8. Lissage spatial

LAURE GENEBES, AURIANE RENAUD ET FRANÇOIS SÉMÉCURBE
Insee

8.1 Lissage spatial 212
8.1.1 Origine et formalisme du lissage spatial 212
8.1.2 Traitement des effets de bord 216
8.1.3 Choix de la bande passante 217
8.2 Lissage géographique 218
8.2.1 Lissage de données pondérées 218
8.2.2 Application utilisant une régression non paramétrique 221
8.2.3 Application utilisant une estimation de densité conditionnelle non paramétrique 221
8.2.4 Application utilisant un lissage quantile 224
8.3 Mise en œuvre avec R 224
8.3.1 Sous R, avec le package spatstat 224
8.3.2 Sous R, avec le package btb 227
8.3.3 Tests de bande passante optimale 230

Résumé

Le lissage spatial est l'une des méthodes essentielles pour analyser les données et l'organisation spatiale. L'idée est de filtrer l'information pour révéler des structures spatiales sous-jacentes.

Du point de vue conceptuel, le lissage spatial est une méthode d'estimation non paramétrique de la fonction d'intensité d'un processus ponctuel à valeurs dans $\mathbb{R}^2$, à partir uniquement d'une de ses réalisations (que l'on observe). La fonction d'intensité théorique en un point x est obtenue en calculant la moyenne des points observés par unité de surface sur des voisinages contenant x, ces voisinages étant de plus en plus petits.

Mais, en pratique, on dispose d'une seule réalisation (observée), et cette approche par passage à la limite n'a plus de sens. Les méthodes non paramétriques à noyau contournent cette limitation, en ne proposant pas directement une estimation de la fonction d'intensité mais une estimation lissée de celle-ci. Au prix de cette approximation, lorsque le paramètre de bande passante est bien choisi, les estimations obtenues sont statistiquement robustes et géographiquement pertinentes, et permettent de déceler si la fonction d'intensité est constante ou variable dans l'espace.

On s'inspire des outils de l'analyse spatiale pour produire des analyses géographiques pertinentes. L'idée est d'obtenir une représentation cartographique simplifiée et lisible, en s'affranchissant de l'arbitraire des découpages territoriaux, et en limitant en partie le "Modifiable Area Units Problem". Dans ce cas, la bande passante s'apparente à un paramètre de généralisation géographique qui conserve ou supprime en fonction des exigences de l'analyse les détails des phénomènes géographiques observés. En pratique, il est possible de lisser des données pondérées d'après BRUNSDON et al. 2002 : chaque point de l'espace est affecté d'une valeur numérique. Plusieurs types de lissage peuvent être menés, notamment un lissage "classique" reposant sur des

calculs locaux de moyennes, ou un lissage "quantile" utilisant des calculs locaux de quantiles (médiane, décile), voir BRUNSDON et al. 2002. De plus, des opérations sur les valeurs lissées permettent notamment de calculer des ratios "lissés", tels que la part d'une sous-population dans l'ensemble de la population.

La mise en œuvre d'un lissage est désormais assez aisée, en particulier avec le logiciel R dont plusieurs packages comportent des fonctions permettant de réaliser des lissages.

8.1 Lissage spatial

La fonction d'intensité théorique en un point x est obtenue en calculant la moyenne des points observés par unité de surface sur des voisinages contenant x (voir chapitre 4 : "Les configurations de points") de plus en plus petits. Le lissage spatial est une méthode d'estimation non paramétrique de la fonction d'intensité d'un processus ponctuel à valeurs dans $\mathbb{R}^2$ à partir uniquement d'une de ses réalisations. Pour obtenir la fonction d'intensité théorique à partir de la connaissance d'une seule réalisation, on n'estime pas directement la fonction d'intensité mais une fonction de la fonction d'intensité.

D'un point de vue pratique, le lissage spatial est une modélisation locale qui repose sur le choix de paramètres.
Le noyau décrit la façon dont le voisinage est appréhendé.
La bande passante est le paramètre fondamental de l'analyse. Elle quantifie la «taille» du voisinage. Ce paramètre résulte d'un arbitrage biais-variance entre la précision spatiale de l'analyse et sa qualité statistique.
Le traitement des effets de bord explicite la façon dont les frontières géographiques et les limites du territoire d'observation sont prises en compte dans l'analyse.

Par ailleurs, on peut définir un ensemble de coordonnées géographiques pour lesquelles les valeurs lissées seront estimées (éventuellement différent de l'ensemble des coordonnées géographiques des données d'origine). La plupart des applications faites par l'Insee lissent les données sur une grille de carreaux (la nouvelle coordonnée étant le centre du carreau).

Dans ce chapitre, nous aborderons dans un premier temps les fondements et le formalisme du lissage spatial puis ses mises en œuvre.

8.1.1 Origine et formalisme du lissage spatial

Historiquement, la première méthode non paramétrique d'estimation de l'intensité repose sur la construction d'intensité territoriale. Elle consiste à calculer, pour chaque unité territoriale, l'intensité de points observée par unité de surface. Dans ce cas, l'intensité est également appelée densité. Au sein de chacune de ces unités territoriales, l'intensité estimée est constante. Par exemple, lorsqu'on calcule la densité d'une région, celle-ci est considérée identique sur tout le territoire.

L'intérêt pratique de l'intensité repose sur la possibilité de représenter les densités territoriales sous la forme de cartes choroplèthes dont les premières réalisations remontent aux travaux du Baron Pierre Charles Dupin, voir PALSKY 1991. Les géographes et statisticiens utilisèrent ensuite cette méthode pour représenter la répartition de la population dans les découpages administratifs. D'un point de vue technique, les cartes de densité généralisent les histogrammes des analyses monodimensionnelles aux espaces géographiques de dimension deux. Un exemple de carte de densité est donné sur la figure 8.1.

Au 20^e siècle, les géographes et les statisticiens se sont mis progressivement à questionner la pertinence statistique et géographique de ce type d'approche. Openshaw a théorisé ses limites sous le nom de *Modifiable Area Units Problem* (MAUP) . Le MAUP (voir figure 8.2) se décompose en

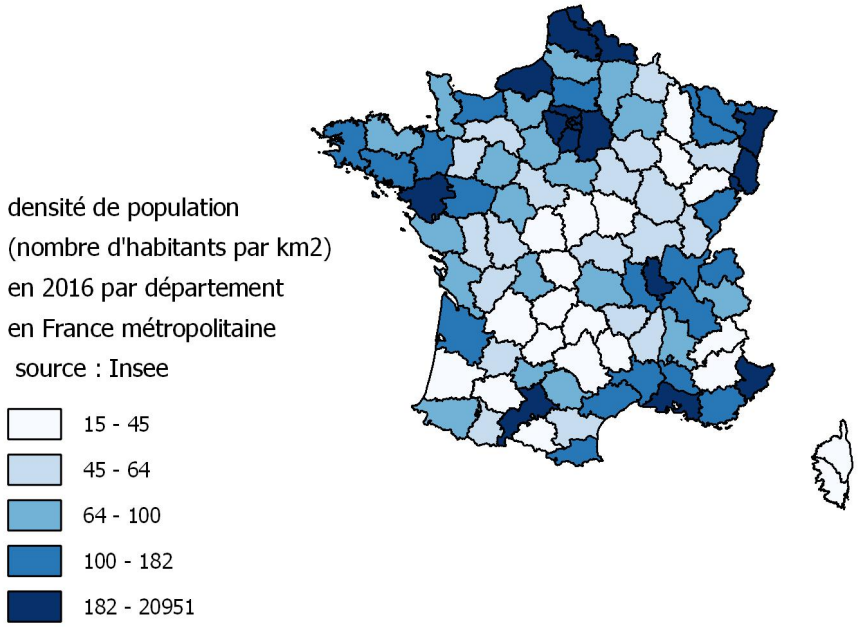


FIGURE 8.1 – Exemple de carte de densité
Source : *Insee*

deux sous-problèmes interdépendants : l'effet d'échelle (*scale effect*) et l'effet de zonage (*zoning effect*). L'effet d'échelle décrit la dépendance du phénomène observé à la taille moyenne des unités spatiales. Plus cette taille est importante et plus les spécificités locales sont réduites et plus les analyses laissent apparaître les structures globales. Au contraire, des petites tailles conservent les spécificités locales et les détaillent, mais en contrepartie les analyses sont sensibles aux bruits statistiques, à la qualité et à la précision des données. L'effet de zonage explicite la dépendance des phénomènes observés à la forme des unités spatiales. La notion de forme inclut la morphologie des unités spatiales mais également leur position dans l'espace. Ainsi, si l'on déplace uniformément le contour des unités spatiales, le phénomène observé est susceptible d'être profondément modifié.

Le lissage spatial hérite de ces réflexions et a pour objectif de s'affranchir de l'arbitraire des découpages territoriaux. Si le lissage spatial trouve une définition rigoureuse dans le cadre de l'analyse spatiale, on détecte des méthodes apparentées en géographie et en statistique dès la fin du 19^{e} siècle avec les travaux de Louis-Leger Gauthier et de Victor Turquan. Cette proximité, voire intrication, entre l'approche des statisticiens spatiaux et des géographes justifie que ce chapitre s'intéresse au lissage à la fois sous un angle d'analyse spatiale pure et d'analyse géographique plus opérationnelle.

