# Kernel density estimation based on Ripley's correction

Arthur Charpentier, Ewen Gallic

# KERNEL DENSITY ESTIMATION BASED ON RIPLEY'S CORRECTION

ARTHUR CHARPENTIER AND EWEN GALLIC

Abstract. In this paper, we investigate a technique inspired by Ripley's circumference method to correct bias of density estimation of edges (or frontiers) of regions. The idea of the method was theoretical and difficult to implement. We provide a simple technique – based of properties of Gaussian kernels – to efficiently compute weights to correct border bias on the frontiers of the region of interest, with automatic selection of an optimal radius for the method. We illustrate the use of that technique to visualize hot spots of car accidents and campsite locations, as well as location of bike thefts.

Keywords: border bias; edge correction; frontier; GIS; kernel density estimation; polygons; Ripley's circumference method; spatial process; visualization

Arthur Charpentier, UQAM, 201, avenue du Président-Kennedy, Montréal (Québec), Canada H2X 3Y7 (corresponding author) charpentier.arthur@uqam.ca, and Ewen Gallic, CREM UMR CNRS 6211, Université de Rennes 1, Campus Centre, CS 86514, 7 Place Hoche, 35065 Rennes Cedex, France.

## 1. Introduction and motivation

Visualizing the density of a spatial process is not only a preliminary step in a spatial analysis, but is also useful for reporting results in a simple and understandable way. Flexible techniques for geographically visualizing data, and occurrences of a spatial process, are necessary. Kernel smoothing has always been a popular technique to estimate a density. Nevertheless, as mentioned in Bailey (1994), "*kernel smoothing over irregular areal units provides difficulties.*" Edge corrections are neceesary to avoid misinterpretations. $K$-functions, introduced in Ripley (1976), can be used to compute quantities with an edge correction, taking into account boundary configurations of a specific area. But as mentioned in Zheng et al. (2004), "*Current algorithms for edge-correction are either difficult to apply or computationally expensive, especially for complex borders.*"

This paper addresses this challenge by developing a simple and efficient correction for kernel density estimation, inspired by Ripley's circumferential method (described in Ripley (1976), Ripley (1977) and Ripley (1981)). In kernel density estimation, we simply count the number of observations in the neighborhood of a given location: the closer an observation, the larger the weight. The shape of the weight function is the kernel, and the length of the neighborhood (also called '*sphere of influence*' in Hearnshaw et al. (1994)) is the bandwidth parameter. Since Epanechnikov (1969) proved that statistical results were not (significantly) affected by the choice of kernel function, most of the authors have emphasized the fact that bandwidth's choice is the crucial issue in this problem. The most popular kernel is the Gaussian one since a dual representation (occurence's locations observed with a random noise) can be used. Nevertheless, if such kernel estimators are easy to compute, and satisfy good statistical properties, Yamada and Rogerson (2003) recall that this methodology suffers a so-called "*edge effect*" also known in statistical literature as "*border bias.*" Yamada and Rogerson (2003) mention Ripley's circumference method (from Ripley (1981)), but claim that "*Ripley's method could be too complicated without proper software or skilled*

*programmers*." In this paper we explain how to use that technique, and provide a way to select the "*optimal*" circumferential parameter.

In this paper, basics on space kernel density estimation are recalled in Section 2. Notations and heuristics of kernel density estimation are given in Section 2.1 and 2.2. In Section 2.3, we discuss the optimal choice of the bandwidth. Section 2.4 provides a discussion on frontiers and space border bias correction. More sepcifically, Ripley's circumferential method (from Ripley (1976) and Ripley (1977)) is described in Section 2.5 and again, heuristics on the interpretation of the weights (in the context of kernel density estimation) are given in Section 2.6. In Section 3, we discuss the link between the radius $r$ used in the circumferential method, and bandwidth $h$ of the kernel smoother. Using Monte Carlo simulation, given a bandwidth $h$, we show that there is an "optimal" radius $r^\star(h)$. Using either a $L_1$ or $L_2$ norm (minimizing either the sum of absolute values of errors or sum of squares of errors) we see that $r^\star(h)$ is linear in $h$. This property (analytically derived for half space regions) allows us to introduce an automatic technique. Sections 4 to 6 feature illustrations of this technique, through three examples (see Figure 1 for a quick visualization of the spatial distribution of observations). Section 4 offers an estimation of the density of the location of bodily injury car accidents, in western France (Morbihan and Finistère). The impacts of the correction are examined (with respect to standard kernel density estimation) and the technique is used to identify hot spots. Section 5 presents an estimation of the density of the location of bike thefts in San Francisco. The interpretation of the density is tackled in that section. An estimation of the density of the locations of campsites is presented in Section 6. Finally, Section 7 concludes this article.[1]

## 2. On kernel density estimation

2.1. **Interpretation of spatial density.** In the context of a spatial process, with $n$ locations $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n$, the interpretation of the value of the density is the following.

[1]Codes used in this article to compute $\widehat{f}$ and visualize it on a map are described (and fully available) on `https://github.com/ripleyCorr/Kernel_density_ripley`.

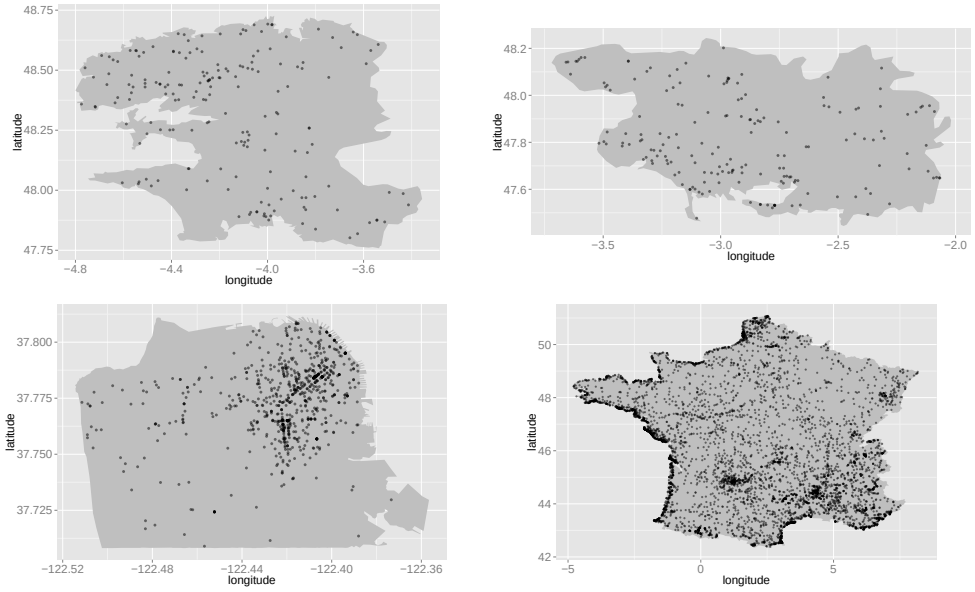



FIGURE 1. Illustrations: car accident locations (Section 4) in Western part of France (Finistère and Morbihan), on top, bike theft locations (Section 5) in San Francisco in the bottom-left panel, and campsite locations (Section 6) in France in the bottom-right panel.

For any region $\mathcal{E}$,

$$\mathbb{P}(\boldsymbol{Z} \in \mathcal{E}) = \int_{\mathcal{E}} f(\boldsymbol{z}) d\boldsymbol{z},$$

where $f(\boldsymbol{z})d\boldsymbol{z}$ is usually interpreted as the probability of $\boldsymbol{Z}$ falling within the infinitesimal (rectangular) region $[\boldsymbol{z}, \boldsymbol{z} + d\boldsymbol{z}]$, which is the area between $(x, y)$ and $(x + dx, y + dy)$, when $d\boldsymbol{z}$ is small. Here, units of the projection coordinates used to locate $\boldsymbol{z}$ are 1° (111.11 km) times 1° (111.11 km on the Equator, but 87.8 km in San Francisco for instance, and more generally 111.11 times $\cos(y)$). In the San Francisco area (discussed in Section 5) the unit corresponds to a $9,758\ km^2$ area, for instance.

It is possible to relate the density $f(\boldsymbol{z})$ to the expected number of observations that should occur in a neighborhood of $\boldsymbol{z}$. At location $\boldsymbol{z} = (x, y)$, the expected number of observations within a distance $r$ of location $\boldsymbol{z}$ (in $km$) is

$$\frac{n \times f(\boldsymbol{z}) \times \pi r^2}{111.11^2 \times \cos(y)},$$

In the context of San Francisco, a density $f(\boldsymbol{z}) = 100$ means that about $n/250$ observations should be within a 500 $m$ distance of $\boldsymbol{z}$. Hence, the density $f(\boldsymbol{z})$ can easily be related to the expected number of observations of the spatial process within a given distance to $\boldsymbol{z}$. See Section 5.2 for a longer discussion on the interpretation of the density.

