8. Spatial smoothing

LAURE GENEBES, AURIANE RENAUD ET FRANÇOIS SÉMÉCURBE
INSEE

Abstract

Kernel smoothing is one of the key methods for analysing data and spatial organisation. The idea consist in filtering information to reveal underlying spatial structures.

From a conceptual point of view, kernel smoothing is a non-parametric estimation method of the intensity function of a point process with values in $\mathbb{R}^2$, based solely on one of its realisations (which has been observed). The theoretical intensity function in one point x is found by calculating the average points observed per unit surface on neighbourhoods containing x, these neighbourhoods being increasingly smaller.

However, in practice, there is only one (observed) realisation, and this approach, consisting of changing to the limit no longer makes sense. The non-parametric kernel methods circumvent this limitation, not by directly suggesting an estimation of the intensity function but by suggesting a smoothed estimation of it. Notwithstanding this approximation, when the bandwidth parameter is well chosen, the resulting estimates are statistically robust and geographically relevant, and make it possible to detect whether the intensity function is constant or variable in space.

Spatial analysis tools are used to produce appropriate geographical analyses. The aim is to develop simplified, clear mapping, relieved of the arbitrariness of territorial boundary lines, as well as partly mitigating the "Modifiable Area Units Problem". In this case, the bandwidth is a geographic generalisation parameter that maintains or deletes, depending on the requirements of the analysis and the details of the geographical phenomena observed. In practice, it is possible to smooth weighted data according to Brunsdon et al. 2002 — each point in the space is assigned a numerical value. Multiple types of smoothing can be carried out, including "classic" smoothing, based on local calculations of averages, or "quantile" smoothing using local calculations of quantiles

(median, decile), see Brunsdon et al. 2002. In addition, operations on smoothed values make it possible, in particular, to calculate "smoothed" ratios, such as the percentage of a sub-population within the population as a whole.

It has become fairly easy to implement smoothing, in particular using R software, for which several packages include functions that make such smoothing possible.

8.1 Kernel smoothing

The theoretical intensity function at point x is found by calculating the average of the points observed per surface on neighbourhoods containing x (see chapter 4: "Point configuration") smaller and smaller point configurations. Kernel smoothing is a non-parametric estimation method for the intensity function of a point process with values in $\mathbb{R}^2$ based solely on one of its realisations. To find the theoretical intensity function based on a single known realisation, it is not the intensity function itself that is estimated, but a function thereof.

From a practical point of view, kernel smoothing is a local modelling based on a selection of parameters.
The kernel describes how the neighbourhood is approached.
The bandwidth is the fundamental parameter in the analysis. It quantifies the «size» of the neighbourhood. This parameter results from a bias-variance trade-off between the spatial accuracy of the analysis and its statistical quality.
The way in which edge effects are handled explains how the geographical boundaries and the limits of observation territory are taken into account in the analysis.

Furthermore, a set of geographic coordinates can be set out, for which the smoothed values will be estimated (possibly different from all the geographical coordinates of the original data). Most of the applications made by INSEE smooth the data on tile grids (the new coordinate being the centre of the tile).

In this chapter, we will start out by discussing the foundations and formalism of kernel smoothing and then proceed to its implementation.

8.1.1 Origin and formalism of kernel smoothing

Historically, the first non-parametric intensity estimation method was based on building territorial intensity. It consist in calculating, for each territorial unit, the point intensity observed per surface unit. In this case, the intensity is also referred to as density. Within each of these territorial units, the estimated intensity is constant. For example, when calculating the density of a region, the said region is considered to be the same throughout the territory.

The practical interest of understanding intensity is based on the possibility of representing territorial densities in the form of choropleth maps, the first versions of which date back to the work of Baron Pierre Charles Dupin, see Palsky 1991. Geographers and statisticians then used this method to represent the distribution of the population within administrative regions. From a technical point of view, density maps generalise the histograms of the monodimensional analyses to the two-dimensional geographical areas. A sample density map is shown on Figure 8.1.

In the 20th century, geographers and statisticians gradually came to question the statistical and geographical relevance of this type of approach. Openshaw theorised its limits under the name *Modifiable Area Units Problem* (MAUP) . MAUP (see Figure 8.2) is broken down into two interdependent sub-problems — the scale effect and the zoning effect. The scale effect describes the dependence of the phenomenon observed on the average size of the spatial units. The larger the size, the smaller the local specificities and the more the analyses show the global structures. In

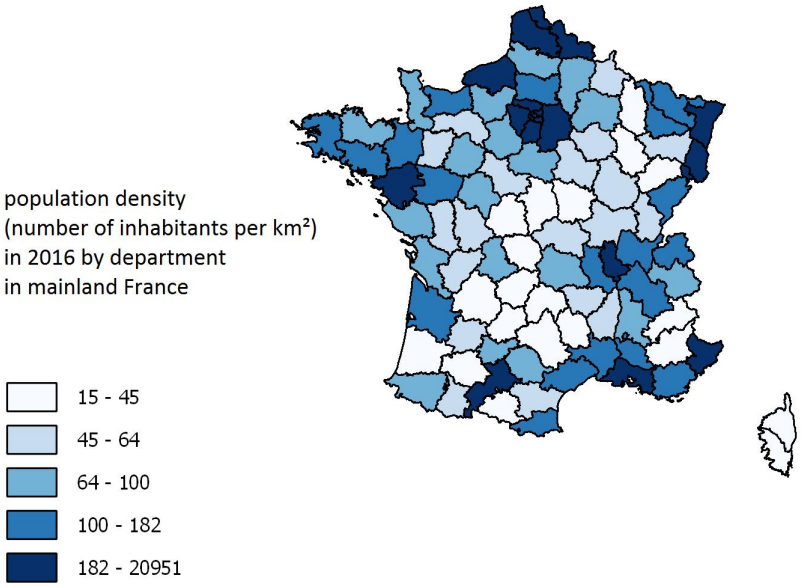


Figure 8.1 – Example of map density
Source: *INSEE*

contrast, small sizes retain local specificities and detail them. However, it should also be noted that the resulting analyses are sensitive to statistical noise, data quality and data accuracy. The zoning effect explains that the phenomena observed are dependant on the form of the spatial units. The concept of form includes the morphology of spatial units but also their position in space. Thus, if the spatial units' contours are displaced uniformly, the phenomenon observed is likely to be profoundly changed.

Kernel smoothing has inherited these reflections and aims to overcome the arbitrariness of territorial divisions. Kernel smoothing is rigorously defined in the context of spatial analysis, but similar methods in geography and statistics can be detected from the end of 19th century with the work of Louis-Leger Gauthier and Victor Turquan. This proximity (or entanglement) between the approach of spatial statisticians and the approach of geographers, justifies this chapter's focus on smoothing both from the standpoints of pure spatial analysis and a more operational geographical analysis.

In practice, the challenge for statisticians lies in observing only a single realisation. In concrete terms, to circumvent the challenge of estimating an intensity function based on a single realisation, kernel smoothing does not directly estimate that realisation, but a smoothed version obtained by convolution with a kernel K_h:

$$(K_h * \lambda)(x) = \int_{\mathbb{R}^2} \lambda(t) K(x-t) dt \tag{8.1}$$

with $K_h(u) = \frac{1}{h^2} K(\frac{u}{h})$

and K a symmetrical function $\mathbb{R}^2$ in $\mathbb{R}$ positive and of integral 1