En pratique, la difficulté à laquelle est confronté le statisticien est celle d'observer une seule réalisation. Concrètement, pour contourner cette difficulté d'estimer une fonction d'intensité à partir d' une seule réalisation, le lissage spatial n'estime pas directement celle-ci mais une version

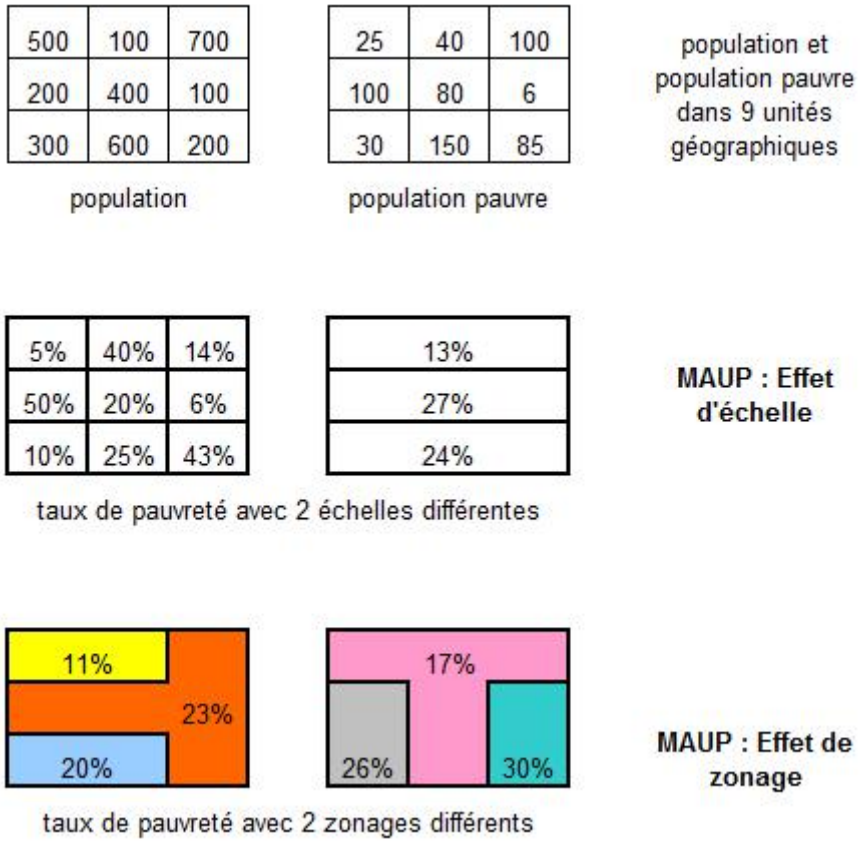


FIGURE 8.2 – Schématisation du MAUP : effet d'échelle et effet de zonage

lissée obtenue par convolution avec un noyau K_h :

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

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

et K une fonction symétrique de $\mathbb{R}^2$ dans $\mathbb{R}$ positive et d'intégrale 1

R Une métaphore simple pour comprendre l'opération de convolution consiste à imaginer que λ représente la répartition de la densité de terriers de lapins dans l'espace. Pour chaque terrier est associé un unique lapin. Chaque lapin, pour subvenir à ses besoins, se déplace à proximité de son terrier de telle sorte que sa probabilité de se retrouver à une position t par rapport à son terrier est $K_h(t)$. La convolution $(K_h * \lambda)(x)$ représente dans ce cas la densité locale de lapins en x. Si h est petit, les lapins se concentrent autour de leur terrier et la fonction d'intensité des lapin diffère peu de celle des terriers. Au contraire si h est important, les lapins ont tendance à se mélanger dans l'espace et la fonction d'intensité des lapins est «floutée» par rapport à celle des terriers.

Pour obtenir un estimateur de $(K_h * \lambda)(x)$ à partir d'un ensemble de points $\{x_i\}$ issu d'une réalisation d'un processus ponctuel, une idée simple consiste à substituer l'intégrale sur $\mathbb{R}^2$ par une somme sur les points observés dans l'équation (8.1).

Définition 8.1.1 — Lissage spatial. Soit K_h un noyau de bande passante h et x un point de $\mathbb{R}^2$, l'intensité lissée estimée en x est définie par :

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

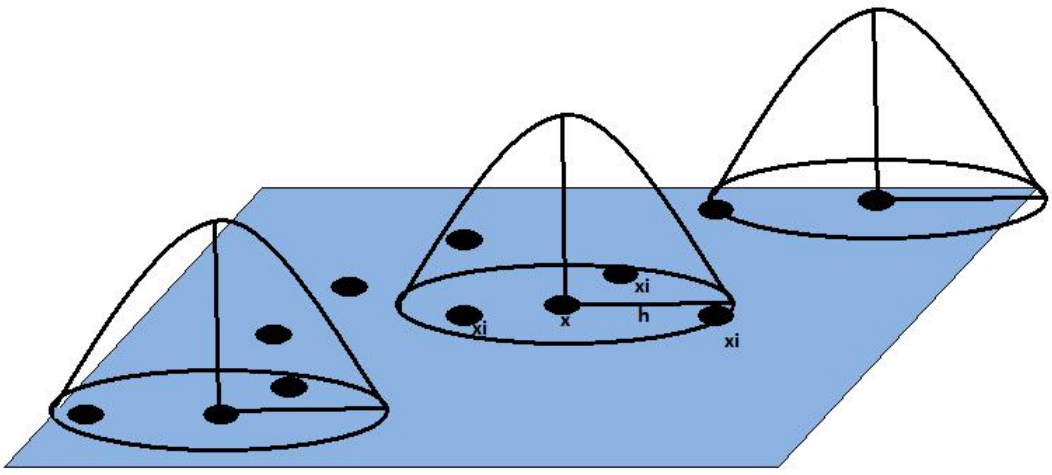


FIGURE 8.3 – Schématisation du lissage spatial avec un noyau
Note : En chaque point d'estimation figure une fonction noyau. La valeur de cette fonction est la plus élevée au niveau du point, et décroît au fur et à mesure qu'on s'en éloigne.

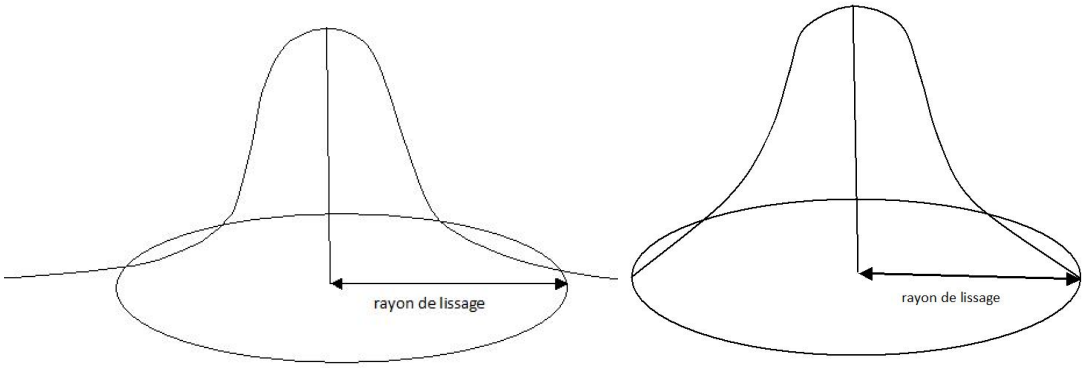
K_h dans cette formule joue un rôle analogue à une unité territoriale centrée sur chaque point de l'espace $\mathbb{R}^2$ de taille h. Contrairement aux analyses reposant sur un découpage géographique, l'estimateur de l'intensité lissée contrôle l'effet de zonage du MAUP, le choix du noyau impacte peu les résultats du lissage. En revanche, l'arbitraire de l'effet d'échelle est conservé au travers

du choix de la bande passante. Différents noyaux ont été proposés dans la littérature. Le lecteur trouvera ci-dessous les noyaux les plus fréquemment utilisés.

Définition 8.1.2 — Noyaux usuels. x est un point de $\mathbb{R}^2$. K^N et K^B sont respectivement appelés noyau gaussien et noyau quadratique :

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

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



(a) Noyau gaussien K^N (b) Noyau quadratique K^B

FIGURE 8.4 – Le noyau quadratique K^B donne un poids plus élevé aux points les plus proches qu'aux points éloignés. Il s'annule au-delà du rayon de lissage. Au contraire, le noyau gaussien K^N prend en compte l'ensemble des points de la zone d'étude.

8.1.2 Traitement des effets de bord

Par rapport à l'estimation par noyau des densités de probabilités, les méthodes de lissage spatial sont impactées par un problème supplémentaire lié à la prise en compte des effets de bord. Dans le cas de l'estimation de densité, l'estimation est réalisée sur $\mathbb{R}^n$. Pour le lissage spatial, en général les points observés sont contenus dans une fenêtre d'analyse W, soit, concrètement, un polygone.

La nature des frontières de la fenêtre peut être de deux sortes. Premièrement, la fenêtre peut résulter du protocole de collecte de l'information. Par exemple, lors d'une fouille archéologique, seule une zone restreinte est fouillée pour des raisons de coûts et d'opportunités. Dans ce cas, les frontières ne sont pas inhérentes au processus observé, et, sans information complémentaire il est raisonnable de postuler la continuité de l'intensité entre l'intérieur et l'extérieur de la fenêtre. *A contrario*, la fenêtre peut être induite par des configurations géographiques qui ont un impact sur le processus sous-jacent générateur de l'ensemble de points observés. En géographie, les fleuves, les reliefs, les côtes maritimes sont autant de frontières qui contraignent l'implantation des activités humaines. À l'extérieur d'une frontière de ce type, l'intensité du phénomène observé est nulle.

La formule (8.2) est la formule d'estimation sans traitement des effets de bord. Elle revient à ignorer la fenêtre d'analyse.

Le traitement de l'effet de bord a pour objectif de prendre en compte, dans l'estimation de l'intensité, l'impact de la frontière. Différentes solutions ont été proposées. Elles se distinguent par leur façon d'appréhender l'extérieur de la zone d'observation et leur rapidité d'exécution (Baddeley, voir BADDELEY et al. 2015a).

