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). We provide a simple technique - based on optimal bandwidth using Gaussian kernels - to compute efficiently weights to correct border bias on frontiers of the region of interest. An illustration on location of bodily-injury car accident in the western part of France is discussed.

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 spa ce (see Taylor (1977)). Spatial (and temporal) patterns along a certain roadway segment is largely determined by its 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 Jun (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 Batt (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* an 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),

Batt (2005) or Borruso (2008)). The goal here is still to obtain a 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. Nevertheless, the time component will not be discussed in this paper. Them in Section 3, we will discuss frontiers and space border bias correction. In this 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 the link between radius computation, and optimal bandwidth. And finally, in Section 4, we illustrate that technique on bodily injury car accidents, in western part of France (Morbihan and Finistère). `R` codes are provides in the Section 5.

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 a local $\boldsymbol{Z} = (X, Y)$ (usually characterized by a latitude and a longitude coordinate) and a time T. A natural assumption is to consider a product kernel, between location and time, as in Brunsdon *et al.* (2007). Hence,

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

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, optimal bandwidth are $h^\star = n^{-1/(3+d)}\sigma$ 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 kernels are then $\widehat{h}^\star = n^{-1/(3+d)}\widehat{\sigma}$. Further,

as mentioned in Härdle *et al.* (2004) bandwidth are rather close, from 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 belong to some specific area $\mathcal{S}$. For instance, for traffic accidents, events have to occur in-land. In that case, 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 $\kappa_{\boldsymbol{z}} \in [0, 1]$.

Hence, in that case, for regions closed to the sea, estimators of density can suffer major drawbacks.

Remark 1. *In standard statistical package, 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 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. Standard techniques for squared areas.

Consider points $\boldsymbol{Z}_i = (X_i, Y_i)$ in the unit square $[0, 1] \times [0, 1]$. The standard kernel estimator for the spatial density at point $\boldsymbol{z} = (x, y)$ is

$$\widehat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}, \frac{y - Y_i}{h}\right) \tag{3.1}$$

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

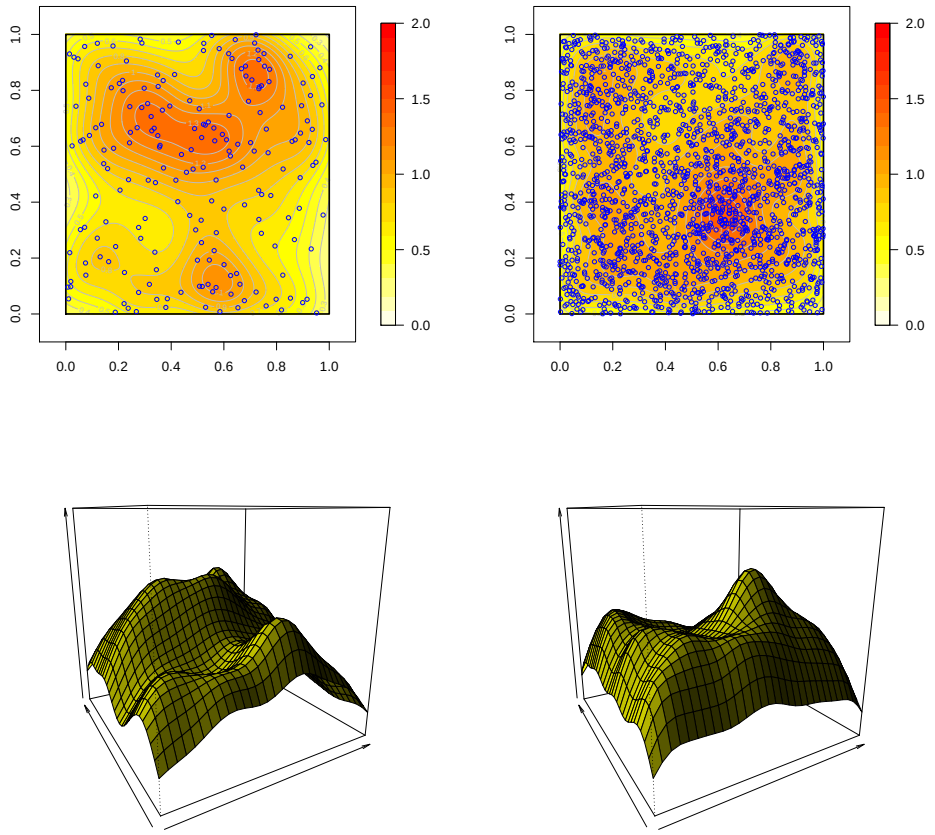

FIGURE 1. 200 and 2,000 uniformly distributed on the unit square, and kernel based estimators of respective densities.