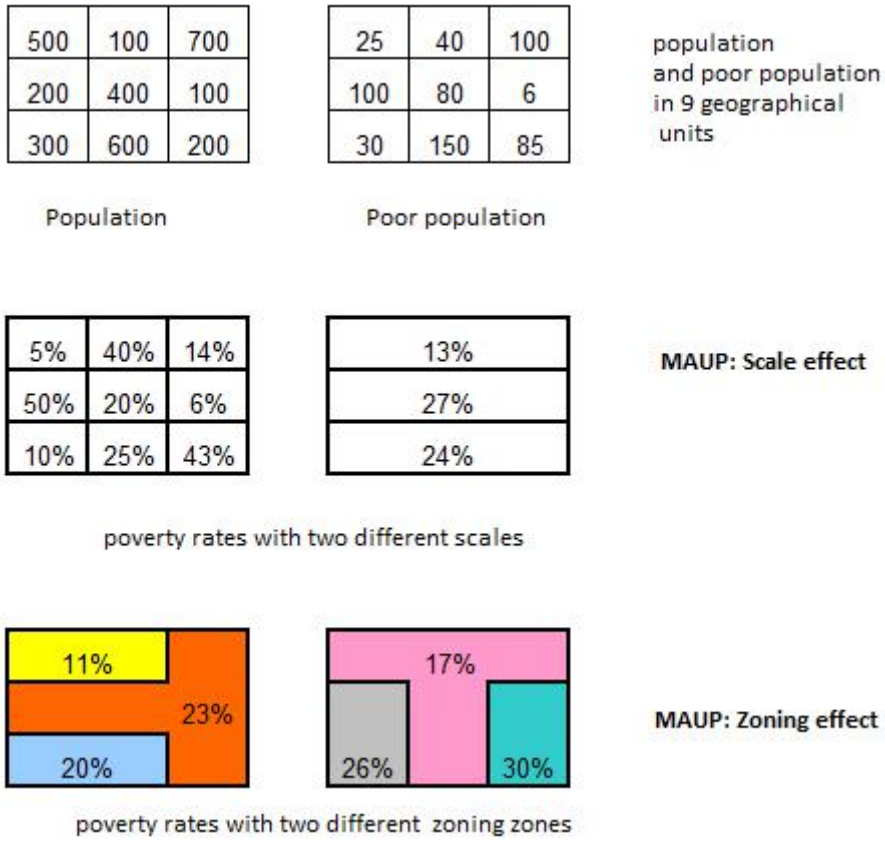


Figure 8.2 – MAUP Diagram: scale effect and zoning effect

R A simple metaphor for understanding the convolution operation consists in imagining that λ represents the distribution of the density of rabbit holes in space. Each hole is associated with a single rabbit. Each rabbit, to satisfy its needs, moves within a given proximity of its hole, so that its probability of being in position t in relation to its hole is $K_h(t)$. Convolution $(K_h * \lambda)(x)$ in this case represents the local density of rabbits in x. If h is small, rabbits concentrate around their holes and the rabbit intensity function differs little from that of the holes. In contrast, if h is high, rabbits tend to mix in space and the rabbit intensity function is «blurred» compared to that of the holes.

To find an estimator for $(K_h * \lambda)(x)$ from a set of points $\{x_i\}$ resulting from the realisation of a point process, a simple idea consists in substituting the integral for $\mathbb{R}^2$ by a sum on the points observed in the Equation (8.1).

Definition 8.1.1 — Kernel smoothing. Given K_h on a bandwidth kernel h and x a point $\mathbb{R}^2$, the estimated smoothed intensity in x is defined by:

$$\widehat{\lambda_h}(x) = \sum_i K_h(x - x_i) \tag{8.2}$$

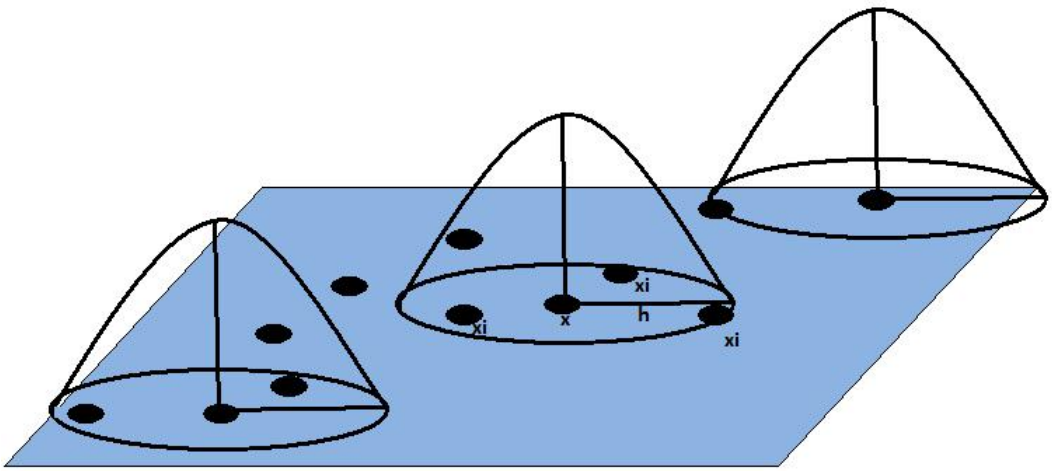


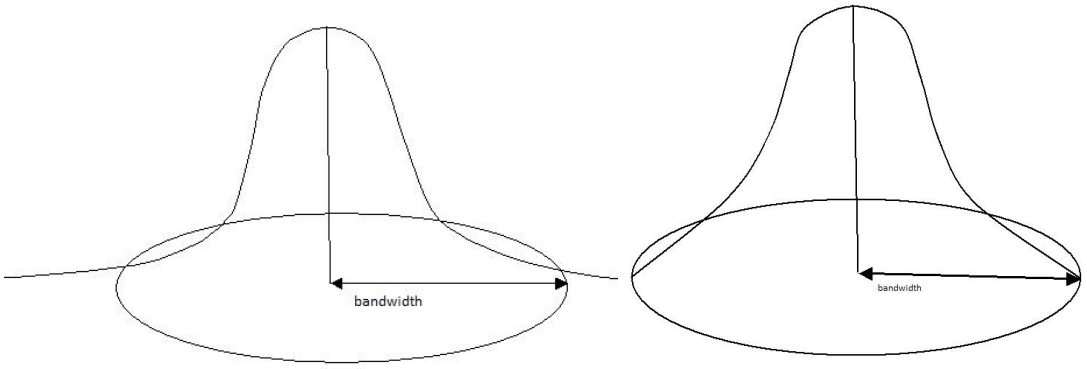
Figure 8.3 – Kernel smoothing scheme
Note: At each point of estimation lies a kernel function. The value of this function is highest at point level and decreases as one moves away.

K_h in this formula plays a role similar to a territorial unit centred on each point in space $\mathbb{R}^2$ with size h. In contrast to analyses based on a geographical split, the estimator in smoothed intensity controls the zoning effect of the MAUP, with the choice of kernel having little impact on the smoothing results. In contrast, the arbitrariness of the scale effect is maintained through the choice of bandwidth. Different kernels have been proposed in the literature. The most frequently used kernels are listed below.

Definition 8.1.2 — Common kernels. x is a point of $\mathbb{R}^2$. K^N and K^B are respectively referred to as the Gaussian kernel and quadratic kernel:

$$K_h^N(x) = \frac{1}{2\pi}\, e^{-||\frac{x}{h}||^2} \tag{8.3}$$

$$K_h^B(x) = \frac{9}{16}\ 1_{||x||<h}\ (1 - ||\frac{x}{h}||^2)^2 \tag{8.4}$$



(a) Gaussian kernel K^N (b) Quadratic kernel K^B

Figure 8.4 – Gaussian and quadratic kernels
Note: The quadratic kernel K^B gives greater weight to the closest points than to the remote points. It cancels itself out beyond the smoothening radius. On the contrary, Gaussian kernel K^N takes all points in the study area into account.

8.1.2 Adjusting to edge effects

As regards kernel estimation of the probability densities, kernel smoothing methods are impacted by an additional problem that arises when edge effects are taken into account. In the case of density estimation, the estimate is made on $\mathbb{R}^n$. Where kernel smoothing is concerned, the observed points are generally contained in an analysis window W, or, concretely, a polygon.

The nature of the window's borders can be of two kinds. Firstly, the window may result from the information collection protocol. For example, during an archaeological excavation, only a restricted area is dug for reasons of costs and opportunities. In this case, the borders are not inherent to the process observed, and without further information it is reasonable to assume continuity of intensity inside and outside the window. Secondly, the window may be induced by geographic configurations that have an impact on the underlying process generating the set of points observed. In geography, rivers, reliefs and coasts are all borders that restrict the settlement of human activities. Beyond such borders, the intensity of the phenomenon observed is null.

The formula (8.2) is the estimation formula, without adjusting for edge effects. In end-effect, the analysis window is ignored.

The purpose of adjusting to edge effects is to take into account the impact of the border when estimating the intensity. A variety of solutions has been suggested to this end. They differ in their approach of the outside ofthe observation area and in how fast they are carried out (Baddeley, see Baddeley et al. 2015a).

Definition 8.1.3 — Adjusting for edge effects. x is a point of $\mathbb{R}^2$, whereby the uniform and Diggle estimates (see Diggle 2013) are found using the following formulas:

$$\text{uniform correction: } \widehat{\lambda_h^U}(x) = \frac{1}{e_{h(x)}} \sum_i K_h(x - x_i) \tag{8.5}$$

$$\text{Diggle's correction: } \widehat{\lambda_h^D}(x) = \sum_i \frac{1}{e_{h(x_i)}} K_h(x - x_i) \tag{8.6}$$

where $e_h(u) = \int_W K_h(u - v)dv$

When the analysis window is independent of the underlying process, the uniform estimate ensures continuity of intensity between the inside and outside of the window. However, if the intensity outside the window is deemed to be null, it is more opportune to use Diggle's estimation method, see Diggle 2013, which is conservative. In this case, the intensity integral estimated in the analysis window exactly matches the number of points observed. From an algorithmic point of view, Diggle's estimation method requires significantly more calculation time than the uniform estimate.