2.2. **Definitions and notations.** Kernel density estimation (see Scott (2009), Silverman (1986), Wand and Jones (1994)) is a standard statistical technique to estimate a smooth probability density function. It has been extended from univariate distributions (on the real line) to multivariate distributions, including spatial and spatio-temporal models. Spatial observations are based on spatial locations $\boldsymbol{Z}_i = (X_i, Y_i)$ (usually characterized by a latitude and a longitude). Based on a sample $\{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n\}$ the estimation of the density at point $\boldsymbol{z} = (x, y)$ is

$$\widehat{f}_{\boldsymbol{H}}(\boldsymbol{z}) = \frac{1}{n}\det(\boldsymbol{H})^{-1}\sum_{i=1}^{n} K\left(\boldsymbol{H}^{-1}(\boldsymbol{z} - \boldsymbol{Z}_i)\right), \tag{2.1}$$

where $K$ is some symmetric (centered) kernel function, and $\boldsymbol{H}$ a bandwidth parameter. A popular kernel is the quadratic one, introduced by Epanechnikov (1969), used *e.g.* in ArcGIS, and defined as

$$K(u, v) = \frac{2}{\pi}(1 - [u^2 + v^2])\mathbf{1}(u^2 + v^2 \in [0, 1)). \tag{2.2}$$

An alternative is to consider Gaussian kernels, *i.e.* $K_Z$ is the density of a Gaussian random vector,

$$K(u, v) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[u^2 + v^2 - 2\rho uv\right]\right). \tag{2.3}$$

For convenience, it is common to consider a kernel with independent components, and a diagonal bandwidth parameter, where values are identical if the spatial process is homogeneous in both directions:

$$\widehat{f}_h(\boldsymbol{z} = (x, y)) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \cdot K\left(\frac{y - Y_i}{h}\right). \tag{2.4}$$

2.3. **On optimal bandwidth.** In the context of product of (symmetric) kernels, one can prove using Taylor's expansion, that

$$\mathbb{E}[\widehat{f}_h(\boldsymbol{z})] \sim f(\boldsymbol{z}) + \alpha_1 \left( \frac{h_X^2}{2} \frac{\partial^2 f}{\partial x^2} f(\boldsymbol{z}) + \frac{h_Y^2}{2} \frac{\partial^2 f}{\partial y^2} f(\boldsymbol{z}) \right) \text{ and } \text{Var}[\widehat{f}_h(\boldsymbol{z})] \sim \frac{\alpha_2}{n h_X h_Y} f(\boldsymbol{z})^2,$$

where $\alpha_2$ and $\alpha_2$ are parameters related to the shape of the kernel function, see Wand and Jones (1994) and Scott (2009) for more details on the exact value of those parameters. The mean integrated squared error is then

$$\text{MISE}(h) = \mathbb{E}\left[ \int [\widehat{f}_h(\boldsymbol{z}) - f(\boldsymbol{z})]^2 d\boldsymbol{z} \right] \sim \alpha_3 h^4 + \frac{\alpha_4}{nh}$$

so $h^\star = \text{argmin}\{\text{MISE}(h)\}$ is $\alpha n^{-1/5}$ for some constant $\alpha$ (function of the kernel as well as the true – unknown – density $f$). In the case where the true density $f$ is a Gaussian distribution, with a diagonal variance matrix $\Sigma$, Silverman's rule of thumb can be used (see Silverman (1986) and Scott (2009) for a discussion)

$$h_i^\star \sim \left( \frac{2}{3} \right)^{\frac{1}{6}} \cdot \sigma_i \cdot n^{-1/6}.$$

2.4. **Frontier and space border bias.** Kernel density estimation is a popular technique to visualize unbounded smoothed densities. Yet in some specific cases, observations have to lie within some specific area $\mathcal{S}$. For instance, for traffic accidents or bike thefts, events have to occur on-land, as discussed in Sections 4 and 5, respectively. $\mathcal{S}$ would stand for some on-land territory. On the contrary, when locating fish or sea animals using GPS trackers, it is known that those animals have to be in the sea. In that case, $\mathcal{S}$ would represent some territorial sea.

In the case where $\mathcal{S}$ is bounded, kernel estimates (with symmetric kernels) have two important drawbacks:

- the total weight is not equal to 1, leading to an incorrect probability distribution function,[2] *i.e.* $\int_{\mathcal{S}} \widehat{f}(\boldsymbol{z}) d\boldsymbol{z} < 1$,

[2]In standard statistical packages, the estimations are usually normalized so that the overall mass (on the area where the density is computed) is equal to 1. A multiplicative coefficient is applied uniformly on the whole area, while a local adjustment is obviously necessary: the density is still underestimated on the edge $\partial\mathcal{S}$.

- close to the frontier $\partial\mathcal{S}$, $\widehat{f}$ has a multiplicative bias, *i.e.* $\mathbb{E}[\widehat{f}(\boldsymbol{z})] = \kappa_{\boldsymbol{z}} \cdot f(\boldsymbol{z})$, where $\kappa_{\boldsymbol{z}} \in [0,1]$ depends on the shape of the border $\partial\mathcal{S}$ in the neighborhood of $\boldsymbol{z}$.

As shown in Section 3, in the context of on-land events for regions close to the sea, estimators of density can behave very defectively. The idea here is to propose a methodology that gives an estimator $\widehat{f}$ that can be associated with a proper probability distribution function, and that does not present border bias.

2.5. **Ripley's circumferential correction.** The (univariate) kernel density estimator for a uniform kernel – also called 'moving histogram' – is defined as

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}(d(x, X_i) \leq h), \text{ where } d(x, X_i) = |x - X_i|.$$

It simply consists in counting the number of observations within a distance $h$ of the arbitrary point $x$. In the context of a two-dimensional spatial density, Ripley (1976) extended the notion above by implementing the $K$ function, where only observations within a distance $h$ of an arbitrary point $(x, y)$ were considered:

$$\widehat{f}_h(\boldsymbol{z}) = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}(d(\boldsymbol{z}, \boldsymbol{Z}_i) \leq h), \text{ where } d^2(\boldsymbol{z}, \boldsymbol{Z}_i) = (x - X_i)^2 + (y - Y_i)^2.$$

Ripley (1977) suggested a "*proper edge correction method*" that we can write – using our own notations for consistency

$$\widehat{f}_h(\boldsymbol{z}) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{\omega_i(\boldsymbol{z})}\mathbf{1}(d(\boldsymbol{z}, \boldsymbol{Z}_i) \leq h), \text{ where } d^2(\boldsymbol{z}, \boldsymbol{Z}_i) = (x - X_i)^2 + (y - Y_i)^2,$$

where the weight, $\omega(\boldsymbol{z})$ is defined as the proportion of a circumference of a circle centered at point $\boldsymbol{z}$ that lies within the study area $\mathcal{S}$. As claimed in Yamada and Rogerson (2003) "*although it is difficult to derive $\omega$'s analytically for an arbitrarily shaped study area, it would still be possible to derive it numerically using GIS.*" This method is called Ripley's circumference method in Cressie (1992) and Bailey and Gatrell (1995).

2.6. **Interpretation of the weight-based correction.** Recall that kernel estimators of densities can be seen as the expected value of the density for sample $\{\widetilde{\boldsymbol{Z}}_i = \boldsymbol{Z}_i + \boldsymbol{\varepsilon}_i\}$ where $\boldsymbol{\varepsilon}_i$'s are i.i.d. random noises, independent of the observations, as in Davis (1975), Tapia and Thompson (1978) or Stefanski and Carroll (1990).Further, the empirical cumulative distribution function is the step function defined as

$$\widehat{F}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\boldsymbol{Z}_i \leq \boldsymbol{z}), \tag{2.5}$$

and the associated empirical measure is

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{Z}_i}(\boldsymbol{z}), \tag{2.6}$$

where $\delta$ denotes the Dirac measure. The idea of Kernel-based estimator is to substitute a continuous distribution for Dirac measures,

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mu_{\boldsymbol{Z}_i}(\boldsymbol{z}), \tag{2.7}$$

where $\mu_{\boldsymbol{Z}_i}$ can be the density of a Gaussian vector, centered in $\boldsymbol{Z}_i$, with variance-covariance matrix $\boldsymbol{H}$. The problem is that if the distribution of $\boldsymbol{Z}$ has a bounded support, then measure $\mu_{\boldsymbol{Z}_i}$ will spread some weight in areas where no observation can be found (outside $\mathcal{S}$). Thus, it might be natural to consider a truncated distribution, restricted to the support $\mathcal{S}$:

$$\mu_{\boldsymbol{Z}_i|\mathcal{S}}(\boldsymbol{z}) = \frac{\mu_{\boldsymbol{Z}_i(\boldsymbol{z})}}{\mu_{\boldsymbol{Z}_i}(\mathcal{S})}. \tag{2.8}$$

Thus, it is natural to consider

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mu_{\boldsymbol{Z}_i|\mathcal{S}}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \omega_i \cdot \mu_{\boldsymbol{Z}_i(\boldsymbol{z})} \text{ where } \omega_i = \mu_{\boldsymbol{Z}_i}(\mathcal{S})^{-1}. \tag{2.9}$$

If we consider a noise with circularly contoured distribution (*e.g.* a Gaussian noise, as mentioned earlier), it is possible to approximate $\mu_{\boldsymbol{Z}_i}(\mathcal{S})$ by

$$\frac{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})}{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r})}, \tag{2.10}$$

where $\mathcal{A}$ denotes the area function, and $\mathcal{D}_{\boldsymbol{Z}_i,r}$ the disk centered in $\boldsymbol{Z}_i$ with radius $r > 0$. This weight is the same as the one used in Ripley's circumference method

(from Ripley (1976)). Note that $r$ should be related to the covariance matrix $\boldsymbol{H}$ (this will be discussed in section 3), since the latter is related to the width of the neighborhood: the wider the neighborhood, the larger the radius. Thus, here, the idea is simply to use *weighted kernel estimators*:

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \omega(\boldsymbol{Z}_i) \cdot \det(\boldsymbol{H})^{-1} K\left(\boldsymbol{H}^{-1}(\boldsymbol{z}-\boldsymbol{Z}_i)\right), \tag{2.11}$$

where weights $\omega(\boldsymbol{Z}_i)$ should reflect the proportion of area around $\boldsymbol{Z}_i$ (within distance $r$) that belongs to $\mathcal{S}$. Those weighted kernel estimators have been used intensively, *e.g.* on censored data, as in Marron and Padgett (1987) (to correct censoring bias) or Gisbert (2003). As mentioned in Hall and Turlach (1999), having weights that depend only on the data ($\boldsymbol{Z}_i$'s) and not on the location ($\boldsymbol{z}$) is interesting from a computational point of view. From this assumption, and since computing the intersection of polygon areas with standard software is extremely simple, Ripley's circumferential technique can easily be implemented.

