EDGE–CORRECTION FOR SPATIAL KERNEL SMOOTHING
METHODS? WHEN IS IT NECESSARY?

P. Zheng¹, P.A. Durr² and P.J. Diggle¹

¹Medical Statistics Unit, Lancaster University, Lancaster, LA1 4YF; ²Veterinary Laboratory Agency (Weybridge), Surrey, KT15 3NB

Abstract

A limitation to the practical implementation of some kernel smoothing methods for spatial data is the need for edge correction. This applies particularly to kernel density estimation. Here we demonstrate by simulation the extent of the bias introduced when edge correction is not applied for realisations of both homogeneous and inhomogeneous Poisson processes. This shows the overwhelming importance of edge-correction. We also argue that edge correction is usually not necessary for kernel regression.

Introduction

Kernel density estimation (KDE) is widely used in spatial epidemiology. Its primary use is for producing smoothed density maps of point patterns. For example, kernel estimation of farm density provides a visual impression of the spatial distribution of farms in a region of interest, whilst acknowledging the uncertainty associated with the geo-referencing of individual farms⁴. Another example is in estimating spatial variation in disease prevalence, in which context it is often reasonable to assume that long-term prevalence varies continuously over the geographical region of interest.

A problem with KDE arises when points of interest are close to the boundary of the study area. Current algorithms for edge-correction are either difficult to apply or computationally expensive, especially for complex borders. Accordingly edge-corrected kernel estimation is not currently implemented in standard GIS software.

Here we describe a more efficient algorithm for the implementation of a particular approach to edge-corrected spatial KDE. We also explain in the discussion section why edge-correction is usually not necessary for the related problem of spatial kernel regression estimation (KRE).

Material and Methods

A non-edge-corrected kernel density estimator, based on data \( x_1, x_2, ..., x_n \) where the points \( x_i \) lie within a spatial region \( A \), takes the form \( \hat{f}(x) = \frac{1}{n} \sum w\left( \frac{x - x_i}{h} \right) \), where \( h \) is the bandwidth and \( w(u) \) is a spatial probability density function. We have conducted a simulation study to investigate the effects of edge-correction on spatial kernel density estimation, using both Gaussian and quartic kernels with five different bandwidths chosen as \( 1/4, 1/2, 1, 2, \) and \( 4 \) times \( h_0 \), where \( h_0 \) is chosen by Scott’s rule in \( \mathbb{R}^2 \). We simulated point patterns as realisations of two different spatial point processes, each of which is conditioned to generate 1000 points in the unit square with the left-bottom vertex at the origin. The first point process is a homogeneous Poisson process, for which the density is spatially constant; the second is an inhomogeneous Poisson process with density proportional to \( 1 + 0.7 \cos[2\pi(\mu - 0.5)] \) where \( \mu \) is the distance of the point to the origin. We generated 100 replications of each process.
For the edge correction, we applied a multiplicative adjustment factor\(^{10}\) to \(\hat{f}(x)\) of the form 
\[
\left(\int w(x-u)du\right)^{-1}.
\]
The integration algorithm used an adaptive triangulation of the simplified polygonal representation of the study region, followed by adaptive numerical integration over each triangle.

For the comparison between the non-edge-corrected and edge-corrected KDE, we defined a measure of the discrepancy \(M_e / M_n\), where \(M\) is the integrated square error
\[
M = \int [\hat{f}(x) - f(x)]^2 dx,
\]
whilst \(M_e\) and \(M_n\) are values of \(M\) with and without edge-correction in KDE, respectively. All simulations were run within R using the packages “Splancs” and “TB”. The latter is an R package specifically coded for the implementation of algorithms in a forthcoming paper\(^ {10}\) and will shortly be submitted to the R project: (www.r-project.org/).

**Results**

The empirical distributions of \(M_e / M_n\) showed very strong effects of applying edge-correction, especially when using large values of the bandwidth, and irrespective of the choice of kernel. For example, using the Gaussian kernel with bandwidth \(h_0\), the \(\frac{1}{4}\) and \(\frac{3}{4}\) quantiles of the empirical distributions of \(M_e / M_n\) in the 100 realisations ranged from 0.18 to 0.26 and 0.26 to 0.34 for the homogeneous and inhomogeneous processes, respectively. The corresponding ranges with the quartic kernel were 0.19 to 0.25 and 0.25 to 0.33. (Figure 1).

**Figure 1.** Box plot of the empirical distributions of \(M_e / M_n\), using five different bandwidths. Point process models and kernels used were (a) Homogeneous Poisson process with Gaussian kernel; (b) Homogeneous Poisson process with quartic kernel; (c) Inhomogeneous Poisson process with Gaussian kernel; (d) Inhomogeneous Poisson process with quartic kernel.
Discussion

The term KDE is sometimes used for the related problem of intensity estimation of a point process. The intensity of a spatial point process is the function \( \lambda(x) \) such that \( \int A \lambda(x) \, dx \) is the expected number of points in the region \( A \). The connection to density estimation is that if the point process is a Poisson point process, then conditional on the number of points in the region the locations of the points form an independent random sample from the distribution with density proportional to \( \lambda(x) \). In practice, KDE of the intensity and density differ only by a constant of proportionality. Thus our conclusions apply equally to intensity estimation.

It is generally accept that the choice of kernel is relatively unimportant for KDE in comparison to the choice of bandwidth\(^9\). We have demonstrated that KDE for two Poisson point process (homogeneous and inhomogeneous) using Gaussian and quartic kernels with equivalent bandwidths, performed very similarly.

A kernel smoothing method closely related to KDE is kernel regression estimation (KRE). One of the most popular kernel regression methods is the Nadaraya-Watson estimate\(^{10}\). In spatial epidemiology a common use of the Nadaraya-Watson kernel regression estimator is to estimate spatial variation in risk from case-control data, treating the case-control labels as a binary response at each of the point locations. In this setting, the edge-correction discussed above cancels algebraically in the Nadaraya-Watson KRE of the risk.

For other versions of KRE\(^5\) in conjunction with other edge-correction methods\(^2\), edge-correction may be beneficial. However, in general edge-corrections for kernel regression estimators are likely to have only a secondary impact on efficiency of estimation and are much less important than that for KDE. Extensions of KRE enable estimation of spatial relative risk\(^6\), determination of covariate effects on risk\(^7\), and estimation of spatial segregation in a multivariate point process\(^9\). We have applied KDE to realisations of a homogeneous Poisson point process purely to illustrate the effects of the edge-correction. For a homogeneous Poisson process, the intensity is a constant and estimated by \( N(A) / S(A) \), where \( N(A) \) is the number of points in \( A \), and \( S(A) \) is the area.

References