Définition 8.1.3 — Traitement des effets de bords. x est un point de $\mathbb{R}^2$, les estimations uniforme et de Diggle (voir DIGGLE 2013) sont obtenues par les formules suivantes :

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

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

où $e_h(u) = \int_W K_h(u - v)dv$

Lorsque la fenêtre d'analyse est indépendante du processus sous-jacent, l'estimation uniforme assure la continuité de l'intensité entre l'intérieur et l'extérieur de la fenêtre. En revanche, si l'intensité en dehors de la fenêtre est jugée nulle, il est plus opportun d'utiliser l'estimation de Diggle, voir DIGGLE 2013, qui est conservative. Dans ce cas, l'intégrale de l'intensité estimée dans la fenêtre d'analyse correspond exactement aux nombres de points observés. D'un point de vue algorithmique, l'estimation de Diggle est sensiblement plus consommatrice en temps de calcul que l'estimation uniforme.

R De façon imagée, le terme $e_{h(x)}$ peut s'interpréter comme une probabilité d'intersection de deux ensembles. Supposons, toujours en prenant notre métaphore des lapins, que l'emprise spatiale de la fenêtre correspond à un enclos. $K_h(x-u)$ décrit approximativement le territoire d'exploration d'un lapin autour d'un terrier situé en x si le lapin ne rencontre pas d'obstacle. $e_h(x) = \int_W K_h(u-x)du$ est la part du territoire exploré par le lapin contenue dans l'emprise spatiale de l'enclos. $e_{h(x)}$ est strictement inférieure à 1 si le point x est à proximité immédiate de la frontière de l'enclos. En revanche, si la zone d'exploration "naturelle" des lapins du terrier est entièrement contenue dans l'enclos, $e_{h(x)}$ est égale à 1.

Dans la formule (8.5) le terme $e_{h(x)}$ est appliqué globalement à l'estimation de densité. À proximité de la frontière de l'enclos, ce terme permet de redresser l'estimation de l'intensité. Plus le point est proche de la frontière et plus $e_{h(x)}$ est faible et la compensation sera forte. La correction uniforme considère qu'à la frontière de la fenêtre, la répartition des terriers est quasi-homogène entre l'intérieur et l'extérieur de la fenêtre. Intuitivement, cette correction revient à postuler que l'enclos n'a aucun effet sur la mobilité des lapins qui la franchisse sans s'en apercevoir. Plus précisément, on considère que les lapins dont les terriers sont situés à l'extérieur de l'enclos participent également au calcul de l'intensité à l'intérieur de l'enclos.

L'estimation de Diggle, formule (8.6), suppose que la fenêtre fait partie intégrante des propriétés du processus sous-jacent. Autrement dit, l'enclos représente une frontière infranchissable pour les lapins et tous les lapins sont contenus dans l'enclos. En divisant $K_h(x-x_i)$ par le terme $e_{h(x_i)}$, on s'assure que le lapin du terrier i a une probabilité égale à 1 de rester dans l'empreinte spatiale de l'enclos.

8.1.3 Choix de la bande passante

Le choix de la bande passante conditionne l'aspect plus ou moins «lissé» de l'estimation de la fonction d'intensité. En analyse spatiale, la bande passante résulte d'un compromis biais-variance. Le biais est induit par le fait que l'estimateur de la fonction d'intensité n'estime pas directement la fonction d'intensité mais une version lissée de celle-ci. Plus la bande passante est importante et plus le biais est important. La variance décroît au contraire en fonction de la bande passante. Plus la bande passante est importante, et plus le nombre de points participant au calcul des estimations locales augmente, ce qui tend à réduire la variance d'estimation.

Plusieurs méthodes sont disponibles pour proposer automatiquement une bande passante qui minimise un critère d'erreur. Ne disposant évidemment pas de la fonction d'intensité recherchée, une partie de ces méthodes repose sur des méthodes de validation croisée. Elles se servent de la distribution de points observée, et supposent qu'elle suit une distribution de Poisson pour

estimer une bande passante optimale. Dans la section 8.3, des exemples exploitant les fonctions de validation croisée du package *spatstat* de R seront proposés. Ces exemples mettent en valeur la grande variabilité des bandes passantes proposées en fonction des critères d'erreurs choisis. Par ailleurs, l'existence d'une unique bande passante pertinente pour toute l'étendue de la zone étudiée est une hypothèse forte. Plusieurs méthodes de lissages adaptatifs ont été proposées pour dépasser cette limite. Le lecteur pourra lire avec intérêt l'ouvrage de Baddeley (BADDELEY et al. 2015a) sur cette thématique qui exploite le package *spatstat* de R.

Finalement, en soi, aucune bande passante n'est optimale : toutes sont susceptibles d'apporter une représentation du monde pertinente conformément au MAUP. Certains géographes conseillent d'adopter une démarche multi-échelle pour appréhender la pluralité des aspects spatiaux d'un même phénomène.

8.2 Lissage géographique

Le lissage géographique s'inspire de l'estimation d'intensité présentée précédemment. Il n'a pas vocation à calculer des intensités, mais à obtenir des représentations cartographiques simplifiées. Le principe de cette utilisation en géographie est de représenter non pas la valeur observée en un point, mais une moyenne pondérée des valeurs observées au voisinage de ce point dans un rayon prédéfini.

R Le lissage peut être interprété comme un outil pouvant assurer une forme de **confidentialité**. Il permet de représenter de manière agrégée des données initialement ponctuelles et confidentielles. Il faut néanmoins rester vigilant sur le nombre de points utilisés pour produire l'estimation lissée.

8.2.1 Lissage de données pondérées

On se place dans le cas où chaque point x_i est affecté d'une valeur numérique w_i. Par exemple, x_i peut représenter un logement et w_i le nombre d'habitants de ce logement. Il suffit (voir BRUNSDON et al. 2002) d'utiliser une version pondérée des estimateurs à noyaux décrits précédemment. Dans la formule (8.2), on multiplie par le poids w_i la contribution d'un point à l'estimateur d'intensité.

Définition 8.2.1 — Estimateurs à noyaux pondérés. Soit K_h un noyau de bande passante h et x_i un point de $\mathbb{R}^2$ affecté d'une pondération w_i, l'intensité lissée estimée en x est définie par :

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

Alors que le choix du noyau K_h a peu d'influence sur les résultats du lissage (voir figure 8.5), le choix de la bande passante h est primordial, bien qu'assez arbitraire.

Comme cela a été souligné plus haut, cette bande passante se comporte comme un paramètre de lissage, contrôlant l'équilibre entre biais et variance. Un rayon élevé conduit à une densité très lissée, avec un biais élevé. Un petit rayon génère une densité peu lissée avec une forte variance. Il est généralement déterminé par l'utilisateur de faire un compromis, en fonction du niveau d'agrégation souhaité. Il est conseillé de tester plusieurs valeurs de bande passante, permettant de révéler des variations locales à différentes échelles. Les cartes de la figure 8.6 sont des exemples de cartes lissées pour Paris et sa banlieue, avec trois rayons de lissage différents.

L'intérêt de l'estimation est de s'intéresser non pas aux points et leur répartition, mais à leur environnement. La bande passante permet ainsi de définir cet environnement.