**Example 1.** *The use of weights is illustrated in Figure 2 in the univariate case: on border, the kernel is no longer the density of a Gaussian distribution centered on $X_i$, but the density of a truncated Gaussian distribution. Thus, those weights have an impact on the border of the support.*

**Example 2.** *The use of weights is illustrated in the top-left graph of Figure 3 on a non-convex polygon $\mathcal{S}$. The weights are the proportion of the disk that lie in the polygon. On the top-left graph of Figure 3 some specific locations are mentioned (they will be used later on, for an intensive simulation study). Points will be uniformly drawn within the region $\mathcal{S}$. The theoretical density can be visualized below, with the three dimensional surface on the bottom-left graph and iso-density curves on the bottom-right graph of Figure 3. An estimation based on* 500 *simulated observations (uniformly drawn on $\mathcal{S}$) can be visualized in Figure 4. Top graphs are the estimation of the density $f$ using a Gaussian kernel technique. Border bias can clearly be observed. Bottom graphs are the estimation of the density $f$ using a Gaussian kernel technique*

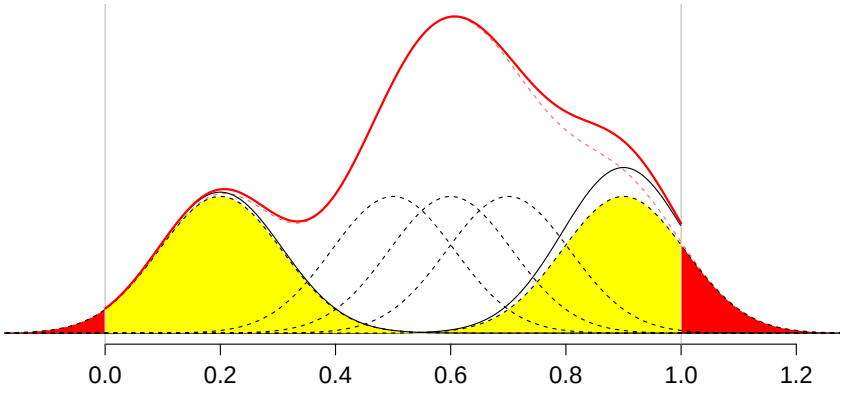



FIGURE 2. Weight correction of a density on $[0,1]$: kernel $K$ is no longer a Gaussian density, but a truncated Gaussian density.

*with Ripley's correction. The estimator is volatile (mainly because of the small sample size), but it seems much better than the previous one.*

## 3. Optimal radius $r$ for circumferential correction

With a Gaussian kernel, in the univariate case, the bandwidth $h$ is the standard deviation of the Gaussian noise $\varepsilon$ (see Chiu (1991)), and in the bivariate case, $\boldsymbol{H}$ is the covariance matrix of the noise, $\boldsymbol{\varepsilon}$. Then the *true* probability $\mu_{\boldsymbol{Z}_i}(\mathcal{S})$ is

$$\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S}) \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}). \tag{3.1}$$

3.1. **Kernel product with identical bandwidth.** A standard assumption in multivariate density estimation is to assume that $K$ is the product of two (univariate) kernels. This assumption can be interpreted as a non-correlated noise $\boldsymbol{\varepsilon}$, *i.e.* $\boldsymbol{H}$ is a diagonal matrix. From the geography of our problem, it is possible to assume further that the two components have the same 'dimension', thus, it might not be a too strong assumption to assume that $\boldsymbol{H}$ is a diagonal matrix with identical terms

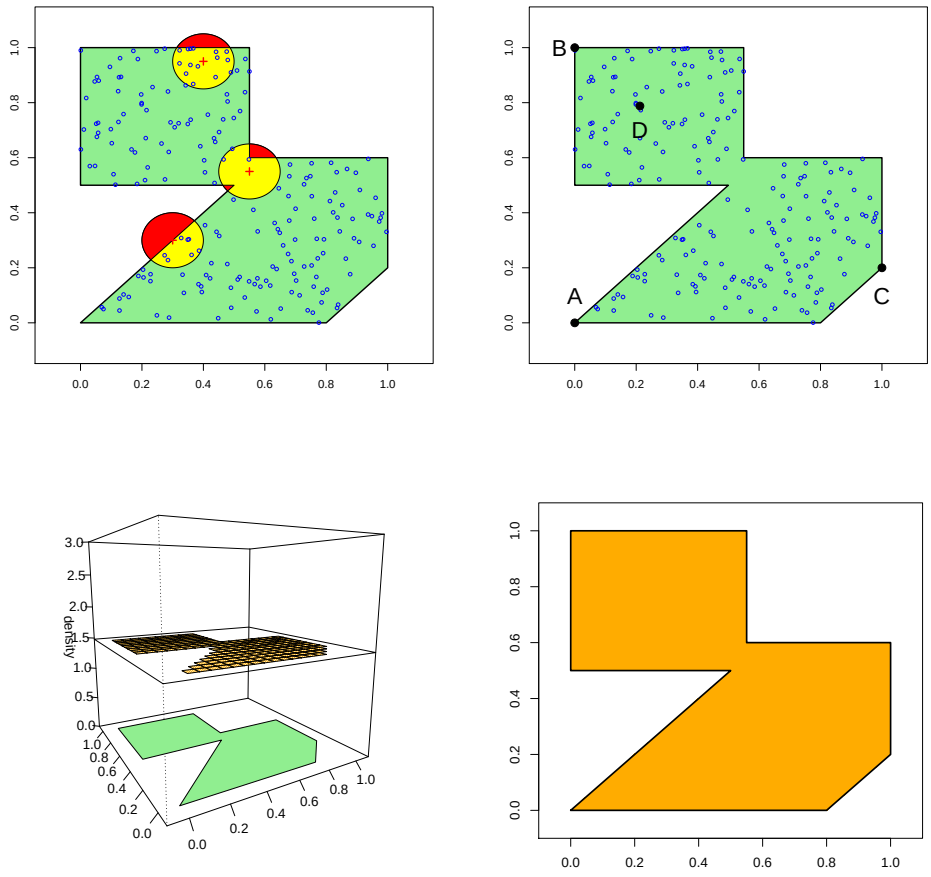



FIGURE 3. Uniform density on polygon $\mathcal{S}$ (■), where $\boldsymbol{Z}_i, r \cap \mathcal{S}$ (■)

on the diagonal. Let $h$ denote this diagonal term (this assumption will be relaxed at the end of this section), so that level curves of the density of $\boldsymbol{Z}$ are circles.

3.1.1. *Analytical computation when $\mathcal{S}$ is a half-plane.* If $\mathcal{S}$ is a half-plane, and if the distance between $\boldsymbol{Z}_i$ and the border is $\alpha$, then

$$\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S}) = 1 - \Phi(-\alpha h^{-1}) = \Phi(\alpha h^{-1}), \tag{3.2}$$

where $\Phi$ denotes the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution (see Figure 5).

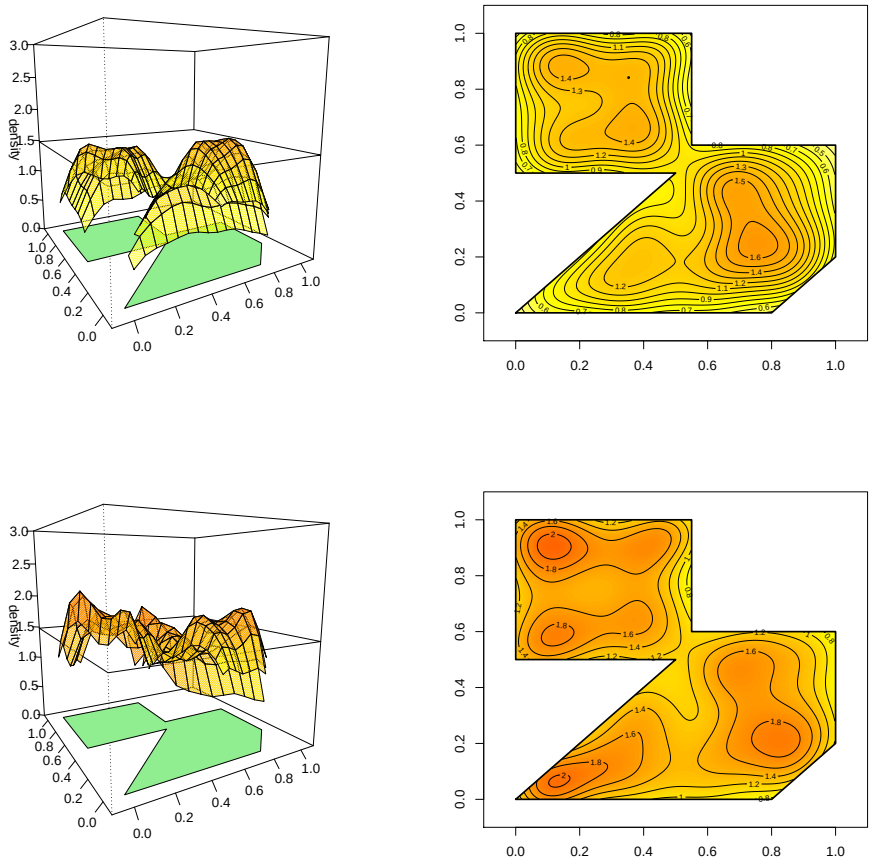



FIGURE 4. Estimation of density $f$ (uniform on polygon $\mathcal{S}$) with a standard Gaussian kernel on top, and the corrected kernel estimate below.

