Visualizing spatial processes using Ripley's correction: an application to bodily-injury car accident location

Arthur Charpentier, Ewen Gallic

VISUALIZING SPATIAL PROCESSES USING RIPLEY'S CORRECTION: AN APPLICATION TO BODILY-INJURY CAR ACCIDENT LOCATION.

ARTHUR CHARPENTIER AND EWEN GALLIC

KEYWORDS: border bias; car accident; frontier; GIS; kernel estimation; polygons; Ripley's circumference method; spatial process

Arthur Charpentier, UQAM, 201, avenue du Président-Kennedy, Montréal (Québec), Canada H2X 3Y7 (corresponding author) charpentier.arthur@uqam.ca, and Ewen Gallic, UQAM, 201, avenue du Président-Kennedy, Montréal (Québec), Canada H2X 3Y7.

Abstract. In this paper, we investigate (and extend) 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 compute efficiently weights to correct border bias on frontiers of the region of interest, with an automatic selection of an optimal radius for the method. An illustration on location of bodily-injury car accident (and hot spots) in the western part of France is discussed, where a lot of accident occur close to large cities, next to the sea.

1. Introduction and motivation

In order to improve road safety and to reduce traffic accidents, public authorities have to understand when and where traffic accident occurred. Analysis of spatial patterns is then a crucial issue, since it is difficult to assume that occurrences of traffic accidents are purely random observations, in space and time. In most cases, traffic accidents form clusters, called '*hot spots*', in geographic space (see Taylor (1977)). 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). Krisp and Durot (2007) mention the case of optimization of *warning sign* placement in southern Finland, while Pulugurtha *et al.* (2007) study sign placement in high pedestrian crash zones in the Las Vegas metropolitan area. Note that analysis of spatial patterns is popular in the study of traffic accident (see also Joly *et al.* (1992), Nguyen (1991), Steenberghen *et al.* (2004), Treno *et al.* (2007), Warden *et al.* (2011), Levine and Kim (1998), Yamada and Thill (2004), Saffet *et al.* (2008), Xie and Yan (2008) or Loo (2006)), similar studies can be conducted in criminology (see Block *et al.* (1995), Eck (1997), Ceccato and Haining (2004) or Nakaya and Yano (2010)) among others.

Detection of '*hot spots*' is based on spatial analysis of point events, or *point pattern analysis* (see Ripley (1981), Bailey and Gatrell (1995), Anselin and Flora (1995) or Batty (2005) and references therein). Quadrat analysis (see Getis (1964), Rogers

(1965) or Thomas (1977)) is one 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 then visualizing the phenomenon via color-based representations of quadrats. Nevertheless, the analysis is then extremely sensitive to the partition considered.

A natural extension is to consider kernel based estimators of densities (see OSullivan and Unwin (2002), Miller (1999), Gatrell (1994), Basawa (1996a), Basawa (1996b), Batty (2005) or Borruso (2008)). The goal here is still to obtain a field representation of the phenomenon (here traffic accidents) by means of a smooth continuous surface, where peaks represent the presence of clusters ('*hot spots*') in the distribution of events. A bandwidth related to the length of the neighborhood (also called '*sphere of influence*' in Gatrell (1994)) is considered, as well as a weighting function (the *kernel*). Since Epanechnikov (1969) proved that statistical results were not (significantly) affected by the choice of the kernel function, most of the authors have emphasized the fact that bandwidths choice is a crucial issue. The most popular kernel is the Gaussian one since a dual representation (accident 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*': on the frontier of the region of interest $\mathcal{S}$. Yamada and Rogerson (2003) mention Ripley's circumference method (from Ripley (1981)), but claims that "*Ripley's method could be too complicated without proper software or skilled programmers*".

In this paper, we recall basics on space and time kernel density estimation, in Section 2. But the time component will not be discussed in this paper. In Section 3, we will discuss frontiers and space border bias correction. In that section, we will present several (standard) techniques when $\mathcal{S}$ is either an half-space, or a rectangular area. Then, we provide a simple method to compute efficiently weights in Ripley's circumference method that can be used for any region $\mathcal{S}$ (characterized as a polynomial).

We will discuss in Section 4 the link between radius r used in the circumferential method, and bandwidth h of the kernel smoother. As we will see using Monte Carlo simulation, given a bandwidth h, there is an optimal radius $r^\star(h)$ that which minimizes the distance between weights used in Ripley's circumference, and weights used by truncating Gaussian densities. Using either a L_1 or L_2 norms (minimizing either sums of absolute values or sums of squares) we will see that $r^\star(h)$ is linear in h. This property (that we will derive analytically for half space regions) will allow us to introduce an automatic technique. Finally, in Section 5, we illustrate that technique on bodily injury car accidents, in western part of France (Morbihan and Finistère). In section 5.2 we will use the boundary correction, and illustrate the estimation of the spatial location of car accident. Since this density should be closely related to road density (more than larger frequency occurrence of car accident), in section 5.2, we will apply this technique on the estimation of road density. Therefore, in section 5.3, we will suggest a technique to produce a map that can be used to identify hot spots. And finally, sketches of `R` codes are provided in the Appendices (Section 6).

2. Space and time kernel density estimation

2.1. **Definitions and notations.** Kernel density estimation (see Silverman (2004), Scott (1992)) 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 temporal models. Spatio-temporal observations are pairs of observations $(\boldsymbol{Z}, T)$, with spatial location $\boldsymbol{Z} = (X, Y)$ (usually characterized by a latitude and a longitude coordinate) and time T. A natural assumption is to consider a product kernel, between location and time, as in Brunsdon *et al.* (2007). Hence,

$$\widehat{f}(x, y, t) = \frac{1}{nh_X h_Y h_T} \sum_{i=1}^{n} K_Z\left(\frac{x - X_i}{h_X}, \frac{y - Y_i}{h_Y}\right) K_T\left(\frac{t - T_i}{h_T}\right) \tag{2.1}$$

is the density estimator at location $\boldsymbol{z} = (x, y)$ at time t, where n denotes the total number of events observed, and h_X, h_Y and h_T are spatial and temporal bandwidth respectively.

Following Epanechnikov (1969), let K_Z and K_T be Epanechnikov kernels (used e.g. in ArcGIS)

$$K_T(\omega) = \frac{3}{4}(1-\omega^2)\mathbf{1}(\omega^2 \in [0,1)) \tag{2.2}$$

and

$$K_Z(u,v) = \frac{2}{\pi}(1-[u^2+v^2])\mathbf{1}(u^2+v^2 \in [0,1)). \tag{2.3}$$

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

$$K_Z(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[u^2+v^2-2\rho uv\right]\right). \tag{2.4}$$

From Silvermnan's rule (see Silverman (2004) or Scott (1992)) for d-dimensional product kernel, and Gaussian observations, the optimal bandwidth is $h^\star = n^{-1/(3+d)}\sigma$ where σ is the standard deviation in the appropriate dimension. E.g. $h_X^\star = n^{-1/(3+d)}\sigma_X$, where $\sigma_X^2 = \mathrm{Var}(X)$. Estimated optimal kernel bandwidth are then $\widehat{h}^\star = n^{-1/(3+d)}\widehat{\sigma}$. Further, as mentioned in Härdle *et al.* (2004) bandwidth are rather close, with those the two kernels. Recall that if the observations are not Gaussian, bandwidth are usually to large, which might cause an excessive smoothing, as discussed in Härdle *et al.* (2004).

3. Frontier and space border bias correction

Kernel density estimation is a popular technique to visualize smoothed densities. But in some specific cases, observations have to be within to some specific area $\mathcal{S}$. For instance, for traffic accidents, events have to occur on-land, as discussed in section . $\mathcal{S}$ would denote some on-land territory. On the contrary, when locating fishes or sea animals using GPS trackers, we know that those animal have to be in the sea. Here, $\mathcal{S}$ can denote some territorial sea.

Remark 1. *One might argue that car crashes cannot occur off the transportation network, as mentioned in Yamada and Thill (2004) and Xie and Yan (2008), and this technique might not be appropriate for constrained to network space. Statistical results may reflect the density of the road network as well as the density of point events on this network space. We will discuss further this point in Section 5.2, where a density of the road network will be estimated. Thus density of car accident location can be visualized with respect to the density of road network.*

In the case where $\mathcal{S}$ is bounded, kernel estimates suffer two important drawbacks,

- the total weight is not equal to 1, so we do not have a proper probability distribution function, i.e. $\int_{\mathcal{S}} \widehat{f}(\boldsymbol{z})d\boldsymbol{z} < 1$,
- 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 $\kappa_{\boldsymbol{z}} \in [0, 1]$.

Example 1. *A generation of* 200 *and* 2,000 *points, respectively, uniformly distributed on the unit square can be visualized on Figure 1.*

As we will see in Section 5, for regions closed to the sea, estimators of density can suffer major drawbacks.

Remark 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 idea here is to propose a methodology which gives an estimator $\widehat{f}$ which could be associated to a proper probability distribution function, and which does not suffer border bias.