R

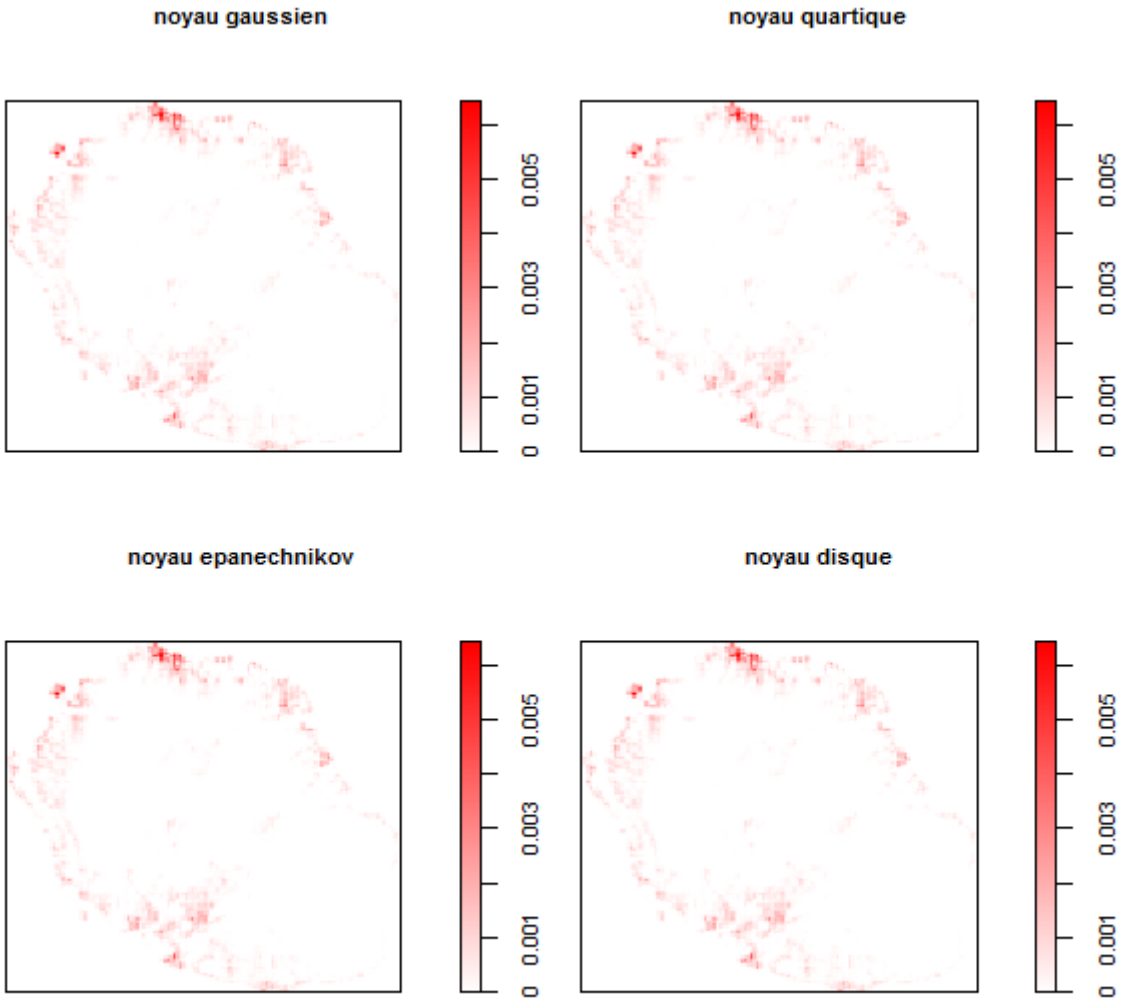
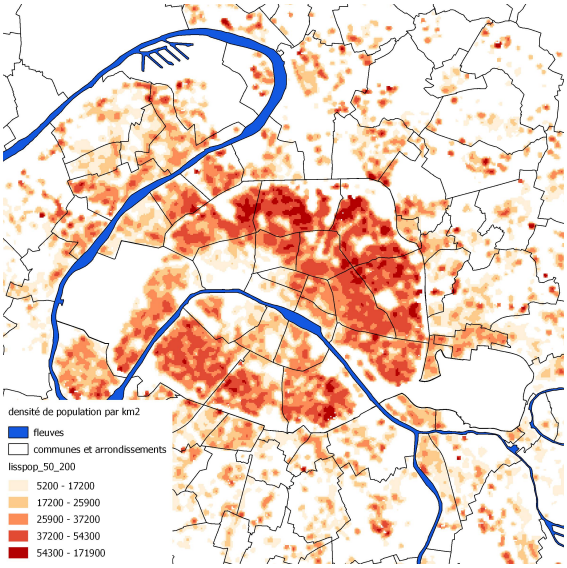
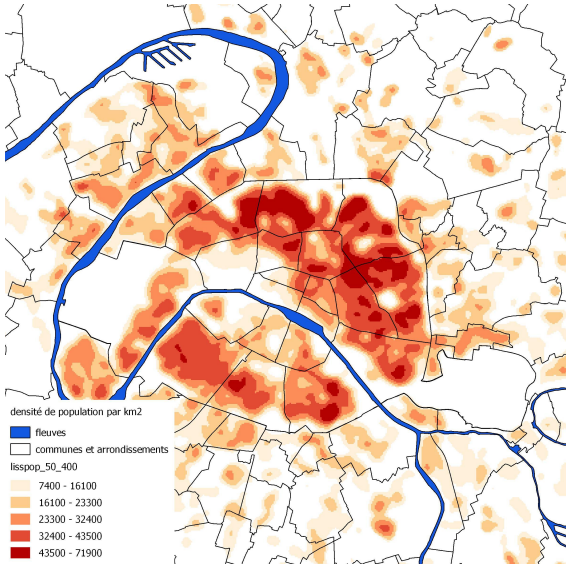


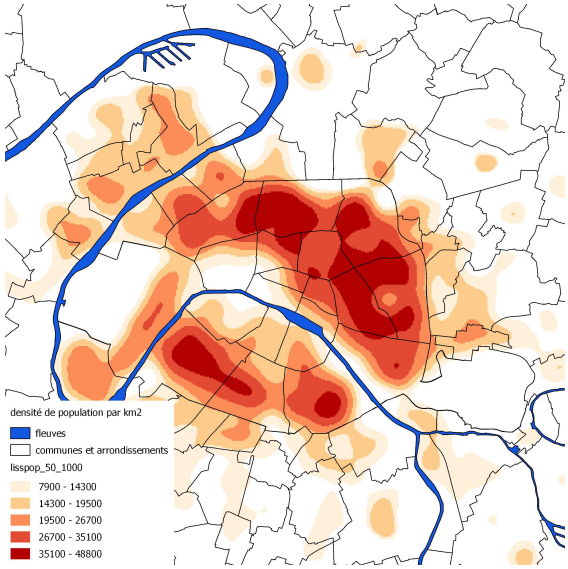
FIGURE 8.5 – Comparaison de résultats obtenus avec quatre noyaux différents à partir de la fonction `density.ppp` du package *spatstat*
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Ile de la Réunion
Note : La variable représentée est le nombre lissé de ménages



(a) Rayon de 200 mètres



(b) Rayon de 400 mètres



(c) Rayon de 1000 mètres

FIGURE 8.6 – Trois rayons de lissage différents pour la densité de population à Paris et sa banlieue : 200 mètres, 400 mètres, 1000 mètres
Source : *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*
Note : Les carreaux représentés contiennent plus de 11 ménages

Plusieurs algorithmes existent pour déterminer un rayon de lissage dit «optimal». Ces tests peuvent donner des résultats variables et parfois très éloignés (voir mise en œuvre) : il est conseillé de ne les utiliser qu'à titre indicatif, il revient à l'utilisateur de choisir le rayon de lissage compte tenu de son expérience des données, et de la problématique.

Des opérations peuvent être effectuées sur les variables lissées, notamment des ratios. La justification théorique se trouve pp. 34 à 37 du document de travail de J.M. Floch (FLOCH 2012). De manière pratique, si l'on souhaite obtenir la valeur lissée du ratio de deux variables, il est primordial de calculer séparément les valeurs lissées du numérateur et du dénominateur, et ensuite de calculer le rapport entre la valeur lissée du numérateur, et la valeur lissée du dénominateur. Ne pas calculer directement une valeur lissée d'un ratio : la carte serait déformée, car on donnerait à tort la même importance aux différents territoires, pourtant inégalement peuplés.

8.2.2 Application utilisant une régression non paramétrique

On s'intéresse au **calcul d'un revenu moyen** par personne. On dispose de deux variables : le revenu, et le nombre de personnes. Le revenu moyen est égal à la somme de l'ensemble des revenus, divisée par la somme du nombre de personnes. On lisse séparément les revenus, et le nombre de personnes. On calcule ensuite le rapport.

On obtient les cartes de la figure 8.7 pour Paris et les communes environnantes faisant partie de la "petite couronne" (*i.e.* les trois départements limitrophes de Paris) :

La carte 8.7a du niveau de vie total des ménages n'a pas beaucoup de sens, il est nécessaire de rapporter ce niveau de vie total par carreau à la population de chaque carreau.
Sur la carte 8.7b du nombre de personnes : la population est très dense au sein de la commune de Paris, essentiellement au Nord-Est de la Seine, et dans une moindre mesure à l'extrême Sud.
Sur la carte 8.7c, le niveau de vie moyen par personne est très élevé en plein cœur de Paris, essentiellement à l'Ouest.

R Ici, on n'est plus dans le cadre théorique. Ce calcul s'inspire d'outils apparentés à une régression non paramétrique. Approximativement, c'est comme si l'on faisait une régression géographique pondérée, qui se limiterait à une seule variable : la constante (voir chapitre 9 : "Régression géographiquement pondérée").

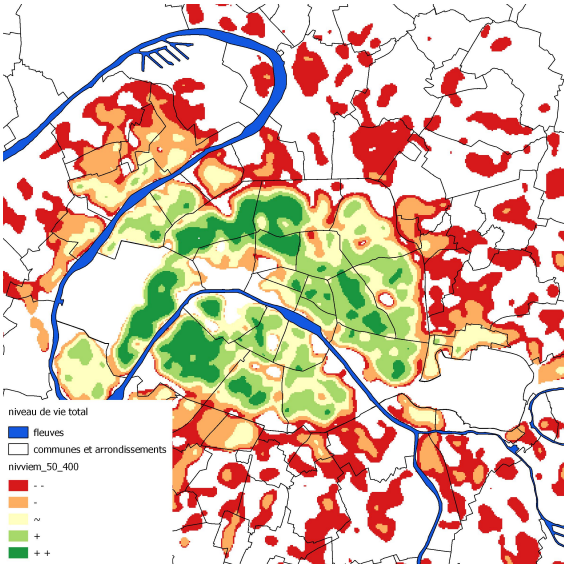
8.2.3 Application utilisant une estimation de densité conditionnelle non paramétrique

On s'intéresse à la **part des ménages pauvres** dans l'ensemble des ménages. On calcule la valeur lissée du nombre de ménages pauvres d'un territoire, et on calcule la valeur lissée du nombre total de ménages sur le territoire. On calcule ensuite le rapport.

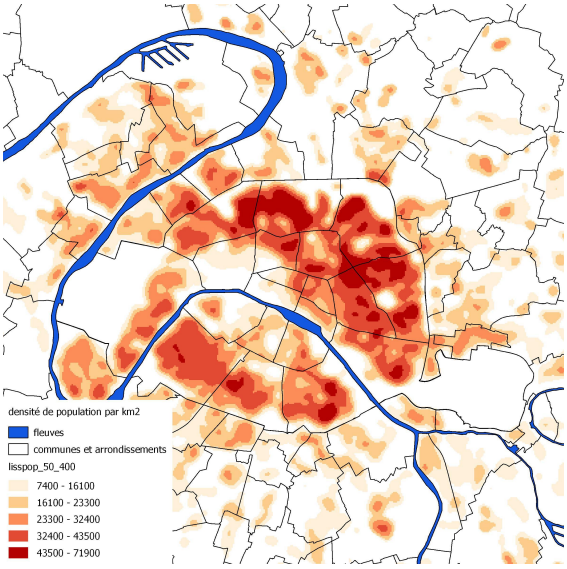
D'après la carte de la figure 8.8a, les zones les plus peuplées sont situées au cœur de Paris : à la fois le quart Nord-Est, et le quart Sud-Ouest. Dans la figure 8.8b, les ménages pauvres sont nombreux à Paris, plutôt dans le quart Nord-Est. Dans la figure 8.8c, la part des ménages pauvres dans l'ensemble des ménages apporte une information complémentaire. Sur cette carte, sont mises en exergue des zones moins densément peuplées, mais pour lesquelles la part de ménages vivant en dessous du seuil de pauvreté est forte. Il s'agit de communes situées au nord de Paris.