The term $e_{h(x)}$ can be interpreted as the intersection probability between two sets. Let us assume, again using our rabbit metaphor, that the window's spatial footprint matches that of an enclosure. $K_h(x-u)$ approximately describes a rabbit's exploration territory around a hole found in x if the rabbit does not encounter any obstacles. $e_h(x) = \int_W K_h(u-x)du$ is the part of the territory explored by the rabbit contained in the spatial footprint of the enclosure. $e_{h(x)}$ is strictly lower than 1 if x immediately neighbours the border of the enclosure. However, if the "natural" exploration zone of the rabbits occupying the whole is entirely contained in the enclosure, $e_{h(x)}$ is equal to 1.

In formula (8.5) the term $e_{h(x)}$ is applied overall to the density estimate. Near the boundary of the enclosure, this term makes it possible to restore the estimated intensity. The closer the point is to the border, the lower $e_{h(x)}$ and the greater the compensation will be. Uniform correction considers that at the window's boundary, the distribution of rabbit holes is almost homogeneous inside and outside the window. Intuitively, this correction amounts to postulating that the enclosure has no effect on the mobility of rabbits, which walk across its boundary without realising. More specifically, it is considered that rabbits whose holes are located outside the enclosure also contribute to the intensity calculated within the enclosure.

Diggle's estimation method, formula (8.6), assumes that the window is an integral part of the properties of the underlying process. In other words, the enclosure forms an insurmountable boundary for the rabbits and all the rabbits are contained in the enclosure. By dividing $K_h(x-x_i)$ with the term $e_{h(x_i)}$, we ensure that the rabbit of hole i has a probability equal to 1 of remaining in the spatial footprint of the enclosure.

8.1.3 Choosing bandwidth

The choice of bandwidth determines the extent to which the estimation of intensity function will be "smoothed". In spatial analysis, bandwidth results from a bias-variance compromise. The bias is caused by the fact that the intensity function estimator does not directly estimate the intensity function but a smoothed version thereof. The greater the bandwidth, the greater the bias. The variance, on the contrary, decreases according to the bandwidth. The greater the bandwidth, the greater the number of points involved in calculating local estimations, which tends to reduce the estimation variance.

Several methods are available to automatically suggest a bandwidth that minimizes an error criterion. As the desired intensity function is obviously not available, some of these methods are based on cross-validation methods. They use the point distribution observed, and assume that it follows a Poisson distribution in order to estimate optimal bandwidth. In section 8.3, examples using the package's cross-validation functions *spatstat* of R will be suggested. These examples highlight the high variability of the proposed bandwidths based on the selected error criteria. Furthermore, the existence of a single bandwidth relevant for the entire extent of the zone studied is a central hypothesis. Several adaptive smoothing methods have been suggested to overcome this limit. Readers will find interesting material in Baddeley's book (Baddeley et al. 2015a) on this topic, which makes use of package *spatstat* in R.

In fact, no bandwidth is optimal: all are capable of providing an apposite depiction of the world as defined in the MAUP. Some geographers recommend adopting a multi-scale approach to study the multiplicity of spatial aspects within a single phenomenon.

8.2 Geographical smoothing

Geographical smoothing is based on the intensity estimation method presented above. It is not intended to calculate intensities, but to come up with simplified mapping representations. The principle of this form of use in geography is to represent not the value observed at a single point, but a weighted average of the values observed in the neighbourhood of this point, within a predefined radius.

Smoothing can be interpreted as a tool capable of guaranteeing a form of **confidentiality**. It makes it possible to represent initially *ad hoc* and confidential data in aggregate form. It is nevertheless important to remain vigilant regarding the number of points used to produce the smoothed estimate.

8.2.1 Smoothing weighted data

In this case, each point x_i is assigned a numerical value w_i. For example, x_i can represent a housing unit, and w_i the number of inhabitants in this housing unit. For this, we need only (see Brunsdon et al. 2002) use a weighted version of the kernel estimators described above. In formula (8.2), weight w_i is multiplied by the contribution of a point to the intensity estimator.

Definition 8.2.1 — Weighted kernel estimators. Given K_h a bandwidth kernel h and x_i one point on $\mathbb{R}^2$ with assigned weighting w_i, the smoothed intensity estimated in x is defined by:

$$\widehat{\lambda_h}(x) = \sum_i w_i K_h(x - x_i) \tag{8.7}$$

While the choice of the kernel K_h has little influence on the smoothing results (see Figure 8.5), the choice of bandwidth h is of fundamental importance, although fairly arbitrary.

As has been noted above, the bandwidth acts like a smoothing parameter, controlling the balance between bias and variance. A high radius leads to a significantly smoothed density and high bias. A small radius generates a density subject to little smoothing, with high variance. It is generally up to the user whether a compromise should be made, depending on the desired level of aggregation. It is advisable to test several bandwidth values, making it possible to reveal local variations at different scales. The maps in Figure 8.6 are examples of smoothed maps for Paris and its suburbs, with three different smoothing radii.

This estimate is valuable in that it focuses attention not on the points and their distribution, but on their environment. The bandwidth thus makes it possible to define this environment.

Multiple algorithms exist to determine a so-called "optimal" smoothing radius. These tests can yield various and sometimes very widely-differing results (see implementation). It is advised that they be used only for indicative purposes, and that the user choose the smoothing radius based on experience with the data and the issue.

Operations can be carried out on smoothed variables, in particular ratios. The theoretical rationale for this can be found on pp. 34 to 37 of Floch's working document (Floch 2012b). In practical terms, to find the smoothed value of the ratio of two variables, it is essential to separately calculate the smoothed values of the numerator and denominator, then to calculate the ratio between the smoothed value of the numerator and the smoothed value of the denominator. Do not directly calculate a smoothed ratio value. The map would be distorted, as the same importance would wrongly be given to all territories, despite their being unequally populated.