3.1. **A weight based correction.** An alternative is to 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

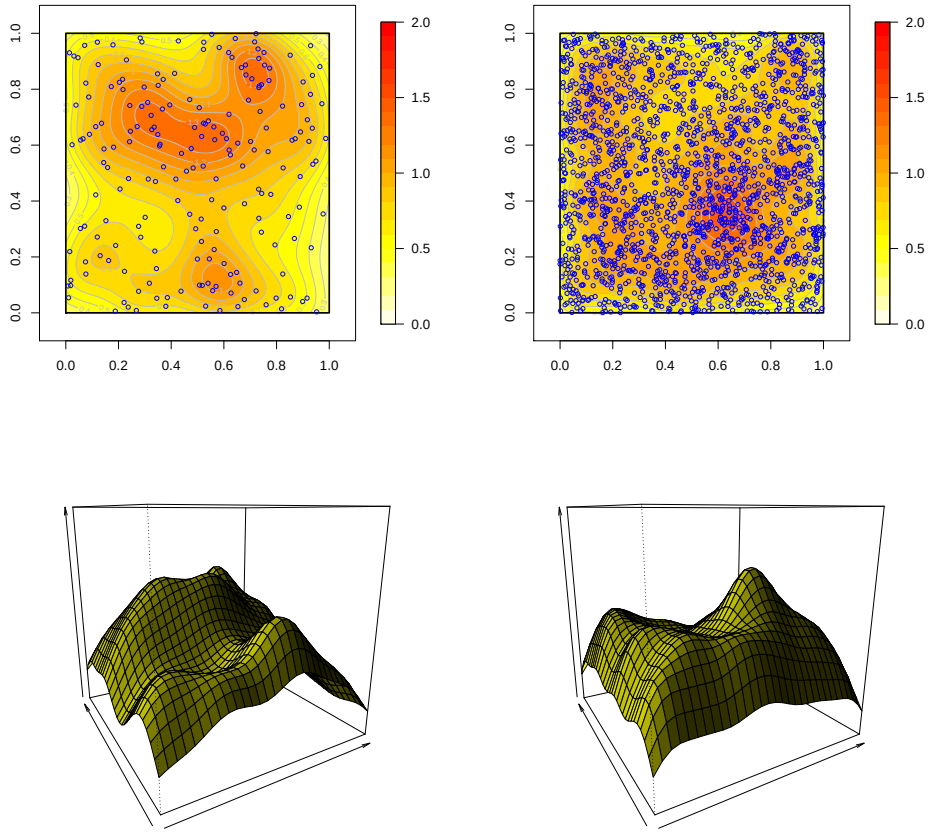

FIGURE 1. 200 and 2,000 uniformly distributed on the unit square, and kernel based estimators of respective densities.

(1975), Tapia and Thompson (1978) or Stefanski and Carrol (1990). Recall that the empirical cumulative distribution function is the step function defined as

$$\widehat{F}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\boldsymbol{Z}_i \leq \boldsymbol{z}), \tag{3.1}$$

and the associated empirical measure is

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{Z}_i}(\boldsymbol{z}), \tag{3.2}$$

where δ denotes the Dirac measure. The idea of Kernel based estimator is to substitute a continuous distribution to Dirac measures,

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mu_{\boldsymbol{Z}_i}(\boldsymbol{z}) \tag{3.3}$$

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 observations can be found. Thus, it might be natural to consider a truncated distribution, restricted to the support $\mathcal{S}$.

$$\mu_{\boldsymbol{Z}_i|\mathcal{S}}(\boldsymbol{z}) = \frac{\mu_{\boldsymbol{Z}_i(\boldsymbol{z})}}{\mu_{\boldsymbol{Z}_i}(\mathcal{S})} \tag{3.4}$$

Thus, it is natural to consider

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mu_{\boldsymbol{Z}_i|\mathcal{S}}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \omega_i \cdot \mu_{\boldsymbol{Z}_i(\boldsymbol{z})} \text{ where } \omega_i = \mu_{\boldsymbol{Z}_i}(\mathcal{S})^{-1}. \tag{3.5}$$

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

$$\frac{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})}{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r})} \tag{3.6}$$

where $\mathcal{A}$ denotes the area function, and $\mathcal{D}_{\boldsymbol{Z}_i,r}$ denotes the disk centered in $\boldsymbol{Z}_i$ with radius $r > 0$ (see Figure 3 for an illustration on observations restricted to the unit square). This method is usually called Ripley's circumference method (from Ripley (1981)). Note that r should be related to the covariance matrix $\boldsymbol{H}$ (this will be discussed in section 4). Thus, here the idea is simply to use *weighted kernel estimators*

$$\widehat{f}(\boldsymbol{z}) = \sum_{i=1}^{n} \omega(\boldsymbol{Z}_i) \cdot \det(\boldsymbol{H})^{-1} K\left(\boldsymbol{H}^{-1}(\boldsymbol{z} - \boldsymbol{Z}_i)\right) \tag{3.7}$$

where weights $\omega(\boldsymbol{Z}_i)$ should reflect the proportion of area around $\boldsymbol{Z}_i$ that belongs to $\mathcal{S}$. Those weighted kernel estimators have been intensively used, 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 intersection of polygon areas with standard softwares is extremely simple, Ripley's correction technique can easily be implemented.

Example 2. *The use of weights is illustrated on 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.*

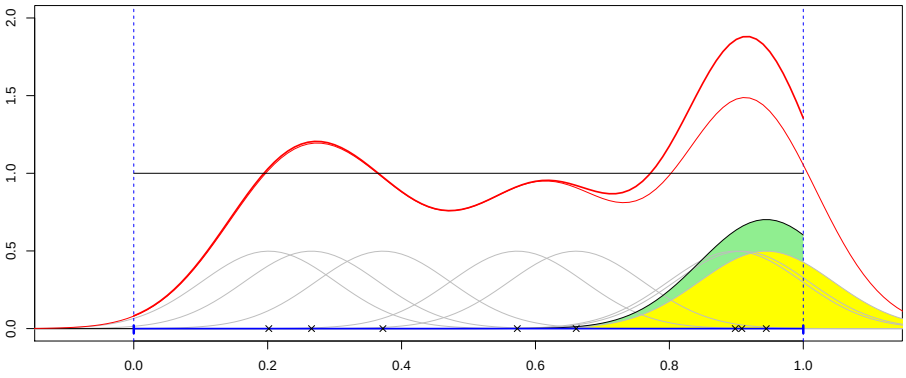

FIGURE 2. Weight correction of a density on $[0, 1]$: the kernel is no longer a Gaussian density, but a truncated Gaussian density.

Example 3. *On the same samples as considered in Example 1, weighted-kernel estimators are considered. On Figure 3 are plotted observations on the unit square, with the circular area around two specific observations. On Figure 4, densities can be visualized. Note that densities now sum to one.*

3.2. **Correction for non-rectangular areas and Monte Carlo study of bandwidth impact.** In order to illustrate that technique, a non-rectangular area is considered here.

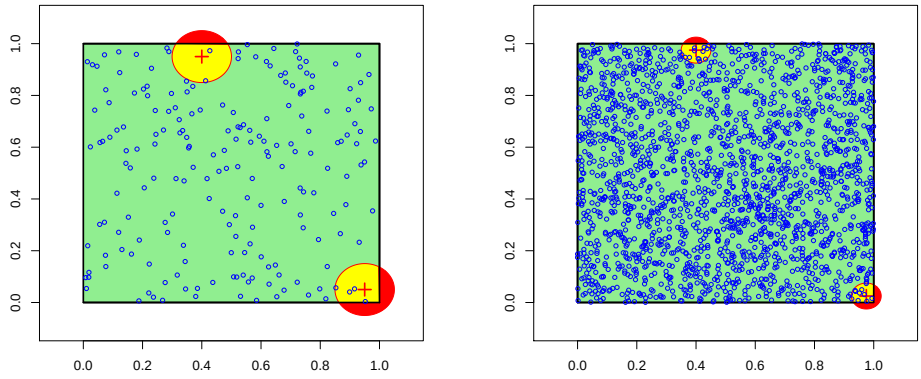

FIGURE 3. 200 and 2,000 points uniformly distributed on the unit square.

Example 4. *Consider the polygon of left of Figure 5, and a sample of points uniformly distributed within the area. In corners and on borders, standard kernel estimators can be strongly biased. For instance, in point* A*, on average, the estimator of* $f(A)$ *will be 1/8-th of the true value, while it will be 1/4-th in* B *and* C *in* C*. Based on sample presented on Figure 5, kernel based densities can be computed. On Figure 6 is presented the output of a Monte Carlo study, when the average density is computed over* $1,000$ *random samples of size* 200*. The distribution of* $\widehat{f}$ *in* A*,* B*,* C *and* D *can be visualized on Figure 7. The choice of the radius will be discussed in the next section.*

4. LINK BETWEEN DISK RADIUS r AND BANDWIDTH h

With a Gaussian kernel, in the univariate case, the bandwidth h is the standard deviation of the Gaussian noise (see mentioned in 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

$$\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S}) \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}). \tag{4.1}$$