Assume for convenience that $h = 1$, and that $a = 1$, then the probability that $\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \notin \mathcal{S}$ is $\Phi(-1) \sim 15\%$. The proxy we suggest for $\mu_{\boldsymbol{Z}_i}(\mathcal{S})$ is to consider the following ratio:

$$\mu^{\circ}_{\boldsymbol{Z}_i}(r, \mathcal{S}) = \frac{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})}{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r})}, \tag{3.3}$$

where $\mathcal{D}_{\boldsymbol{Z}_i,r}$ is a disk centered in $\boldsymbol{Z}_i$ with radius $r$. Again, if $\mathcal{S}$ is a half-plane, it is possible to derive an analytical expression, because it will be related only to the *circular segment* (the region bounded by a chord and the arc subtended by the chord,

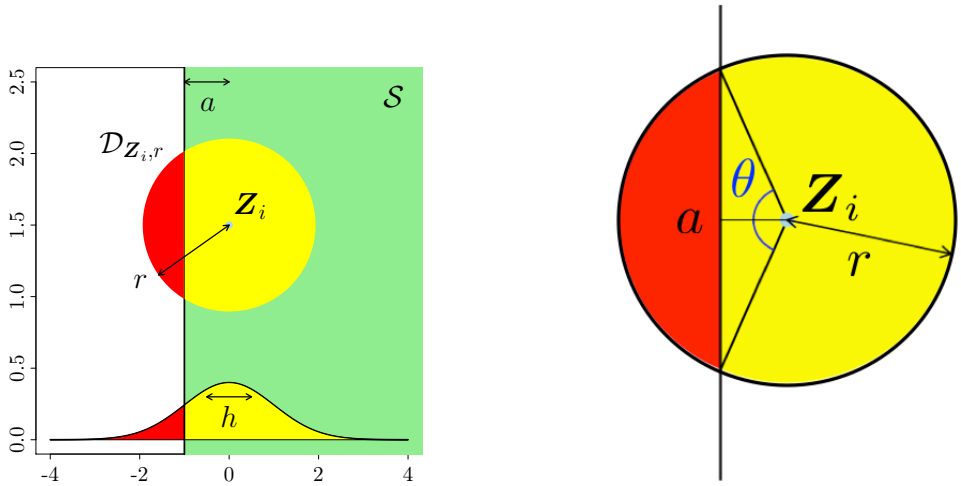



FIGURE 5. Link between $\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S})$ and $\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})$ where $\mathcal{S}$ is a half-plane.

see Figure 5). The area of the circular segment is equal to the area of the circular sector minus the area of the triangular portion:

$$\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S}) = \underbrace{\frac{\theta}{2\pi}\pi r^2}_{\text{sector area}} - \underbrace{\frac{r^2\sin(\theta)}{2}}_{\text{triangle area}} \text{ where } \cos\left(\frac{\theta}{2}\right) = \frac{a}{r}. \tag{3.4}$$

Thus,

$$\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S}) = \begin{cases} \frac{r^2}{2}[\theta - \sin(\theta)] \text{ if } a < r \\ 0 \text{ if } a > r \end{cases}. \tag{3.5}$$

From the previous computation, we would like to find $r^\star$ (or $\theta^\star$) such that $\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})$ is 15% of $\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r})$, when $a$ is equal to 1, *i.e.*

$$\frac{r^2}{2\pi r^2}[\theta - \sin(\theta)] = \frac{1}{2\pi}[\theta - \sin(\theta)] = 15\% \ (= \Phi(-1)), \tag{3.6}$$

or equivalently,

$$\theta^\star - \sin(\theta^\star) = 2\pi\Phi(-1) \sim 1,$$

thus, $\theta^\star = 1 + u$ where $u$ is the root of $\sin(1+u) = u$, which is numerically equal to 0.93. Since $\theta = 2\arccos(r^{-1})$, then $r^\star \sim 1/cos(1.93/2)$, which is numerically equal to

1.76. Therefore, with a disk with radius 1.76, the area of the circular segment located at 1 from the center of the disk is 15% of the area of the disk.

More generally (with any $a$ and $h$), if $r^\star = \beta^\star h$, the ratio of the area of the circular segment is

$$\frac{\theta^\star - \sin(\theta^\star)}{2\pi} = \frac{1}{2\pi}\left[2\arccos\left(\frac{a}{\beta^\star h}\right) - \sin\left(2\arccos\left(\frac{a}{\beta^\star h}\right)\right)\right]. \tag{3.7}$$

Let $x = ah^{-1}$ and $b = 1/\beta^\star$, then the ratio is

$$x \mapsto \frac{1}{2\pi}\left[2\arccos(bx) - \sin(2\arccos(bx))\right]. \tag{3.8}$$

Taylor's expansion (when $x$ is closed to 0) is

$$x \mapsto \frac{1}{2} - \frac{2b}{\pi}x + \frac{b^3}{3\pi}x^3 + \frac{b^5}{20\pi}x^5 + O(x^7).$$

Following Shah (1985) and Bryc (2002), Taylor's expansion of $\Phi(-x)$ is

$$\Phi(-x) \sim \frac{1}{2} - 0.368929x - 0.037758x^3 + O(x^5).$$

Therefore, linear terms are equal when $\beta^\star = 2/(0.3689\pi) \sim 1.725$. The use of a linear relationship, with a proportionality factor around 1.76 seems to be legitimate.

The intuition is that $r^\star$ might be a (linear) function of $h$, $r^\star = \beta^\star h$ where $\beta^\star \sim 1.76$, with half-plane domains. This relationship might also be a good approximation on more general spaces $\mathcal{S}$.

3.1.2. *Monte Carlo study for more complex areas $\mathcal{S}$.* To illustrate the general case, two regions are considered in this section: the polygon of Figure 3, and the contour of Finistère (the French region). In those two regions, 10,000 points are drawn uniformly $\boldsymbol{Z}_i$ (1,000 are plotted in Figure 6).

Given that $h > 0$,

- *theorerical* weights $\omega_i(h)$ are numerically computed, based on $\mu_{\boldsymbol{Z}_i}(\mathcal{S}) = \mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S})$, using Monte Carlo simulations, since $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, h\mathbb{I})$,
- $\omega_i^\circ(h)$ are computed, based on $\mu^\circ_{r,\boldsymbol{Z}_i}(\mathcal{S})$ for different values of $r$,

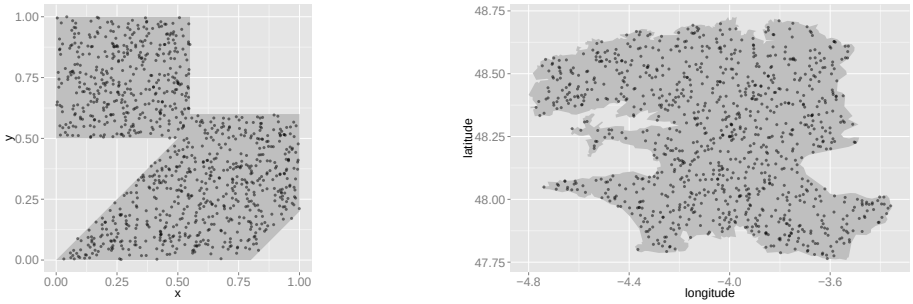



FIGURE 6. Polygon $\mathcal{S}$ on the left and the Finistère region on the right, where $1,000$ points are uniformly drawn.

- for some norm $\|\cdot\|$, the optimal radius $r^\star$ is the solution of

$$r^\star(h) = \text{argmin}\left\{\sum_{i=1}^{n} \|\omega_i^\circ(h) - \omega_i(h)\|\right\},$$

(two norms are considered in this study $\|x\|_1 = |x|$ and $\|x\|_2 = x^2$).

In Figure 7, $h \mapsto r^\star(h)$ is plotted on top, where a linear relationship can easily be identified, and below the slope, *i.e.* $h \mapsto r^\star(h)/h$. The horizontal dashed line is the 1.76 value obtained empirically in the computations of the previous section (using a half-plane region).

Thus, from bandwidth $h$, it is possible to approximate weights using

$$\omega_i^\circ(h) = \frac{\mathcal{A}(\mathcal{D}_{\boldsymbol{z}_i,r^\star})}{\mathcal{A}(\mathcal{D}_{\boldsymbol{z}_i,r^\star} \cap \mathcal{S})} \text{ where } r^\star = \beta^\star h \text{ and } \beta^\star \sim 1.76.$$

3.2. **Comparison with other corrections.** A method for edge correction of an intensity estimator was introduced in Diggle (1985), including a discussion of bandwidth estimation (see also Berman and Diggle (1989) and the estimation of relative risk (see Kelsall and Diggle (1995)). In Section 2.3, we explained how to estimate densities at various points $\boldsymbol{z}$, and we mentioned that a standard global measure to assess the quality of the fit was to use the mean integrated squared error (where the

