymbol{x} \cap W_j)/|W_j|$ on $W_j$. In contrast, the fixed-bandwidth kernel estimates with a continuous kernel, using the uniform correction or Jones-Diggle correction, converge as $\sigma \to \infty$ to a uniform density with constant value over $W$, namely $\lambda = n(\boldsymbol{x})/|W|$.

Appendix B.1 examines the behaviour of the heat kernel near the boundary of the window, in a special case.

### 3.4 Statistical properties of the diffusion estimator

Statistical properties of the diffusion estimator can now be derived using basic theorems for point processes [24, 25].

**Lemma 3** *If the true point process intensity is $\lambda(x)$, then the expectation of the diffusion estimator* (7) *is*

$$\mathbb{E}[\widehat{\lambda}_t(x)] = \int_W \kappa_t(x \mid y)\lambda(y)\,\mathrm{d}y. \tag{11}$$

This is a simple application of Campbell's theorem [25, p. 163] to (7). Note that the right hand side of (11) is the solution of the heat conduction problem with initial condition $f_0(x) = \lambda(x)$, $x \in W$.

**Lemma 4** *If the true intensity is uniform, $\lambda(x) \equiv \lambda > 0$, then the diffusion estimator is unbiased, $\mathbb{E}[\widehat{\lambda}_t(x)] \equiv \lambda$ for all $t$.*

To prove this, substitute $\lambda(x) \equiv \lambda$ in (11), invoke the symmetry property of the heat kernel, and apply conservation of mass.

The preceding results serve to highlight a unique and powerful practical consequence of the diffusion estimator—that it simultaneously satisfies both conservation of mass *and* unbiasedness for uniform intensities. In the classical fixed-bandwidth kernel estimator these desiderata are incompatible and we use a different edge correction to achieve each of them ('uniform' correction (2) for the latter and 'Jones-Diggle' correction (3) for the former).

**Lemma 5** *For a Poisson point process with true intensity $\lambda(x)$, the pointwise variance of the diffusion estimator is*

$$v_t(x) = \mathsf{var}[\widehat{\lambda}_t(x)] = \int_W \kappa_t(x \mid y)^2\lambda(y)\,\mathrm{d}y. \tag{12}$$

*An unbiased estimator of this variance is*

$$\widehat{v}_t(x) = \sum_{i=1}^{n} \kappa_t(x \mid x_i)^2. \tag{13}$$

Equation (12) follows from the formula for the variance of a sum over a Poisson process [24, p. 188]. The unbiasedness of (13) follows from Campbell's theorem. These results can also be established from first principles.

Edge effects can be ignored when bandwidth is small with respect to the size of the domain. For any points $x, y$ in the interior of $W$, as $t \to 0$, the heat kernel $\kappa_t(y \mid x)$ is asymptotically equivalent to the Gaussian density with the same bandwidth. Consequently, the asymptotics of bias and variance as $t \to 0$ are the same for the diffusion estimator as they are for the fixed-bandwidth Gaussian kernel estimator. This argument is familiar from other contexts; cf. [12].

Appendix B.2 analyses the statistical performance of the diffusion estimator when $W$ is a square.

For any point process (not necessarily Poisson), the following formula gives the covariance of the diffusion estimator at any spatial lag, and hence the variance.

**Lemma 6** *Consider a point process on $W$ with intensity function $\lambda(u)$, $u \in W$ and second moment intensity $\lambda_2(u,v)$, $u, v \in W$, so that the pair correlation function is $g(u,v) = \lambda_2(u,v)/(\lambda(u)\lambda(v))$. Then*

$$\begin{aligned} \mathsf{cov}[\widehat{\lambda}_t(u), \widehat{\lambda}_t(v)] &= \int_W \kappa_t(u \mid x)\kappa_t(v \mid x)\lambda(x)\,\mathrm{d}x \\ &\quad + \int_W \int_W \kappa_t(u \mid x)\kappa_t(v \mid y)\,[g(x,y) - 1]\,\lambda(x)\lambda(y)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned} \tag{14}$$

This is a consequence of the second moment Campbell theorem [3, pp. 242, 250–251]. The variance of $\widehat{\lambda}_t(u)$ is obtained by setting $u = v$ in (14). If $\boldsymbol{X}$ is a Poisson process, then $g \equiv 1$ and the double integral term in (14) vanishes, and we recover (12). In general, the double integral term could be either positive or negative. For large values of diffusion bandwidth, the mean square error of $\widehat{\lambda}(x)$ will be dominated by the squared bias due to smoothing, rather than the variance due to smaller-scale clustering.

### 3.5 Methods for bandwidth selection

The bandwidth $\sigma$ for the diffusion estimator must be chosen to avoid over- or under-smoothing. Data-driven procedures for bandwidth selection in kernel estimation include cross-validation and asymptotically efficient methods [63, 17, 70, 45]. These have been extended from univariate to multivariate data [57, 36, 37, 38, 72, 29], but may require further modification for spatial data. Bandwidth selection for spatially-adaptive smoothers is even more challenging [1, 27, 28].

Asymptotically efficient bandwidth selection is based on a large-sample limit, in which the point process intensity increases and the bandwidth decreases at a rate justifying a normal approximation to the distribution of the kernel estimate. The bandwidth is chosen to minimise the asymptotic mean integrated square error of the kernel estimator. For the diffusion estimator, the asymptotic mean and variance are *identical* to those obtained for the Gaussian kernel estimator, so the asymptotically optimal bandwidth selection rule is the same for the two estimators. This justifies using Silverman's [63, eq. 3.31, p. 48] and Scott's [60, eq. 6.42, p. 152] rules of thumb for bandwidth selection for the diffusion estimator. More complicated alternatives include plug-in methods [36, 38].

Cross-validation methods of bandwidth selection minimise a data-based estimate of disagreement between data and estimator [48, Sec. 5.3, pp. 87–95]. For the

estimation of point process intensity $\lambda(x)$, the *likelihood cross-validation* criterion is [48, eq. (5.12), p. 90]

$$\mathrm{LCV}(\sigma) = \sum_{i=1}^{n} \log \hat{\lambda}_\sigma^{[-i]}(x_i) - \int_W \hat{\lambda}_\sigma(x)\,\mathrm{d}x \tag{15}$$

where $\hat{\lambda}_\sigma^{[-i]}(x_i)$ is the "leave-one-out" estimate of intensity at the data point $x_i$, computed by applying the estimator to the point pattern with $x_i$ removed. Typically the integral in (15) is omitted since it is usually approximately equal to the total number of data points (and therefore not dependent on bandwidth). The bandwidth is chosen to maximise $\mathrm{LCV}(\sigma)$.

The likelihood cross-validation criterion (15) can be applied to the diffusion estimator. A drawback is the cost of calculating the leave-one-out estimates $\hat{\lambda}_\sigma^{[-i]}(x_i)$. For the Gaussian kernel estimator, the leave-one-out estimates satisfy $\hat{\lambda}_\sigma^{[-i]}(x_i) = \hat{\lambda}_\sigma(x_i) - k_\sigma(0)$ and can be computed very quickly from the intensity estimate $\hat{\lambda}_\sigma(x)$. No such short-cut seems to exist for the diffusion smoother in general. Except for rectangular and circular domains $W$ where there is a computable expression for the heat kernel, the only available option seems to be brute-force computation. That is, $\hat{\lambda}_\sigma^{[-i]}(x_i)$ must be computed by applying the diffusion algorithm to the point pattern $\boldsymbol{x} \setminus \{x_i\}$. The diffusion algorithm will be executed $n+1$ times to compute $\mathrm{LCV}(\sigma)$. On the other hand, an advantage of the diffusion estimator is that the computation of $\widehat{\lambda}_\sigma(x)$ involves the computation of $\widehat{\lambda}_\tau(x)$ for a sequence of smaller bandwidths $\tau \le \sigma$. Total computation time is proportional to $n\sigma_{\max}^2$, where $\sigma_{\max}$ is the largest bandwidth value under consideration.

Alternatives to likelihood cross-validation include least-squares cross-validation [9, 49] and the Cronie-Van Lieshout moment method [23].

The high computational costs of the diffusion estimator at large bandwidths (and for large $n$) could be circumvented by using the Gaussian kernel estimator for bandwidth selection, then simply using the resulting bandwidth in the diffusion smoother. Alternatively, one could use the faster Gaussian kernel estimator to first obtain a reasonable value for $\sigma_{\max}^2$, since this is the main control on the computational cost, and then calibrate the diffusion-based bandwidth selection procedure thereafter. These shortcuts come with obvious drawbacks; their appropriateness in practice will be partially dependent on the window geometry and the data at hand. They are viable for exploratory purposes, as shown in Section 8.

## 4. DISCRETE APPROXIMATION AND IMPLEMENTATION

In this section we describe a discrete approximation to the diffusion estimator, which forms the basis of our numerical implementation as described in Appendix C.1. This generalises the results of Barry and McIntyre [6], and serves to unify their approach with the theory of Botev *et al.* [12]. It clearly highlights the connection between spatial smoothing bandwidth and diffusion "time". The details are useful for expository purposes and aid intuitive interpretation. Our generalisation is applicable to lattices commonly encountered in image processing software and is readily extensible to adaptive smoothing as we investigate in Section 6. We also measure the discretisation error and show how it can be vastly reduced using numerical techniques.

### 4.1 Discretisation of the Gaussian kernel estimator

First we establish the principle and the notation by discretising the Gaussian kernel estimator of intensity on the infinite plane.

Consider a time-homogeneous Markov chain $(Y_\tau)$ in discrete time $\tau = 0, 1, 2, \ldots$ with discrete (finite or countable) state space $C$ and transition probability matrix $\boldsymbol{P} = \big[p_{ab}\big]_{a,b\in C}$ where $p_{ab} = \Pr\{Y_{\tau+1} = b \mid Y_\tau = a\}$. Let $\boldsymbol{P}^{(\tau)} = \big[p_{ab}^{(\tau)}\big]_{a,b\in C}$ be the matrix of $\tau$-step transition probabilities $p_{ab}^{(\tau)} = \Pr\{Y_{\tau+t} = b \mid Y_t = a\}$. The Chapman-Kolmogorov equations state that $\boldsymbol{P}^{(\tau+1)} = \boldsymbol{P}^{(\tau)}\,\boldsymbol{P}$ by considering the $(\tau+1)$th step ("forward equation"), and that $\boldsymbol{P}^{(\tau+1)} = \boldsymbol{P}\,\boldsymbol{P}^{(\tau)}$ by considering the first step ("backward equation"). Either of these equations can be used to evaluate $\boldsymbol{P}^{(\tau)}$ recursively for $\tau = 2, 3, \ldots$.