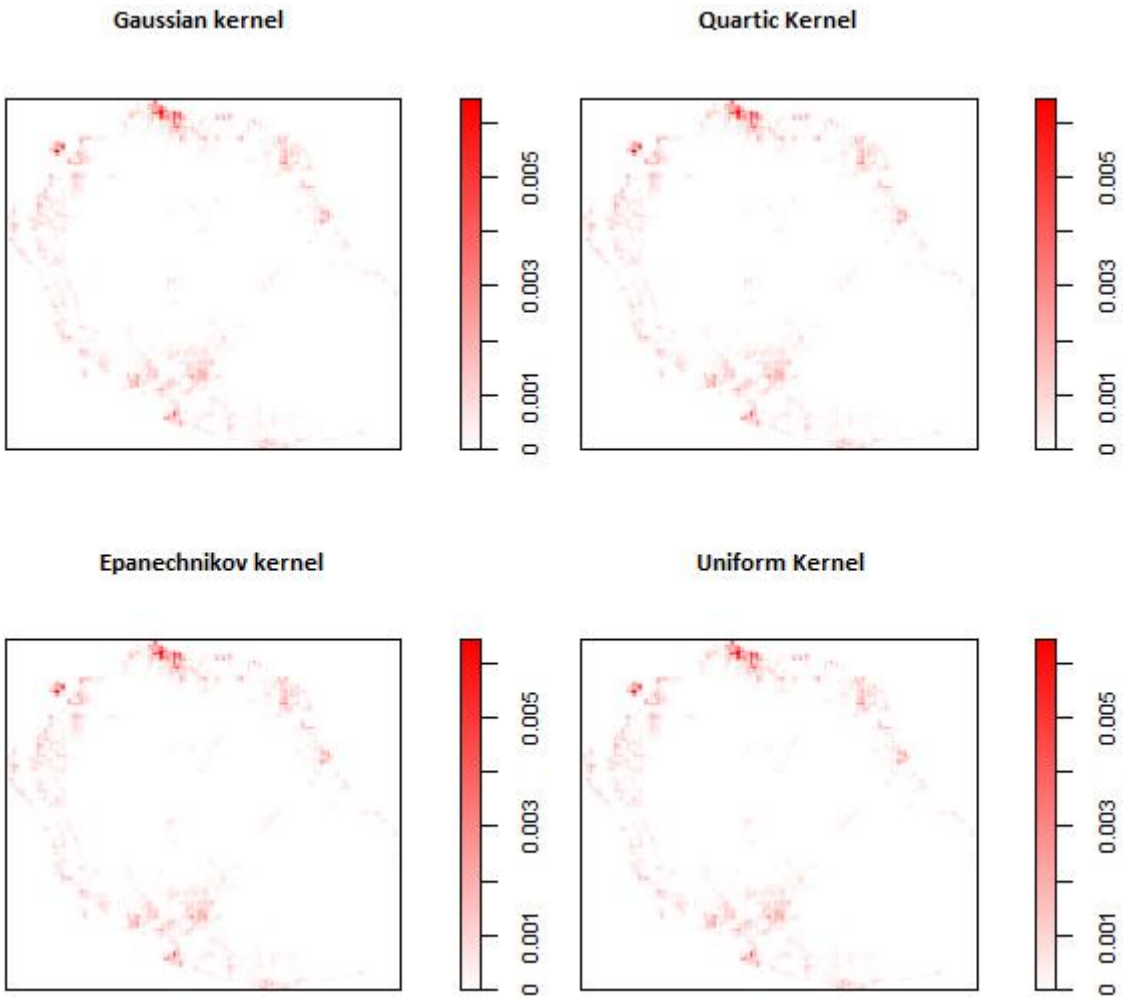
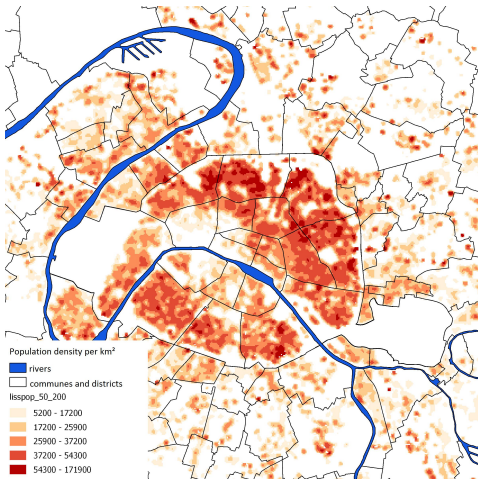
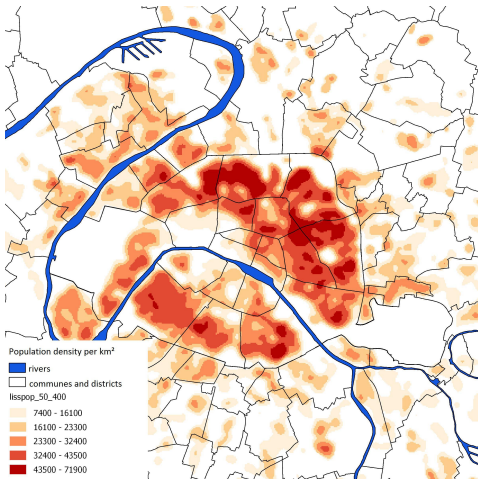


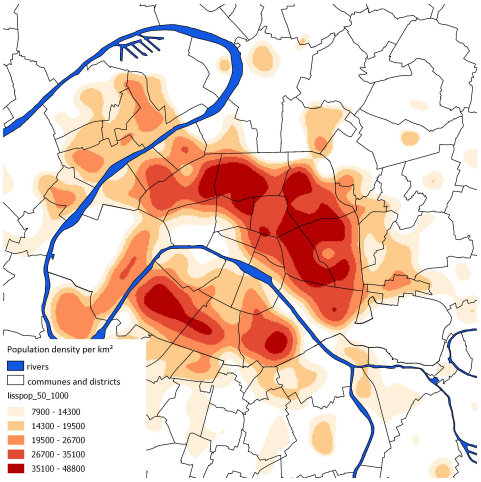
Figure 8.5 – Comparison of results found using four different kernels from the function `density.ppp` of package *spatstat*
Source: *INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011*
Note: The variable shown is the smoothed number of households on Reunion Island.



(a) 200 meter radius



(b) 400 meter radius



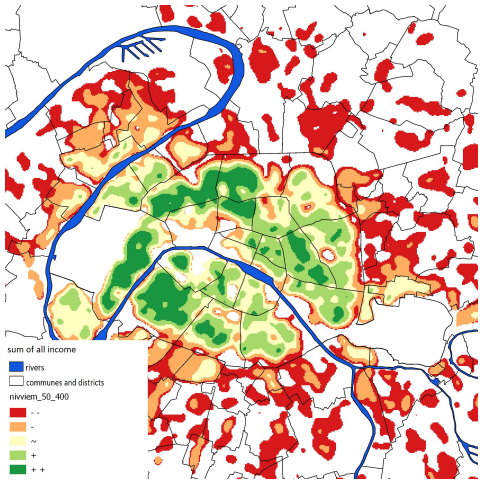
(c) 1000 meter radius

Figure 8.6 – Three different smoothing radii for population density in Paris and its suburbs — 200 meters, 400 meters, 1000 meters
Source: *INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised social and tax file 2012*
Note: the tiles depicted contain more than 11 households.

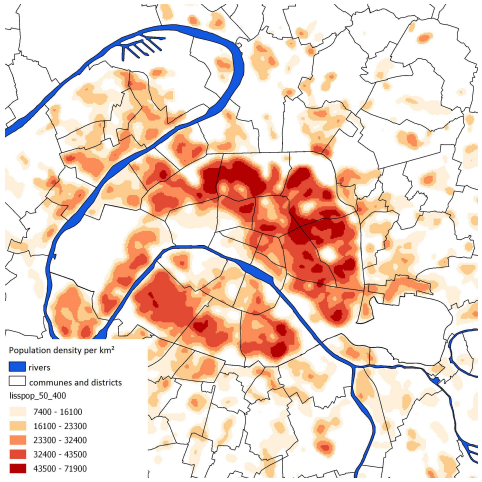
8.2.2 Application using non-parametric regression

In this example, the focus is on **average income calculation** per person. We have two variables — income, and number of people. Average income is equal to the sum of the total income, divided by the sum of the number of people. Income and the number of people are smoothed separately. The ratio can then be calculated.

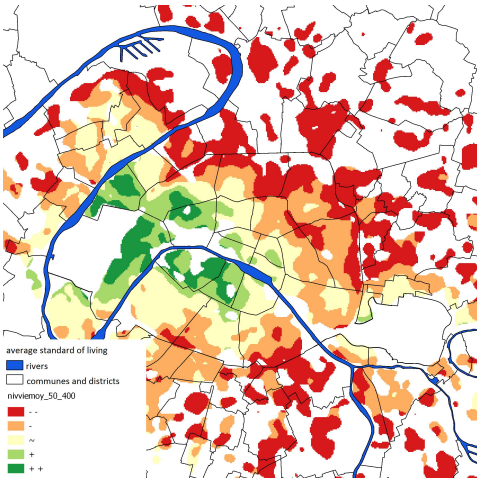
The maps are derived from Figure 8.7 as regards Paris and the surrounding municipalities in the immediately-surrounding suburbs (*i.e.* the three departments bordering Paris).



(a) Total standard of living of all households (in euros)



(b) Household population (density per km2)



(c) Average standard of living

Figure 8.7 – Calculation of a smoothed average standard of living
Source: *INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised social and tax file 2012*
Note: The tiles shown contain more than 11 households. As to maps depicting levels of income, the markers "++", "+" and "∼", "-" and "–" respectively reflect very high, high, average, low or very low values for the indicator in question. They were used for questions of non- profiling of the population.

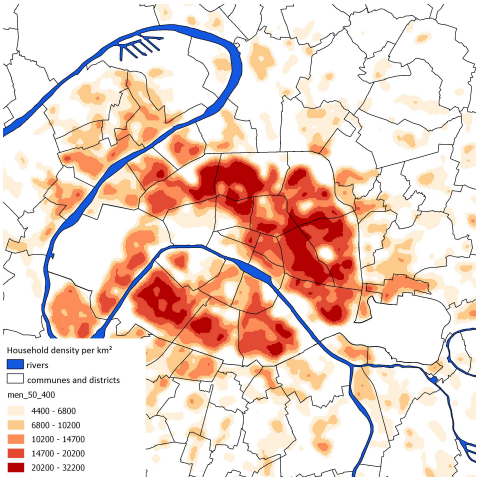
Map 8.7a of the total standard of living of households does not give much information on its own. The total standard of living per square needs to be compared to population in each tile. On map 8.7a showing the number of people, the population is very dense within the municipality

of Paris, mainly north-east of the Seine, and to a lesser extent deeply southward.
On map 8.7a, the average standard of living per person is very high in the heart of Paris, essentially in the west.