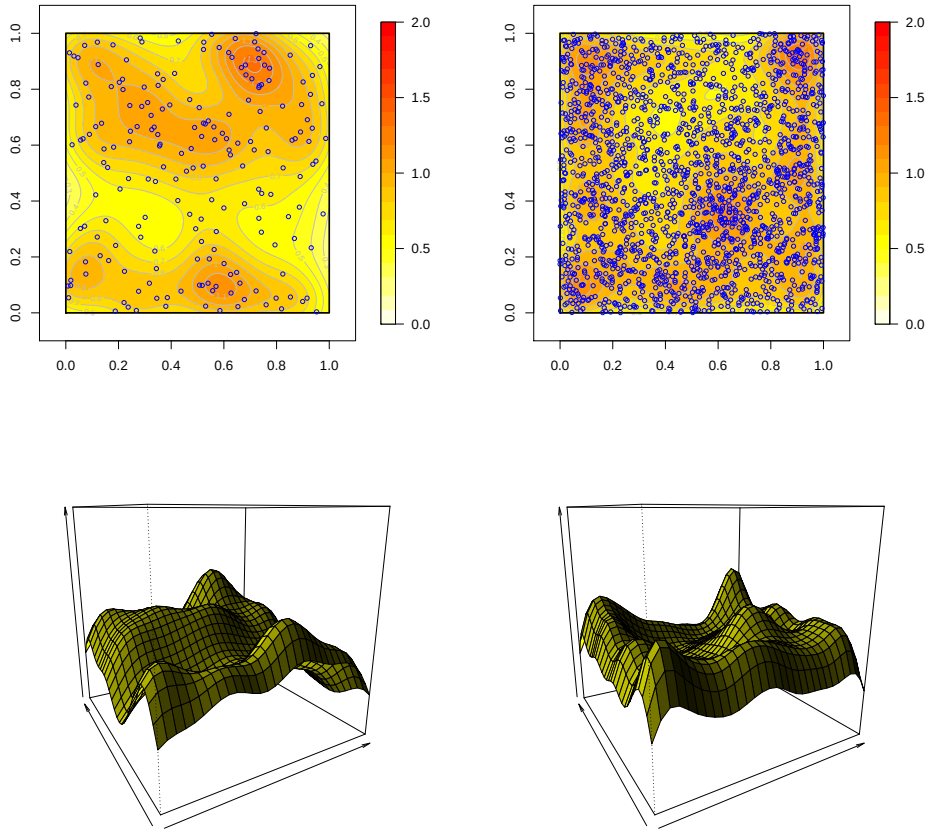

FIGURE 4. 200 and 2,000 points uniformly distributed on the unit square, and kernel based estimators of respective densities, with weight correction.

4.1. **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 on the

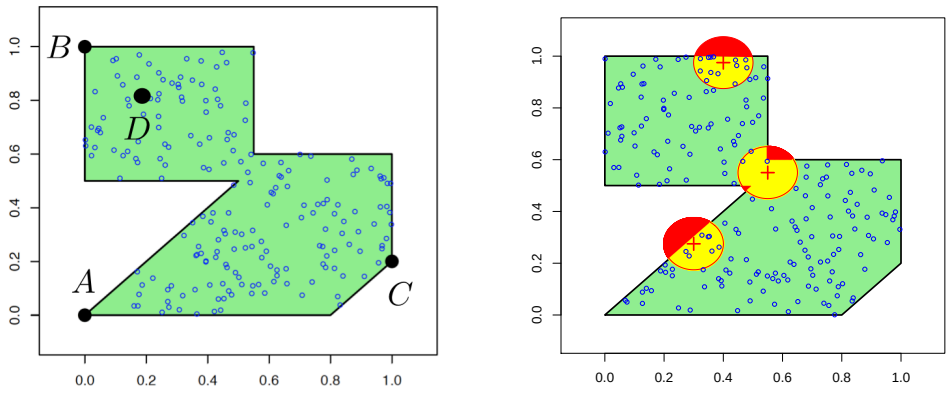

FIGURE 5. 200 points uniformly distributed on $\mathcal{S}$

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.

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

$$\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S}) = 1 - \Phi(-ah^{-1}) = \Phi(ah^{-1}) \tag{4.2}$$

where Φ denotes the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution (see Figure 8).

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

$$\mu^0_{\boldsymbol{Z}_i}(r, \mathcal{S}) = \frac{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})}{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r})} \tag{4.3}$$

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-space, it is possible to derive an analytical expression, since it will just be related to the *circular segment* (the region bounded by a chord and the arc subtended by the chord, see Figure 8). The area of the circular segment is equal to the area of the circular sector

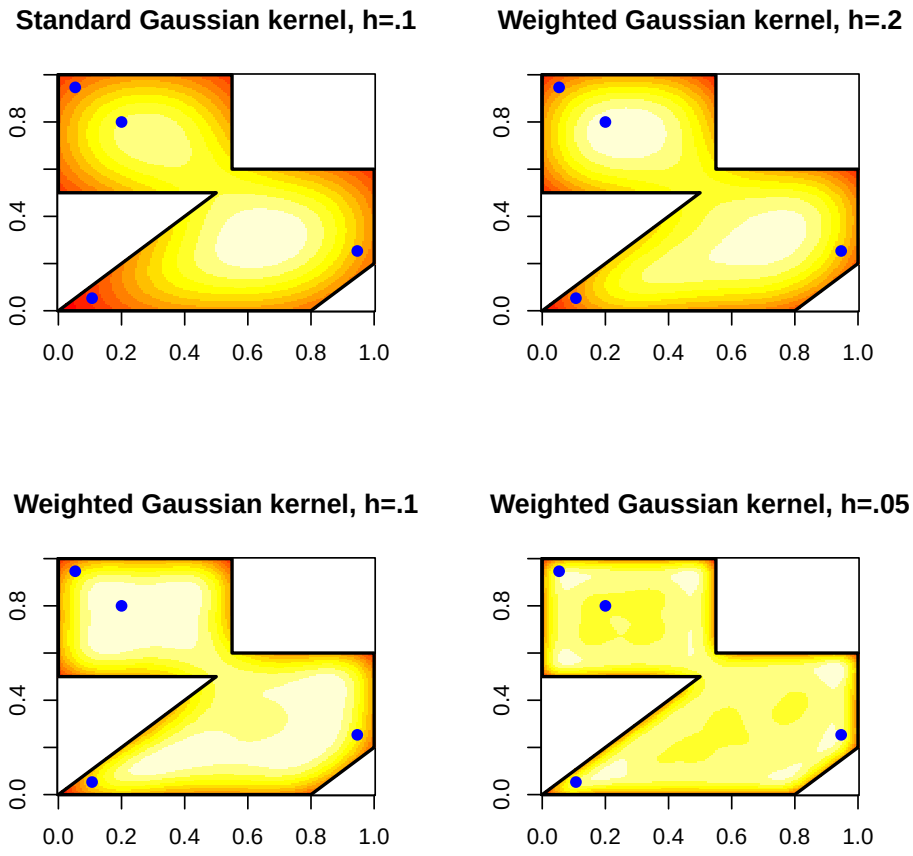

FIGURE 6. Average level curves of $\widehat{f}$ over $1,000$ simulated samples of size 200, on the left. *Optimal* (standard) bandwidth is $h = 0.1$. Weighted kernels have been calculated with 3 different bandwidth $h = 0.2$, $h = 0.1$ and $h = 0.05$, and using then $r^\star = \beta^\star h$

minus the area of the triangular portion.

$$\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S}) = \underbrace{\frac{\theta}{2\pi}\pi r^2}_{\text{sector area}} - \underbrace{\frac{r^2 \sin(\theta)}{2}}_{\text{triangle area}} \text{ where } \cos\left(\frac{\theta}{2}\right) = \frac{a}{r}. \tag{4.4}$$

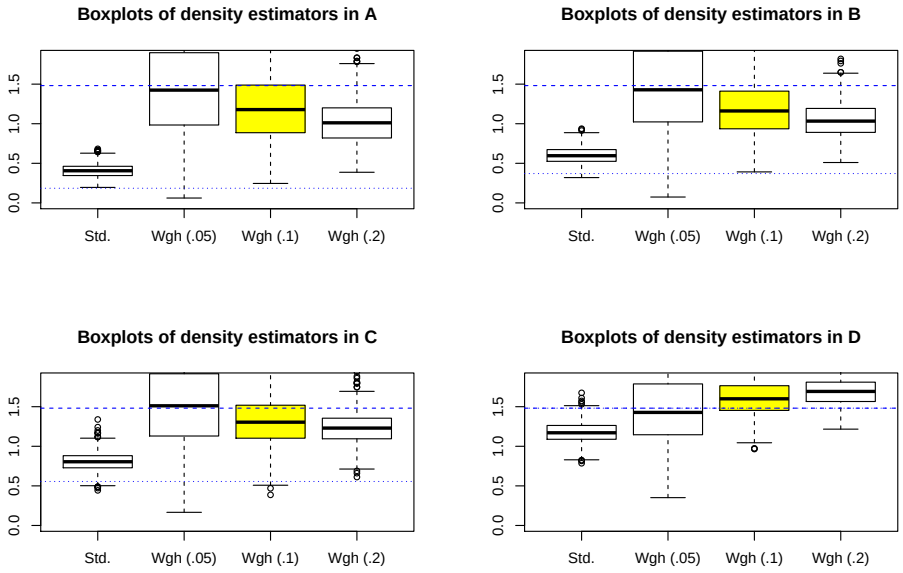

FIGURE 7. Average level curves of $\widehat{f}$ over $1,000$ simulated samples of size 200, on the left. *Optimal* (standard) bandwidth is $h = 0.1$. Weighted kernels have been calculated with 3 different bandwidth $h = 0.2$, $h = 0.1$ and $h = 0.05$, and using then $r^\star = \beta^\star h$

Thus,

$$\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S}) = \begin{cases} \dfrac{r^2}{2}[\theta - \sin(\theta)] \text{ if } a < r \\ 0 \text{ if } a > r \end{cases} \tag{4.5}$$

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.

$$\frac{r^2}{2\pi r^2}[\theta - \sin(\theta)] = \frac{1}{2\pi}[\theta - \sin(\theta)] = 15\% \ (= \Phi(-1)) \tag{4.6}$$

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

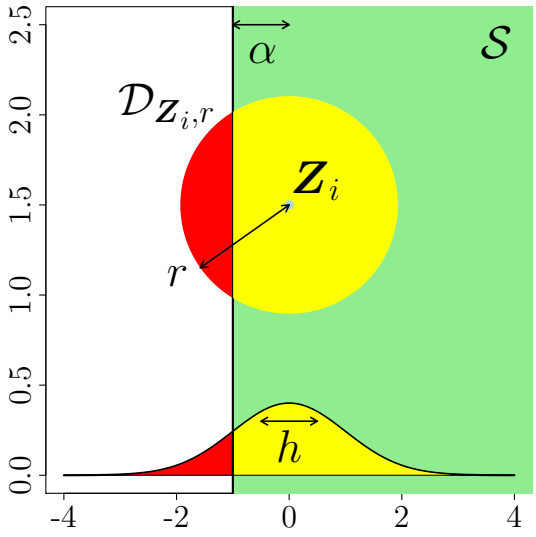

FIGURE 8. Link between $\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S})$ and $\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})$ where $\mathcal{S}$ is a half-space.