A first idea, introduced by Devroye and Györfi (1981), is to consider a transformation of the variates. Let $\psi : [0,1] \to \mathbb{R}$ a (known) continuous increasing function (in order to preserve clusters and neighbors), and consider the transformed sample $\tilde{\boldsymbol{Z}}_i = (\psi(X_i), \psi(Y_i))$. If $\tilde{f}$ denotes the density of $\tilde{\boldsymbol{Z}}$, then

$$f(x,y) = \frac{\tilde{f}(\psi(x), \psi(y))}{(\psi^{-1\prime}(\psi(x)) \cdot \psi^{-1\prime}(\psi(x))}.$$

A natural idea is to consider ψ as a quantile function taking values on the real line, e.g. the Gaussian distribution $\psi = \Phi^{-1}$, and the standard kernel estimator on transformed observations $\tilde{\boldsymbol{Z}}_i$'s,

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

where ϕ is the density of the standard normal distribution.

A second idea, that was introduced in Chen (1999) in the univariate case and Charpentier *et al.* (2006) for the extension in higher dimension, is to consider products of beta kernels,

$$\widehat{f}(x,y) = \frac{1}{n} \sum_{i=1}^{n} K\left(X_i, \frac{x}{b} + 1, \frac{1-x}{b} + 1\right) \cdot K\left(Y_i, \frac{y}{b} + 1, \frac{1-y}{b} + 1\right),$$

where $K(\cdot, \alpha, \beta)$ denotes the density of the Beta distribution with parameters α and β. Note that this idea of finding a kernel family having a support which fits exactly with the support of the observations has been intensively used, for positive valued observations, see Scaillet (2004) for the univariate case and Bouezmarni and Rombouts (2010) for the multivariate case.

Those two techniques are extremely popular, but unfortunately, not appropriate to the study of geographic patterns, where the support is a region $\mathcal{S}$ identified as a polygon, but not necessarily a rectangle. An alternative is to recall that kernel estimators of densities can be seen as the expected value of then 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 Carrol (1990). Recall that the empirical cumulative distribution function is the step function defined as

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

and the associated empirical measure is

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

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

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mu_{\boldsymbol{Z}_i}(\boldsymbol{z}) \tag{3.4}$$

where $\mu_{\boldsymbol{Z}_i}$ can be the density of a Gaussian vector, centred 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 spead some weight in areas where no observations can be found. Thus, it might be natural to consider a truncated distribution, restricted to the support $\mathcal{S}$.

$$\mu_{\boldsymbol{Z}_i|\mathcal{S}}(\boldsymbol{z}) = \frac{\mu_{\boldsymbol{Z}_i(\boldsymbol{z})}}{\mu_{\boldsymbol{Z}_i}(\mathcal{S})} \tag{3.5}$$

Thus, it is natural to consider

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mu_{\boldsymbol{Z}_i|\mathcal{S}}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \omega_i \cdot \mu_{\boldsymbol{Z}_i(\boldsymbol{z})} \text{ where } \omega_i = \mu_{\boldsymbol{Z}_i}(\mathcal{S})^{-1}. \tag{3.6}$$

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

$$\frac{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})}{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i})} \tag{3.7}$$

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 3.3). Thus, here the idea is simply to use *weighted kernel estimators*

$$\widehat{f}(\boldsymbol{z}) = \sum_{i=1}^{n} \omega(\boldsymbol{Z}_i) \cdot \det(\boldsymbol{H})^{-1} K\left(\boldsymbol{H}^{-1}(\boldsymbol{z} - \boldsymbol{Z}_i)\right) \tag{3.8}$$

where weights $\omega(\boldsymbol{X}_i)$ 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.*