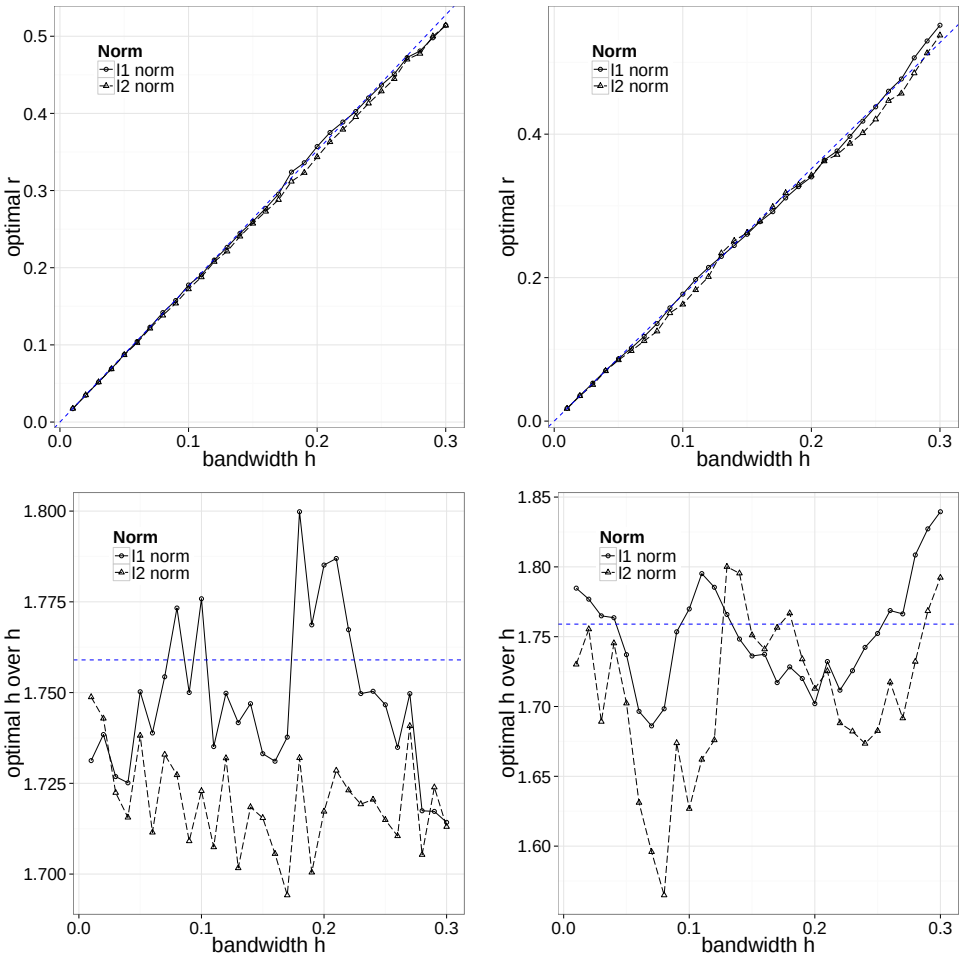



FIGURE 7. Optimal $r^\star(h)$ as a function of $h$ on the polygon shape (on the left) and the Finistère region (on the right), from Figure 6, on top, and below, ratio of $r^\star(h)$ over $h$, as a function of $h$.

mean squared error is integrated on the whole area) with

$$\text{MSE}(\boldsymbol{z},h)=\mathbb{E}\left[\widehat{f}_h(\boldsymbol{z})-f(\boldsymbol{z})\right]^2=\underbrace{\left(\mathbb{E}\left[\widehat{f}_h(\boldsymbol{z})-f(\boldsymbol{z})\right]\right)^2}_{\text{bias}^2}+\underbrace{\text{Var}\left[\widehat{f}_h(\boldsymbol{z})\right]}_{\text{standard deviation}^2}.$$

Here we estimate that function using simulations, and the two components (bias and variance) are reported in Figure 8. $n_s=1,000$ samples of $n$ points uniformly drawn on the polygon of Figure 3 are generated. The density is displayed on the diagonal

of the upper-left corner $[B, D]$ (as defined in Figure 3). For any point $\boldsymbol{z}$, estimators $\widehat{f}_1(\boldsymbol{z}), \cdots, \widehat{f}_{n_s}(\boldsymbol{z})$ (with the two techniques) are obtained, based on those $n_s = 1,000$ samples. The average value of those estimators, $\bar{f}(\boldsymbol{z})$ (which is a approximation of $\mathbb{E}[\widehat{f}_h(\boldsymbol{z})]$), as well as the variance,

$$\widehat{\mathrm{Var}}[\widehat{f}(\boldsymbol{z})] = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}\left[\widehat{f}_i(\boldsymbol{z}) - \bar{f}(\boldsymbol{z})\right]^2 \text{ with } \bar{f}(\boldsymbol{z}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\widehat{f}_i(\boldsymbol{z}),$$

(which is an approximation of $\mathrm{Var}[\widehat{f}_h(\boldsymbol{z})]$) are plotted in Figure 8, for all $\boldsymbol{z} \in [B, D]$.

Figure 8, shows that both estimators have a similar behavior on average, but the variance (and therefore the mean-squared error) is much smaller with Ripley's correction, especially on small samples. Figure 9 exhibits much more volatility for Diggle's correction (on the right) compared to our estimate (with Ripley's correction, on the left).

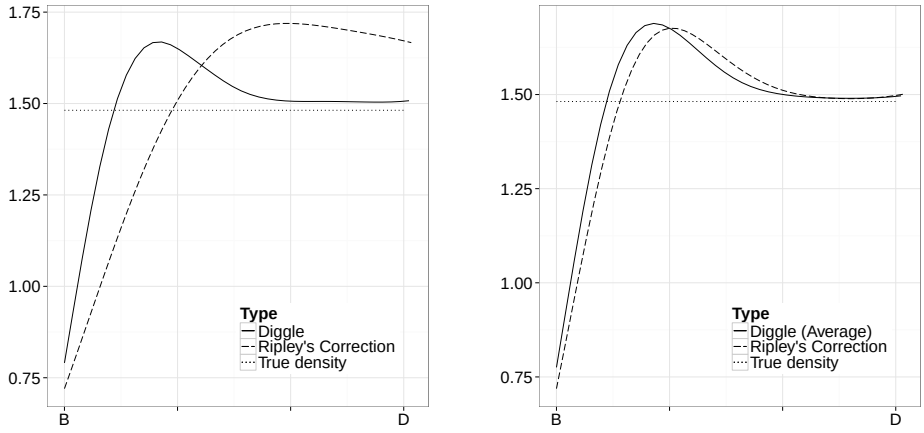
## 4. Visualizing locations of car accidents

Car accident concentration is usually identified as *black spots*, as in Nguyen (1991) or Joly (1992). Those zones suggest that there might exist some spatial dependence between individual occurrences, as suggested by Steenberghen et al. (2004). Detecting clustering (in time and space) might be an important issue, to improve road safety and to reduce traffic accidents. We consider here the dataset of traffic accidents that occurred in 2008 in France and involved bodily injuries. The BAAC dataset (*bulletins d'analyse d'accident corporel*) is filed by police forces, and most accidents have a specific location. In 2008, the dataset contains 10,854 accidents with a location.

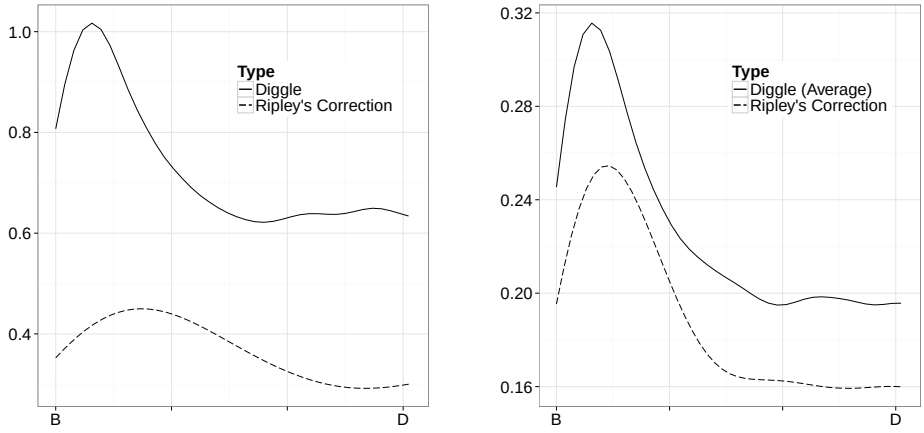
### 4.1. Spatial location of bodily injury car accidents in two regions.

To illustrate border issues, we focus here on two specific regions, Finistère and Morbihan,[3] where major cities (Brest in Finistère and Lorient, or Vannes in Morbihan are by the sea). There are 186 observations for the first region, and 180 for the other one.

[3]Note that islands were removed, namely Belle-Ile, Ile de Groix, Ile de Hoëdic and Ile d'Houat because no traffic accident occurred on those islands in 2008.



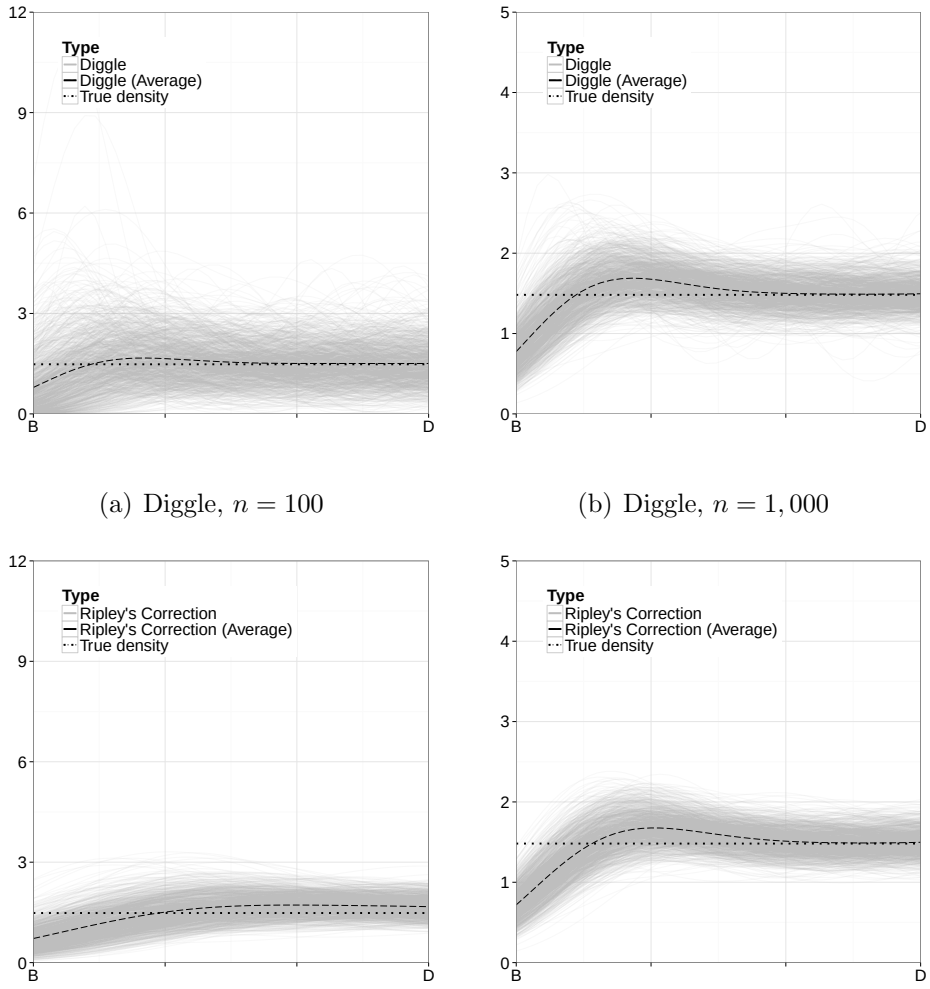

(a) $\mathbb{E}(\hat{f}), n = 100$ (b) $\mathbb{E}(\hat{f}), n = 1000$




(c) $\sqrt{Var(\hat{f})}, n = 100$ (d) $\sqrt{Var(\hat{f})}, n = 1000$

FIGURE 8. Estimation of the density in the upper left corner of polygon $\mathcal{S}$, on interval $[B, D]$, with $n = 100$ and $n = 1,000$ points, on the left and on the right, respectively, with the average density (on 1,000 samples), and the standard deviation.