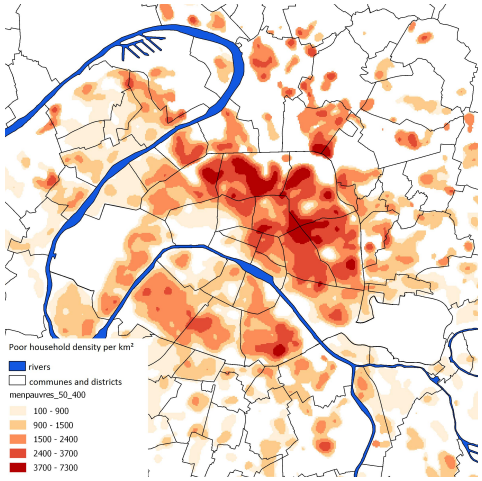
R This is no longer a theoretical framework. This calculation is based on tools comparable to a non-parametric regression. Roughly speaking, it is as if a weighted geographical regression were carried out, limited to a single variable — the constant (see chapter 9: "Geographically Weighted Regression").

8.2.3 Application using a non-parametric conditional density estimate

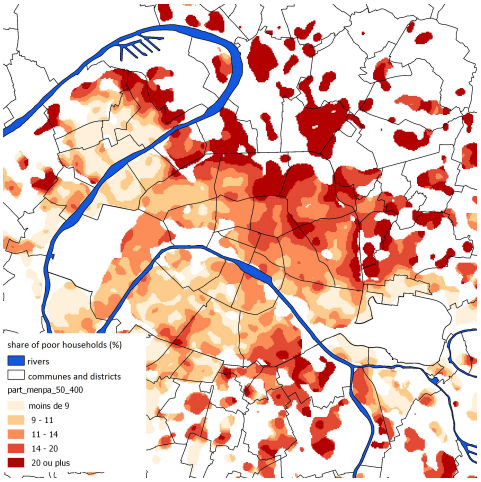
The focus is on the **proportion of poor households** across all households. The smoothed value of the number of poor households in a territory is calculated, and the smoothed value of the total number of households in the territory is calculated. The ratio is then calculated.



(a) Households (density per km2)



(b) Poor households (density per km2)



(c) Proportion of poor households as a whole

Figure 8.8 – Calculating the smoothed proportion of poor households
Source: *INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised social and tax file 2012*
Note: The tiles depicted contain more than 11 households.

According to the map from Figure 8.8a, the most populated areas are found in the heart of Paris, both in the north-east quarter and the south-west quarter. In Figure 8.8b, there are many poor

households in Paris, that tend to be located in the north-east quarter. In Figure 8.8c, the proportion of poor households in all households provides additional information. The map highlights the less densely populated areas, where the share of households living below the poverty line is nonetheless high. These are municipalities located north of Paris.

Thus, depending on the map produced, the messages derived can be different. When analysing the rates, it is also of fundamental importance to analyse the distribution of the number of people alone (population density, for example), in order to verify the robustness of the calculated rates and their representativeness.

R This calculation can be likened to a conditional probability calculation. The result is a map of poverty rates at the local level, which is close to the idea of identifying the likelihood of a household's being poor, on the basis of its having settled in a given place.

R **Important!** In theory, it is always possible to calculate the ratio of two smoothed variables. In practice, it is important to pay attention to the small numbers. In the example where the proportion of poor people was calculated, areas with small numbers could wrongly be shown separately in the smoothed map. Having a low population, they would not appear on a map showing raw data. Thus, by failing to take this phenomenon into account, we could mechanically give the - skewed - impression that all territories are populated.

8.2.4 Application using quantile smoothing

The smoothing described up to this point is an average smoothing, in the sense that it is based on local calculations of averages. In the article by Brunsdon et al. 2002, the authors extend this concept, in order to define local statistics based on quantiles (median, deciles, etc.). These indicators are deemed, in the exploratory analysis of "traditional" data, to be less sensitive to extreme values. Quantile smoothing makes it possible above all to calculate indicators that considerably enrich the analysis of certain variables, in particular income variables.

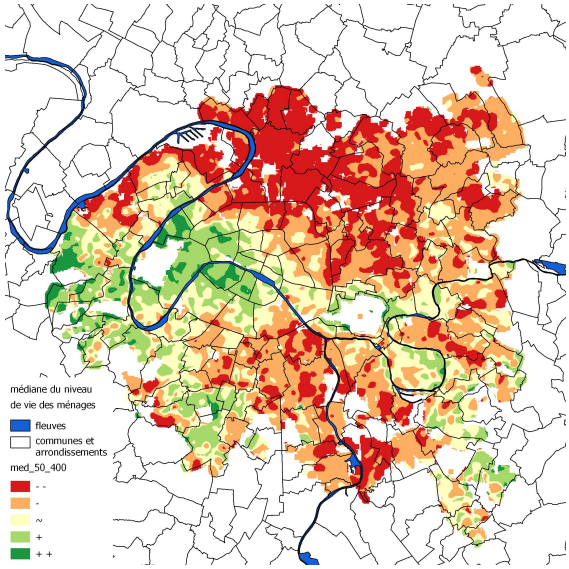
The four thumbnails in Figure 8.9 represent multiple smoothed indicators calculated based on standard of living (source *INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised Social and Tax File 2012*), *i.e.* the disposable income of a household divided by the number of consumer units in the household.

The maps in Figure 8.9 are centred on Paris, and include the immediately surrounding suburbs. Smoothing is performed on 50-meter tiles, with a 400-meter bandwidth. Only the tiles for which the number of observations (of households) contributing to the estimate is strictly greater than 50 have been used for viewing.

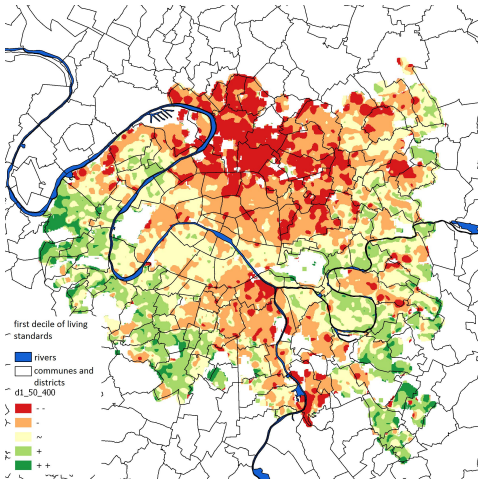
Map 8.9a represents the median standard of living. Zones can be seen in the west where residents are much more affluent. Maps 8.9b and 8.9c show the 1st decile and 9th deciles of living standard. Map 8.9d depicts the interdecile ratio (ratio between the 9th and 1st deciles) and provides additional insight. Interdecile ratio is stated without units. It shows the minimum standard of living of the richest 10% relative to the maximum standard of living of the poorest 10% and brings out the gap between the top and bottom of the distribution. This is one of the measures of inequality in this distribution. In the aforementioned zones in the west, the interdecile ratios are very high. These neighbourhoods are home to populations with a very high standards of living, as well as populations with much lower standards of living.

8.3 Implementation with R

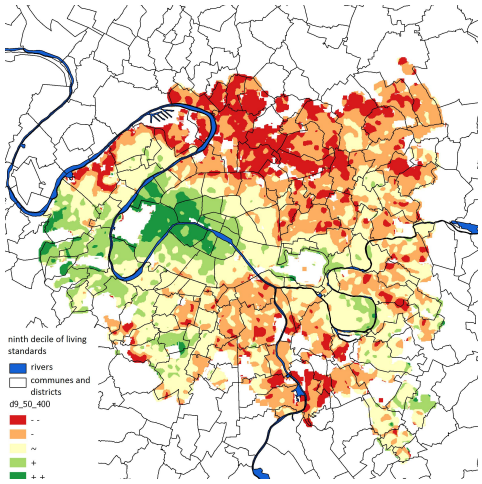
R offers multiple packages that can be used to perform smoothing. The practical implementation process is detailed below, using packages *spatstat* and *btb*, applied to data pertaining to Reunion



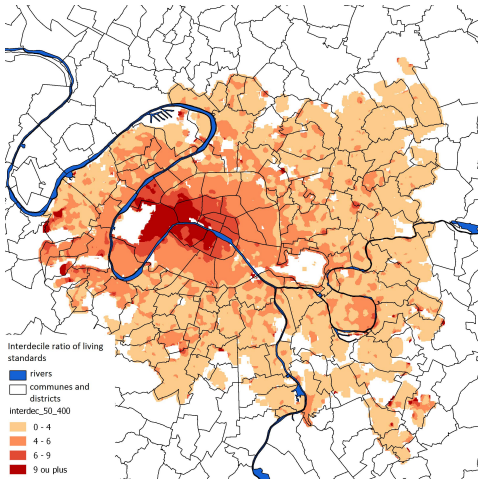
(a) Median standard of living



(b) 1st decile of standard of living



(c) 9th decile of standard of living



(d) Interdecile ratio of standard of living

Figure 8.9 – Standard of living distribution
Source: *INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised social and tax file 2012*
Note: The tiles shown contain more than 11 households. As to maps depicting levels of income, the markers "++", "+" and "∼", "-" and "–" respectively reflect very high, high, average, low or very low values for the indicator in question. They were used for questions of non- profiling of the population.