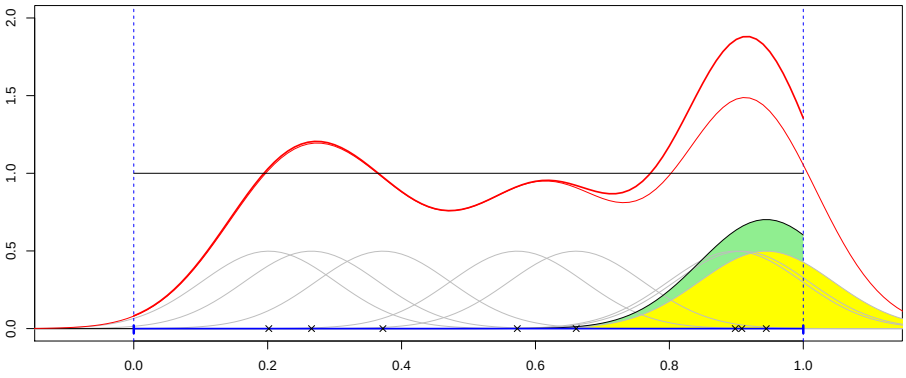

FIGURE 2. Weight correction of a density on $[0, 1]$: the kernel is no longer a Gaussian density, but a truncated Gaussian density.

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.

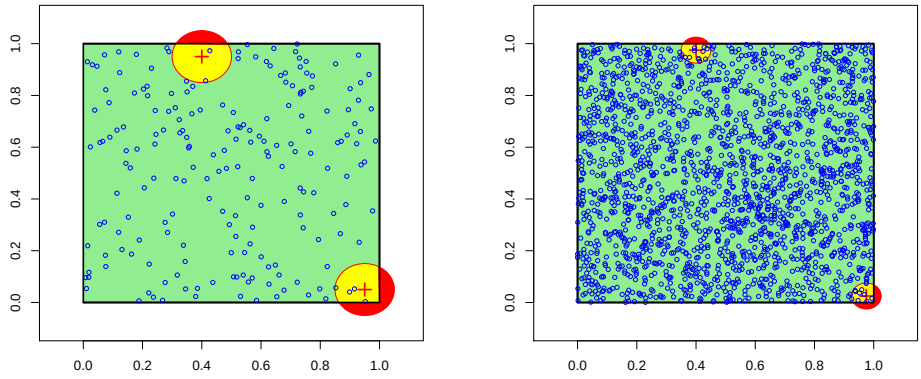

FIGURE 3. 200 and 2,000 uniformly distributed on the unit square.

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* **??** *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 is can be visualized on Figure 7. The choice of the radius will be discussed in the next section.*

3.3. **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

$$\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S}) \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}). \tag{3.9}$$

Assume for convenience that $\boldsymbol{H}$ is a diagonal matrix with identical terms on the diagonal, denoted h (this assumption will be relaxed at the end of this section), so