Results of the estimations for Finistère can be seen in Figure 10. When the standard kernel is used, we can think of at least two *black spots*: the one in the north is bigger than the other one on the south coastline. When the correction is used, the two spots still show up, but another locale stands out on the lower tip of Finistère. The area




(a) Diggle, $n = 100$

(b) Diggle, $n = 1,000$

(c) Ripley's Correction, $n = 100$

(d) Ripley's Correction, $n = 1,000$

FIGURE 9. Estimation of the density in the upper left corner of polygon $\mathcal{S}$, on interval $[B, D]$, with $n = 100$ and $n = 1,000$ points, on the left and on the right, respectively, with the average density (on 500 samples), and the standard deviation.

of this third place is surrounded by water, thus the estimation with standard kernel fails to highlight it.

The same happens in Morbihan, as seen in Figure 11. The density estimation at the north-west border differs greatly depending on the use of weight corrections. Once weights are applied to correct the border bias, one can easily detect a *black spot*.

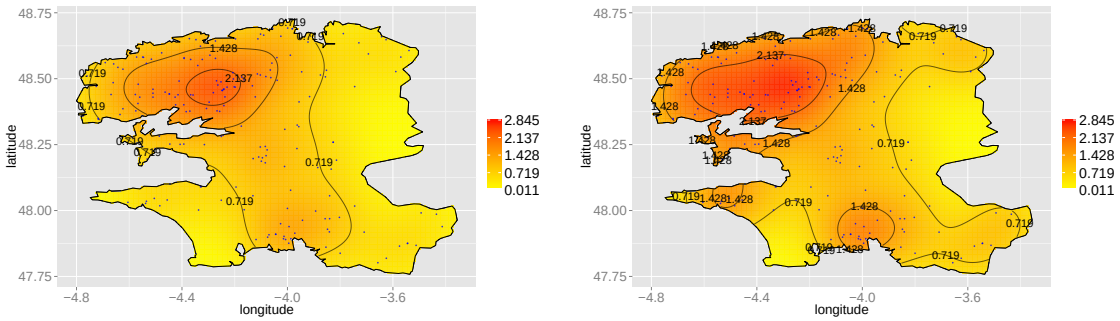



FIGURE 10. Locations of car accidents in Finistère standard kernel on the left, and corrected one on the right.

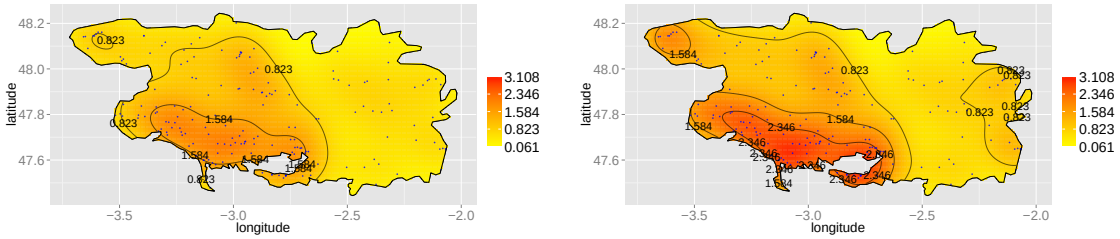



FIGURE 11. Locations of car accidents in Morbihan standard kernel on the left, and corrected one on the right.

4.2. **Detecting hot spots.** To improve road safety and reduce traffic accidents, public authorities have to understand where traffic accidents occur. Analysis of spatial patterns is thus crucial, because it is difficult to assume that occurrences of traffic accidents are purely random observations in space. In most cases, traffic accidents form clusters, called "*hot spots*", in geographic space (see Taylor (1977) or Steenberghen et al. (2004)). Spatial (and temporal) patterns along a certain roadway segment are largely determined by their traffic volume, but also physical environment (slopes

and angles) or weather (see Black (1991), Noland and Quddus (2004) and references therein). Detecting spatial patterns and clusters of car accidents is a recurrent problem (see Yamada and Thill (2004), Erdogan et al. (2008), Xie and Yan (2008), Loo (2006) and reference therein). The so-called "*quadrat*" analysis (see Getis (1964), Rogers (1965) or Thomas (1977) for a description) is a popular technique to analyse the pattern of a distribution of events within a given region $\mathcal{S}$. The idea is to divide region $\mathcal{S}$ into sub-regions $\mathcal{S}_i$'s having equal (and homogeneous) areas, called *quadrats*, and to study histograms on this partition of $\mathcal{S}$. GIS packages allow one to visualize the phenomenon via color-based representations of quadrats. However, the analysis is extremely sensitive to the partition considered.

As described in Chapters 7 of Levine (2010) (see also Levine (2008) for additional motivations and Everitt et al. (2011) for more technical discussions), it is possible to use the density estimation to identify and visualize hot spots. In Levine (2010), a "*Single Kernel Density method*" is considered. It is employed here on the accidents data, and referred to as "*the estimate without border correction.*" Ripley's correction is also applied to visualize hot spots. Figure 12 displays the convex hull of the hot-spot regions for car accidents, for both estimation techniques.

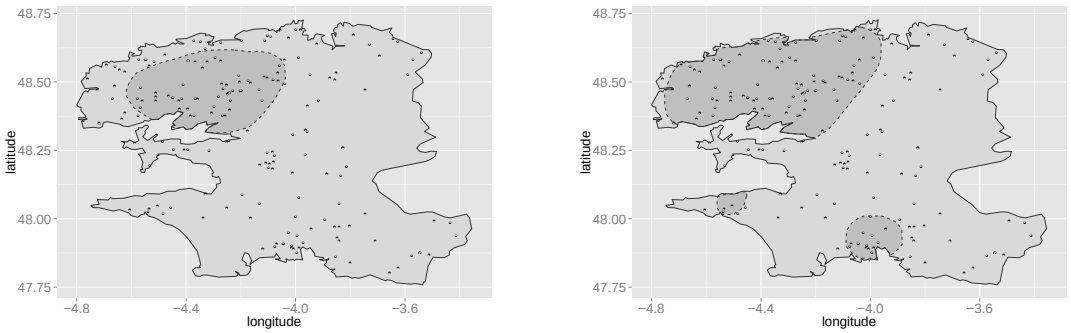



FIGURE 12. Convex hull of hot spot areas, without border correction (on the left) and with Ripley's correction (on the right).

Further, following Chainey et al. (2002) and Van Patten et al. (2009), the Predictive Accuracy Index (PAI) can be computed as the ratio of the hit rate percentage in the

hot spot to the area percentage (area in the hotspot in relation to the study's total area). PAIs using Ripley's correction are reported in the last column of Table 1.

| | hot spot $n$ | area percent | hit rate | PAI |
|---|---|---|---|---|
| non correction | 55 | 15.09 | 29.57 | 1.9596 |
| correction | 102 | 28.50 | 54.84 | 1.9242 |

TABLE 1. Predictive Accuracy Index (PAI) for car accidents.

## 5. Visualizing bike theft locations

5.1. **Density estimation of bike theft locations.** Another popular area of applications, where visualizing spatial densities is also a crucial step, is criminology (see Block et al. (1995), Eck (1997), Ceccato and Haining (2004), Levine (2010) or Nakaya and Yano (2010) among others). In order to illustrate Ripley's correction technique, another application on bike thefts in San Francisco is considered. Data about reported crimes in San Francisco are available on `https://data.sfgov.org/`. Density estimates are computed as a first step, on the 794 reported bicycle thefts from 2013. These estimates are used in a second step to compute an estimation of the number of stolen bikes per year within a 500 m radius.

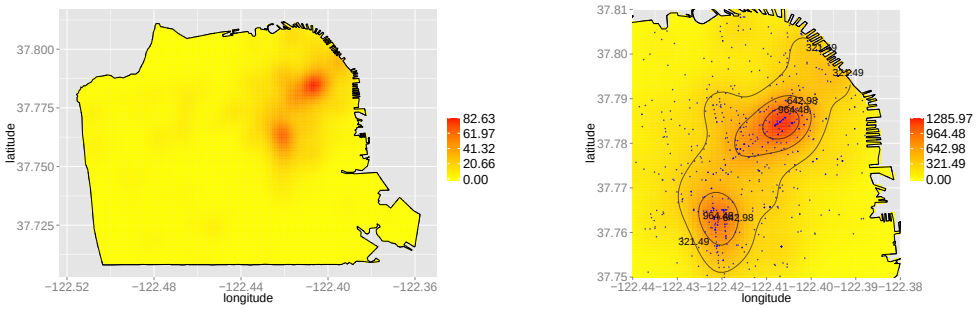



FIGURE 13. Estimates of expected bike thefts per year within a 500 $m$ radius on the left, and corrected kernel density estimates on the right (zoom).