Let the probability distribution of the state $Y_\tau$ at time $\tau$ be represented by the vector $\boldsymbol{v}_\tau = \big[\Pr\{Y_\tau = a\}\big]_{a\in C}$. Then we have the forward recursion $\boldsymbol{v}_{\tau+1} = \boldsymbol{v}_\tau \boldsymbol{P}$ with solution $\boldsymbol{v}_\tau = \boldsymbol{v}_0 \boldsymbol{P}^\tau$. Note that successive steps of the chain correspond to right-multiplication by $\boldsymbol{P}$.

Now consider a graph whose nodes (vertices) are the points of $C$, with $a \sim b$ denoting that $a, b \in C$ are joined by an edge of the graph. The degree of a node $a$ is the number of neighbours, $\deg(a) = \#\{b \in C : a \sim b\}$. Suppose the maximum degree $v = \max_{a\in C} \deg(a)$ is finite. Let $0 < q < 1/v$ and consider the *quasi-symmetric random walk* on the graph, with transition probabilities

$$p_{ab} = \Pr\{Y_{\tau+1} = b \mid Y_\tau = a\} = \begin{cases} q & \text{if } a \sim b \\ 1 - q\deg(a) & \text{if } a = b \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Importantly this ensures that the transition probabilities are symmetric, $p_{ab} = p_{ba}$, and $p_{ab}^{(k)} = p_{ba}^{(k)}$. The chain is aperiodic because there is a positive probability of "staying put", $p_{aa} > 0$ at any vertex $a$. If $C$ is finite, the random walk is time-reversible and converges in distribution to the uniform distribution on $C$.

We caution that many probabilists use the term "random walk on a graph" exclusively for the chain that, from a given vertex $a$, always jumps to one of the neighbouring vertices, with equal probability $1/\deg(a)$ for each neighbour [35]. This chain has undesirable properties: it may be periodic, and it does not converge to the uniform distribution on a finite graph; it is not useful here.

Next let $C$ be the infinite square grid consisting of all points $(i, j)$ in two-dimensional space with integer coordinates $i$ and $j$. Make $C$ a graph by joining every pair of *horizontal neighbours* $(i, j) \sim (i+1, j)$ and joining every pair of *vertical neighbours* $(i, j) \sim (i, j+1)$ so that every vertex has degree 4. Consider a quasi-symmetric random walk on this graph, so that the transition probabilities between two sites $a, b \in C$ are $p_{ab} = q$ if $a \sim b$, $p_{aa} = 1 - 4q$, and $p_{ab} = 0$ if $a \not\sim b$, where $0 < q < 1/4$ is fixed. The transition matrix is symmetric. The position of the random walk is the sum of independent and identically distributed random vector increments $\Delta Y_\tau = Y_\tau - Y_{\tau-1}$ for $\tau = 1, 2, \ldots$, with mean $\mathbb{E}[\Delta Y_1] = \boldsymbol{0}$ and variance-covariance matrix $\mathsf{var}[\Delta Y_1] = 2q\boldsymbol{I}_2$. Hence, $\mathbb{E}[Y_\tau] = \tau\mathbb{E}[\Delta Y_1]$ and $\mathsf{var}[Y_\tau] = \tau\,\mathsf{var}[\Delta Y_1]$.

Now let us rescale space and time so that the grid spacing is $\Delta c$ and the time step is $\Delta t$. The left panel of Figure 2 sketches the possible transitions from a

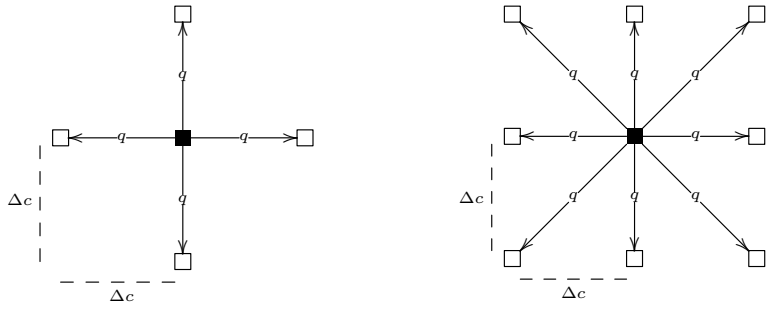


FIG 2. *Transition diagrams for a quasi-symmetric random walk on the infinite square grid.* Left: *4-connected grid.* Right: *8-connected grid. Arrows show permitted transitions from the currently-occupied node (■) to other nodes that are reachable in one step (□) with the transition probability shown.*

given state. A single step occurs in time $\Delta t$ and has variance-covariance matrix $2q(\Delta c)^2 \boldsymbol{I}_2$. Let $\Delta c \to 0$ and $\Delta t \to 0$ with $\Delta t \sim 2q(\Delta c)^2$. In brief, the Central Limit Theorem implies that the probability distribution of the rescaled particle location at time $t$ is approximately Gaussian with mean $\mathbf{0}$ and variance $t\boldsymbol{I}_2$. The rescaled random walk converges weakly to standard Brownian motion. This provides the key connection between the approaches of Barry and McIntyre [6] and Botev *et al.* [12].

The forward equation for the random walk on the integer lattice, $\boldsymbol{v}_{\tau+1} = \boldsymbol{v}_\tau \boldsymbol{P}$, can be rewritten as a difference equation $\boldsymbol{v}_{\tau+1} - \boldsymbol{v}_\tau = \boldsymbol{v}_\tau(\boldsymbol{P} - \boldsymbol{I}_N)$. In the rescaled limit, this yields the classical Fourier time-dependent heat equation: for a grid location $x = (u, v) = (i\Delta c, j\Delta c)$ with $i, j \in \mathbb{Z}$, write $f_\tau(u, v)$ for the probability of occupying site $x$ at time $t = \tau\Delta t$. The rescaled difference equation is

$$\begin{aligned} f_{\tau+1}(u,v) - f_\tau(u,v) &= q\, f_\tau(u+\Delta c, v) + q\, f_\tau(u-\Delta c, v) \\ &\quad + q\, f_\tau(u, v+\Delta c) + q\, f_\tau(u, v-\Delta c) - 4q\, f_\tau(u,v) \\ &= q\,[f_\tau(u+\Delta c, v) + f_\tau(u-\Delta c, v) - 2f_\tau(u,v)] \\ &\quad + q\,[f_\tau(u, v+\Delta c) + f_\tau(u, v-\Delta c) - 2f_\tau(u,v)]\,. \end{aligned} \tag{17}$$

On the right-hand side of (17), the first and second brackets contain the discrete second differences of the function $f(u, v)$ in the horizontal and vertical directions, respectively. Dividing both sides by $\Delta t = 2q(\Delta c)^2$ yields a discrete approximation to the heat equation (4). This also determines the correspondence between the number of iterations $\tau$ of the discrete approximation and the bandwidth $\sigma$ of the Gaussian kernel, namely $\tau = \sigma^2/\Delta t$.

The results above apply to the distribution of the location of a single particle undergoing a random walk. For the diffusion estimator of point process intensity we may simply consider $n$ particles, at initial positions $x_1, \ldots, x_n$. Particle $j$ executes a random walk $(Y_{jt}, t \geq 0)$ as described above. Then the intensity of the point process $\{Y_{1t}, \ldots, Y_{nt}\}$ at time $t$ is the discrete approximation of the Gaussian kernel smoother with bandwidth $\sigma = \sqrt{t}$.

The derivation above used the "4-connected" rectangular grid. One could equally use the "8–connected grid" discussed by Barry and McIntyre [6]. This graph also joins *diagonal neighbours*, $(i, j) \sim (i+1, j+1)$ and $(i, j) \sim (i-1, j+1)$ for all integers $i, j$ [22, p. 383 ff.]. The rescaled 8-connected graph transitions are

depicted in the right panel of Figure 2. The random walk increments $\Delta Y_\tau$ are i.i.d. random vectors with mean $\mathbf{0}$ and variance-covariance matrix $6q\boldsymbol{I}_2$. The 4-connected and 8-connected grids lead to essentially equivalent estimators, which can be reconciled by matching the covariance structures as explained above.

### 4.2 Discretisation of the heat kernel

The previous results can be adapted to a given, bounded spatial domain $W \subset \mathbb{R}^2$ by simply restricting the random walk to remain inside $W$ at all times. The state space $C$ is replaced by $C \cap W$, and we restrict the graph edges to those which join vertices inside $W$. An illustrative example is shown in Figure 3. Vertices near the boundary of $W$ have few neighbours than the maximum possible degree $v$.