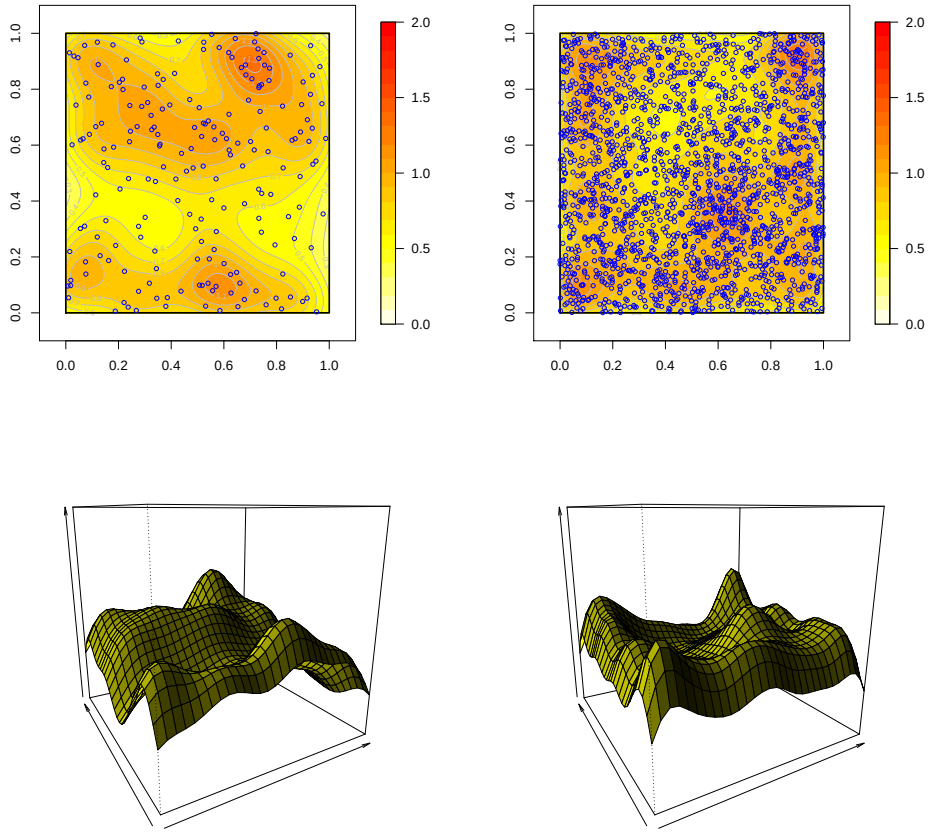

FIGURE 4. 200 and 2,000 uniformly distributed on the unit square, and kernel based estimators of respective densities, with weight correction.

that level curves of the density of $\boldsymbol{Z}$ are circles. If $\mathcal{S}$ is a half-space, and if the distance between $\boldsymbol{Z}_i$ and the border is a, then

$$\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S}) = 1 - \Phi(-ah^{-1}) = \Phi(ah^{-1}) \tag{3.10}$$

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

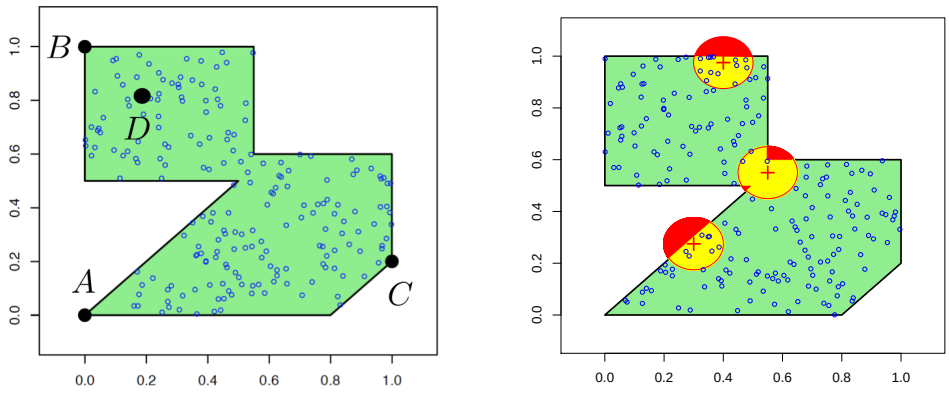

FIGURE 5. 200 points uniformly distributed on $\mathcal{S}$

The proxy we suggest for $\mu_{\boldsymbol{Z}_i}(\mathcal{S})$ is to consider the following ratio

$$\frac{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})}{\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i})} \tag{3.11}$$

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 *segment of a circle* (the region bounded by a chord and the arc subtended by the chord, see Figure 8).

$$\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S}) = \underbrace{\frac{2\pi - \theta}{2\pi}\pi r^2}_{\text{sector area}} - \underbrace{a\sqrt{r^2 - a^2}}_{\text{triangle area}} \text{ where } \cos\left(\frac{\theta}{2}\right) = \frac{a}{r}. \tag{3.12}$$

Thus,

$$\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S}) = \begin{cases} (\pi - \operatorname{acos}(-ar^{-1}))r^2 + a\sqrt{r^2 - a^2} \text{ if } a < r \\ 0 \text{ if } a > r \end{cases} \tag{3.13}$$

If we want those two quantities to be close, we should have

$$\Phi(ah^{-1}) \approx \frac{(\pi - \operatorname{acos}(ar^{-1}))r^2 + a\sqrt{r^2 - a^2}}{\pi r^2}\mathbf{1}(a < r), \tag{3.14}$$