Island. The data used in the example are the `dataframe` *reunion.Rdata* provided in package *btb*. An overview of this `dataframe` is provided in Figure 8.10.

	x	y	houhold	phouhold
1	359500	7634300	5.0693069	2.37623762
2	359500	7634500	26.9306931	12.62376238
3	355900	7634500	15.0000000	4.00000000
4	356100	7634500	39.0000000	20.00000000
5	356300	7634500	41.6428571	15.14285714
6	356500	7634500	2.3571429	0.85714286
7	359700	7634500	11.4210526	0.00000000
8	359700	7634700	2.5789474	0.00000000
9	359900	7634500	12.0000000	6.00000000
10	355700	7634700	1.0243902	0.00000000
11	355700	7635100	1.3658537	0.00000000
12	355700	7635300	11.6097561	0.00000000
13	355900	7634700	20.0000000	7.00000000
14	356100	7634700	131.0000000	71.00000000
15	356300	7634700	110.0000000	58.00000000

Figure 8.10 – The first 15 lines of the `data.frame`*reunion.Rdata* in package *btb*
Source: *INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011*
Scope: Reunion Island

This is 200-meter grid data, downloadable on `insee.fr`. The source is *INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011*.

The variables are thus defined:

— x: longitude (projection system: WGS 84 / UTM zone 40S, EPSG code: 32740)
— y: latitude (projection system: WGS 84 / UTM zone 40S, EPSG code: 32740)
— houhold: number of households
— phouhold: number of poor households (poverty definition at 60%)

8.3.1 Under R, with package *spatstat*

The R package known as *spatstat* is a comprehensive package dedicated to analysing spatial point processes. It is available on the CRAN website at the following address: `https://CRAN.R-project.org/package=spatstat`

The function `density.ppp` available in package *spatstat* makes it possible to run data smoothing. The use of this function requires the use of an object in format *.ppp*upon entry. To use this function, the x and y coordinates of the `data.frame` must be converted to *.ppp*.

```
#smoothing of the houhold variable (number of households) with Spatstat

library(spatstat)
```

```
library(btb)#only for the meeting dataframe
data(reunion)

# duplicate coordinate aggregation and deletion
base_temp <- aggregate(houhold ~ x+y, reunion, sum)

# x,y transformation into .ppp objects
base.ppp = spatstat::ppp(base_temp$x, base_temp$y,
c(min(base_temp$x), max(base_temp$x)),
c(min(base_temp$y), max(base_temp$y)) )

#density.ppp function call
#the sigma parameter is h/2 with h the bandwidth
densite <- spatstat::density.ppp (base.ppp, sigma = 200, weights=base_temp$
    houhold )

#map display
plot(densite, main = "Spatstat smoothing, user defined radius")
```

8.3.2 Under R, with package *btb*

Package *btb* ("beyond the border") [1] is online on the CRAN website, at the following address: `https://CRAN.R-project.org/package=btb`. It offers functions dedicated to urban analysis and implements a density estimate using the KDE method (`kernel density estimator`), *i.e.* a kernel method. The kernel used is a quadratic kernel.

In the estimation produced by the package, the edge effect is taken into account in the smoothing function `kernelSmoothing` *via* Diggle's correction (Diggle 2013). This correction makes it possible in particular to deal with the case of boundaries depicting geographical limits (coasts, for example). Within the observation area, the intensity is non-null. Outside the observation area, it is null. The method implemented is conservative (thanks to standardisation). Before and after smoothing, the number of points observed is the same.

Calculation times have been significantly reduced, in several ways:
- by coding in C++ all the most time-consuming methods;
- by limiting, for each point, to an observation window around this point, making it possible to limit the number of operations (calculations of distances) to be carried out.

```
#smoothing with btb: calculating the proportion of poor households
#the numerator (number of poor households), and the denominator (total
    number of households) are smoothed separately

library(btb)

#data loading
data(reunion)

#smoothing
```

1. There will soon be a version of package *btb* adapted to new package *sf* (simple features).

```
#parameter setting
pas <- 200 #200-meter square
rayon <- 400 #400-meter bandwith

#smoothing function call
#the function automatically smoothes all the variables contained in the
    database
#here, phouhold and houhold are smoothed

dfLisse <- btb::kernelSmoothing(dfObservations = reunion,  iCellSize = pas,
                                iBandwidth = rayon, sEPSG="32740")

#rate of poor households : ratio of smoothed variables
dfLisse$txmenpa = 100 * dfLisse$phouhold / dfLisse$houhold

#overview in R
library(sp)
library(cartography)
#map display
cartography::choroLayer(dfLisse, var = "txmenpa", nclass = 5, method = "
    fisher-jenks", border = NA, legend.pos = "topright", legend.title.txt =
     "txmenpa (%)")

#title and outline added
cartography::layoutLayer(title = "Reunion Island : rate of poor households"
    ,
                         sources = "",
                         author = "",
                         scale = NULL,
                         frame = TRUE,
                         col = "black",
                         coltitle = "white")
```

The resulting map is shown in Figure 8.11.

The user can also export the result in shapefile format, then rework it in a GIS.

```
#export in shapefile format

rgdal::writeOGR(as(dfLisse, 'Spatial'), "txmenpauvre.shp", "txmenpauvre",
    driver = "ESRI Shapefile")
```

Package *btb* also enables the use of quantile smoothing, described above. The user need only specify as a parameter `vQuantiles` the quantile vector to be calculated. For example *c(0.1, 0.25, 0.5)* will return the first decile, the first quartile and the median of each of the variables of the input `data.frame` .

```
# quantile smoothing
library(btb)
data(reunion)
```

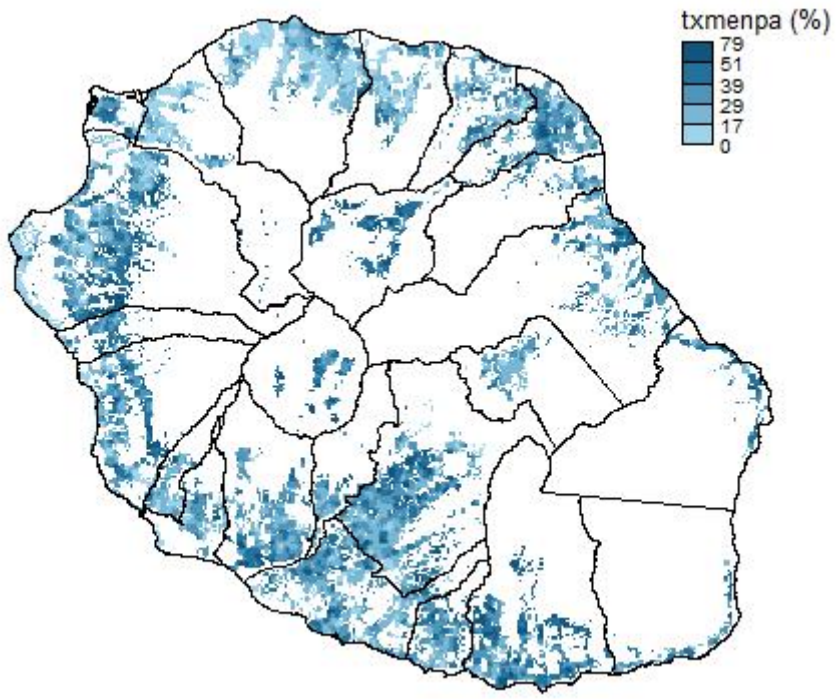


Figure 8.11 – Rate of poor households on Reunion Island after smoothing
Source: *INSEE, Localised Tax Revenues as at 31/12/2010 and Housing Tax as at 01/01/2011*
Note: Ratio between number of smoothed poor households and the total number of smoothed households. The black outlines show the municipal boundaries.

```
#parameter setting
pas <- 200
rayon <- 400

#smoothing function call
dfLisse_quantile<- btb::kernelSmoothing(dfObservations = reunion,
                                        iCellSize = pas,
                                        iBandwidth = rayon,
                                        vQuantiles = c(0.1, 0.5, 0.9),
                                        sEPSG="32740")

#export to QGIS
rgdal::writeOGR(as(dfLisse_quantile, 'Spatial'), "lissage_quantile.shp", "
    lissage_quantile",
                driver = "ESRI Shapefile")
```

Package *btb* defaults to an automatic tile grid. The smoothing function can also be queried using a grid of the user's choice. In this case, the user must have a `data.frame` consisting of two columns x and y, which match up with the desired centroid coordinates.

```
#smoothing function with grid, optional to user
kernelSmoothing(dfObservations, iCellSize, iBandwidth, dfCentroids)
```

8.3.3 Optimal bandwidth tests

In R, multiple methods proposing to calculate an "optimal" bandwidth can be implemented, based on different criteria. The aim is generally to minimise an error measurement. In *spatstat* for example, the following four functions are found: `bw.diggle`, `bw.ppl`, `bw.frac` and `bw.scott`.