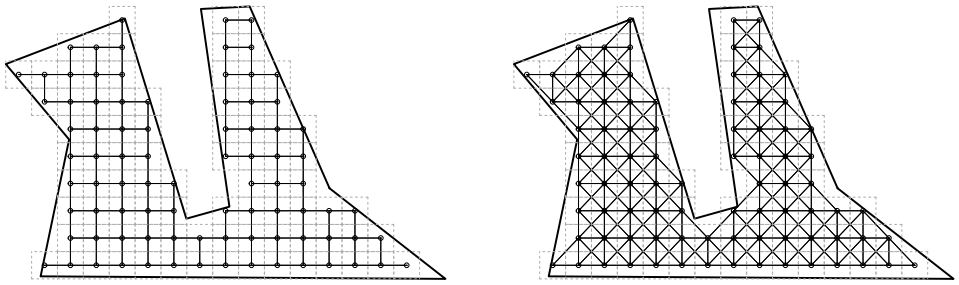

FIG 3. *Illustrative example of regular discretisation of an irregular polygonal domain (bold solid line) using a perfectly square lattice of nodes (open dots). The neighbour networks defining possible jumps in a random walk from node to node are given as solid lines (4-connected network* left*; 8-connected network* right*), and the equally-sized square pixels formed by taking each node as a centroid are visible as a grid of grey dashed lines.*

The quasi-symmetric random walk on $C \cap W$ has transition probabilities $\Pr\{Y_{\tau+1} = b \mid Y_\tau = a\} = q$ if $a \sim b$ and $\Pr\{Y_{\tau+1} = a \mid Y_\tau = a\} = 1 - q\deg_W(a)$, where $\deg_W(a) = \#\{b \in C \cap W : b \sim a\}$ is the degree of node $a$ in the graph restricted to $W$. For nodes $a$ near the boundary of $W$, the stayput probability $\Pr\{Y_{\tau+1} = a \mid Y_\tau = a\}$ is increased, compared to the walk on the infinite grid.

This construction ensures that the transition matrix $\boldsymbol{P}$ is symmetric, and that the equilibrium distribution of the chain is the uniform distribution on $C \cap W$. In brief, the Kolmogorov forward equation at an interior point gives the discrete analogue of the heat equation (4), and at a boundary point gives a discrete analogue of the Neumann boundary conditions (5).

Given an observed point pattern $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ in $W$, we effectively assume that each point follows a random walk. For $i = 1, \ldots, n$ let $(Y_{i\tau})$ be a quasi-symmetric random walk on $C \cap W$. Defining the total counting measure $Z_\tau(a) = \sum_i \mathbb{1}\{Y_{i\tau} = a\}$ for these random walks on $a \in (C \cap W)$, we consider the expected number of individuals at each site $a$,

$$s_\tau(a) = \mathbb{E}[Z_\tau(a)] = \sum_i \Pr\{Y_{i\tau} = a\}, \quad a \in C \cap W.$$

The vector $\boldsymbol{s}_\tau = \big[s_\tau(a)\big]_{a \in C \cap W}$ satisfies the forward recursion

$$\boldsymbol{s}_{\tau+1} = \boldsymbol{s}_\tau \boldsymbol{P}. \tag{18}$$

The solution of the discretised heat equation is computed by initialising $\boldsymbol{s}_0$ to the counting measure of the point pattern $\boldsymbol{x}$, then iteratively applying (18) for the required number of steps.

For the rescaled Markov chain with mesh size $\Delta c$ and time step $\Delta t$, we would convert the expected counting measure $\boldsymbol{s}_\tau$ to an intensity $\boldsymbol{\lambda}_\tau = a^{-1}\boldsymbol{s}_\tau$ where $a = (\Delta c)^2$ is the area of one grid cell ("pixel"). Full details are given in Appendix C.1.

The iterative procedure is numerically stable, because $\boldsymbol{P}$ is a stochastic matrix, so that its eigenvalues all have magnitude less than or equal to 1. Similar calculations are performed for the 8-connected grid.

### 4.3 General regular grid

For practical purposes it is important to deal with pixel grids which are rectangular but not square, such as camera image rasters with a 3 : 2 aspect ratio, and with hexagonal pixel grids used in image analysis [61].

This can be achieved with minor modifications to the calculations above. To determine the appropriate scaling and transition probabilities, one simply needs to determine the variance of each increment in the random walk. The numerical stability argument still holds.

Figure 4 sketches the case of a general rectangular lattice with horizontal and vertical step sizes $\Delta x$ and $\Delta y$ respectively, using either the 4-connected or 8-connected graph.

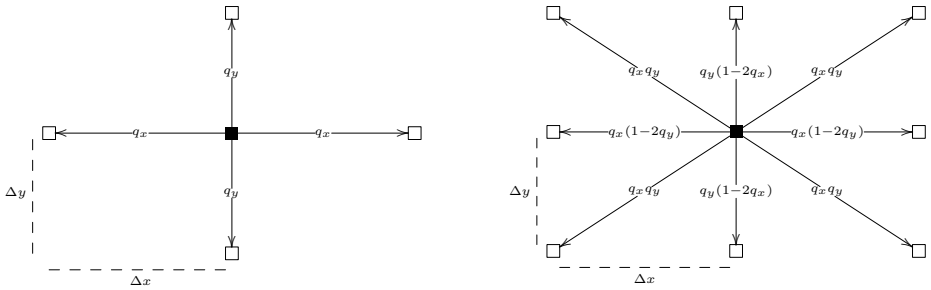


FIG 4. *Transition diagrams for quasi-symmetric random walks on a non-square rectangular lattice. Left: A 4-connected network. Right: An 8-connected network.*

For the 4-connected grid, transitions between horizontal neighbours will be assigned probability $q_x$, and transitions between vertical neighbours have probability $q_y$, where $0 < q_x, q_y < 1$ are fixed numbers to be determined. We constrain $2q_x + 2q_y < 1$ so that there is nonzero probability of staying put, $p_{aa} > 0$.

In the infinite rectangular grid, the vector increments $\Delta Y_\tau$ are easily calculated to have mean $\mathbf{0}$ and variance-covariance matrix

$$\mathsf{var}\,\Delta Y_1 = \begin{pmatrix} 2q_x(\Delta x)^2 & 0 \\ 0 & 2q_y(\Delta y)^2 \end{pmatrix} \tag{19}$$

in the notation of Section 4.1, where the off-diagonal covariances are zero because horizontal and vertical jumps are mutually exclusive. For an isotropic covariance matrix, we should set $q_x(\Delta x)^2 = q_y(\Delta y)^2$. The remainder of the derivation is identical to that in the previous section.

For an 8-connected rectangular grid, we let the horizontal and vertical jumps be *independent*. The vector increment $\Delta Y_\tau$ has coordinates $(\Delta U_\tau, \Delta V_\tau)$ which are

independent random variables, $\Delta U_\tau$ taking values $\Delta x, 0, -\Delta x$ with probabilities $q_x, 1-2q_x, q_x$ respectively, and $\Delta V_\tau$ taking values $\Delta y, 0, -\Delta y$ with probabilities $q_y, 1-2q_y, q_y$ respectively. Here $0 < q_x, q_y < 1/2$ so that there is nonzero probability of staying put. The transition diagram is shown in the right panel of Figure 4. The vector increment $\Delta Y_\tau$ has mean $\mathbf{0}$ and variance-covariance matrix of the same form (19), in which the off-diagonal entries are zero because of *independence*. Thus we should again set $q_x(\Delta x)^2 = q_y(\Delta y)^2$ to obtain an isotropic covariance matrix.

For a regular *hexagonal* grid, in which each node is connected to 6 neighbours with distance $\Delta c$, the random walk with equal transition probability $q$ between any pair of neighbours (where $q < 1/6$) has $\mathbb{E}[\Delta Y_1] = \mathbf{0}$ and $\mathsf{var}\, \Delta Y_1 = 3q(\Delta c)^2 I_2$.

Algorithm 2 of Appendix C.1 gives a detailed specification of the preceding algorithm for computing the fixed-bandwidth diffusion estimator. Software is available (Section 9).

### 4.4 Accuracy of discrete approximation

The discrete random walk calculation presented above is an instance of the *Euler scheme* for numerical approximation of differential equations [16, Sections 20–21, pp. 55–89]. More accurate approximations are available, but are typically more complicated to implement and more computationally demanding than the Euler scheme. We use the Euler scheme because it provides intuitive parallels to the underlying theory, and clarifies connections to existing work. We also describe a simple technique for improving accuracy.

The approximation error of the Euler scheme is of the same order of convergence as the grid step size [16, loc. cit.]. In Appendix D we use the Berry-Esséen Theorem to obtain the more specific result that, for a quasi-symmetric random walk on the 8-connected grid, the maximum absolute error in the bivariate cumulative distribution function is less than $\Delta c/\sigma$. A good rule of thumb is that the approximation is adequate when $\Delta c < \sigma/20$.

Table 1 shows the maximum absolute error in the discrete approximation to the heat kernel itself (that is, to the density value rather than the cumulative probability) in one example. The window is the unit square and the source point is at the centre of the window. The bandwidth is $\sigma = 0.1$ and the maximum value of the (true) heat kernel is 15.53. The discrete approximation is computed using our software implementation and the exact value is computed as described in Appendix A. The error is roughly halved when the grid step $\Delta c = \delta$ is halved, which is consistent with the expected order of convergence $O(\delta^1)$.

Since the approximation error in the Euler scheme has a known convergence rate, it can be improved using *Richardson extrapolation*, a classical technique of numerical analysis ([55, 56], [15, p. 72 ff.]). For the Euler scheme with step size $\Delta c = \delta$, let $A(\delta)$ be the calculated value of the heat kernel at a given location for a

TABLE 1

*Maximum absolute pointwise error of discrete approximation to heat kernel. Window is the unit square. Single source point at* $(0.5, 0.5)$. *Bandwidth* $\sigma = 0.1$.

| | Grid size | | | | |
|---|---|---|---|---|---|
| | 32 | 64 | 128 | 256 | 512 |
| 4-connected | 2.08 | 1.07 | 0.53 | 0.27 | 0.13 |
| 8-connected | 2.15 | 1.07 | 0.53 | 0.27 | 0.13 |

TABLE 2
*Maximum absolute pointwise error of discrete approximation to heat kernel* ***using Richardson extrapolation****. Window is the unit square. Single source point at* $(0.5, 0.5)$*. Bandwidth* $\sigma = 0.1$*.*

| | FINEST GRID SIZE | | | | |
|---|---|---|---|---|---|
| | 32 | 64 | 128 | 256 | 512 |
| 4-connected | 1.00 | 0.57 | 0.15 | 0.04 | 0.01 |
| 8-connected | 1.31 | 0.41 | 0.10 | 0.03 | 0.01 |

given bandwidth. Performing the calculation again for the grid with coarser spacing $r\delta$ where $r > 1$, we combine the two estimates in the Richardson extrapolant of order $k \geq 1$

$$R_{r,k}(\delta) = \frac{r^k A(\delta) - A(r\delta)}{r^k - 1} = A(\delta) + \frac{A(\delta) - A(r\delta)}{r^k - 1}. \tag{20}$$

When $A(\delta) = O(\delta^k)$ as $\delta \to 0$, the Richardson extrapolant (20) with the same exponent $k$ converges at the faster rate $O(\delta^{k+1})$ or faster. For the Euler scheme, $A(\delta) = O(\delta^1)$ and we expect $R_{r,1}(\delta)$ to converge at rate $O(\delta^2)$.