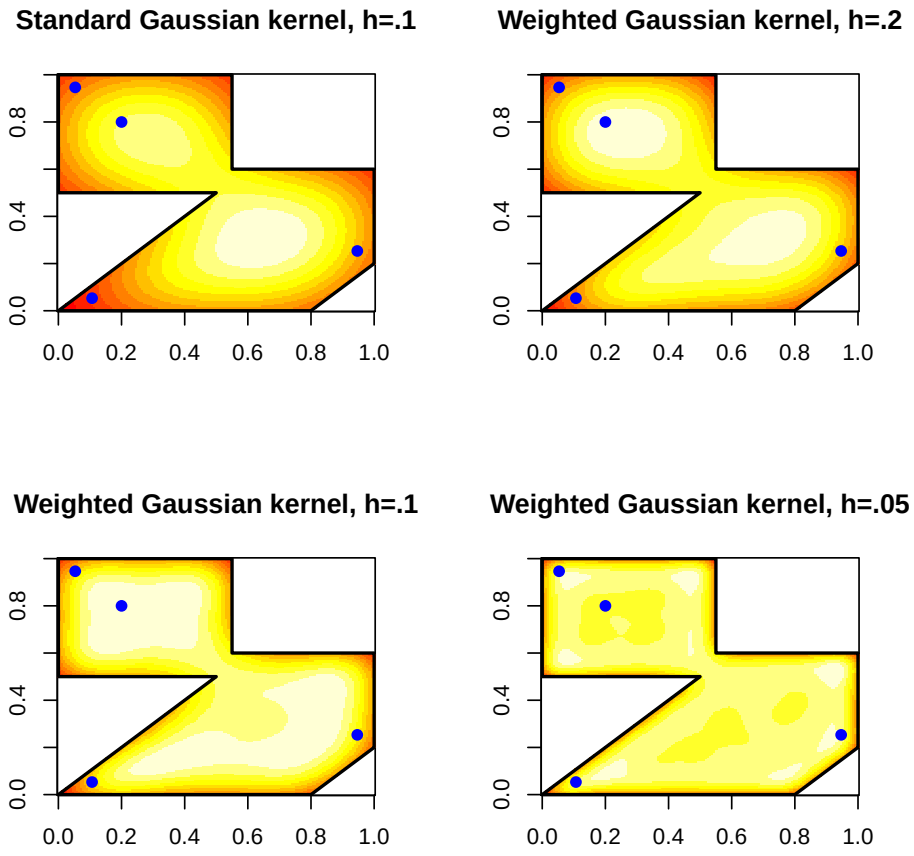

FIGURE 6. Average level curves of $\widehat{f}$ over $1,000$ simulated samples of size 200, on the left. *Optimal* (standard) bandwidth is $h = 0.1$. Weighted kernels have been calculated with 3 different bandwidth $h = 0.2$, $h = 0.1$ and $h = 0.05$, and using then $r^\star = \pi h/5$

i.e. it is possible to relate h and r, when a is fixed. As shown on Figure 9, on the left, a good *optimal* value might be

$$r^\star = \frac{\pi}{5} h, \tag{3.15}$$

which is the slope when h is large enough.