5.2. **On the interpretation of density.** As discussed in Section 2.1, for any region $\mathcal{E}$,

$$\mathbb{P}(\boldsymbol{Z} \in \mathcal{E}) = \int_{\mathcal{E}} f(\boldsymbol{z}) d\boldsymbol{z},$$

where $f(\boldsymbol{z})d\boldsymbol{z}$ is usually interpreted as the probability of $\boldsymbol{Z}$ falling within the infinitesimal region $[\boldsymbol{z}, \boldsymbol{z} + d\boldsymbol{z}]$. Here, units of the projection coordinates used to locate $\boldsymbol{z}$ are 1° (111.11 km) times 1° (111.11 km on the Equator but 87.8 $km$ in San Francisco), which has an area of $9,758$ $km^2$. Therefore, the whole area $\mathcal{S}$ of San Francisco (120.11 $km^2$) is 1/81 of the total 1° times 1° area: a uniform distribution over San Francisco would be $f_{\perp}(\boldsymbol{z}) \sim 81$.

To interpret the density in terms of the number of bikes stolen per year within a given area $\mathcal{D}$ – say a 500 $m$ distance to location $\boldsymbol{z}$ – then $\mathbb{P}(\boldsymbol{Z} \in \mathcal{D}) \sim f(\boldsymbol{z})\mathcal{A}(\mathcal{D} \cap \mathcal{S})$, where $\mathcal{D}$ is the disk of radius $r$ centered in $\boldsymbol{z}$. If the distance from $\boldsymbol{z}$ to the sea exceeds $r$, then $\mathbb{P}(\boldsymbol{Z} \in \mathcal{D}) \sim f(\boldsymbol{z}) \cdot \mathcal{A}(\mathcal{D})$. A circle with a radius of radius 500 $m$ is 0.785 $km^2$; and since the whole area $\mathcal{S}$ of San Francisco is 120.11 $km^2$, $\mathcal{A}(\mathcal{D})$ is 1/153 of the total San Francisco area. If the distance from $\boldsymbol{z}$ to the sea is below $r$, there is a multiplicative factor $\mathcal{A}(\mathcal{D} \cap \mathcal{S})/\mathcal{A}(\mathcal{D})$ (proportion of the disk inland). This ratio is the same as the one when computing weights for the correction.

Figure 14 shows those computations.

Observe that the interpretation of the number of stolen bikes can be related to the standard kernel density estimation without the correction. The correction is necessary when computing a density (which can be related to a probability of occurrence with respect to some unit) but not when computing the number of events that should occur within a given time frame.

## 6. Visualizing the density of campsites

The third and last example to apply Ripley's circumference method concerns campsite locations in France. More generally, from an economic perspective, getting a better geographical perception of the location of accommodation facilities is extremely important, as explained in Hsueh and Tseng (2013). In this example, we

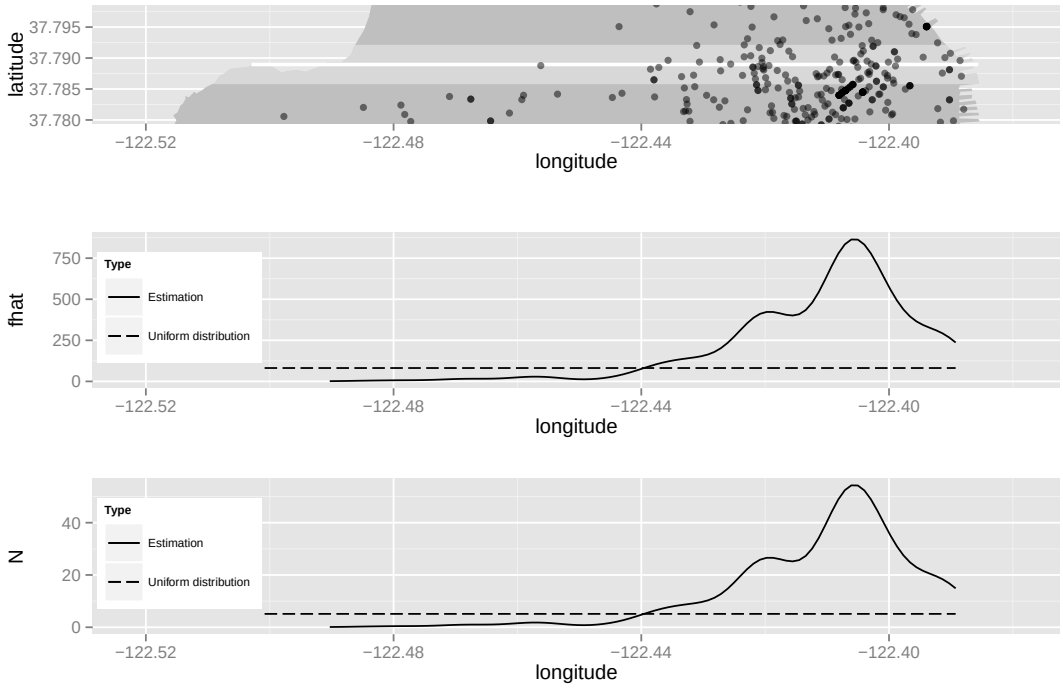



FIGURE 14. Density estimation $\widehat{f}(\boldsymbol{z})$ - in the middle - at latitude 37.788°, and estimation of the expected number of bikes stolen, per year, within a 500 $m$ distance to location $\boldsymbol{z}$ - bottom graph. On top, observations in the neighborhood of the latitude can be visualized. A $\pm 500$ $m$ tube was added.

will discuss locations of campgrounds in France. Data about French lodgings were downloaded at https://www.classement.atout-france.fr, and only observation concerning campgrounds were kept. A total of 5494 camping pitches were geolocated using the Google Maps API.[4] The density estimates can be visualized in Figure 17, with and without applying the correction. Figures 17 and 19 provide zooms on two regions where a large number of campgrounds can be found near the coastline. To highlight the difference between the two methods, using the technique described in

[4]See https://developers.google.com/maps/.

Section 4, convex hulls of hot spot areas are also plotted (see Figure 16 for France, Figure 18 for the Mediterranean side and Figure 20 for the Atlantic side).

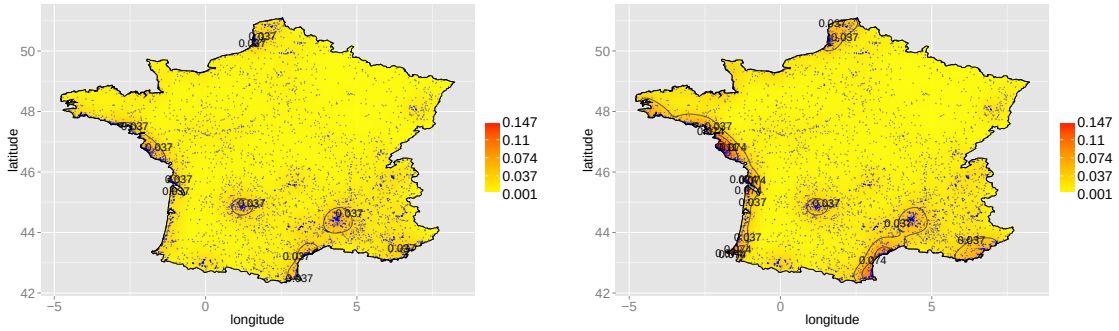



FIGURE 15. Density estimates of French campground locations, standard kernel on the left, and corrected one on the right.

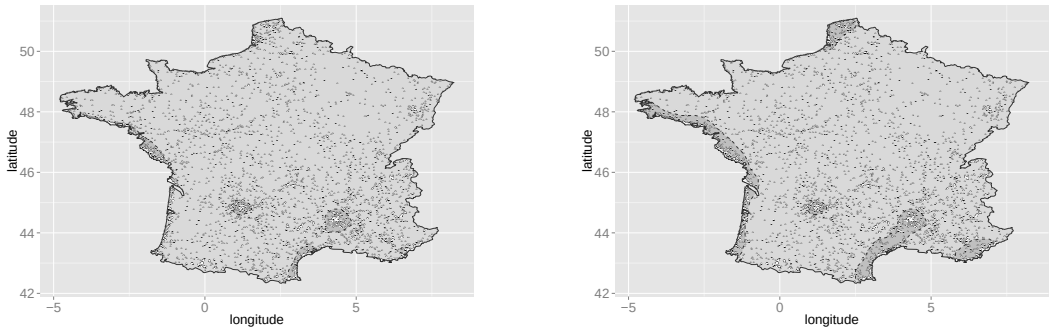



FIGURE 16. Convex hull of hot spot areas, without border correction (on the left) and with Ripley's correction (on the right).

## 7. Conclusion

In this article, a technique relating kernel density estimation and Ripley's circumferential technique was discussed, including a practical technique to select the optimal radius of Ripley's correction technique. This correction is necessary to provide adequate visualization of the density. Nevertheless, when interpreting the density as an expected number of occurrences, this correction might be misleading. Computations are fast, and our estimate provided a less volatile estimation of the density, compared

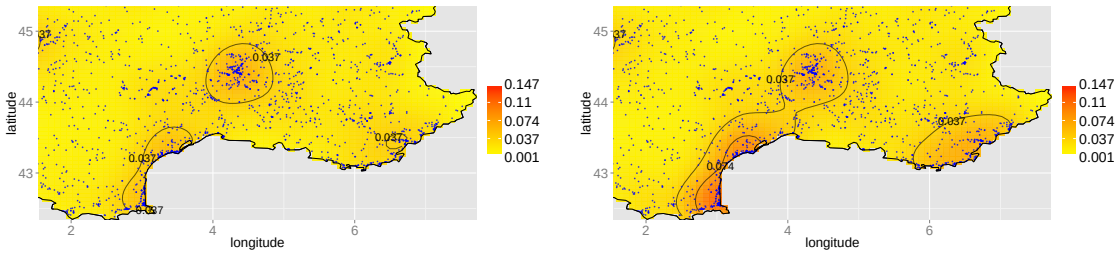



FIGURE 17. Density estimates of French campground locations (zoom on Southern France), standard kernel on the left, and corrected one on the right.