Table 2 shows the counterpart of Table 1 when the estimates are improved using Richardson extrapolation. To obtain estimates on the $n \times n$ grid, the Euler scheme was applied to the $n \times n$ ("fine") and $n/2 \times n/2$ ("coarse") grids. The results on the coarse grid were bilinearly interpolated to the fine grid. The Richardson extrapolant (20) with $r = 2$ and $k = 1$ was computed to obtain estimates on the $n \times n$ grid. In Table 2 each halving of the grid step reduces the errors by roughly a factor of 4, which is consistent with the expected $O(\delta^2)$ error rate.

Tables 1 and 2 also show that the Richardson-extrapolated Euler scheme at step size $\delta$ is often more accurate than the un-extrapolated Euler scheme at the finer step size $\delta/2$. The former is also faster to compute; since the number of grid points is quadrupled when the step size is halved, computation of the former requires a fraction $(1 + 1/4)/4 = 31\%$ of the computation time for the latter.

Tables 1 and 2 measure performance by the maximum pointwise discrepancy in the density estimate. Similar results are found when performance is measured by the total variation distance, or by the maximum discrepancy between bivariate cumulative distribution functions, as reported in the supplementary material.

Although other methods for solving the heat equation may be preferred on theoretical grounds, they can be less satisfactory in practice. Implicit-solution methods are more difficult to implement for sparse matrices. Increasing the grid resolution will increase computation time, and may paradoxically increase error, due to numerical underflow. We recommend the use of Richardson extrapolation here, because it avoids these problems and is simple to implement. We recommend using the 4-connected grid, which requires less time than the 8-connected grid, and achieves similar accuracy.

## 5. APPLICATION: NEW ZEALAND PĀ SITES

Figure 5 shows the recorded locations of 854 historic Māori sites called *pā* in a region of the North Island of New Zealand (the present-day city of Auckland lies in the bottom-right quadrant of the dashed rectangle). Pā are loosely defined as sites that are enclosed or possess defensive features such as earthworks. These observations are part of a larger dataset documented for archaeological research into Māori land use by the University of Otago [64, 65].

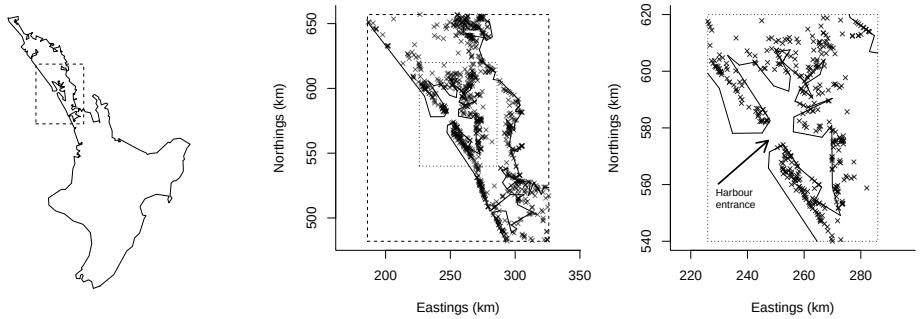


FIG 5. *New Zealand pā sites.* Left: *simplified polygon representing the North Island of New Zealand showing the region of interest as a dashed box.* Middle: *locations of pā sites in the region of interest. A smaller, dotted box delineates a region for closer inspection, centred on the water body of Kaipara Harbour.* Right: *the zoomed-in area around Kaipara Harbour, with harbour entrance indicated. Data provided by the Department of Anthropology and Archaeology, University of Otago, with thanks to Baylee Smith and Tim Thomas.*

Figure 6 shows estimates of the spatial point process intensity, using both the fixed-bandwidth Gaussian kernel estimator with Jones-Diggle correction (3) and the diffusion estimator (7). The estimates were computed for the full region of interest in the middle panel of Figure 5 using the full dataset, but are shown only within the zoomed-in region around Kaipara Harbour. In both estimates we purposely choose the relatively generous bandwidth of $\sigma = 12.6$, calculated using the oversmoothing principle of [68] for the Gaussian kernel estimator.

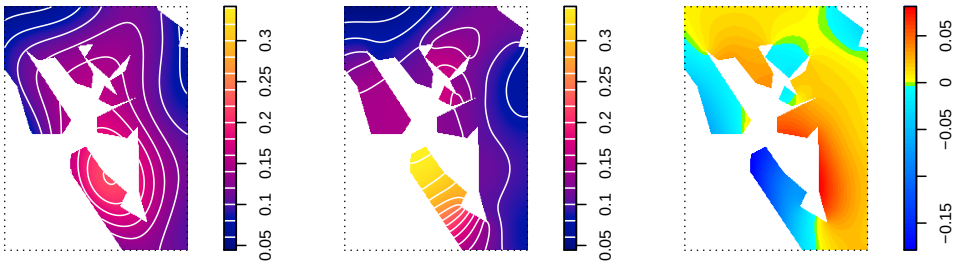


FIG 6. *Estimates of the inhomogeneous spatial intensity of pā sites around Kaipara Harbour.* Left: *fixed-bandwidth Gaussian kernel estimator.* Middle: *fixed-bandwidth diffusion smoother.* Right: *pointwise difference, Gaussian minus diffusion. Intensity values are numbers per square km.*

There is a compelling contrast between the two estimates. The Gaussian kernel estimate exhibits the tunnelling artefact discussed in the Introduction (regardless of the choice of edge correction) and has the appearance of a smooth function clipped to an irregular domain. The diffusion estimate, on the other hand, is highly responsive to the coastline, and assigns very different masses to the individual peninsulas around the harbour. The discrepancy is particularly striking on the southern side of the entrance to the Kaipara Harbour, where the diffusion estimate is twice as high as the Gaussian kernel estimate. Additional commentary and analysis is provided in the supplementary materials.

## 6. ADAPTIVE SMOOTHING

In some applications, data are so severely inhomogeneous that a fixed-bandwidth kernel estimate is unsatisfactory, exhibiting both under- and over-smoothing in different areas. *Adaptive* kernel estimation [63, Chap. 5], [70, 41] mitigates this problem by applying different amounts of smoothing to different areas.

### 6.1 Sample-point-adaptive Gaussian kernel estimate

In *sample-point adaptive* kernel smoothing [63, Chap. 5], each data point $x_i$ is assigned its own individual bandwidth $\sigma_i$. The adaptive estimate takes the general form

$$\widehat{\lambda}(x) = \sum_{i=1}^{n} \frac{k_{\sigma_i}(x - x_i)}{e(x, x_i, \sigma_i)}, \tag{21}$$

where $k_\sigma$ is typically the Gaussian kernel, and $e(x, x_i, \sigma_i)$ is an edge-correction term. Edge corrections are not discussed in the early literature, but by analogy with the fixed-bandwidth case, one could use the "uniform" correction $e(x, x_i, \sigma_i) = c_W(\sigma_i, x)$ analogous to (2), or the "Jones" correction $e(x, x_i, \sigma_i) = c_W(\sigma_i, x_i)$ analogous to (3). See [50, 26].

The typical procedure for assigning adaptive bandwidths is the rule of Abramson [1]. Starting with a pilot estimate $\tilde{f}$ of the normalised probability density $f(x) = \lambda(x)/\int_W \lambda(y)\,\mathrm{d}y$, we calculate bandwidth factors $b_i = \tilde{f}(x_i)^{-1/2}$, compute the geometric mean $\gamma = (\prod_i b_i)^{1/n}$, and take bandwidths

$$\sigma_i = \sigma_0 \min\left\{\frac{b_i}{\gamma}, B\right\}, \tag{22}$$

where $\sigma_0 > 0$ is the *global bandwidth* and $B > 1$ is a truncation constant [41]. This reduces the task of selecting the adaptive bandwidths $\sigma_i$ to the one-dimensional problem of choosing the global bandwidth $\sigma_0$.

### 6.2 Adaptive smoothing with a non-uniform diffusion

*6.2.1 Concept* A spatially non-uniform version of diffusion smoothing was proposed by Botev *et al.* [12]. Brownian motion $\boldsymbol{B}_t$ is replaced by another diffusion $\boldsymbol{X}_t$ whose properties are spatially varying. Correspondingly, the discrete random walk (Section 4.1) would be modified so that its transition probabilities depend on spatial location. The classical Fourier heat equation (4) is replaced by the more general Fokker-Planck heat equation for a physical material with spatially-varying thermal properties.

Spatially-varying transition probabilities lead to a non-uniform equilibrium distribution. Equivalently, in a physical material with non-uniform thermal properties, the equilibrium distribution of heat is non-uniform.

Write $p_t$ for the transition kernel of the diffusion $\boldsymbol{X}_t$. That is, $p_t(\cdot \mid y)$ is the probability density of $\boldsymbol{X}_t$ given $\boldsymbol{X}_0 = y$. Then the non-uniform diffusion estimator of intensity is $\widehat{\lambda}_t(x) = \sum_{i=1}^{n} p_t(x \mid x_i)$, the analogue of (7) using the non-uniform kernel $p_t$. Here the elapsed time $t$ is a smoothing parameter, which determines the overall amount of smoothing, analogous to the squared global bandwidth $\sigma_0^2$ in Section 6.1.

The approach of [12] is to nominate a target density $f$ for which it is desired that the estimator should be unbiased. The properties of the diffusion are then chosen so that the equilibrium distribution of the diffusion is $f$. In fact the diffusion will satisfy *detailed balance*

$$f(y)p_t(x \mid y) = f(x)p_t(y \mid x). \tag{23}$$

Botev *et al.* [12] developed this technique mainly in one dimension and showed that it enjoys the same good properties as the original diffusion smoother associated with the heat equation.