Ainsi, selon la carte produite, les messages obtenus peuvent être différents. Lorsque l'on analyse des taux, il est indispensable d'analyser également la répartition des simples effectifs (densités de population par exemple), pour vérifier la robustesse des taux calculés, et leur représentativité.

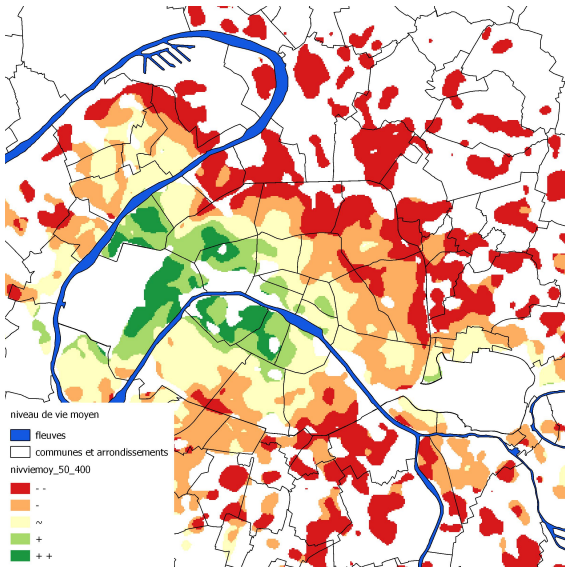
R Ce calcul s'apparente à un calcul de probabilité conditionnelle. On obtient une carte représentant des taux de pauvreté au niveau local, ce qui est proche de l'idée d'obtenir la probabilité qu'un ménage soit pauvre sachant qu'il habite à un endroit donné.



(a) Total du niveau de vie des ménages (en euros)

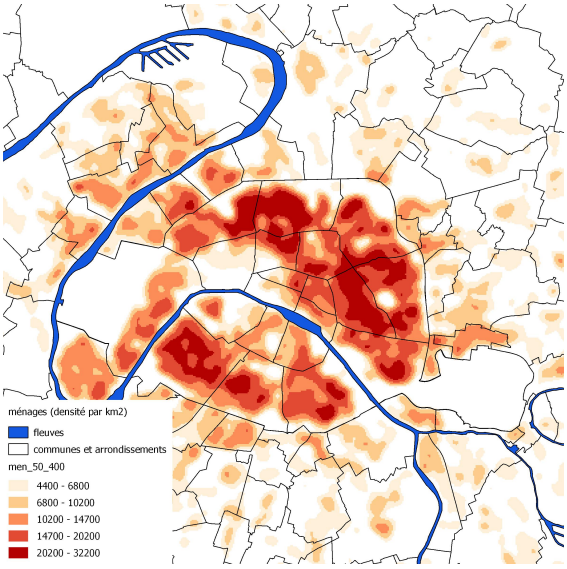


(b) Population des ménages (densité par km2)

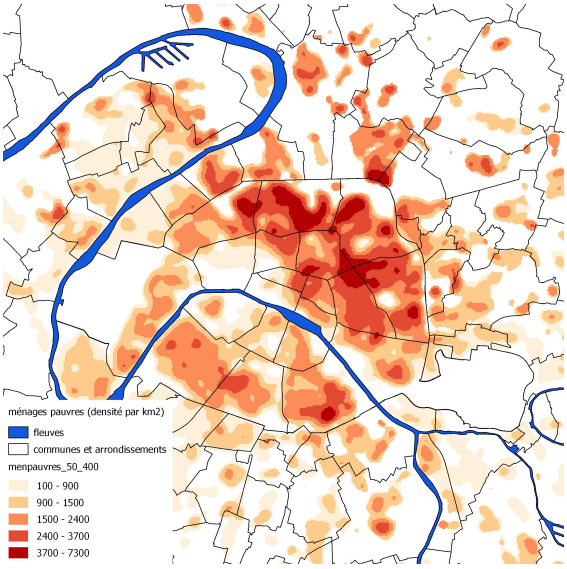


(c) Niveau de vie moyen

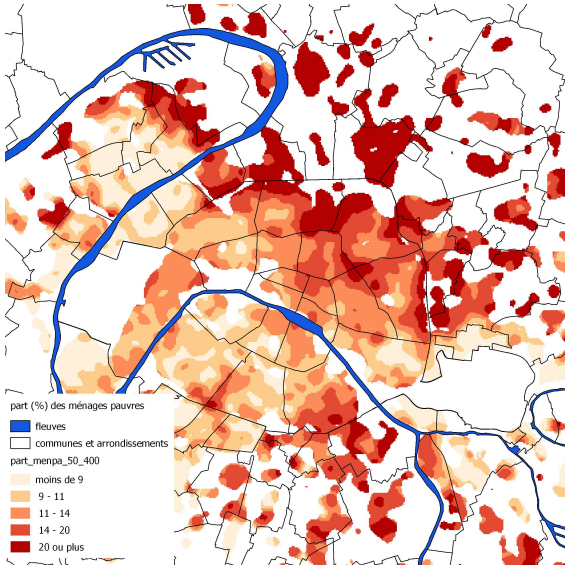
FIGURE 8.7 – Calcul d'un niveau de vie moyen lissé
Source : *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*
Note : Les carreaux représentés contiennent plus de 11 ménages. Pour les cartes représentant des niveaux de revenus, les marqueurs "++", "+", "∼", "-" et "- -" correspondent à des valeurs respectivement très élevées, élevées, moyennes, basses ou très basses pour l'indicateur considéré. Ils ont été utilisés pour des questions de non-profilage de la population



(a) Ménages (densité par km2)



(b) Ménages pauvres (densité par km2)



(c) Part des ménages pauvres dans l'ensemble des ménages

FIGURE 8.8 – Calcul de la part lissée des ménages pauvres
Source : *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*
Note : Les carreaux représentés contiennent plus de 11 ménages

Attention ! En théorie, il est toujours possible de calculer le rapport de deux variables lissées. En pratique, il est nécessaire de faire attention aux petits effectifs. Dans l'exemple du calcul de la part de population pauvre, les zones comportant des petits effectifs pourraient à tort apparaître distinctement dans la carte lissée. Peu peuplées, elles n'apparaîtraient pas sur une carte représentant les données brutes. Ainsi, en ne prenant pas garde à ce phénomène, on pourrait mécaniquement donner l'impression, faussée, que tous les territoires seraient peuplés.

8.2.4 Application utilisant un lissage quantile

Le lissage décrit jusqu'à présent est un lissage moyen, dans le sens où il est fondé sur des calculs locaux de moyennes. Dans l'article de BRUNSDON et al. 2002, les auteurs étendent cette notion, pour définir des statistiques locales fondées sur des quantiles (médiane, déciles ...). Ces indicateurs sont réputés, dans l'analyse exploratoire des données "classique", pour être moins sensibles aux valeurs extrêmes. Le lissage quantile permet surtout de calculer des indicateurs qui enrichissent considérablement l'analyse de certaines variables, notamment les variables de revenus.

Les quatre vignettes de la figure 8.9 représentent plusieurs indicateurs lissés calculés à partir du niveau de vie (source *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*), c'est-à-dire le revenu disponible d'un ménage divisé par le nombre d'unités de consommation de ce ménage.

Les cartes de la figure 8.9 sont centrées sur Paris, et incluent la "petite couronne". Le lissage est réalisé sur des carreaux de 50m, avec une bande passante de 400m. N'ont été retenus pour la visualisation que les carreaux pour lesquels le nombre d'observations (de ménages) ayant contribué à l'estimation est strictement supérieur à 50.

La carte 8.9a représente le niveau de vie médian. On retrouve, à l'Ouest, des zones où les habitants sont beaucoup plus aisés. Les cartes 8.9b et 8.9c représentent le 1er décile et le 9e décile du niveau de vie. La carte 8.9d représente le rapport interdécile, ratio entre le 9e et le 1er décile, et apporte un éclairage complémentaire. Exprimé sans unité, il s'agit du niveau de vie minimal des 10 % les plus riches rapporté au niveau de vie maximal des 10 % les plus pauvres. Il met en évidence l'écart entre le haut et le bas de la distribution : c'est une des mesures de l'inégalité de cette distribution. Dans les zones situées à l'Ouest, les rapports interdéciles sont très élevés : dans ces quartiers, cohabitent des populations dont le niveau de vie est très élevé, avec des populations dont le niveau de vie est beaucoup plus faible.

8.3 Mise en œuvre avec R

Avec R, plusieurs packages permettent de réaliser des lissages. Nous détaillerons ci-dessous la mise en œuvre pratique en utilisant les packages *spatstat* et *btb*, appliqués à des données concernant l'Île de la Réunion. Les données utilisées en exemple sont le `dataframe` *reunion.Rdata* fourni dans le package *btb*. Un aperçu de ce `dataframe` est donné sur la figure 8.10.