1.76. Thus, 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

$$\frac{\theta^\star - \sin(\theta^\star)}{2\pi} = \frac{1}{2\pi}\left[2\text{acos}\left(\frac{a}{\beta^\star h}\right) - \sin\left(2\text{acos}\left(\frac{a}{\beta^\star h}\right)\right)\right] \tag{4.7}$$

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

$$x \mapsto \frac{1}{2\pi}\left[2\text{acos}\left(bx\right) - \sin\left(2\text{acos}\left(bx\right)\right)\right]. \tag{4.8}$$

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

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

Thus, 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-space domains. And this relationship might also be a good approximation on more general spaces $\mathcal{S}$.

4.1.2. *Monte Carlo study for more complex areas $\mathcal{S}$.* In order to illustrate the general case, two regions will be considered in this section: the polygon of Figure 5, and the contour of Finistère (the French region). In those two regions, $10,000$ points were drawn uniformly $\boldsymbol{Z}_i$ (1,000 are plotted on Figure 9)

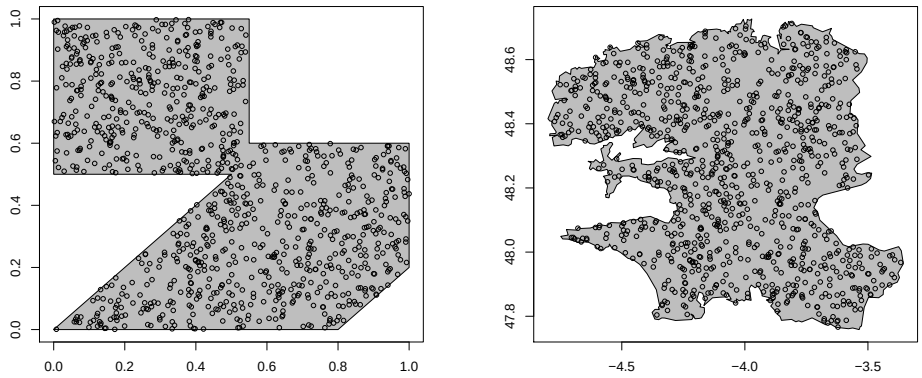

FIGURE 9. Random generation of $1,000$ points uniformly located on two different areas.

Given $h > 0$,

- we compute numerically *theorerical* weights $\omega_i(h)$ based on $\mu_{\boldsymbol{Z}_i}(\mathcal{S}) = \mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S})$, using Monte Carlo simulations, since $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, h\mathbb{I})$,
- we compute $\omega_i^0(h)$ based on $\mu^0_{r,\boldsymbol{Z}_i}(\mathcal{S})$ for different values of r,
- for some norm $\|\cdot\|$, the optimal radius $r^\star$ is solution of

$$r^\star(h) = \text{argmin}\left\{\sum_{i=1}^{n} \|\omega_i^0(h) - \omega_i(h)\|\right\},$$

(two norms will be consider in this study $\|x\|_1 = |x|$ and $\|x\|_2 = x^2$).

On Figure 10 is plotted $h \mapsto r^\star(h)$ on top, where a linear relationship can easily be identified, and below the slope, i.e. $h \mapsto r^\star(h)/h$. The horizontal doted line is the 1.76 value obtained empirically in the computations of the previous section (using a half-space region).

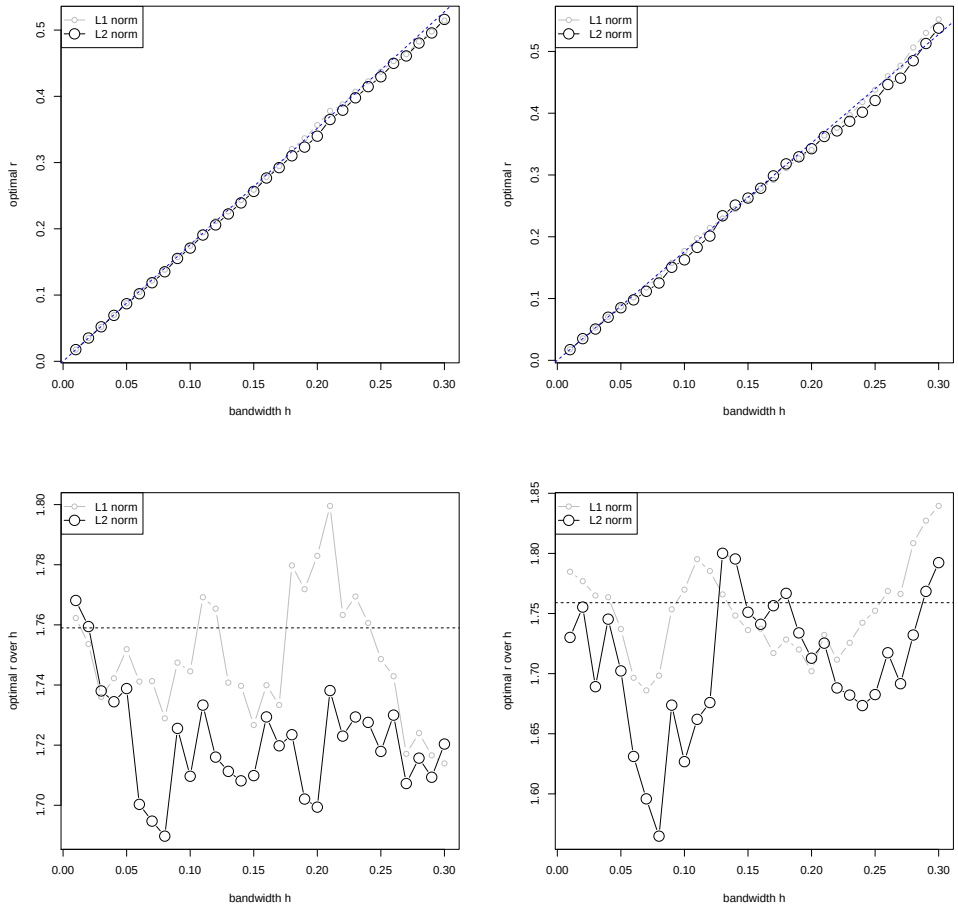

FIGURE 10. Optimal $r^\star(h)$ as a function of h on the polygon shape (on the left) and the Finistère region (on the right), from Figure 9, on top, and below, ratio of $r^\star(h)$ over h, as a function of h.

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

$$\omega_i(h)^0 = \frac{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r^\star})}{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r^\star} \cap \mathcal{S})} \text{ where } r^\star = \beta^\star h \text{ and } \beta^\star \sim 1.76.$$

4.2. **An elliptical correction.** So far, a correction using a *circular* distribution was considered, since we assumed that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H})$ were $\boldsymbol{H}$ was a diagonal covariance matrix with identical terms on the diagonal. But It is possible to consider a non-diagonal matrix $\boldsymbol{H}$ as bandwidth. In that case, level curves of the density of $\boldsymbol{\varepsilon}$ are *ellipses*. The link between covariance matrices, Cholesky decomposition and ellipses is discussed in sections of conics in matrix forms in Banchoff and Wermer (1991)

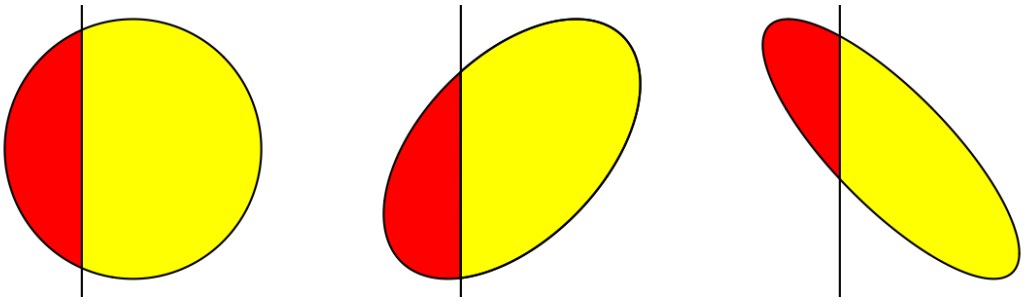

FIGURE 11. $\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i} \cap \mathcal{S})$ for half-space area $\mathcal{S}$, when $\mathcal{D}_{\boldsymbol{Z}_i}$ is an ellipse centered in $\boldsymbol{Z}_i$.

As for the circular-based correction, on average this technique provide proper approximation of weights.

4.3. **Comparison with other corrections.** A method for edge correction of an intensity estimator was introduced in **?** (**?**), including a discussion on the bandwidth estimation (see also **?** (**?**) and the estimation of relative risk (see Kelsall and Diggle (1995)). The algorithm can be found in the library `Spatstat` of `R`.