We use the term "non-uniform smoothing" when the target equilibrium density $f$ is fixed and chosen by the researcher in advance, and "adaptive smoothing" when the target density $f$ is a data-based pilot estimate of the intensity. Botev *et al.* [12] note that many previous studies of the performance of adaptive smoothers actually involve non-uniform rather than adaptive smoothing.

*6.2.2 Non-uniform diffusion equations* Following [12] we consider a diffusion in the two-dimensional plane, a stochastic process $\boldsymbol{X}_t$ indexed by one-dimensional time $t \geq 0$, whose states are spatial locations in $\mathbb{R}^2$, with Itô stochastic differential equation

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t)\,\mathrm{d}t + \sigma(\boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{B}_t, \tag{24}$$

where at any location $x \in \mathbb{R}^2$, the numerical value $\sigma(x)$ is the instantaneous variance or "speed", and the vector $\boldsymbol{b}(x) = (b_1(x), b_2(x))$ is the instantaneous bias or "drift".

The discrete random walk which approximates the diffusion (24) is a modification of the random walk described in Section 4. If the time step is $\Delta t$ and the current state is $x$, the next vector increment of the random walk has mean value $\boldsymbol{b}(x)\Delta t$ and variance-covariance matrix $\sigma(x)^2(\Delta t)\boldsymbol{I}$.

In physical terms, the diffusion (24) describes thermodynamics in a material with spatially-varying thermal diffusivity $\sigma(x)^2$, such as an alloy with spatially-varying composition, and advection gradient $\boldsymbol{b}(x)$, such as a fluid (or a solid which is melting) which transports heat as it flows.

The classical Fourier heat equation (4) is replaced by the Fokker-Planck heat equation (Kolmogorov forward equation)

$$\frac{\partial}{\partial t}p_t(x \mid y) = -\nabla_x{\cdot}(p_t(x \mid y)\boldsymbol{b}(x)) + \frac{1}{2}\nabla_x^2(\sigma(x)^2 p_t(x \mid y)) \tag{25}$$

(cf. equation (11) of [12]), and the Kolmogorov backward equation

$$\frac{\partial}{\partial t}p_t(x \mid y) = \nabla_y{\cdot}(p_t(x \mid y)\boldsymbol{b}(x)) + \frac{\sigma(x)^2}{2}\nabla_y^2\, p_t(x \mid y), \tag{26}$$

(cf. equation (10) of [12]). Here $\nabla_x{\cdot}$ is the divergence operator, and $\nabla_x^2$ the Laplacian operator, with respect to the coordinates of $x$, and similarly $\nabla_y{\cdot}$ and $\nabla_y^2$ are the divergence and Laplacian with respect to $y$. Technical details and references are given in Appendix E.

There are now two "heat equations" instead of one. Analogously the discrete random walk satisfies two recurrence equations, corresponding to left and right multiplication by the transition matrix $P$, because $P$ is no longer symmetric.

**Definition 3** *The* non-uniform diffusion estimate *of the intensity function* $\lambda(x)$ *based on the diffusion* (24) *is*

$$\widehat{\lambda}_t(x) = \sum_{i=1}^{n} p_t(x \mid x_i), \tag{27}$$

*where the kernel* $p_t$ *is the solution of* (25) *and* (26)*, and* $t > 0$ *is the global smoothing parameter.*

Given a target density $f$ for which it is desired that (27) should be unbiased, the goal is to choose the diffusion characteristics $\sigma(x)$ and $\boldsymbol{b}(x)$ such that the resulting kernel $p_t$ satisfies detailed balance (23).

Identifying valid choices of $\sigma(x)$ and $\boldsymbol{b}(x)$ is not simple, because $p_t$ depends on $\sigma(x)$ and $\boldsymbol{b}(x)$ through the differential equations (25) and (26). This problem is familiar from Markov chain Monte Carlo methods, where the goal is to construct a Markov chain sampler for a given target density $f$. Typically there will be no comprehensive characterisation of all possible valid choices for the sampler; instead there will be a handful of recipes for valid choices, involving a tradeoff between computational complexity and statistical efficiency.

Botev *et al.* [12] described several valid choices for the diffusion terms. In the remainder of this section, we consider the two simplest choices.

*6.2.3 Adaptive smoothing by variable-speed diffusion* A simple choice for the non-stationary diffusion is that which has zero drift, $\boldsymbol{b}(x) \equiv 0$, and spatially-varying speed $\sigma(x)$. The Itô equation of this diffusion is

$$\mathrm{d}\boldsymbol{X}_t = \sigma(\boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{B}_t. \tag{28}$$

The Fokker-Planck-Kolmogorov forward equation (25) reduces to

$$\frac{\partial}{\partial t} p_t(x \mid y) = \frac{1}{2}\,\nabla_x^2 \left(\sigma(x)^2 p_t(x \mid y)\right) \tag{29}$$

and the backward equation (26) reduces to

$$\frac{\partial}{\partial t} p_t(x \mid y) = \frac{\sigma(x)^2}{2}\,\nabla_y^2\, p_t(x \mid y). \tag{30}$$

The resulting diffusion estimate (27) is called the *variable-rate diffusion estimate.*

Define $f(x) = 1/\sigma(x)^2$; then instead of the symmetry property (1) of Lemma 1, the kernel satisfies detailed balance (23), so that the equilibrium density of the diffusion is proportional to $f(x)$ (cf. equation (15) of [12]). In Lemma 2, properties (1) and (2) remain true, while the convergence property (3) is modified so that $\widehat{p}_t(x)$ converges uniformly to $c\,f(x)$ where $c = n/\int_W f(x)\,\mathrm{d}x$. Lemma 3 remains true. Lemma 4 is modified so that the diffusion estimator is unbiased for any intensity proportional to $f(x)$. Lemma 5 remains true.

In order to implement the estimator with kernel $p_t$ satisfying (29) and (30), we need to discretise one of these equations. The generator (29) is locally just a rescaled version of the Laplacian, so the discretisation is very similar to that sketched in Section 4.

Algorithm 3 in Appendix C.2 describes our implementation of the variable-bandwidth smoother. It is similar to the implementation of the simpler fixed-bandwidth estimator described in Algorithm 2 of Appendix C.1, with the main difference being the dependence of the transition probabilities on spatial location.

*6.2.4 Langevin diffusion* The other simple choice is a diffusion with constant speed and nonzero drift, the *Langevin diffusion*

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t. \tag{31}$$

This has been studied in one dimension by Botev *et al.* [12]. They show that detailed balance (23) is achieved if $\boldsymbol{b}(x) = \frac{1}{2}\nabla \log f(x)$. The approximating random walk would have increments $(\Delta U, \Delta V)$ satisfying $\mathbb{E}[\Delta U] = \frac{1}{2}\Delta t(\partial/\partial u)\log f(x)$, $\mathbb{E}[\Delta V] = \frac{1}{2}\Delta t(\partial/\partial v)\log f(x)$, $\mathsf{var}((\Delta U, \Delta V)) = \Delta t\ I_2$. The forward equation is $(\partial/\partial t)p_t(x \mid y) = \frac{1}{2}\nabla_x^2(p_t(x \mid y) - f(x))$. Botev *et al.* show that the corresponding estimator $\widehat{\lambda}(x)$ has good statistical properties, and recommend the use of this estimator.

The Langevin diffusion is also a mainstay of modern Markov chain Monte Carlo methods, where it often produces very efficient samplers. This deserves further investigation in spatial applications. Implementation is slightly more complicated, and for lack of space, we do not consider this further, apart from some comments in the Discussion.

### 6.3 Adaptive diffusion smoothing — observation-specific bandwidths

The variable-speed diffusion smoother of Section 6.2.3 is mathematically elegant, and has theoretical advantages, but practical drawbacks remain. In applications, the spatially-varying rate $\sigma(x)$ will usually be determined from a pilot estimate of intensity, $\widetilde{f}(x)$. The diffusion is constructed so that its equilibrium density is $\widetilde{f}(x)$ by setting $\sigma(x) \propto (\widetilde{f}(x))^{-1/2}$. The overall amount of smoothing is determined by the elapsed time $t$. However, larger values of $t$ do not lead to greater over-smoothing, but to greater conformity with the pilot estimate, since the diffusion converges to its equilibrium distribution $\widetilde{f}$. This is undesirable when the pilot estimate is poor.

These weaknesses occur because the variable-rate diffusion smoother is not a direct counterpart of the sample-point adaptive kernel estimator (21), in which individual data points $x_i$ are assigned different bandwidths $\sigma_i$.

Here we explore a new alternative for spatially adaptive diffusion smoothing, which may be useful even for one-dimensional kernel estimation. The key idea is to modify the fixed-bandwidth diffusion smoother so that each data point $x_i$ enters the diffusion process at a different *starting time*. Consequently, different data points are subjected to different amounts of smoothing, and the result closely resembles (21). This estimator is less sensitive to misspecification of the pilot density than is the variable-rate smoother.

**Definition 4** *The* lagged-arrival *adaptive diffusion estimate of intensity* $\lambda(x)$ *is*

$$\widehat{\lambda}(x) = \sum_{i=1}^{n} \kappa_{t_i}(x \mid x_i), \tag{32}$$

*where* $t_i = \sigma_i^2$ *is the observation-specific smoothing variance for data point* $x_i$*, and* $\kappa_t(x \mid y)$ *is the classical Fourier heat kernel corresponding to standard Brownian motion, that is, with unit speed and zero drift, given in Theorem 1.*

This is the formal analogue of (21). In the estimator (32) all data points undergo diffusion smoothing according to the same diffusion process but for *different*

durations of time $t_i$. This is equivalent to assigning to each observation a different "starting time" in the overall diffusion process: points associated with larger bandwidths will be introduced earlier than those associated with less smoothing. More precisely, let $t_{\max} = \max t_i = \max \sigma_i^2$. Suppose that, for each data point $x_i$, a standard Brownian motion $\boldsymbol{B}_t^{(i)}$ begins at time $s_i = t_{\max} - t_i$ from the initial location $x_i$. The expected total density of the Brownian motions $\boldsymbol{B}_t^{(1)}, \ldots, \boldsymbol{B}_t^{(n)}$ at time $t$ is $\ell(t, x) = \sum_{i=1}^n \mathbb{1}\{t > s_i\}\kappa_{t-s_i}(x \mid x_i)$ so that, at time $t = t_{\max}$, we have $\ell(t_{\max}, x) = \sum_{i=1}^n \kappa_{t_i}(x \mid x_i)$, identical to (32). In line with this interpretation we call (32) the lagged-arrival adaptive estimator.