With function `bw.diggle` in *spatstat*

Function `bw.diggle` in *spatstat* chooses a bandwidth that minimises a criterion M(σ) based on average quadratic error (MSE for *Mean Square Error*) in the estimator.

The chart in Figure 8.12 represents criterion M(σ), which must be minimised. To find value σ, we will need to identify the value on the x-axis, which matches up with the minimum value on the y-axis.

For more details, see `https://www.rdocumentation.org/packages/spatstat/versions/1.49-0/topics/bw.diggle`.

```
#the base.ppp created above is used again
# bw.diggle test for optimal bandwidth
bw_diggle <- spatstat::bw.diggle(base.ppp)
plot(bw_diggle, main = "cross validation")

#density.ppp call with the automatically calculated bandwidth
densite_optim <- spatstat::density.ppp(base.ppp,bw_diggle, weights=base_
    temp$houhold)
```

The result is:

```
bw_diggle
##    sigma
## 141.9445
```

Using the default settings, the proposed value for σ is 142 metres or 284 metres for bandwidth h (σ = h/2 , see package documentation).

With function `bw.ppl` of *spatstat*

The bandwidth is chosen by calculating a maximum likelihood estimator, using a cross-validation method (*likelihood cross-validation criterion*). The calculations are then iterated. Each time, we work only on $n-1$ observations, then validate the model on the observation that had been discarded. We repeat this n times.

The graph below shows the CV criterion(σ) that we wish to minimise. To find value σ, we will need to identify the value on the x-axis, which matches up with the maximum value on the y-axis.

For more details, see `https://rdrr.io/cran/spatstat/man/bw.ppl.html`.

```
#the base.ppp created above is used again

# bw.ppl test for optimum bandwidth
bw_ppl <- spatstat::bw.ppl(base.ppp)
plot(bw_ppl, main = "bw.ppl")
```

The result is:

```
bw_ppl
##    sigma
## 286.0097
```

Using the default settings, the proposed value for the σ value is 286 metres.

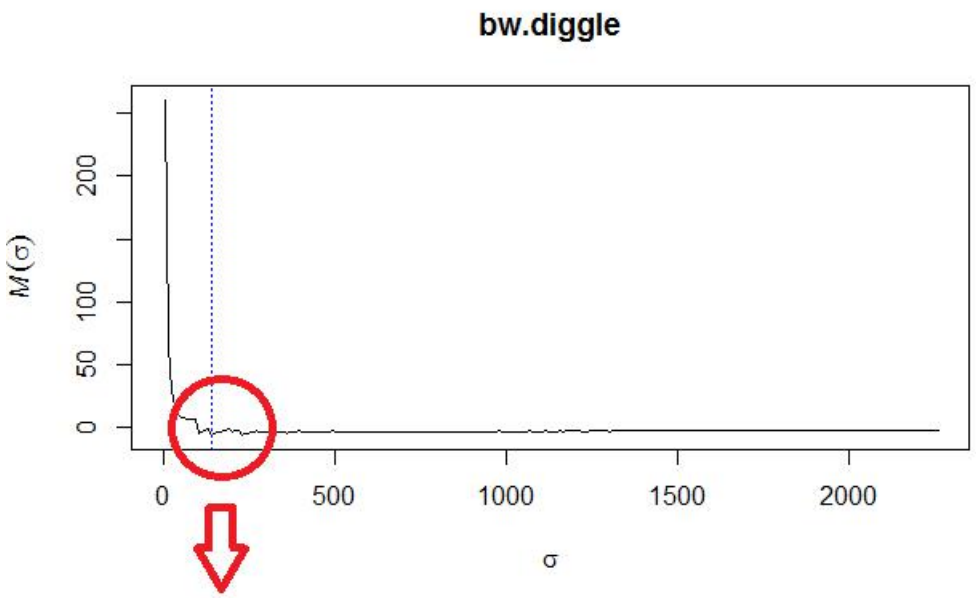
With function `bw.frac` of *spatstat*

This method selects a bandwidth based solely on the geometry of the observation window.

The bandwidth is a quantile (specified by the user) of the distance between two independent points, chosen randomly in the window. By default, the first distribution quartile is used. If CDF(r) is used to denote the cumulative distribution function of the distance between two independent points randomly and uniformly distributed in the window, then the value that is returned is the quantile with probability f. The bandwidth is then the r value, such that CDF(r)=f. First, the algorithm calculates the cumulative CDF(r) distribution function with the function `distcdf` of package *spatstat*. This function makes it possible to calculate the function CDF(r) = P(T<=r) of the ECU T=|X_1-X_2| between two randomly-chosen independent points X_1 and X_2. Then we look for the smallest number r such that $CDF(r) >= f$.

The chart below shows the CDF(r) function. To obtain the bandwidth, we read the value of x-axis r such as CDF(r) = 0.25 (by default, the first quartile is used).

For more details see `https://www.rdocumentation.org/packages/spatstat/versions/1.48-0/topics/bw.frac`.



By zooming in on the proposed value:

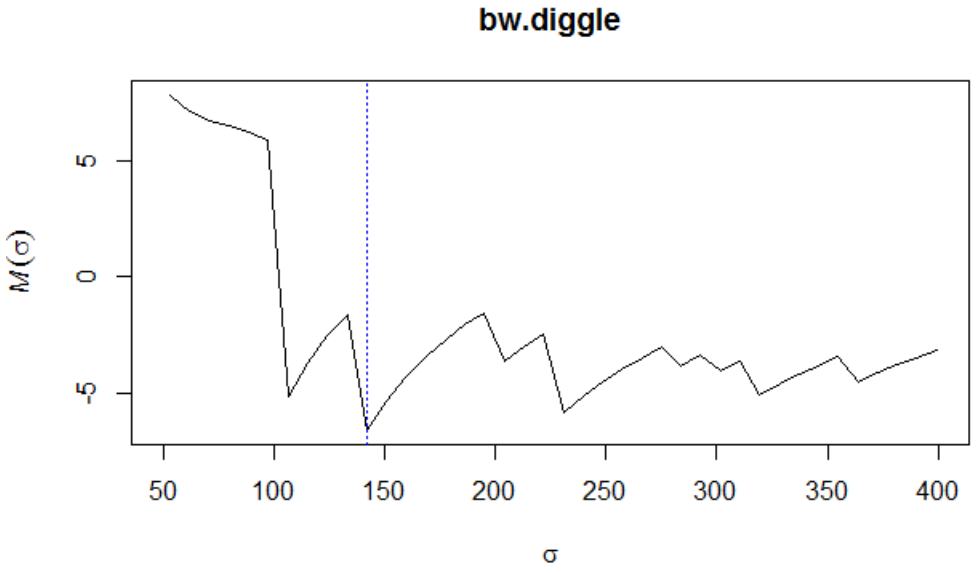
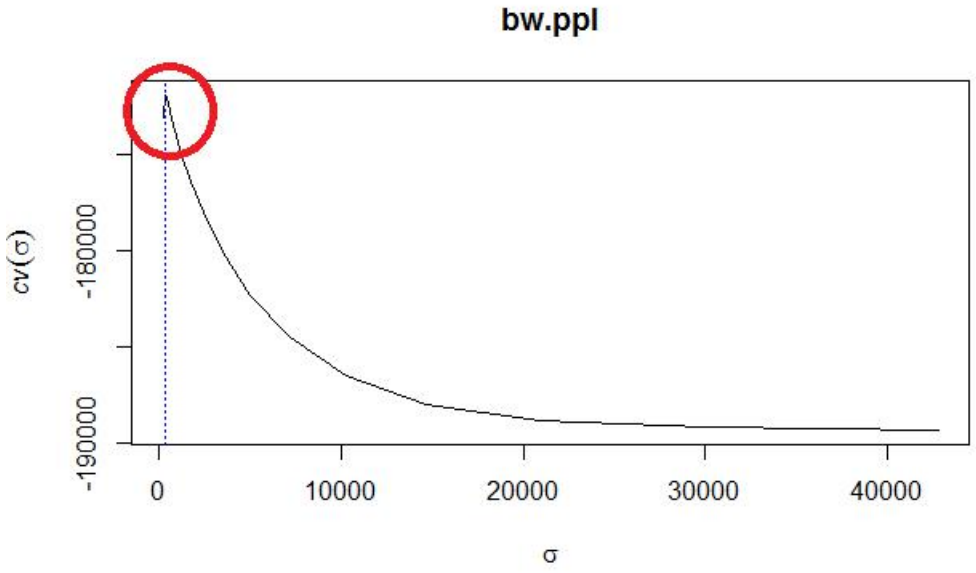