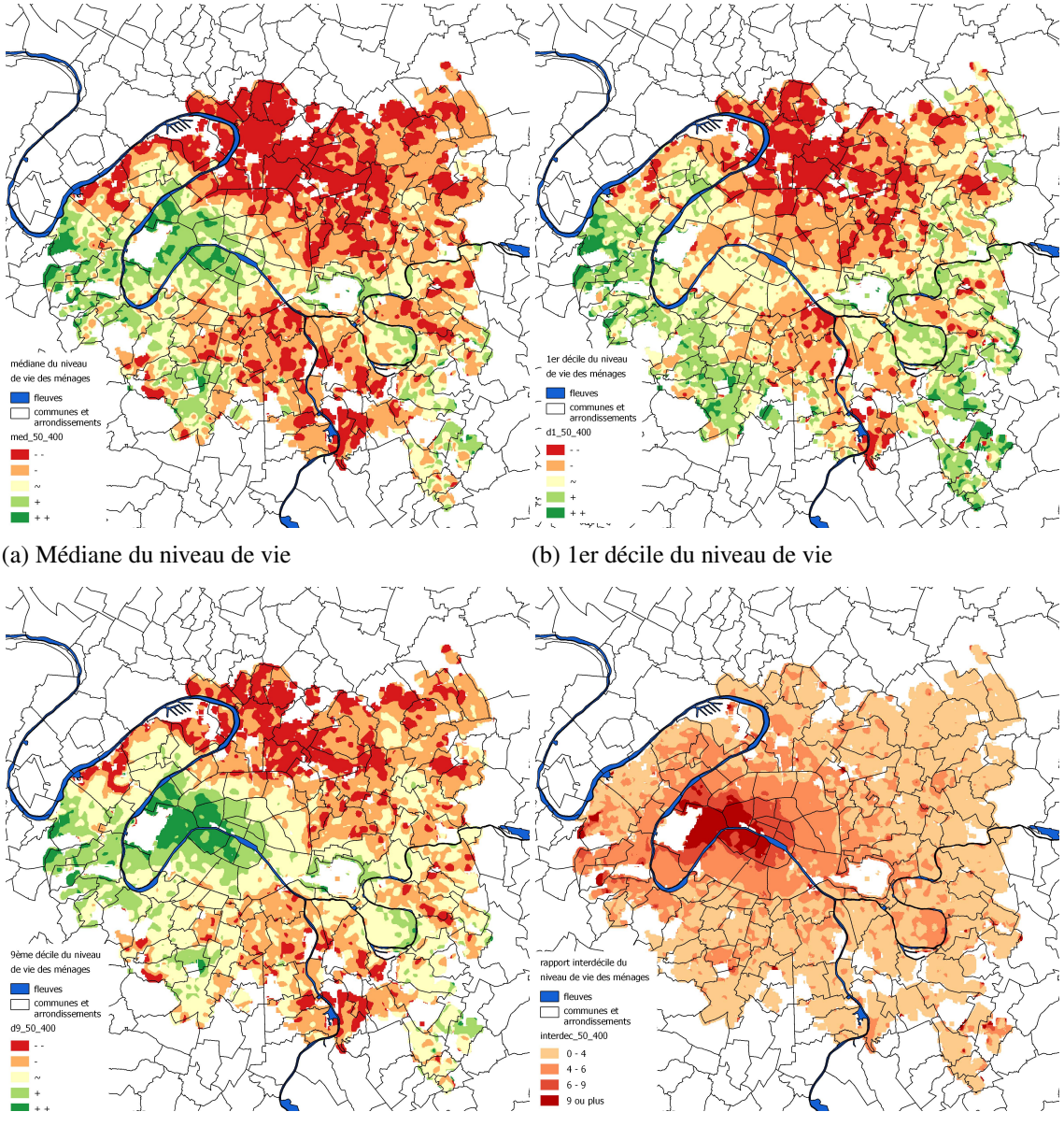
Il s'agit de données carroyées à 200 mètres, téléchargeables sur `insee.fr`. La source est *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*.

Les variables sont ainsi définies :

- x : longitude (système de projection : WGS 84 / UTM zone 40S, code EPSG : 32740)
- y : latitude (système de projection : WGS 84 / UTM zone 40S, code EPSG : 32740)
- houhold : nombre de ménages
- phouhold : nombre de ménages pauvres (définition pauvreté à 60 %)

8.3.1 Sous R, avec le package spatstat

Le package R appelé *spatstat* est un package très complet dédié à l'analyse des processus de points spatiaux. Il est disponible sur le site du CRAN à l'adresse suivante : `https://CRAN`.



(a) Médiane du niveau de vie

(b) 1er décile du niveau de vie

(c) 9ème décile du niveau de vie

(d) Rapport interdécile du niveau de vie

FIGURE 8.9 – Distribution du niveau de vie
Source : *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*
Note : Les carreaux représentés contiennent plus de 11 ménages. Pour les cartes représentant des niveaux de revenus, les marqueurs "++", "+", "∼", "-" et "- -" correspondent à des valeurs respectivement très élevées, élevées, moyennes, basses ou très basses pour l'indicateur considéré. Ils ont été utilisés pour des questions de non-profilage de la population

	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 – Les 15 premières lignes du `data.frame` *reunion.Rdata* du package *btb*
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion

`R-project.org/package=spatstat`

La fonction `density.ppp` disponible dans le package *spatstat* permet d'effectuer le lissage des données. L'utilisation de cette fonction nécessite en entrée l'utilisation d'un objet au format *.ppp*. Afin d'utiliser cette fonction, les coordonnées x et y du `data.frame` de départ doivent être converties au format *.ppp*.

```
#lissage de la variable houhold (nombre de ménages) avec Spatstat

library(spatstat)
library(btb) #uniquement pour le dataframe reunion
data(reunion)

# agrégation et suppression des doublons sur les coordonnées
base_temp <- aggregate(houhold ~ x+y, reunion, sum)

# transformation des x,y en objet .ppp
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)) )

#appel de la fonction density.ppp
#le parametre sigma correspond à h/2 avec h la bande passante
densite <- spatstat::density.ppp (base.ppp, sigma = 200, weights=base_temp$
    houhold )

#affichage de la carte
plot(densite, main = "Lissage spatstat, rayon défini par utilisateur")
```

8.3.2 Sous R, avec le package btb

Le package *btb* ("beyond the border") [1], en ligne sur le site du CRAN à l'adresse suivante : `https://CRAN.R-project.org/package=btb`, propose des fonctions dédiées à l'analyse urbaine. Il met en œuvre une estimation de densité par la méthode KDE (*kernel density estimator*), c'est-à-dire une méthode par noyau. Le noyau utilisé est un noyau quadratique.

Dans l'estimation réalisée par le package, l'effet des frontières est pris en compte dans la fonction de lissage `kernelSmoothing` *via* la correction de Diggle (DIGGLE 2013). Cette correction permet notamment de traiter le cas de frontières qui représentent des limites géographiques (des côtes maritimes par exemple). À l'intérieur de la zone d'observation, l'intensité est non nulle. À l'extérieur de la zone d'observation, elle est nulle. La méthode implémentée est conservative (grâce à une normalisation) : avant et après lissage, le nombre de points observés est identique.

Les temps de calcul ont été fortement réduits, et ce de plusieurs manières :
- en codant en C++ les méthodes les plus chronophages ;
- en se restreignant, pour chaque point, à une fenêtre d'observation autour de ce point, permettant de limiter le nombre d'opérations (calculs de distances) à effectuer.

1. Il y aura prochainement une version du package *btb* qui sera adaptée au nouveau package *sf* (simple features)

```
#lissage avec btb : calcul de la part de ménages pauvres
#on lisse séparément le numérateur (nombre de ménages pauvres), et le dé
    nominateur (nombre total de ménages)

library(btb)

#chargement des données
data(reunion)

#lissage
#définition des paramètres
pas <- 200 #carreau de 200m de cote
rayon <- 400 #bande passante de 400m

#appel de la fonction de lissage
#la fonction lisse automatiquement l'ensemble des variables contenues dans
    la base
#ici on lisse phouhold et  houhold
dfLisse <- btb::kernelSmoothing(dfObservations = reunion, iCellSize = pas,
                                          iBandwidth = rayon, sEPSG="32740")

#taux de ménages pauvres : ratio des variables lissées
dfLisse$txmenpa = 100 * dfLisse$phouhold / dfLisse$houhold

#aperçu dans R
library(sp)
library(cartography)
#affichage de la carte
cartography::choroLayer(dfLisse, var = "txmenpa", nclass = 5, method = "
    fisher-jenks", border = NA, legend.pos = "topright", legend.title.txt =
     "txmenpa (%)")

#ajout du titre et d'un contour
cartography::layoutLayer(title = "La Réunion : taux de ménages pauvres",
                         sources = "",
                         author = "",
                         scale = NULL,
                         frame = TRUE,
                         col = "black",
                         coltitle = "white")
```

La carte obtenue est représenté sur la figure 8.11.