The observation-specific variances $\sigma_i^2$ can be chosen in the same way as for the Gaussian sample-point-adaptive kernel estimator (Section 6.1), using Abramson's rule (22) applied to a pilot estimate $\tilde{f}$ of the normalised probability density.

Our implementation of the lagged-arrival adaptive diffusion estimator is described in Algorithm 4 in Appendix C.3. This resembles the implementation of the fixed-bandwidth estimator described in Algorithm 2, except that data points are introduced progressively during the iteration sequence. Numerical stability of the iterative procedure is easily established.

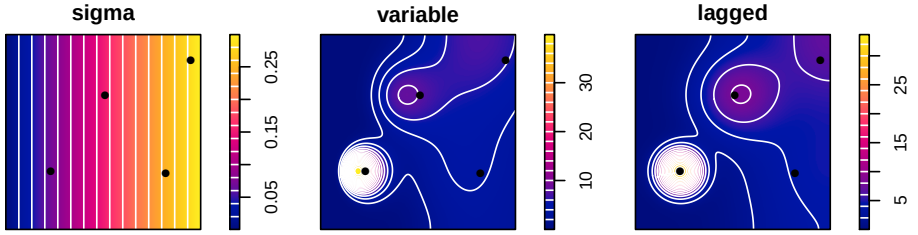


FIG 7. *Comparison of adaptive smoothers in a synthetic example.* Left: *bandwidth surface is a linear ramp.* Middle: *variable-rate smoother.* Right: *lagged-arrival smoother. Data points are superimposed on each panel.*

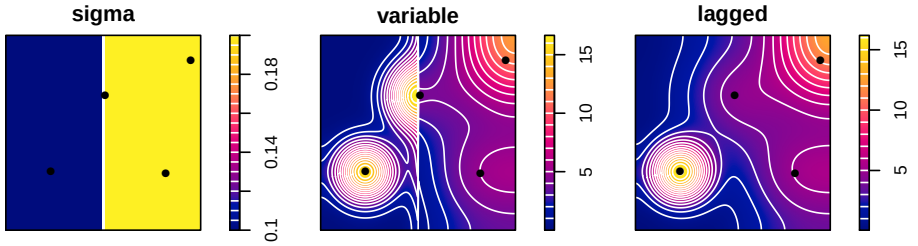


FIG 8. *Comparison of adaptive smoothers in a synthetic example.* Left: *bandwidth surface with sharp cliff.* Middle: *variable-rate smoother.* Right: *lagged-arrival smoother. Data points are superimposed on each panel.*

Figures 7 and 8 compare the variable-rate adaptive and lagged-arrival adaptive smoothers on synthetic examples where the bandwidth is a function of spatial location. If the bandwidth function is smooth, as in Figure 7, the two methods give similar results. However if the bandwidth function has a sharp discontinuity, as in Figure 8, the discontinuity will remain clearly visible in the variable-rate smoother, but not the lagged-arrival smoother. Conversely, the output of the lagged-arrival smoother will be more sensitive to small changes in the input data

coordinates, if these changes give rise to large changes in smoothing bandwidth.

For greater computational efficiency, one could follow the partitioning strategy of [26] in which the $n$ individual bandwidths $\sigma_i$ are grouped into $m = O(\sqrt{n})$ quantiles and the contribution from each quantile is computed in a single instance of the heat equation solver. For further efficiency one could use coarser time step sizes and coarser grid spacings for larger bandwidths.

## 7. APPLICATION: UNITED KINGDOM PBC CASES

Figure 9 displays the domicile locations of 761 cases of primary biliary cholangitis (PBC, formerly known as primary biliary cirrhosis) recorded between 1987 and 1994 in a region of northeast England comprising six adjacent health districts. A primary aim of the original presentation of these data in [53] was to understand the spatial variation in PBC cases across the study region. The eastern border of the study region is the North Sea coastline, beyond which there are no cases, so that the diffusion estimator is likely to perform better than the edge-corrected Gaussian kernel estimators. There is a heavy concentration of cases in the urban area of Newcastle close to the coastline, and a far lower density of cases elsewhere, so that this dataset is likely to require adaptive smoothing.

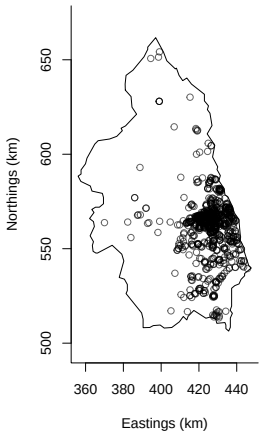


FIG 9. *Cases of Primary Biliary Cholangitis/Cirrhosis in a region of northeast England, first presented and analysed by Prince et al. [53]. Data are available as* `pbc` *in the* `sparr` *package [29] with thanks to Peter Diggle.*

Adaptive estimates of the spatially-varying intensity of cases are shown in Figure 10. The three estimates are Gaussian-adaptive (GA) with Jones-Diggle edge-correction; diffusion-variable-rate (VR); and diffusion-lagged-arrival (LA). We employ Abramson adaptation [1], using Terrell's oversmoothing rule-of-thumb [68]—an asymptotic result providing the maximal amount of smoothing compatible with the estimated scale of the data—to set both global and pilot bandwidths to 3.2 km (the pilot densities for each estimate are found using the corresponding fixed-bandwidth estimators). Recent work in [28] showed that good practical performance is obtained if the global and pilot bandwidths are chosen to be equal.

The top row of Figure 10 shows the three estimates of intensity of PBC cases, using a common, logarithmic colour scale. The bottom row shows the ratios of each pair of estimates, again on a common, logarithmic scale. We note an overall similarity between the lagged-arrival diffusion and the Gaussian estimates,

attributable to their similarity in structure. The three estimates differ most at the region boundary, with the lagged-arrival estimate exhibiting a generally higher intensity close to the edges. This is a natural consequence of the behaviour of the diffusion, which spreads mass along the boundary rather than losing mass. Given that human settlements tend to spread along coastlines, this may be more realistic than the GA estimate.

The variable-rate estimate is markedly different from the other estimates, due to its tendency to reproduce the appearance of the pilot density. This highlights a critical practical consideration: The lagged-arrival adaptive estimator, based only on the values of the pilot estimate at the observation locations, is less prone than the variable-rate estimator to adverse effects arising from misspecification of the pilot density. We amplify this finding in the supplementary materials.

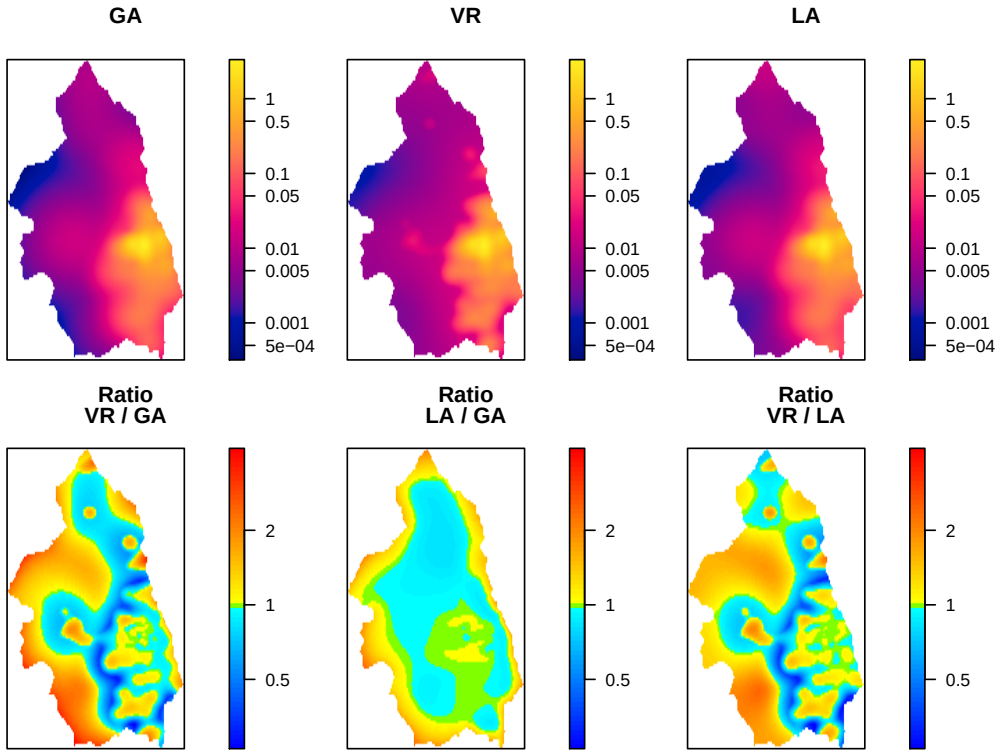


FIG 10. *Adaptive intensity estimates of the PBC case data (top row) and pairwise difference surfaces (bottom row); logarithmic colour scales; common colour scales in each row.*

## 8. SIMULATION EXPERIMENTS ON BANDWIDTH SELECTION

We conducted simulation experiments to measure the performance of likelihood cross-validation (15) for bandwidth selection. The spatial domain $W$ was taken to be the Kaipara Harbour region in the right panel of Figure 5, because this produced such striking differences between the Gaussian and diffusion kernel estimates of the pā data.