We have compared two estimators of the density on the the polygon of Figure 5, denoted $\mathcal{S}$, using Monte Carlo simulations,

- Diggle's estimator, obtained from , using

```
density(as.ppp(U,owin(poly=P)), diggle=TRUE)
```

- our estimator based on Ripley's Correction, with optimal bandwidth

$n_s = 500$ samples of size $n = 100$ and $n = 1,000$ were generated, and densities were compared. On Figure 12, we can visualize the densities, when the n points are generated uniformly on polygon $\mathcal{S}$. The density is displayed on the diagonal of the upper left corner $[B, D]$ (as defined on Figure 5). Based on those $n_s = 500$ samples, we have, for any point $\boldsymbol{z}$, estimators $\widehat{f}_1(\boldsymbol{z}), \cdots, \widehat{f}_{n_s}(\boldsymbol{z})$ (with the two techniques). On Figure 13, we plot $\bar{f}(\boldsymbol{z}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\widehat{f}_i(\boldsymbol{z})$, the average value of those estimators, as well as the mean squared error,

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

Those quantities are ploted for all $\boldsymbol{z} \in [B, D]$.

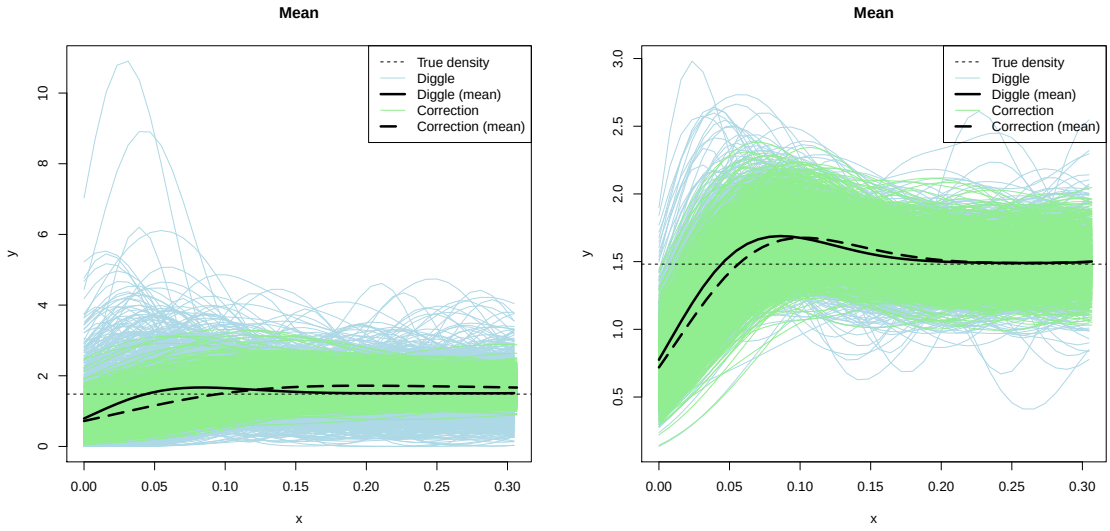

FIGURE 12. Estimation of the density in the upper left corner, on interval $[B, D]$, with $n = 100$ and $n = 1,000$ points, on the left and on the right, respectively .

On Figure 13, we can observe that the two estimators have a similar behavior, on average, but the mean-squared error is much smaller, especially on small samples.

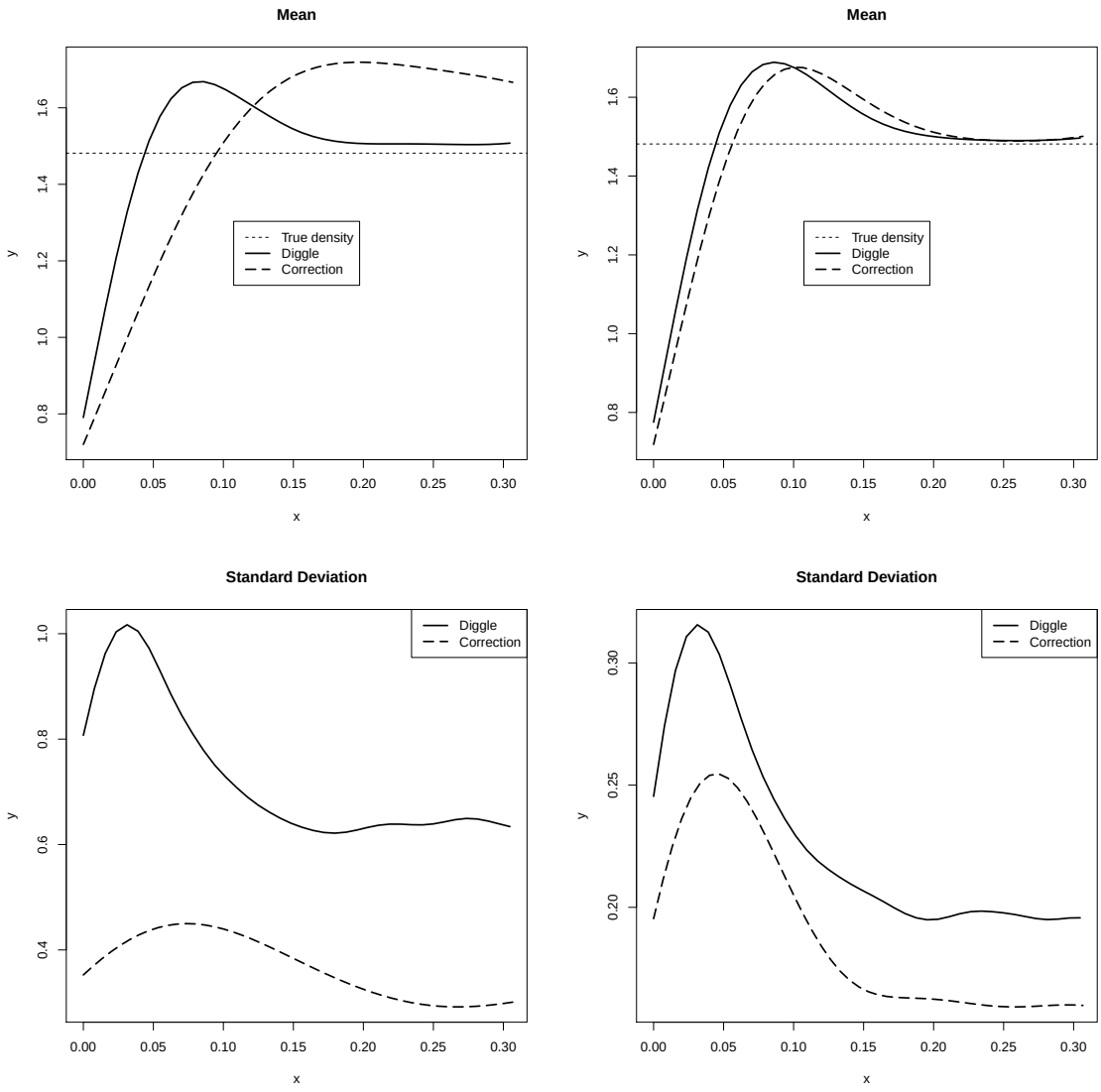

FIGURE 13. Estimation of the density in the upper left corner, on interval $[B, D]$, with $n = 100$ and $n = 1,000$ points, on the left and on the right, respectively, with the average density (on 500 samples), and the standard deviation.

5. Application to bodily injury car accidents, in France

Car accident concentration is usually identified as *black spots*, as in Nguyen (1991) or Joly *et al.* (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 accident, occurred in 2008 in France that involved bodily injuries. The BAAC dataset (*bulletins danalyse daccident corporel*) is filed by police forces, and most accident have a specific location. In 2008, we have 10854 accidents with a location.

5.1. **Spatial location of bodily injury car accidents two regions.** In order to illustrate border issues, we focus here on two specific regions, Finistère and Morbihan[1], where major cities (Brest in Finistère and Lorient, or Vannes in Morbihan are next to the sea). We have 186 observations for the first region, and 180 for the other one.

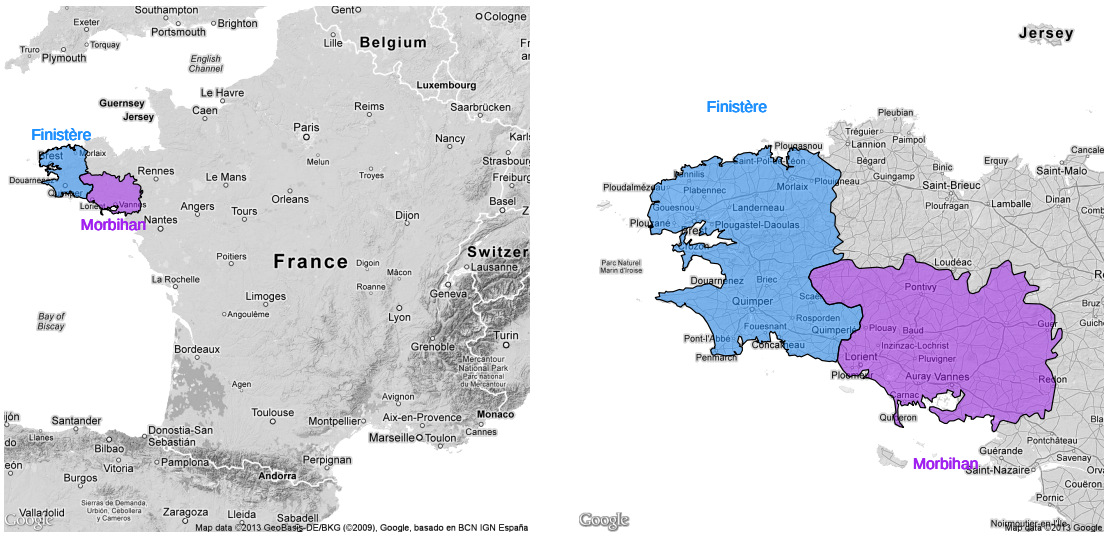

FIGURE 14. The two regions of interest, in Brittany: Finistère and Morbihan.