L'utilisateur peut également exporter le résultat au format shapefile pour le retravailler ensuite dans un SIG.

```
#export au format shapefile

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

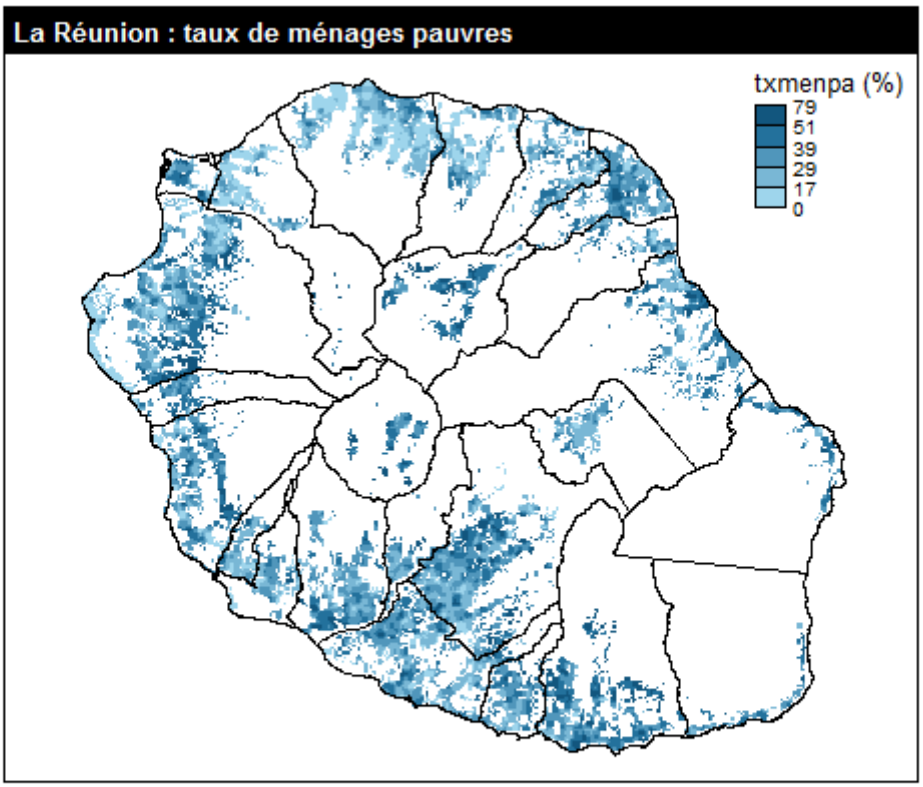


FIGURE 8.11 – Taux de ménages pauvres à la Réunion après lissage
Source : *Insee, Revenus Fiscaux Localises au 31/12/2010 et Taxe d'habitation au 01/01/2011*
Note : ratio du nombre de ménages pauvres lissé et du nombre total de ménages lissé. Les contours noirs représentent les limites communales

Le package *btb* permet également d'utiliser le lissage quantile, décrit plus haut. Il suffit de spécifier comme paramètre `vQuantiles` le vecteur des quantiles à calculer. Par exemple *c(0.1, 0.25, 0.5)* retournera le premier décile, le premier quartile et la médiane de chacune des variables du `data.frame` en entrée.

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

# définition des paramètres
pas <- 200
rayon <- 400

# appel de la fonction de lissage
dfLisse_quantile<- btb::kernelSmoothing(dfObservations = reunion,
                                        iCellSize = pas,
                                        iBandwidth = rayon,
                                        vQuantiles = c(0.1, 0.5, 0.9),
                                        sEPSG="32740")

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

R Le package *btb* propose par défaut une grille automatique de carreaux. Il est également possible d'appeler la fonction de lissage en utilisant une grille au choix de l'utilisateur. Dans ce cas, l'utilisateur doit disposer d'un `data.frame` composé de deux colonnes x et y, qui correspondent aux coordonnées des centroïdes souhaités.

```
# fonction de lissage avec grille au choix de l'utilisateur
kernelSmoothing(dfObservations, iCellSize, iBandwidth, dfCentroids)
```

8.3.3 Tests de bande passante optimale

Dans R, plusieurs méthodes proposant de calculer une bande passante "optimale" sont implémentées, selon différents critères. L'objectif est généralement de minimiser une mesure d'erreur. Dans *spatstat* par exemple, il existe les quatre fonctions suivantes : `bw.diggle`, `bw.ppl`, `bw.frac` et `bw.scott`.

Avec la fonction `bw.diggle` de *spatstat*

La fonction `bw.diggle` de *spatstat* choisit une bande passante qui minimise un critère M(σ) basé sur l'erreur quadratique moyenne (en anglais MSE pour *Mean Square Error*) de l'estimateur.

Le graphique de la figure 8.12 représente le critère M(σ) que l'on souhaite minimiser. Pour obtenir la valeur σ, il s'agit de repérer sur l'axe des abscisses la valeur qui correspond à la valeur minimale en ordonnées.

Pour plus de détails, voir `https://www.rdocumentation.org/packages/spatstat/versions/1.49-0/topics/bw.diggle`.

```
#on réutilise base.ppp créée plus haut
# test bw.diggle de bande passante optimale
bw_diggle <- spatstat::bw.diggle(base.ppp)
plot(bw_diggle, main = "cross validation")

#appel de density.ppp avec la bande passante calculée automatiquement
densite_optim <- spatstat::density.ppp(base.ppp,bw_diggle, weights=base_
    temp$houhold)
```

On obtient :

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

Avec les paramètres par défaut, la valeur proposée pour σ est de 142 mètres soit 284 mètres pour la bande passante h (σ = h/2 , voir la documentation du package).

Avec la fonction `bw.ppl` de *spatstat*

La bande passante est choisie en calculant un estimateur de maximum de vraisemblance, en utilisant une méthode de validation croisée (*likelihood cross-validation criterion*). On itère les calculs : à chaque fois, on ne travaille que sur $n-1$ observations puis on valide le modèle sur l'observation qui avait été écartée. On répète cela n fois.

Le graphique ci-dessous représente le critère CV(σ) que l'on souhaite minimiser. Pour obtenir la valeur σ, il s'agit de repérer sur l'axe des abscisses la valeur qui correspond à la valeur maximale en ordonnées.

Pour plus de détails, voir `https://rdrr.io/cran/spatstat/man/bw.ppl.html`.