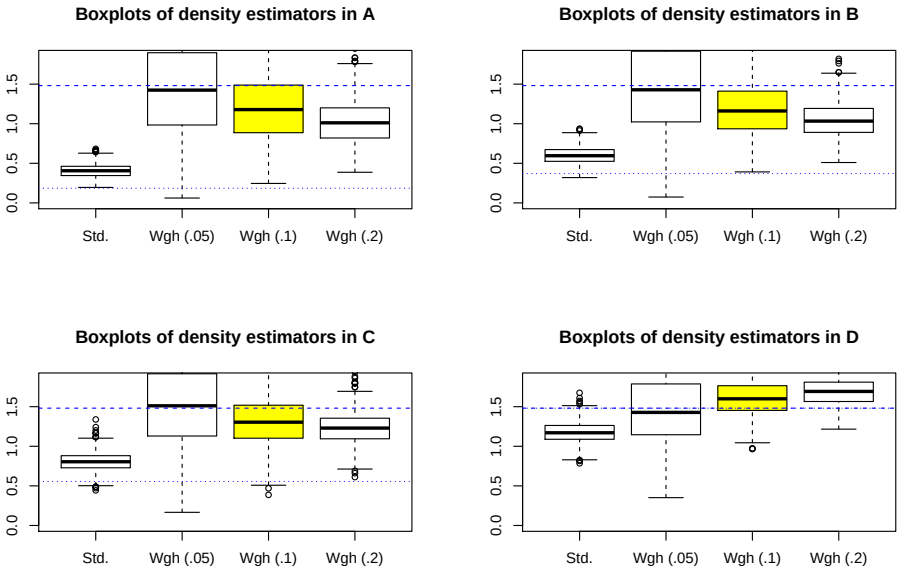

FIGURE 7. Average level curves of $\widehat{f}$ over $1,000$ simulated samples of size 200, on the left. *Optimal* (standard) bandwidth is $h = 0.1$. Weighted kernels have been calculated with 3 different bandwidth $h = 0.2$, $h = 0.1$ and $h = 0.05$, and using then $r^\star = \pi h/5$

This relationship has been derived in the context of half-space domains. In order to visualize the goodness of that approximation, we can compare theoretical weights (obtained by Monte Carlo simulation of Gaussian random vectors) and approximated one on the previous sample with 200 points uniformly distributed (see Figure 9, on the right).

Remark 2. *Using Taylor's expansion, when a tends to 0, since $acos(ar^{-1}) \sim \pi/2 - ar^{-1}$, we have*

$$\frac{a}{h} \sim \Phi^{-1}\left(\frac{1}{2} + \frac{a}{\pi r} + a\sqrt{1 - \frac{a}{r}}\right) \sim \Phi^{-1}\left(\frac{1}{2} + \frac{a}{\pi r}\right) \tag{3.16}$$

Then we use the expansion $\Phi^{-1}(1/2 + u) = 5u/2$ and we have that

$$h \sim \frac{5}{\pi} r \text{ or alternatively } r \sim \frac{\pi}{5} h. \tag{3.17}$$

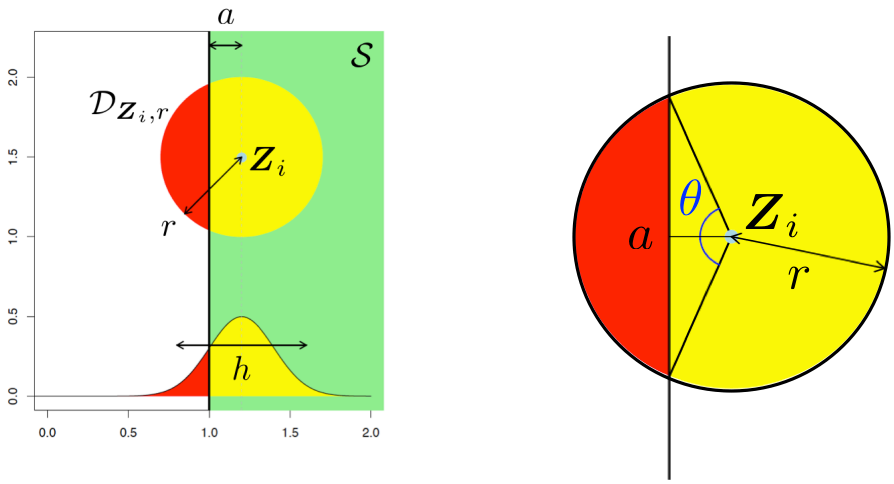

FIGURE 8. Link between $\mathbb{P}(\boldsymbol{Z}_i + \boldsymbol{\varepsilon} \in \mathcal{S})$ and $\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i,r} \cap \mathcal{S})$ where $\mathcal{S}$ is a half-space.