Results of the estimations for Finistère can be seen on Figure 15. When the standard kernel is used, we can think of at least two *black spots*, with one in the North being more important than the other one in the South coastline. When the correction is used, the two spots still show up, but another locale stands out on the lower tip

[1]Note that we have removed island, namely Belle-Ile, Ile de Groix, Ile de Hoëdic and Ile d'Houat since no traffic accident occurred on those islands in 2008.

of Finistère. The area 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 on Figure 16. The density estimation at the North-West frontier is really different depending if we use, or not, weight corrections. Once weights are applied to correct the border bias, one can easily detect a *black spot*.

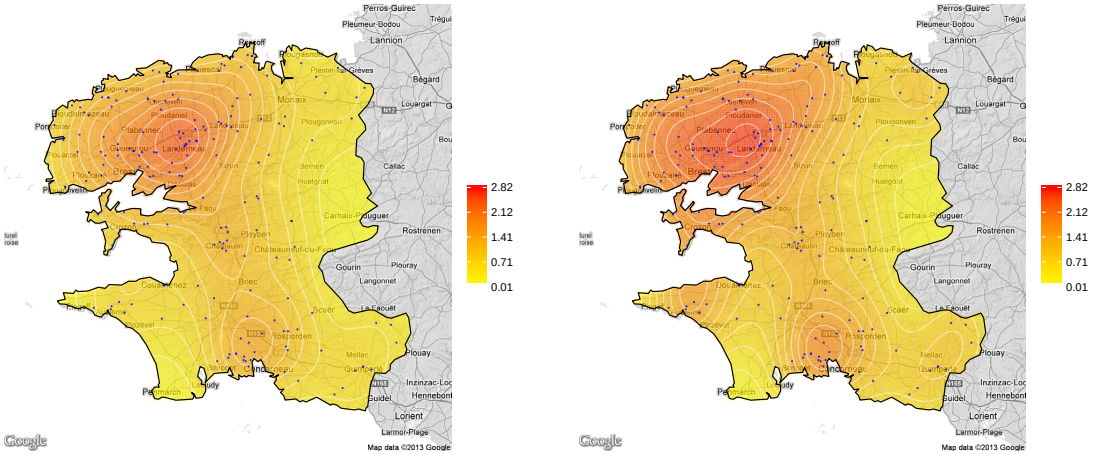

FIGURE 15. Location of car accidents, in Finistère, standard kernel on the left, and corrected one on the right.

5.2. **Density of the road network.** In the previous section, we have spotted regions or areas with a high concentration of car accident. But in order to detect *hot spots* it is necessary to have also an estimation of road density (see Anselin et al. (2012) for a discussion). As explained in the technical Appendices (Section 6), we have coordinates of road segments, with a distinction between motorway, trunk, primary, secondary and territory roads, as well as residential streets. Weights have been used to take into account traffic intensity in all areas. In those two areas (Finistère and Morbihan) we had more than one million points. The two estimators of road densities can be visualized on Figure 17.

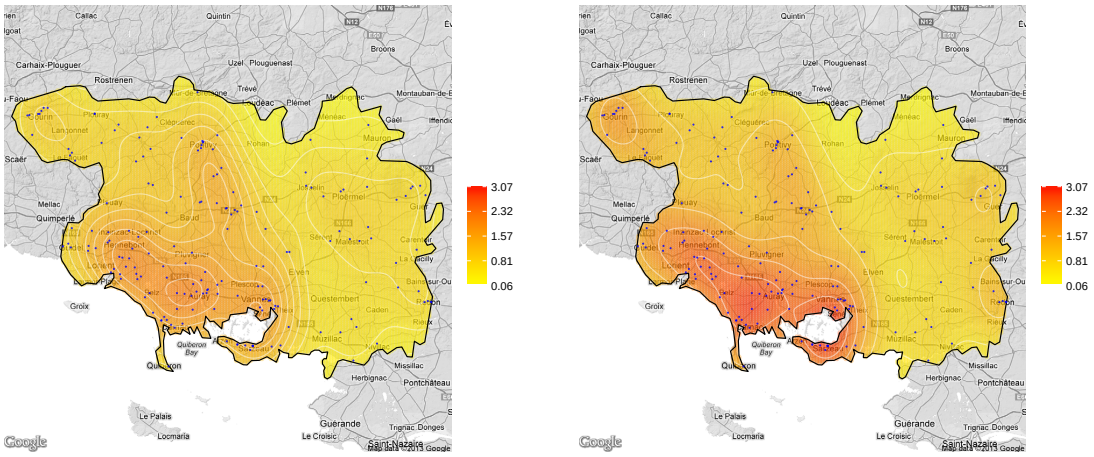

FIGURE 16. Location of car accidents, in Morbihan, standard kernel on the left, and corrected one on the right.

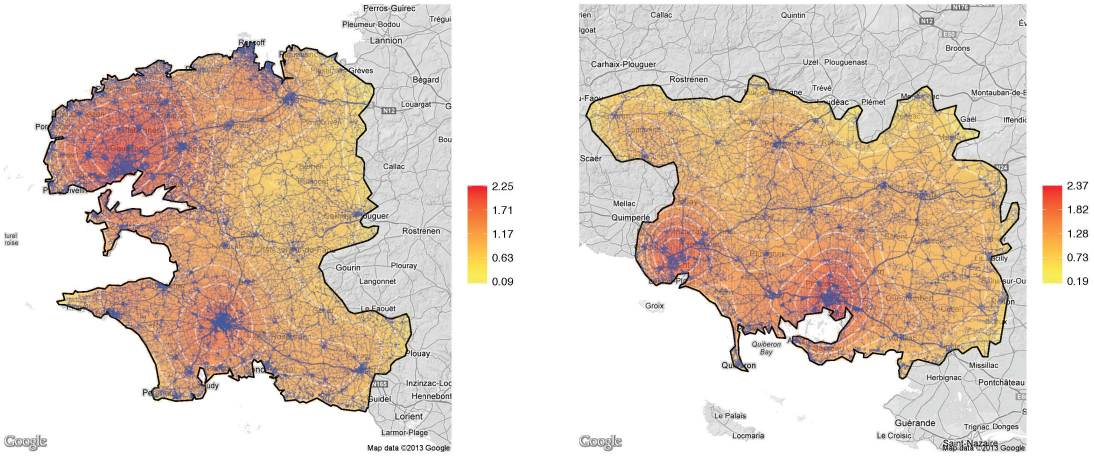

FIGURE 17. Roads in Finistère (on the left) and Morbihan (on the right)

5.3. **Detecting hot spots.** Based on the estimation of the density of car accident, and the density of road density, we have been able to isolate *hot spot* regions, using the

ratio of those two densities. On Figure 18 we can visualize location of car accident, on top, and a correction based on road density below. Some areas clearly appear are risky areas, such as the North-East part of Finistère (between Morlaix and Lannion), or the North-West part of Morbihan (near Gourin) which did not show up initially. Those two areas are on the boundary of the two regions, and we can clearly see here the impact of use a border correction to estimate properly densities.

Note that the estimation of that ratio was not very sensitive to weights chosen in the estimation of road density. Using weights proportional to the intensity of the traffic (at least a proxy of that quantity based on national statistics) gave similar maps.