Figure 8.12 – Criteria M(σ) found using function `bw.diggle` of package *spatstat* in R
Source: *INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011*
Scope: Reunion Island



By zooming in on the proposed value:

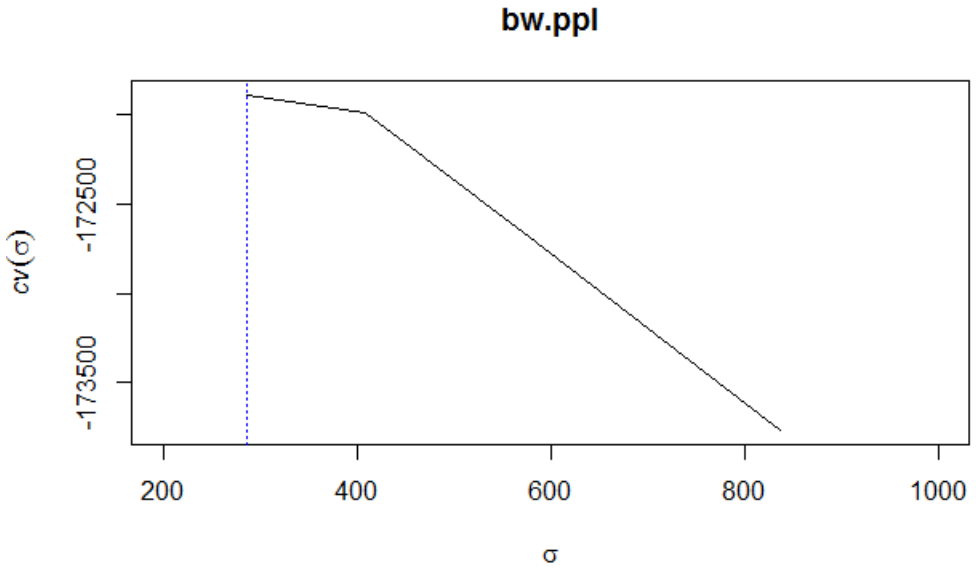


Figure 8.13 – The CV criterion(σ) derived using function `bw.ppl` of package *spatstat* in R
Source: *INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011*
Scope: Reunion Island

```
#base.ppp created above is used again

# bw.frac test for optimum bandwidth
bw_frac<-  spatstat::bw.frac(base.ppp)
plot(bw_frac, main = "bw.frac")
```

The result is:

```
bw_frac
## [1] 19747.02
```

Using the default settings, the proposed value for the σ value is 19,747 metres.

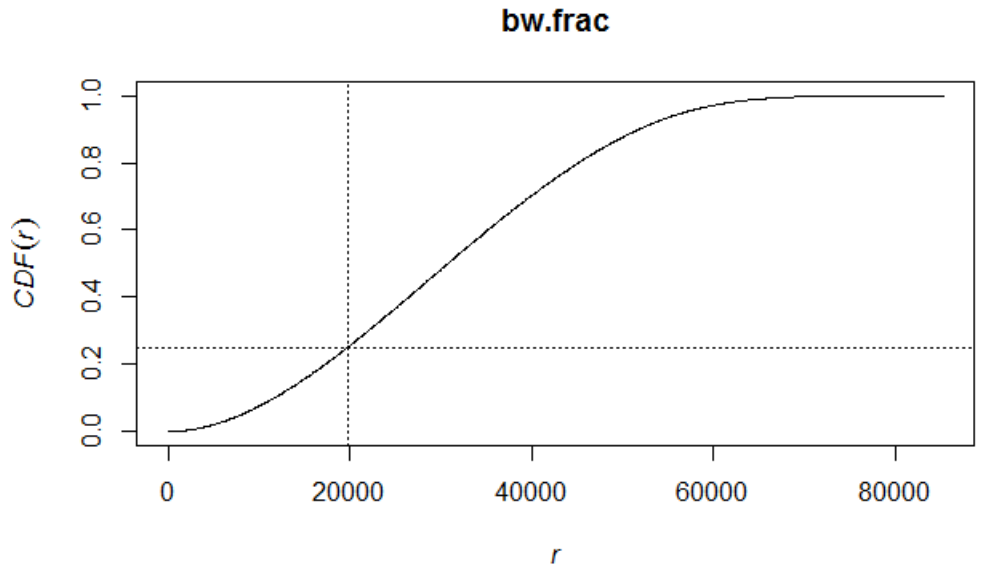


Figure 8.14 – The cumulative CDF(r) distribution function obtained by function `bw.frac` of package *spatstat* in R
Source: *INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011*
Scope: Reunion Island

With function `bw.scott` of *spatstat*

This function is based on the "Scott rule" (see Scott 1992). The idea consist in assuming that the sample is distributed according to a normal law. In this case, a bandwidth estimator is derived, minimizing an error called "mean integrated squared error". The estimator formula includes in particular the sample's standard deviation.

The result is a vector composed of two values — the bandwidths suggested in the direction of the x and y values.

```
#the base.ppp created above is used again

#bw.scott test for optimal bandwidth
bw_scott<-  spatstat::bw.scott(base.ppp)
```

The result is:

```
bw_scott
## [1] 2973.548 3455.256
```

With the default parameters, the proposed value is the couple (2974; 3455): 2,974 metres in the direction of the x and 3,455 in the direction of y values. According to the package documentation, the value suggested by this test is generally higher than that provided by `bw.diggle`.

Summary of results found

Function	σ (in metres)
bw.diggle	142
bw.ppl	286
bw.frac	19747
bw.scott	2974 (x) et 3455 (y)

Table 8.1 – "Optimal" bandwidths ($h = 2\sigma$) found using the functions of package *spatstat*
Source: *INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011*
Scope: Reunion Island

In this example, the results sometimes vary widely depending on the methods. The disparities are exacerbated by the unusual distribution of the population on Reunion Island, almost exclusively located on the coast.

Conclusion

Behind the aesthetic quality of the smoothed maps, however, lies a major trap. By construction, smoothing methods mitigate breakdowns and borders and induce continuous representation of geographical phenomena. The smoothed maps therefore show the spatial autocorrelation locally. Two points close to the smoothing radius have mechanically comparable characteristics in this type of analysis. As a result, there is little point in drawing conclusions from a smoothed map of geographical phenomena whose spatial scale is of the order of the smoothing radius. Intuitively, this amounts to commenting on the homogeneity observed within the spatial units of a choropleth map. In other words, the smoothing radius (the bandwidth) implicitly defines a minimum level of information restitution. As a corollary to these remarks, it is essential to comment only on the phenomena whose order of magnitude is much higher than the smoothing radius.

References - Chapter 8

Baddeley, A. et al. (2015a). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press.

Brunsdon, C. et al. (2002). « Geographically weighted summary statistics : a framework for localised exploratory data analysis ». *Computers, Environment and Urban Systems*.

Diggle, Peter J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press.

Floch, J.M. (2012b). « Détection des disparités socio-économiques - L'apport de la statistique spatiale ». *Documents de Travail INSEE*.

Palsky, Gilles (1991). « La cartographie statistique de la population au XIXe siecle ». *Espace, populations, sociétés* 9.3, pp. 451–458.

Scott, D.W. (1992). *Multivariate Density Estimation : Theory, Practice, and Visualization.* New York, Chichester : Wiley.