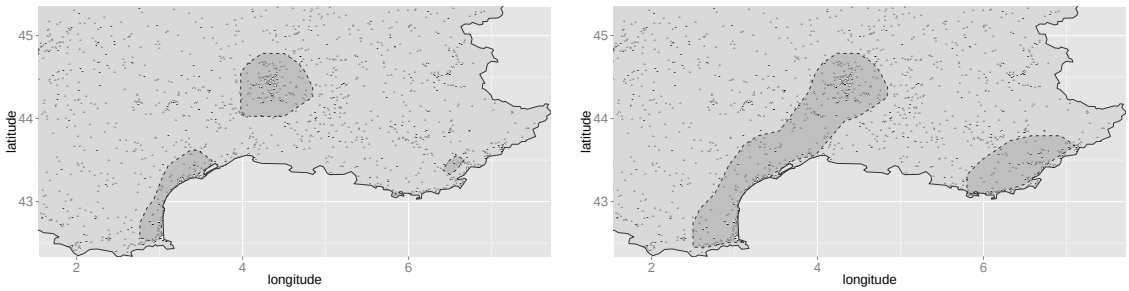



FIGURE 18. Convex hull of hot spot areas (zoom on Southern France), without border correction (on the left) and with Ripley's correction (on the right).

with the popular estimate introduced by Diggle (1985). That estimate was used on three different applications, when regions have different shapes, and with different sample size.

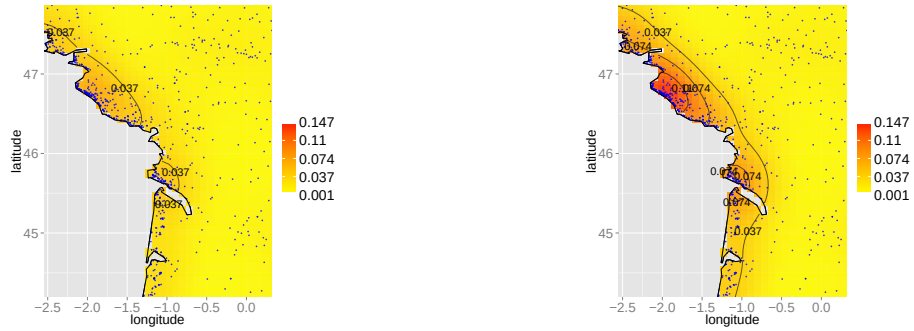



FIGURE 19. Density estimates of French campground locations (zoom on the Atlantic coast), standard kernel on the left, and corrected one on the right.

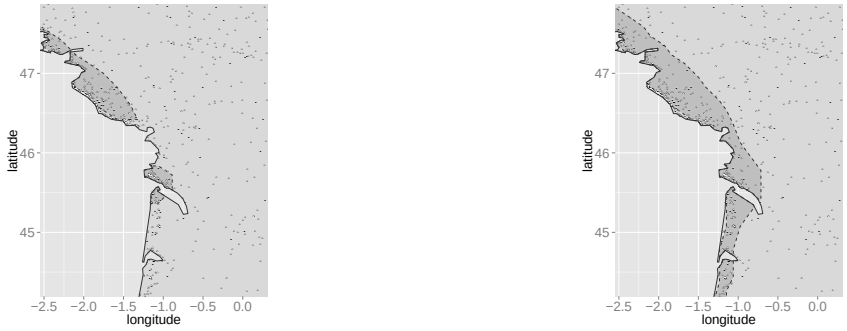



FIGURE 20. Convex hull of hot spot areas (zoom on the Atlantic coast), without border correction (on the left) and with Ripley's correction (on the right).

ACKNOWLEDGEMENTS. The authors would like to thank Olivier Scaillet and John Wilson for stimulating comments, and helping us to improve the paper, as well as three anonymous reviewers.

## REFERENCES

Bailey, T. C. (1994). A review of statistical spatial analysis in geographical information systems. *Spatial analysis and GIS*, pages 13–44.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis*, volume 413. Longman Scientific & Technical Essex.

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 (Methodological)*, pages 81–92.

Black, W. R. (1991). Highway accidents: a spatial and temporal analysis. *Transportation Research Record*, (1318).

Block, C. R., Dabdoub, M., and Fregly, S. (1995). Crime analysis through computer mapping. Police Executive Research Forum Washington, DC.

Bryc, W. (2002). A uniform approximation to the right normal tail integral. *Applied mathematics and computation*, 127(2):365–374.

Ceccato, V. and Haining, R. (2004). Crime in border regions: The scandinavian case of öresund, 1998–2001. *Annals of the Association of American Geographers*, 94(4):807–826.

Chainey, S., Reid, S., and Stuart, N. (2002). *When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime*. Taylor & Francis, London, England.

Chiu, S.-T. (1991). Bandwidth selection for kernel density estimation. *The Annals of Statistics*, pages 1883–1905.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613–617.

Davis, K. B. (1975). Mean square error properties of density estimates. *The Annals of Statistics*, pages 1025–1030.

Diggle, P. (1985). A kernel method for smoothing point process data. *Applied Statistics*, pages 138–147.

Eck, J. E. (1997). What do those dots mean? mapping theories with data. *Crime mapping and crime prevention*, 8:379–406.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158.

Erdogan, S., Yilmaz, I., Baybura, T., and Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: city of afyonkarahisar. *Accident Analysis & Prevention*, 40(1):174–181.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley series in probability and statistics. Wiley.

Getis, A. (1964). Temporal land-use pattern analysis with the use of nearest neighbor and quadrat methods. *Annals of the Association of American Geographers*, 54(3):391–399.

Gisbert, F. J. G. (2003). Weighted samples, kernel density estimators and convergence. *Empirical Economics*, 28(2):335–351.

Hall, P. and Turlach, B. A. (1999). Reducing bias in curve estimation by use of weights. *Computational statistics & data analysis*, 30(1):67–86.

Hearnshaw, H. M., Unwin, D. J., et al. (1994). *Visualization in geographical information systems*. John Wiley & Sons Ltd.

Hsueh, Y.-H. and Tseng, H.-Y. (2013). Exploring the clustering location of accommodation units through the tourism development in the cing jing area of taiwan. *International Journal of Basic & Applied Sciences*, 13(4):34–39.

Joly, M.-F. (1992). Analytical approach to the identification of hazardous road locations: A review of the literature.

Kelsall, J. E. and Diggle, P. J. (1995). Kernel estimation of relative risk. *Bernoulli*, pages 3–16.

Levine, N. (2008). The "hottes" part of a hotspot: comments on "the utility of hotspot mapping for predicting spatial patterns of crime". *Security journal*, 21(4):295–302.

Levine, N. (2010). *CrimeStat*: A spatial statistics program for the analysis of crime incident locations (v 3.3).

Loo, B. P. (2006). Validating crash locations for quantitative spatial analysis: a gis-based approach. *Accident Analysis & Prevention*, 38(5):879–886.

Marron, J. and Padgett, W. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *The Annals of Statistics*, pages 1520–1535.

Nakaya, T. and Yano, K. (2010). Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3):223–239.

Nguyen, T. (1991). *Identification of accident blackspot locations: an overview*. Number DP/91-4.

Noland, R. B. and Quddus, M. A. (2004). A spatially disaggregate analysis of road casualties in england. *Accident Analysis & Prevention*, 36(6):973–984.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability*, pages 255–266.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212.

Ripley, B. D. (1981). *Spatial statistics*. Wiley series in probability and mathematical statistics. Wiley.

Rogers, A. (1965). A stochastic analysis of the spatial clustering of retail establishments. *Journal of the American Statistical Association*, 60(312):1094–1103.

Scott, D. W. (2009). *Multivariate density estimation: theory, practice, and visualization*, volume 383. John Wiley & Sons.

Shah, A. K. (1985). A simpler approximation for areas under the standard normal curve. *The American Statistician*, 39(1):80–80.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Steenberghen, T., Dufays, T., Thomas, I., and Flahaut, B. (2004). Intra-urban location and clustering of road accidents using gis: a belgian example. *International Journal of Geographical Information Science*, 18(2):169–181.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics*, 21(2):169–184.

Tapia, R. and Thompson, J. (1978). *Nonparametric probability density estimation*. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press.

Taylor, P. (1977). *Quantitative Methods in Geography: An Introduction to Spatial Analysis*. Waveland Press.

Thomas, R. W. (1977). *An introduction to quadrat analysis*. Geo Abstracts Limited.

Van Patten, I. T., McKeldin-Coner, J., and Cox, D. (2009). A microspatial analysis of robbery: Prospective hot spotting in a small city. *Crime Mapping: A journal of research and practice*, 1(1):7–32.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*, volume 60. Crc Press.

Xie, Z. and Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32(5):396–406.

Yamada, I. and Rogerson, P. A. (2003). An empirical comparison of edge effect correction methods applied to k-function analysis. *Geographical Analysis*, 35(2):97–109.

Yamada, I. and Thill, J.-C. (2004). Comparison of planar and network k-functions in traffic accident analysis. *Journal of Transport Geography*, 12(2):149–158.

Zheng, P., Durr, P., and Diggle, P. J. (2004). Edge-correction for spatial kernel smoothing methods; when is it necessary. In *Proceedings of the second international conference on the applications of GIS and spatial analysis to veterinary science. University of Guelph, June*, pages 23–25.