The top row of Figure 11 shows five synthetic intensity functions on $W$, each scaled to integrate to exactly 500. Scenario **S1** is a mixture of Gaussian densities with a small uniform constant. Scenario **S2** is a function taking large values near the region boundary, and defined as a quadratic function of distance to

the boundary. Scenario **S3** is a rescaled version of the Gaussian kernel density estimate (with Jones-Diggle edge-correction) of the original pā observations in this window. Scenario **S4** is a rescaled version of the diffusion kernel density estimate of the same data, with bandwidth equivalent to that used in S3. The final scenario **S5** is a single realisation of a stationary and isotropic log-Gaussian Random Field with exponential correlation. This realisation is held fixed for the entire experiment — that is, a new intensity is not generated at each iteration of the simulations, only a new dataset is sampled. Full details of these functions are given in the online supplementary material.

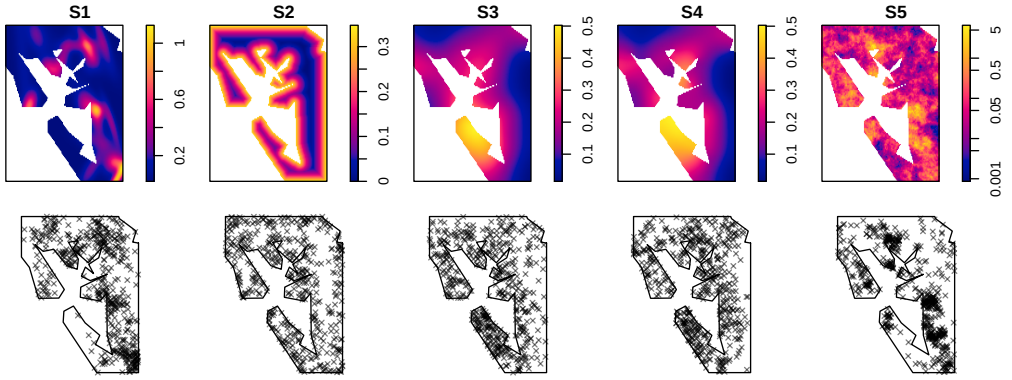


FIG 11. *Synthetic scenarios S1 to S5 for the simulation study; S5 is shown on a log-colour scale for visibility.* Top row: *Intensity functions.* Bottom row: *Correspondingly generated example datasets.*

We run the simulation study for 1000 iterations for each scenario. The LCV bandwidth selector is deployed to select two bandwidths: one using the Gaussian kernel smoother (with Jones-Diggle correction), and the other using the diffusion smoother. Using these two bandwidths we compute three estimates of the target density: the Gaussian estimate using the bandwidth selected by LCV applied to Gaussian estimates (G); a diffusion estimate using the bandwidth selected by LCV applied to diffusion estimates (D); and a diffusion estimate using the bandwidth selected by LCV applied to Gaussian estimates (Dg). Integrated squared error (ISE) with respect to the true scenario is computed.

One reason for considering the hybrid scheme Dg is pragmatic: if the bandwidths selected by likelihood cross-validation are similar whether we use the Gaussian kernel or the (far more computationally expensive) diffusion kernel, then we might recommend using the Gaussian kernel for bandwidth-selection purposes. This would be justifiable at least for small bandwidths, for which the diffusion kernel is approximately Gaussian, but it remains to be seen whether this works for larger bandwidths.

Figure 12 shows the distributions of optimal bandwidths for both versions of the selector for each scenario, as well as the ISEs for the three density estimates. Examining the top row of selected bandwidths, we see those selected based on the leave-one-out diffusion estimates are quite comparable to their Gaussian counterparts. That said, in S2, the 'edge-heavy' scenario, the diffusion-based selections appear slightly larger on average, suggesting edge effects can indeed play a role in optimising the bandwidth; it would seem the diffusion-based LCV procedure does not "shy away" from selecting larger bandwidths if necessary, due to the

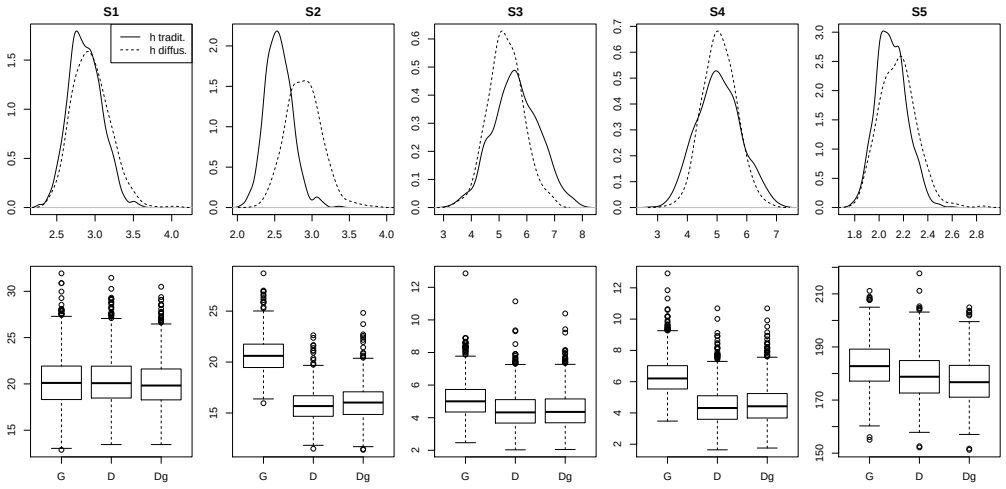


FIG 12. *Results of the simulation study.* Top row: *Distributions of the selected bandwidths under the LCV criterion for both Gaussian and diffusion estimators.* Bottom row: *ISEs relative to the true density, for the Gaussian estimate with corresponding LCV bandwidth (*G*); the diffusion estimate with diffusion bandwidth (*D*); and diffusion estimate with Gaussian bandwidth (*Dg*).*

unique style of reflective edge correction inherent in the heat kernel.

Turning to the ISEs, we see the diffusion estimates (both D and Dg) significantly outperform the Gaussian estimate (G) where we might expect it to (i.e. S2 and S4). We also see the diffusion estimates outperform the Gaussian kernel estimator in S3 and S5, albeit to a lesser degree, with results more comparable in the results for S1. Comfortingly, the performances of Dg still result in good ISE performance relative to the other estimates, which indicates using a Gaussian-selected bandwidth in a diffusion estimate is an acceptable strategy when computational expense might prohibit a diffusion-selected bandwidth.

Overall, the simulations point to the diffusion estimator performing well against the Gaussian estimator. This may be particularly evident in situations where the study region is highly irregular.

## 9. DISCUSSION

Arguably the most attractive feature of diffusion smoothing is that it is intrinsically adapted to the study domain. It can be regarded as a more rigorous version of existing *ad hoc* approaches, including reflected-kernel corrections for edge effects for kernel density estimates on the positive half-line ([11, 42, 59, 32, 40], [63, p. 26]) and inverse-path-weighted-distance methods [66]. Diffusion smoothing does not require edge correction, and the resulting estimates are usually plausible in the application context. The diffusion estimate has desirable sample properties such as conservation of mass, and desirable statistical properties such as unbiasedness for the uniform density.

By design, the diffusion estimate always satisfies the Neumann boundary condition (5) that, along the boundary of the study region, the estimate has zero slope in the direction normal to the boundary. This may be unrealistic in some applications, and may lead to statistical under-performance when the true density or intensity does not satisfy the same condition. Unlike Gaussian kernel estimates, diffusion estimates are also highly sensitive to errors in the connectivity of the

study domain. This is especially relevant when a study region is made up of highly irregular polygons, which may include islands, holes and thin peninsulas.

In the density-estimation literature, it is common to assume that the point process is Poisson (or at least that, conditional on the number of points, the locations are i.i.d.). This is not necessary for any of the results stated here. We obtained explicit formulas for the mean and variance of the diffusion estimator for a general point process. It is well-known that it can be difficult or impossible to distinguish inferentially between clustering and inhomogeneity from a single realisation of a point process [7]. However, this paper concerns moment estimation rather than inference. Unbiased estimation in the presence of correlation is commonplace in many fields. In spatial statistics, it is common to estimate the $K$-function or pair correlation function in the presence of inhomogeneous first-order intensity [2, 34]. This is theoretically justified in the case of a Cox process [3, Chap. 12]. Finally, statistical inference for general point processes is fully supported when replicated point pattern data are available [8, 33].

We investigated two kinds of adaptive diffusion smoothing: the "variable-rate" diffusion proposed by Botev *et al.* [12], and our new "lagged-arrival" approach. Variable-rate diffusion estimates tend to reproduce the appearance of the pilot estimate, and they converge to the pilot as the global bandwidth increases. In practice, the pilot density is often computed using a fixed-bandwidth estimate; in such a case, while it is tempting to think of the variable-rate diffusion estimate as 'adaptive', its final features borrow strongly from a fixed-bandwidth estimate. An analogy is to think of the pilot density estimate for variable-rate diffusions as a strong, informative Bayesian prior.

Lagged-arrival diffusion, based only on the values of the pilot estimate *at the observation locations*, appears to be more robust against potential errors in the pilot, whereas *all* weaknesses of the pilot are inherited by a variable-rate diffusion. On the other hand, the lagged arrival estimator is sensitive to errors in point locations if these would cause substantial changes in individual bandwidths.

Computation time for a diffusion estimate is much slower than for the Gaussian kernel estimator, because the latter can be computed rapidly using the Fast Fourier Transform [62]. In this paper we used the Euler scheme, because it is convenient for exposition, easy to implement, and corresponds in special cases to the algorithm of Barry and McIntyre [6]. The Euler scheme is known to perform poorly on the class of "stiff" partial differential equations, which includes the heat equation. Performance is greatly improved by using Richardson extrapolation. Alternatives to the Euler scheme should also be explored; they include Galerkin methods [14] and the method of lines [58]. Diffusion methods are well-established in computer image analysis and in medical imaging; existing algorithms in those fields could be helpful, especially for bandwidth selection.

Diffusion smoothing in other spaces is worth attention. Extensions to three-dimensional space and space-time are theoretically straightforward (indeed they are covered by the theory in Section 6), but their practical application would involve further computational challenges, and methodological questions about the treatment of smoothing along the different coordinate axes. Diffusion smoothing on linear networks has already been developed [51, 52]; diffusion smoothing of point patterns observed on the surface of a sphere is important.