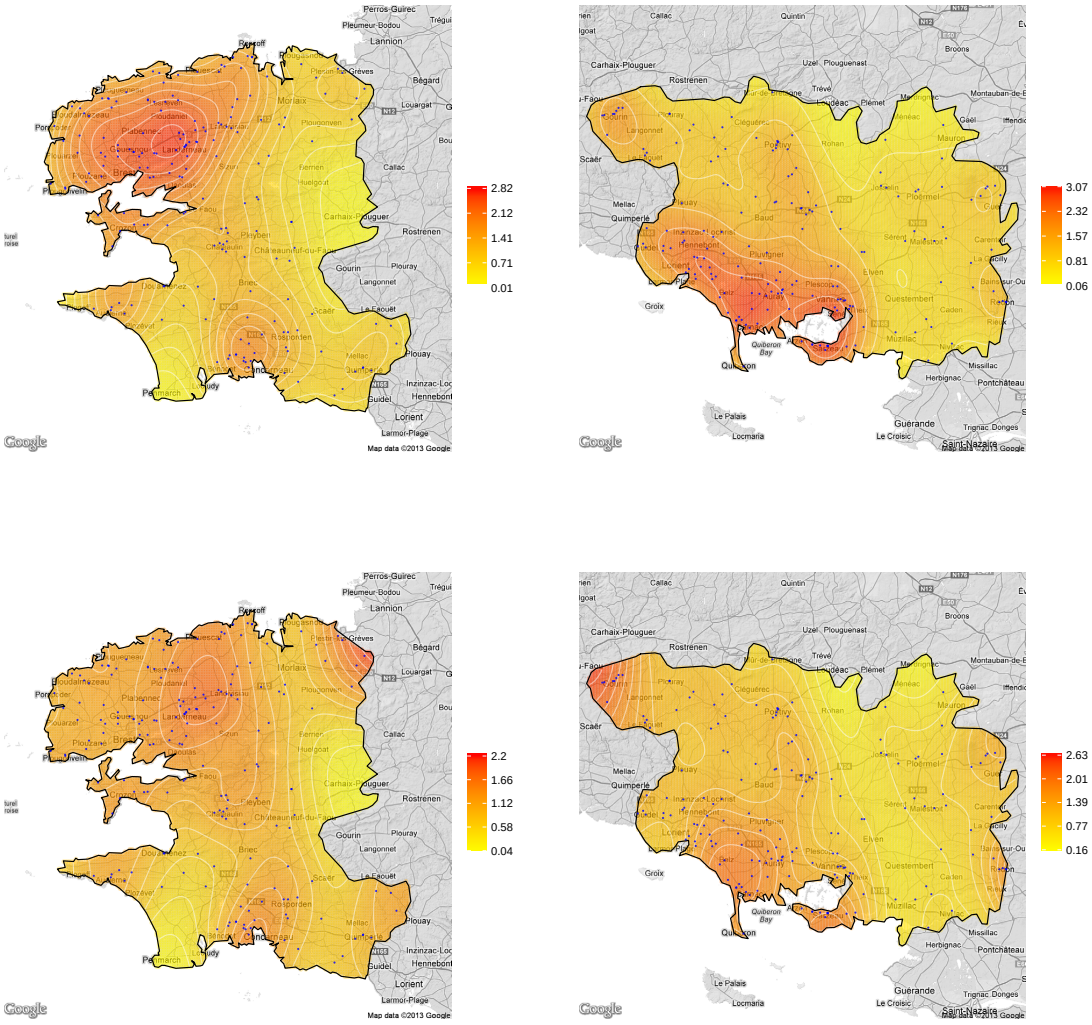

FIGURE 18. Estimated density car accident, on top, and hot spots (accident w.r.t. to road density) blow, in Finistère (on the left) and Morbihan (on the right).

6. Appendix and R code

6.1. **Density estimation.** Computation in R is extremely simple, from functions `area.poly` and `intersect` from `rgeos` and `base` packages, which allow to compute

areas of intersections of polygons. If `pol` is a polygon (latitude and longitude of knots), if `sCircle` is a function that computes the circle polygon centered in `x`, i.e.

```
sCircle <- function(n = 100, center = c(0, 0), radius) {
  theta <- seq(0, 2 * pi, length = n)
  m <- cbind(cos(theta), sin(theta)) * radius
  m[, 1] <- m[, 1] + center[1]
  m[, 2] <- m[, 2] + center[2]
  colnames(m) <- c("x", "y")
  return(m)
}
```

Then the weight function associated to observation `x` is

```
sWeights <- function(x, h, pol) {
  theCircle <- sCircle(center = x, radius = 5/pi * h)
  circlePol <- as(theCircle[-nrow(theCircle), ], "gpc.poly")
  return(area.poly(intersect(pol, circlePol))/area.poly(
     circlePol))
}
```

More generally, it is possible to consider an `ellipse` function

```
sCircle <- function(n = 200, center = c(0, 0), radius,
   correlation) {
  theta <- seq(0, 2 * pi, length = 100)
  MAT <- chol(matrix(c(1,correlation,correlation, 1), nrow=2))
  m <- cbind(cos(theta), sin(theta)) %*% MAT * radius
  m[, 1] <- m[, 1] + center[1]
  m[, 2] <- m[, 2] + center[2]
  names(m) <- c("x", "y")
  return(m)
}
```

Then, the code to compute kernel estimate is based on the `kde` function from `ks` package: the first step is to compute the optimal (standard) bandwidth, `H=Hpi(X)`. In the case of circular wieghts, let

```
H = matrix(c(sqrt(H[1,1]*H[2,2]),0,0,sqrt(H[1,1]*H[2,2])),2,2)
```

Then, weights need to be computed for all the observations, i.e.

```
wHelp <- function(i) {
  sWeights(x = as.numeric(X[i, ]), h = sqrt(H[1, 1]),
    pol = as(pol, "gpc.poly"))
}
OMEGA <- Vectorize(wHelp)(1:nrow(X))
```

Then, the weighted kernel estimator has to be computed (and normalized) on a rectangular grid which contains all the observations

```
fhat <- kde(X, H, w = 1/OMEGA, xmin = c(min(pol[, 1]), min(pol
   [, 2])), xmax = c(max(pol[, 1]), max(pol[, 2])))
fhat$estimate <- fhat$estimate * sum(1/OMEGA)/nrow(X)
```

Since we compute the density outside the region, we shall put *empty* values outside the region of interest,

```
vx <- unlist(fhat$eval.points[1])
vy <- unlist(fhat$eval.points[2])
VX <- cbind(rep(vx, each = length(vy)))
VY <- cbind(rep(vy, length(vx)))
VXY <- cbind(VX, VY)
Ind = matrix(point.in.polygon(VX, VY, pol[, 1], pol[, 2]),
  length(vy), length(vx))
f0 <- fhat
f0$estimate[t(Ind) == 0] <- NA
```

Finally, let us put the result in a list

```
result = list(
  X = fhat$eval.points[[1]],
  Y = fhat$eval.points[[2]],
  Z = fhat$estimate,
  ZNA = f0$estimate,
  H = fhat$H,
  W = fhat$w
)
```

6.2. **Maps.** In order to display the results of the estimation of the density on a map, we provide two ways. The first one is extremely simple, and produces a quick visualization. The other one, based on the `ggmap` function from `ggmap` package, is a bit more sophisticated and requires a few more steps but offers an aesthetic result.

First, one needs to define a color scale and associated breaks

```
breaks <- seq(min(result$ZNA, na.rm = TRUE) * 0.95, max(result
   $ZNA, na.rm = TRUE) * 1.05, length = 21)
col <- rev(heat.colors(20))
```

to finally plot the estimation

```
image.plot(result$X, result$Y, result$ZNA, xlim = range(pol[,
  1]), ylim = range(pol[, 2]), breaks = breaks, col = col,
  xlab = "", ylab = "", xaxt = "n", yaxt = "n", bty = "n",
  zlim = range(breaks), horizontal = TRUE)
```

It is possible to add a contour, the observations, and the border of the polygon

```
contour(result$X, result$Y, result$ZNA, add = TRUE, col = "
   grey")
points(X[, 1], X[, 2], pch = 19, cex = 0.2, col = "dodger blue
   ")
polygon(pol, lwd = 2)
```

Now, if one wants to improve the aesthetics of the map, by adding a Google Maps base map, the first thing to do – after loading ggmap package – is to get the base map

```
theMap <- get_map(location=c(left=min(pol[,1]), bottom=min(pol
   [,2]), right=max(pol[,1]), top=max(pol[,2])), source="
   google", messaging=F, color="bw")
```

Data need to be put in the right format

```
getMelt <- function(smoothed){
  res <- melt(smoothed$ZNA)
  res[,1] <- smoothed$X[res[,1]]
  res[,2] <- smoothed$Y[res[,2]]
  names(res) <- list("X","Y","ZNA")
  return(res)
}

smCont <- getMelt(result)
```

Breaks and labels should be prepared

```
theLabels <- round(breaks,2)
indLabels <- floor(seq(1,length(theLabels),length.out=5)) #
   keep only 5 items
indLabels[length(indLabels)] <- length(theLabels) # make sure
   to have the max value
theLabels <- as.character(theLabels[indLabels])
theLabels[theLabels=="0"] <- "0.00"
```

Now, the map can be built

```
P <- ggmap(theMap)
P <- P + geom_point(aes(x=X, y=Y, col=ZNA), alpha=.3, data=
   smCont[!is.na(smCont$ZNA),], na.rm=TRUE)
```

It is possible to add a contour

```
P <- P + geom_contour(data=smCont[!is.na(smCont$ZNA),] ,aes(x=
   X, y=Y, z=ZNA), alpha=0.5,  colour="white")
```

Colors need to be updated

```
P <- P + scale_colour_gradient(name="", low="yellow", high="
   red", breaks=breaks[indLabels], limits=range(breaks),
   labels= theLabels)
```

To remove the axis legends and labels, the theme should be updated

```
P <- P + theme(panel.grid.minor=element_line(colour=NA), panel
   .grid.minor=element_line(colour=NA), panel.background=
   element_rect(fill=NA, colour=NA), axis.text.x=element_blank
   (), axis.text.y=element_blank(), axis.ticks.x=element_blank
   (), axis.ticks.y=element_blank(), axis.title=element_blank
   (), rect=element_blank())
```

The final step, in order to draw the border of the polygon

```
polDF <- data.frame(pol)
colnames(polDF) <- list("lon","lat")
(P <- P + geom_polygon(data=polDF, mapping=(aes(x=lon, y=lat))
   , colour="black", fill=NA))
```

6.3. **Road data.** Road data used in the application were obtained from Geofabrik website (http://www.geofabrik.de/index.html), which provides files based on OpenStreetMap data. Each observation is a section of a road, and contains a few points identified by their geographical coordinates that allow to draw lines. It is obvious that less points are needed for straight roads than winding path. Hence, if one were to estimate the density only with these knots, smaller roads would weight more than highways. Therefore, it is necessary to interpolate points between the extremities of

each road section. Since the type of road is provided in almost all cases[2], we propose to use it as a proxy for traffic. The distance between interpolated points is reduced for busier roads. Weights chosen in the application are displayed in table 1.

TABLE 1. Road types

Road type	Weight	Finsitère		Morbihan	
		No. Points	%	Nb. Points	%
motorway	10	0	0	0	0
motorway_link	10	0	0	0	0
trunk	9	20634	3.15	25973	5.82
trunk_link	9	3651	0.56	4439	0.99
primary	8	35595	5.44	13004	2.91
primary_link	8	669	0.10	473	0.11
secondary	7	63724	9.74	64569	14.46
secondary_link	7	182	0.03	231	0.05
tertiary	6	70037	10.71	70850	15.86
tertiary_link	6	202	0.03	224	0.05
living_street	5	1199	0.18	1359	0.30
residential	4	111099	16.98	67307	15.07
unclassified	3	318470	48.69	178030	39.86
service	2	25050	3.83	17865	4.00
road	1	3602	0.55	2313	0.52
Sum		654114	100	446637	100

Note : all duplicated points have been removed.