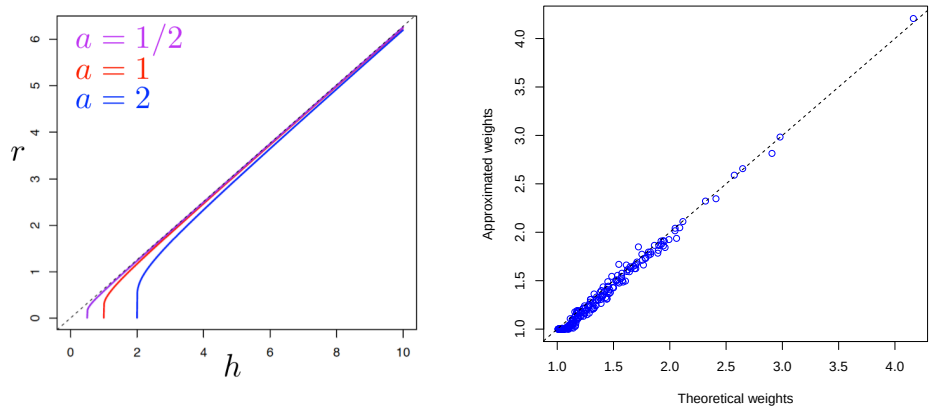

FIGURE 9. On the right, theoretical weights, $\Phi(-a_i h^{-1})^{-1}$, where a_i denotes the distance to the border, and approximated wights $(\pi^{-1}\text{acos}(a_i r^{-1}) - a\sqrt{1-(a_i/r)^2})\mathbf{1}(a_i < r))$ where $r^\star = \pi h^\star/5$.

3.4. **An elliptical correction.** So far, a correction using a *circular* distribution was considered, since we assumed that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{H})$ 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)

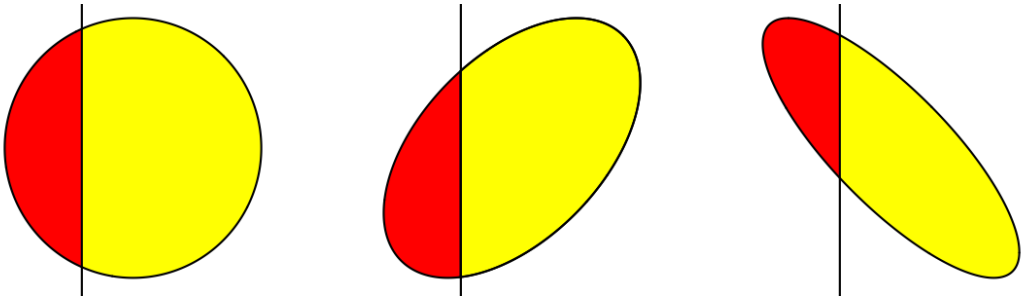

FIGURE 10. $\mathcal{A}(\mathcal{D}_{\boldsymbol{Z}_i} \cap \mathcal{S})$ for half-space area $\mathcal{S}$, when $\mathcal{D}_{\boldsymbol{Z}_i}$ is an ellipse centered in $\boldsymbol{Z}_i$.

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

4. 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 SAAC dataset (*bulletins danalyse daccidents corporels*) is filed by police forces, and most accident have a specific location. In 2008, we have 10854 accidents with a location.

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

[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

Morbihan are next to the sea). We have 186 observations for the first region, and 180 for the other one.

FIGURE 11. The two regions of interest, in Brittany: Finistère and Morbihan.

Results of the estimations for Finistère can be seen on Figure 12. 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 how up, but another locale stands out on the lower tip 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 13. 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*.

5. APPENDIX AND R CODE

Computation in `R` is extremely simple, from functions `area.poly` and `intersect` in package `R`, which allows to compute areas of intersections of polygons. If `region` is a polygon (latitude and longitude of knots), if `circle` is a function that compute the circle polygon centered in `x`, i.e.

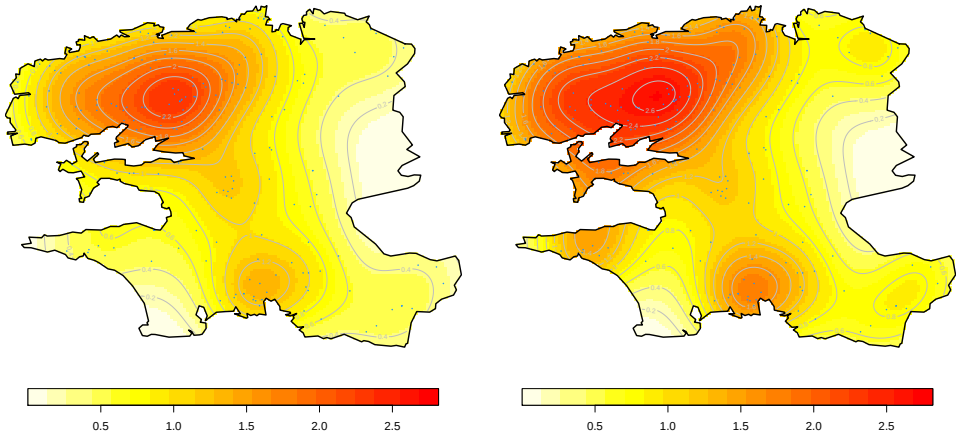

FIGURE 12. Location of car accidents, in Finistère, standard kernel on the left, and corrected one on the right.