Density estimators can be extended to estimators of spatially-varying relative

risk, spatial segregation and smooth regression. Diffusion estimators for these tasks will be studied in a forthcoming article.

Other potential research topics include bandwidth selection, Choi-Hall data sharpening [19] and the use of Langevin diffusions (Section 6.2.4).

*Acknowledgements* Data analysis was performed in the R language using the contributed packages `spatstat` [4, 3] and `sparr` [29]. Software implementations of the algorithms described here are now included in `spatstat`. The simulations also used the packages `spagmix` [54] and `doParallel` [21]. Code scripts to perform all the calculations in this paper are available as an online supplement.

Funding was received from the Australian Research Council discovery grants DP130104470 (Baddeley) and DP130102322 (Baddeley, Rakshit, Nair); the Grains Research and Development Corporation and the University of Western Australia (Rakshit); and Royal Society of New Zealand Marsden Fund grants 15-UOO-192 and 19-UOO-191 (Davies).

## REFERENCES

[1] Abramson, I. S. (1982). On bandwidth estimation in kernel estimates – a square root law. *Annals of Statistics* **10** 1217–1223.

[2] Baddeley, A., Møller, J. and Waagepetersen, R. (2000). Non- and semiparametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* **54** 329–350.

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

[4] Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* **12** 1–42. URL: `www.jstatsoft.org`, ISSN: 1548-7660.

[5] Baddeley, A., Turner, R. and Rubak, E. (2016). Adjusted composite likelihood ratio test for spatial Gibbs point processes. *Journal of Statistical Computation and Simulation* **86** 922–941.

[6] Barry, R. P. and McIntyre, J. (2011). Estimating animal densities and home range in regions with irregular boundaries and holes: a lattice-based alternative to the kernel density esimator. *Ecological Modelling* **222** 1666–1672.

[7] Bartlett, M. S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika* **51** 299–311.

[8] Bell, M. and Grunwald, G. (2004). Mixed models for the analysis of replicated spatial point patterns. *Biostatistics* **5** 633–648.

[9] Berman, M. and Diggle, P. (1989). Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society, Series B* **51** 81–92.

[10] Bithell, J. F. (1990). An application of density estimation to geographical epidemiology. *Statistics in Medicine* **9** 691–701.

[11] Boneva, L. I., Kendall, D. G. and Stefanov, I. (1971). Spline transformations: three new diagnostic aids for the statistical data-analyst (with discussion). *Journal of the Royal Statistical Society, Series B* **33** 1–70.

[12] Botev, Z. I., Grotowski, J. F. and Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics* **38** 2916–2957.

[13] Boyce, W. E. and DiPrima, R. C. (1969). *Elementary Differential Equations and Boundary Value Problems*, Second ed. Wiley, New York.

[14] Brenner, S. and Scott, R. (2007). *The Mathematical Theory of Finite Element Methods*. Springer, New York.

[15] Brezinski, C. and Zaglia, M. R. (1991). *Extrapolation Methods. Theory and Practice*. North-Holland, Amsterdam.

[16] Butcher, J. C. (2003). *Numerical Methods for Ordinary Differential Equations*, Third ed. Wiley, Chichester.

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

[18] Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *Annals of Statistics* **28** 408–428.

[19] Choi, E. and Hall, P. (2001). Nonparametric analysis of earthquake point-process data. In *State of the Art in Probability and Statistics: Festschrift for Willem R. van Zwet* (M. de Gunst, C. Klaassen and A. van der Vaart, eds.) 324–344. Institute of Mathematical Statistics, Beachwood, Ohio.

[20] Chung, M. K., Qiu, A., Seo, S. and Vorperian, H. K. (2015). Unified heat kernel regression for diffusion, kernel smoothing and wavelets on manifolds and its application to mandible growth modeling in CT images. *Medical Image Analysis* **22** 63–76.

[21] Microsoft Corporation and Weston, S. (2019). doParallel: Foreach Parallel Adaptor for the 'parallel' Package R package version 1.0.15.

[22] Cressie, N. A. C. (1993). *Statistics for Spatial Data*, Second ed. John Wiley and Sons, New York.

[23] Cronie, O. and van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika* **105** 455–462.

[24] Daley, D. J. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York.

[25] Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods*, Second ed. Springer-Verlag, New York.

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

[27] Davies, T. M., Flynn, C. R. and Hazelton, M. L. (2018). On the utility of asymptotic bandwidth selectors for spatially adaptive kernel density estimation. *Statistics and Probability Letters* **138** 75–81.

[28] Davies, T. M. and Lawson, A. B. (2019). An evaluation of likelihood-based bandwidth selectors for spatial and spatiotemporal kernel estimates. *Journal of Statistical Computation and Simulation* **89** 1131–1152.

[29] Davies, T. M., Marshall, J. C. and Hazelton, M. L. (2018). Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk. *Statistics in Medicine* **37** 1191-1221.

[30] Diggle, P. J. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **34** 138–147.

[31] Diggle, P. J. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, Third ed. Chapman and Hall/CRC, Boca Raton, FL.

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

[33] Diggle, P. J., Mateu, J. and Clough, H. E. (2000). A comparison between parametric and non-parametric approaches to the analysis of replicated spatial point patterns. *Advances in Applied Probability (SGSA)* **32** 331–343.

[34] Diggle, P. J., Rowlingson, B. and Su, T. L. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* **16** 423–434.

[35] Doyle, P. G. and Snell, J. L. (1984). *Random Walks and Electric Networks*. American Mathematical Society.

[36] Duong, T. and Hazelton, M. L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics* **15** 17–30.

[37] Duong, T. and Hazelton, M. L. (2005). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis* **93** 417–433.

[38] Duong, T. and Hazelton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics* **32** 485–506.

[39] Fukushima, M. (1967). A construction of reflecting barrier Brownian motions for bounded domains. *Osaka Journal of Mathematics* **4** 183–215.

[40] Ghosh, B. K. and Huang, W. M. (1992). Optimum bandwidths and kernels for estimating certain discontinuous densities. *Annals of the Institute of Statistical Mathematics* **44** 563–577.

[41] Hall, P. and Marron, J. S. (1988). Variable window width kernel estimation of probability densities. *Probability Theory and Related Fields* **80** 37–49.

[42] Hominal, P. and Deheuvels, P. (1979). Estimation non paramétrique de la densité compte-tenu d'informations sur la support. *Revue Statistique Appliqué* **27** 47–68.

[43] Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley and Sons, Chichester.

[44] Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing* **3** 135–146.

[45] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91** 401–407.

[46] Koenderink, J. J. (1984). The structure of images. *Biological Cybernetics* **50** 363–370.

[47] Lindeberg, T. (1994). *Scale Space Theory in Computer Vision*. Kluwer, Boston, USA.

[48] Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

[49] Loader, C. (1999). Bandwidth selection: classical or plug-in? *Annals of Statistics* **27** 415–438.

[50] Marshall, J. C. and Hazelton, M. L. (2010). Boundary kernels for adaptive density estimators on regions with irregular boundaries. *Journal of Multivariate Analysis* **101** 949–963.

[51] McSwiggan, G., Baddeley, A. and Nair, G. (2016). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics* **44** 324–345.

[52] McSwiggan, G., Baddeley, A. and Nair, G. (2019). Estimation of relative risk for events on a linear network. *Statistics and Computing* **30** 469–484. Published online 24 august 2019.

[53] Prince, M. I., Chetwynd, A., Diggle, P., Jarner, M., Metcalf, J. V. and James, O. F. W. (2001). The geographical distribution of primary biliary cirrhosis in a well-defined cohort. *Hepatology* **34** 1083–1088.

[54] Redmond, A. K. and Davies, T. M. (2018). spagmix: Artificial Spatial and Spatiotemporal Densities on Bounded Windows R package version 0.3-1.

[55] Richardson, L. F. (1911). The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London, Series A* **210** 307–357.

[56] Richardson, L. F. and Gaunt, J. A. (1927). The deferred approach to the limit. Part I. Single lattice. Part II. Interpenetrating lattices. *Philosophical Transactions of the Royal Society of London, Series A* **226** 299–361.

[57] Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association* **89** 807–817.

[58] Schiesser, W. E. (1991). *The Numerical Method of Lines*. Academic Press, New York.

[59] Schuster, E. F. (1985). Incorporating support constraints into non-parametric estimators of densities. *Communications in Statistics — Theory and Methods* **14** 1123–1136.

[60] Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization*. John Wiley and Sons, New York.

[61] Serra, J. (1982). *Image Analysis and Mathematical Morphology*. Academic Press, London.

[62] Silverman, B. W. (1982). Kernel density estimation using the fast Fourier transform. *Applied Statistics* **31** 93–99.

[63] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[64] Smith, B. (2018). Fantastic Pā and Where to Find Them: A Spatial Analysis of the pre-European Pā of Aotearoa, Master's thesis, University of Otago, Dunedin, New Zealand Retrieved from http://hdl.handle.net/10523/8524.

[65] Smith, B. A., Thomas, T. and Davies, T. M. (2021). Statistical approaches to spatial variation in conflict and social organisation: Assessing the emergent distribution of fortification works (pā) in New Zealand/Aotearoa. Submitted for publication.

[66] Stachelek, J. and Madden, C. J. (2015). Application of inverse path distance weighting for high-density spatial mapping of water quality patterns. *International Journal of Geographical Information Science* **29** 1240–1250.

[67] Stroock, D. W. and Varadhan, S. R. S. (1971). Diffusion processes with boundary conditions. *Communications in Pure and Applied Mathematics* **24** 147–225.

[68] Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association* **85** 470–476.

[69] Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation. *Annals of Statistics* **20** 1236–1265.

[70] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall.

[71] Weickert, J. (1997). *Anisotropic Diffusion in Image Processing*. Teubner, Stuttgart.

[72] Zhang, X., King, M. L. and Hyndman, R. J. (2006). A Bayesian approach to bandwidth selection for multivariate kernel estimation. *Computational Statistics and Data Analysis* **50** 3009–3031.