Let `listroad` be a list of length n, whose $(n-1)$ first elements correspond to road sections, the n^{th} element being the type of road, stored as a character. Let `types.weights` be a data frame in which every line gives the type of road and the

[2]more details can be found one the OpenStreetMap : http://wiki.openstreetmap.org/wiki/Map_Features

associated weight needed for the interpolation. Then, a way to derive a set of points according to the importance of the road is given by the `splitroad` function:

```
splitroad <- function(listroad, h = 0.0025) {
  pts = NULL
  weights <- types.weights[match(unique(listroad$type), types.
     weights$type), "weight"]
  for (i in 1:(length(listroad) - 1)) {
    d = diag(as.matrix(dist(listroad[[i]]))[, 2:nrow(listroad
       [[i]])])
    for (j in 1:(nrow(listroad[[i]]) - 1)) {
      pts = rbind(pts, cbind(seq((listroad[[i]])[j, 1], (
         listroad[[i]])[j + 1, 1], length = weights * d[j]/h),
        seq((listroad[[i]])[j, 2], (listroad[[i]])[j + 1, 2],
           length = weights * d[j]/h)))
    }
  }
  return(pts)
}
```

Acknowledgements. The authors would like to thank Olivier Scaillet and John Wilson for stimulating comments, and helping us to improve the paper.

References

Ang, Q., Baddeley, A.J., Nair, G. 2012, Geometrically Corrected Second Order, with Applications to Ecology and Criminology. *Scandinavian Journal of Statistics*, **39**:4, 591–617.

Anselin, L. and Florax, R.J.G.M. 1995. *New Directions in Spatial Econometrics*, Springer, Berlin.

Banchoff , T. and Wermer, J. 1991. Linear Algebra Through Geometry. Springer Verlag.

Basawa, I.V. 1996a. Special Issue on Spatial Statistics, Part I, *Journal of Statistical Planning and Inference*, 50: 311–411.

Basawa, I.V. 1996b. Special Issue on Spatial Statistics, Part II, *Journal of Statistical Planning and Inference*, 51:1–97.

Bailey, T.C. and Gatrell, A.C. 1995. Interactive Spatial Data Analysis. Harlow, Longman

Batty, M. 2005. Network geography: Relations, interactions, scaling and spatial processes in GIS. *in* Unwin and Fisher (eds.) *Re-presenting Geographical Information Systems*. Chichester, John Wiley and Sons: 14970

Black, W. R. 1991. Highway Accidents: A Spatial and Temporal Analysis. *Transportation Research Record*, 1318, 7582.

Block, C.R., M. Dabdoub, and S. Fregly, eds. 1995. Crime Analysis Through Computer Mapping. Washington, DC: Police Executive Research Forum

Borruso, G. 2008. Network Density Estimation: A GIS Approach for Analysing Point Patterns in a Network Space *Transactions in GIS* 12:3, 377402.

Bouezmarni, T. and Rombouts, J.V.K., 2010 Nonparametric Density Estimation for. Multivariate Bounded Data, *Journal of Statistical Planning and Inference*, 140, 139–152.

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.

Brunsdon, C., Corcoran, J., and Higgs, G. 2007 Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems* 31: 52–75.

Bryc, W., 2002 A Uniform Approximation to the Right Normal Tail Integral, *Applied Mathematics and Computation*, **127**, 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:

80726

Charpentier, A., Fermanian, J-D. & Scaillet, O. 2006 The Estimation of Copulas: Theory and Practice, *in* Copulas, from theory to application in Finance, J. Rank (editor), Risk Books.

Chen, S.X., 1999. Beta kernel estimators for density functions, *Computational Statistics & Data Analysis*, 31(2), 131–145

Chiu, S.T. 1991 Bandwidth Selection for Kernel Density Estimation, *The Annals of Statistics*, 19(4) 1883–1905

Davis, M 1975 Mean square error properties of density estimates. *Annals of Statistics*, 3, 1025–1030.

Devroye, L. & Györfi, L. 1981 Nonparametric density estimation: the $L1$ view. John Wiley & Sons.

Diggle, P. 1985. A Kernel Method for Smoothing Point Process Data. *Applied Statistics*, **34** (2), 138–147.

Diggle, P., Heagerty, P., Liang, K.Y. & Zeger, S. 2002. Analysis of Longitudinal Data. Oxford University Press.

Eck, J.E. 1997. What do those dots mean? Mapping theories with data. *in* Weisburd and McEwen (eds.) *Crime Mapping and Crime Prevention*. Monsey, NY: Criminal Justice Press, pp. 379-406.

Epanechnikov, V A 1969 Nonparametric estimation of a multivariate probability density. *Theory of Probability and Its Applications* 14: 153–158.

Gatrell, A. 1994. Density estimation and the visualisation of point patterns. *in* Hearnshaw and Unwin (eds.) *Visualisation in Geographical Information Systems*. Chichester, John Wiley and Sons: 6575.

Getis, A. 1964. Temporal Land Use Pattern Analyses with the Use of the Nearest Neighbor and Quadrat Methods. *Annals of the Association of American Geographers*, 54, 391-98.

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

Hall, P. and Turlach, B. 1999. Reducing bias in curve estimation by use of weights. *Computational Statistics and Data Analysis*, 30, 6786

Härdle W, Müller M, Sperlich S, and Werwatz A 2004 Nonparametric and Semiparametric Models. Berlin, Springer.

Joly, M-F Bourbeau, R Bergeron, J & Messier, S 1992 Analytical Approach to the Identification of Hazardous Road Locations: A Review of the Literature *Centre de recherche sur les transports, Universit de Montral*, CRT publication No. 815

Kelsall, J.E. and Diggle, P.J. 1995. Kernel estimation of relative risk. *Bernoulli*, **1**, 3–16.

Krisp, J.M. and Durot, S. 2007. Segmentation of lines based on point densities : an optimisation of wildlife warning sign placement in southern Finland. *Accident Analysis and Prevention*, 39(1), 38-46.

Levine, N. K.E. Kim 1998 The location of motor vehicle crashes in Honolulu: a methodology for geocoding intersections *Computers, Environment and Urban Systems*, 22(6) 557–576.

Loo, B.P.Y. 2006 Validating Crash Locations for Quantitative Spatial Analysis: A GIS-Based Approach *Accident Analysis & Prevention*, 38(5), 879-886,

J.S. Marron and W.J. Padgett. 1987 Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples, *The Annals of Statistics*, 15, 1520–1535.

Miller H J 1999 Measuring space-time accessibility benefits within transportation networks: Basic theory and computational methods. Geographical Analysis 31: 187212

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. N 1991 Identification of Accident Blackspot Locations, an Overview VIC Roads/Safety Division, Research and Development Department, Australia. VIC Discussion Paper (DP/91/4)

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

OSullivan, D. and Unwin, D. J. 2002. Geographic Information Analysis. John Wiley, Hoboken, New Jersey.

Pulugurtha, S.S., Krishnakumar, V. K., and Nambisan, S. S. 2007. New methods to identify and rank high pedestrian crash zones: An illustration. *Accident Analysis and Prevention*, 39(4,) 800-811.

Ripley, B. 1981. Spatial Statistics, Wiley, New York.

Rogers, A. 1965. A Stochastic Analysis of the Spatial Clustering of Retail Establishments. *Journal of the American Statistical Association*, 60, 1094-1103.

Saffet E., Ibrahim, Y., Tamer, B., Mevlut G. 2008. Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. *Accident, analysis and prevention* 40(1) 174-181.

Scaillet, O. 2004. Density estimation using inverse and reciprocal inverse Gaussian kernels *Nonparametric Statistics*, 16, 217–226.

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

Shah, A.K. 1985 A Simpler Approximation for Areas under the Standard Normal Curve, *The American Statistician*, 39(1), 80.

Silverman B W 1986 Density Estimation for Statistics and Data Analysis. London, Chapman & Hall.

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. and Carrol,R.J. 1990. Deconvoluting Kernel Density Estimators. *Statistics*, 21, 2, 169–184.

Tapia, R. A. and Thompson, J. R. 1978, Nonparametric Probability Density Estimation, John Hopkins Univiverity Press, Baltimore.

Taylor, P.J. 1977. Quantitative Methods in Geography: An Introduction to Spatial Analysis. Boston, MA: Houghton Mifflin Company.

Thomas, R.W. 1977. An introduction to Quadrat analysis. Concepts and Techniques in Modern Geography, Institute of British Geographers.

Treno, A.J., Johnson, F.W., Remer, L.G. , and Gruenewald, P.J. (2007) The impact of outlet densities on alcohol-related crashes: A spatial panel approach *Accident Analysis and Prevention*, 39:5, 894-901.

Warden, Craig R., Duh, J-D., Lafrenz, M., Chang, H. and Monsere, C. 2011. Geographical analysis of commercial motor vehicle hazardous materials crashes on the Oregon state highway system *Environmental Hazards*, 10(2) 171-184

Xie, Z. and Yan, J. 2008. Kernel Density Estimation of Traffic Accidents in a Network Space. *Computers, Environment, and Urban Systems*, 35(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: 97109

Yamada, I. and Thill, J. 2004. Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography*, 12: 14958.