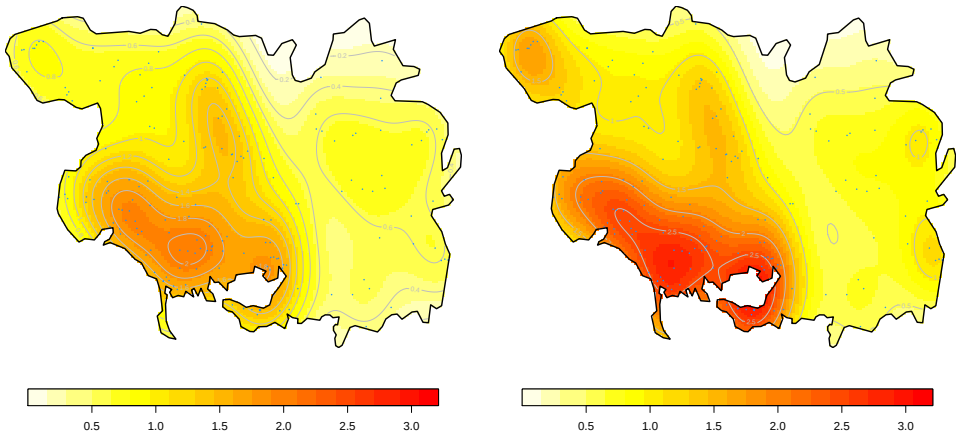

FIGURE 13. Location of car accidents, in Morbihan, standard kernel on the left, and corrected one on the right.

```
circle=function(n=200,centre=c(0,0),radius){
theta=seq(0,2*pi,length=100)
m=cbind(cos(theta),sin(theta))*radius
m[,1]=m[,1]+centre[1]
m[,2]=m[,2]+centre[2]
```

```
names(m)=c("x","y")
return(m)}
```

Then the weight function associated to observation `x` is then

```
weight=function(x,h,region){
POL=as(region, "gpc.poly")
POLcircle=as(circle(centre=x,radius=5/pi*h), "gpc.poly")
return(area.poly(intersect(POL,POLcircle))/area.poly(POLcircle))}
```

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

```
circle=function(n=200,centre=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]+centre[1]
m[,2]=m[,2]+centre[2]
names(m)=c("x","y")
return(m)}
```

Then, the code to compute kernel estimate is based on the `kde` function: 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 compute weights for all the observations, i.e.

```
W=function(i){weight(x=X[i,],h=sqrt(H[1,1]),region=polygon)}
omega=1/Vectorize(W)(1:n)
```

Then, we have to compute the weighted kernel estimator (and to renormalize it) on a rectangular grid that contain all the observations

```
fat=kde(U,H,w=omega,xmin=c(min(X[,1]),min(X[,2])),xmax=c(max(X[,1]),max(X[,2])))
fat$estimate=fhat$estimate*sum(1/OMEGAU)/n
```

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, polygon[,1],polygon[,2]),length(vy),length(vx))
fhat$estimate[t(Ind)==0]=NA
```

References

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

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

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

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

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

[6] 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

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

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

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

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

[11] 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.

[12] 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

[13] 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.

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

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

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

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

[18] 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.

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

[20] 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.

[21] 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.

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

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

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

[25] 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

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.

[32] 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)

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

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

[35] 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.

[36] Ripley, B. 1981. Spatial Statistics, Wiley, New York.

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

[38] 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.

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

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

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

[42] 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.

[43] Stefanski, L. and Carrol,R.J. 1990. Deconvoluting Kernel Density Estimators. *Statistics*, 21, 2, 169–184.

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

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

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

[47] 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.

[48] 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

[49] 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.

[50] Yamada, I. and Rogerson, P.A. 2003. An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical Analysis* 35: 97109

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