```
#on réutilise base.ppp créée plus haut

# test bw.ppl de bande passante optimale
bw_ppl <- spatstat::bw.ppl(base.ppp)
plot(bw_ppl, main = "bw.ppl")
```

On obtient :

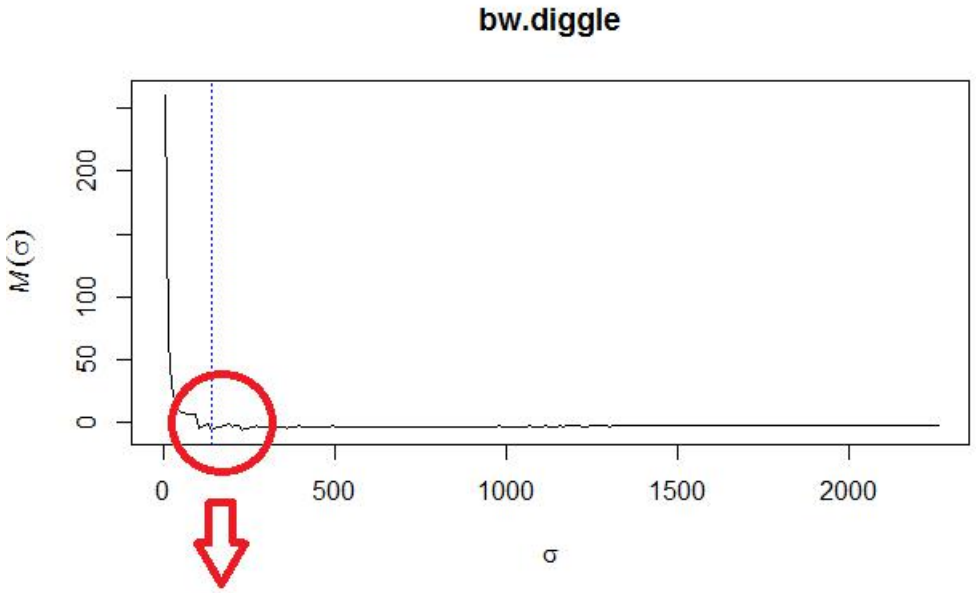
```
bw_ppl
##    sigma
## 286.0097
```

Avec les paramètres par défaut, la valeur proposée pour la valeur de σ est de 286 mètres.

Avec la fonction `bw.frac` de *spatstat*

Cette méthode sélectionne une bande passante qui repose uniquement sur la géométrie de la fenêtre d'observation.

La bande passante est un quantile (que l'on spécifie) de la distance entre deux points indépendants, pris au hasard dans la fenêtre. Par défaut, c'est le premier quartile de la distribution qui est utilisé. Si on note CDF(r) la fonction de distribution cumulée de la distance entre deux points indépendants pris au hasard et uniformément distribués dans la fenêtre, alors la valeur qui



En zoomant autour de la valeur proposée :

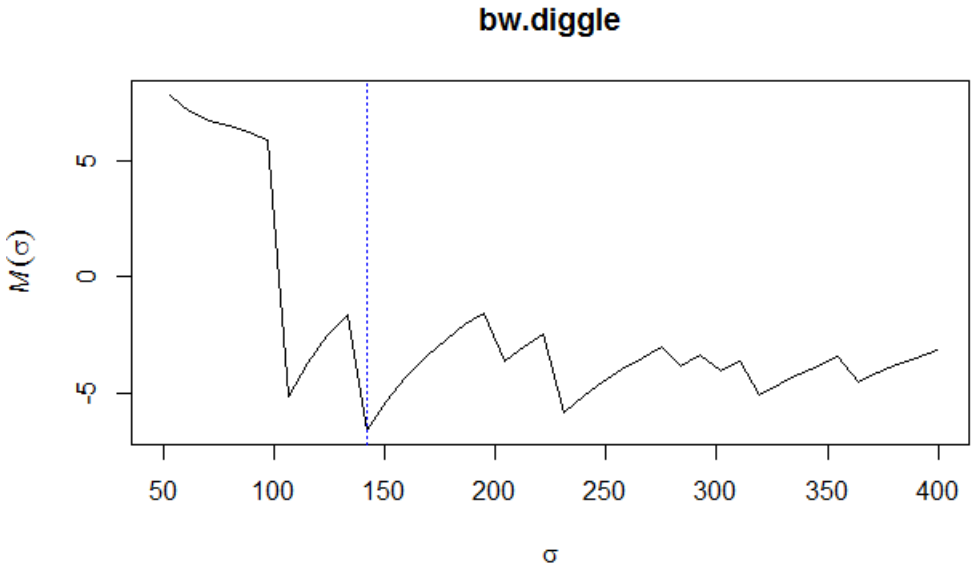
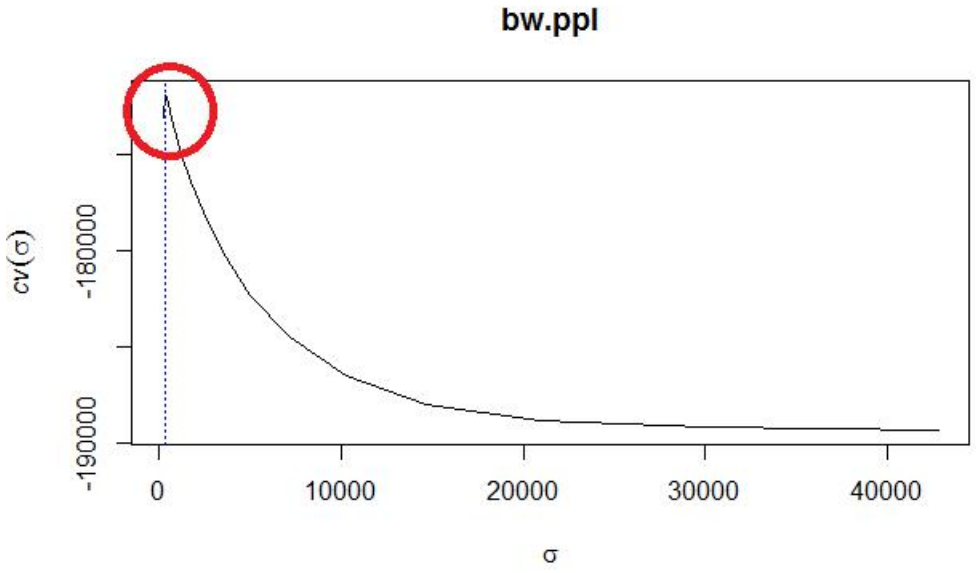


FIGURE 8.12 – Le critère M(σ) obtenu par la fonction `bw.diggle` du package *spatstat* de R
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion



En zoomant autour de la valeur proposée :

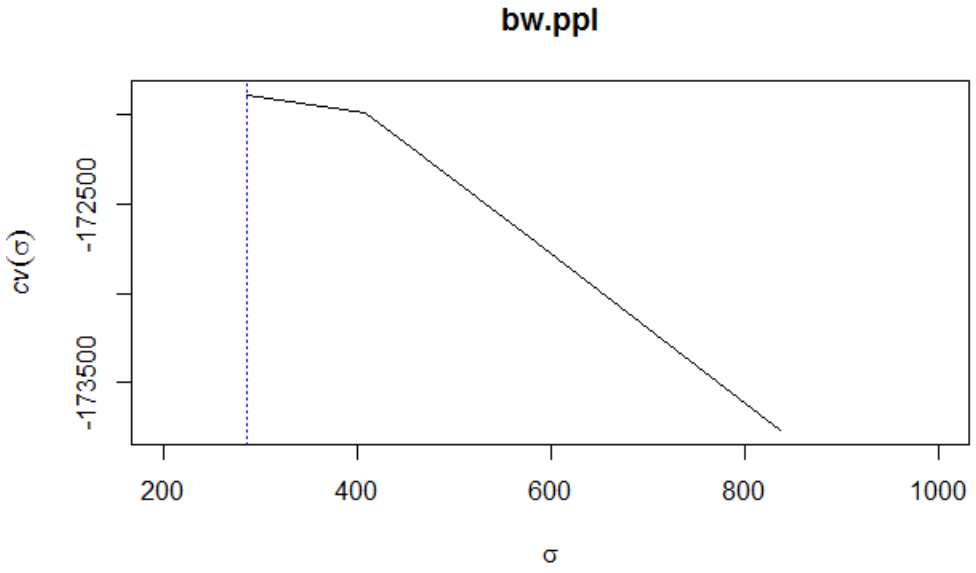


FIGURE 8.13 – Le critère CV(σ) obtenu par la fonction `bw.ppl` du package *spatstat* de R
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion

est retournée est le quantile avec la probabilité f. Alors la bande passante est la valeur r telle que CDF(r)=f. En premier, l'algorithme calcule la fonction de distribution cumulée CDF(r) avec la fonction `distcdf` du package *spatstat*. Cette fonction permet de calculer la fonction CDF(r) = P(T<=r) de la distance euclidienne T=|X_1-X_2| entre deux points indépendants pris au hasard X_1 et X_2. Ensuite, on cherche le plus petit nombre r tel que CDF(r) $\geq$ f.

Le graphique ci-dessous représente la fonction CDF(r). Pour obtenir la bande passante, on lit la valeur des abscisses r telle que CDF(r) = 0.25 (par défaut c'est le premier quartile qui est utilisé).

Pour plus de détails voir `https://www.rdocumentation.org/packages/spatstat/versions/1.48-0/topics/bw.frac`.

```
#on réutilise base.ppp créée plus haut

# test bw.frac de bande passante optimale
bw_frac<-  spatstat::bw.frac(base.ppp)
plot(bw_frac, main = "bw.frac")
```

On obtient :

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

Avec les paramètres par défaut, la valeur proposée pour la valeur de σ est de 19747 mètres.

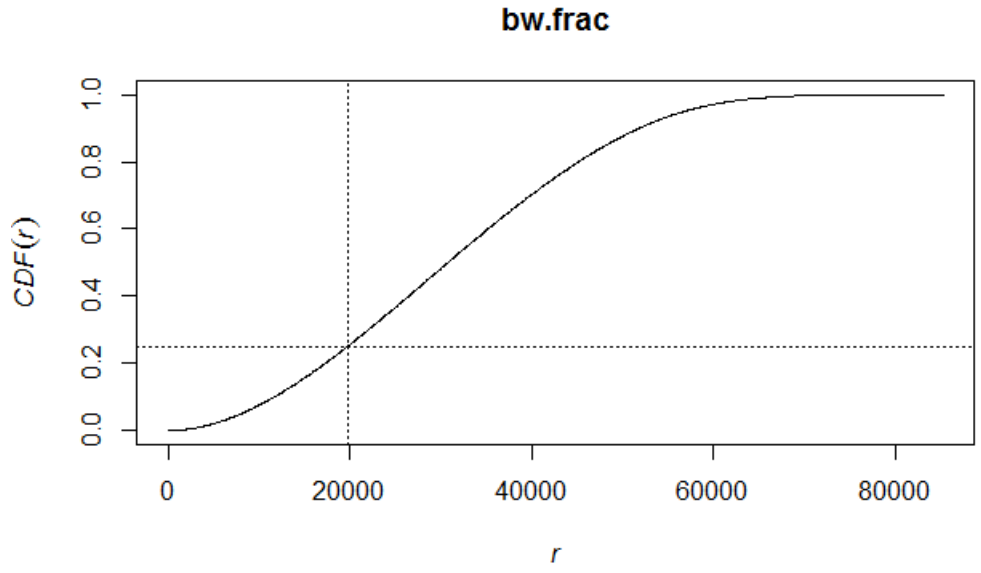


FIGURE 8.14 – La fonction de distribution cumulée CDF(r) obtenue par la fonction `bw.frac` du package *spatstat* de R
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion

Avec la fonction `bw.scott` de *spatstat*

Cette fonction est basée sur la «règle de Scott» (voir SCOTT 1992). Il s'agit de supposer que l'échantillon est distribué selon une loi normale. Dans ce cas, on obtient un estimateur de la bande passante, en minimisant une erreur appelée «erreur moyenne quadratique intégrée». Dans la formule de l'estimateur intervient notamment l'écart-type de l'échantillon.

Le résultat obtenu est un vecteur composé de deux valeurs : les bandes passantes proposées dans la direction des x et des y.

```
#on reutilise base.ppp créée plus haut

#test bw.scott de bande passante optimale
bw_scott <- spatstat::bw.scott(base.ppp)
```

On obtient :

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

Avec les paramètres par défaut, la valeur proposée est le couple (2974 ; 3455) : 2974 mètres dans la direction des x et 3455 dans la direction des y. D'après la documentation du package, la valeur proposée par ce test est généralement plus élevée que celle fournie par `bw.diggle`.

Synthèse des résultats obtenus

fonction	σ(en mètres)
bw.diggle	142
bw.ppl	286
bw.frac	19747
bw.scott	2974 (x) et 3455 (y)

TABLE 8.1 – Bandes passantes "optimales" ($h = 2\sigma$) obtenues par les fonctions du package *spatstat*
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion

Dans cet exemple, les résultats diffèrent parfois beaucoup selon les méthodes. Les écarts sont exacerbés par la répartition atypique de la population sur l'île de la Réunion, quasi-exclusivement située sur le littoral.

Conclusion

Derrière la qualité esthétique des cartes lissées se cache néanmoins un piège majeur. Par construction, les méthodes de lissage atténuent les ruptures et les frontières et induisent des représentations continues des phénomènes géographiques. Les cartes lissées font donc apparaître localement de l'autocorrélation spatiale. Deux points proches par rapport au rayon de lissage ont mécaniquement des caractéristiques comparables dans ce type d'analyse. De ce fait, commenter à partir d'une carte lissée des phénomènes géographiques dont l'ampleur spatiale est de l'ordre du rayon de lissage n'a guère de sens. Intuitivement, cela revient à commenter l'homogénéité observée au sein des unités spatiales d'une carte choroplèthe. Autrement dit, le rayon de lissage (la bande

passante) définit implicitement une maille minimale de restitution de l'information. En corollaire de ces remarques, il est primordial de commenter uniquement des phénomènes dont l'ordre de grandeur est très supérieur au rayon de lissage.

Références - Chapitre 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. (2012). « 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, p. 451–458.

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