9. Régression géographiquement pondérée

MARIE-PIERRE DE BELLEFON, JEAN-MICHEL FLOCH
Insee

Résumé

La régression géographiquement pondérée (RGP) répond au constat qu'un modèle de régression estimé sur l'ensemble d'un territoire d'intérêt peut ne pas appréhender de façon adéquate les variations locales. Son principe, assez simple, consiste en l'estimation de modèles locaux par les moindres carrés, chaque observation étant pondérée par une fonction décroissante de sa distance au point d'estimation. La réunion de ces modèles locaux permet la construction d'un modèle global aux propriétés spécifiques. La RGP permet, notamment à l'aide de représentations cartographiques associées, de repérer où les coefficients locaux s'écartent le plus des coefficients globaux, de construire des tests permettant d'apprécier si le phénomène est non stationnaire et de caractériser la non stationnarité. La méthode est présentée à partir de l'exemple d'un modèle des prix hédoniques (prix des logements anciens à Lyon). Nous montrons comment déterminer de façon optimale le rayon du disque sur lequel seront effectuées les régressions locales et présentons les résultats d'estimation, les méthodes d'estimation robustes et les tests de non-stationnarité des coefficients. En complément de cette utilisation descriptive, nous présentons une approche plus prédictive, montrant comment la prise en compte de la non-stationnarité permet d'améliorer un estimateur sur un domaine spatial. L'exemple est construit à partir d'un modèle liant la population pauvre et le nombre de bénéficiaires de la couverture maladie universelle complémentaire (CMU-C) à Rennes.

R La lecture préalable du chapitre 3 : "Indices d'autocorrélation spatiale" est recommandée.

9.1 Pourquoi utiliser une régression géographiquement pondérée ?

Pour identifier la nature des relations entre les variables, la régression linéaire modélise la variable dépendante y comme une fonction linéaire des variables explicatives $x_1, ..., x_p$. Si l'on dispose de n observations, le modèle s'écrit :

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \tag{9.1}$$

avec β_0, β_1, ..., β_p les paramètres et ε_1, ε_2,..., ε_n les termes d'erreur. Dans ce modèle, les coefficients β_k sont considérés comme identiques sur toute la zone d'étude. Cependant, cette hypothèse d'uniformité spatiale de l'effet des variables explicatives sur la variable dépendante est souvent irréaliste (BRUNSDON et al. 1996). Si les paramètres varient significativement dans l'espace, un estimateur global occultera la richesse géographique du phénomène étudié.

L'hétérogénéité spatiale correspond à cette variabilité spatiale des paramètres ou de la forme fonctionnelle du modèle. Lorsque l'on dispose d'une bonne connaissance du territoire d'intérêt, elle est souvent traitée dans la littérature empirique en ajoutant des indicatrices de zones géographiques dans le modèle (éventuellement croisées avec chaque variable explicative), et en estimant le modèle pour différentes zones ou en conduisant des tests de stabilité géographique des paramètres (dits de Chow). Lorsque le nombre de ces zones géographiques augmente, ce traitement diminue néanmoins le nombre de degrés de liberté et donc la précision des estimateurs.

On peut également utiliser des régressions locales dont l'application spatiale est la régression géographique pondérée (GWR, Geographically Weighted Regression, BRUNSDON et al. 1996). À travers l'exemple de l'étude des prix de l'immobilier à Lyon, nous présentons l'intérêt d'effectuer une régression géographiquement pondérée (exemple 9.1) et la façon dont elle peut être mise en œuvre (exemple 9.2).

Des méthodes plus complexes venues du monde de la géographie ont été développées (LE GALLO 2004), mais elles restent en grande partie descriptives et exploratoires (notamment à travers des représentations graphiques), car leur comportement théorique n'est pas complètement connu, notamment en ce qui concerne la convergence et la prise en compte des ruptures géographiques.

■ **Exemple 9.1 — Utilisation d'un modèle hédonique pour étudier les prix de l'immobilier lyonnais.** Cartographier les variations des prix de l'immobilier permet de déduire de façon générale que les prix ont tendance à être plus élevés dans le centre qu'en périphérie (figure 9.1). Cependant, ces prix élevés s'expliquent peut-être par une meilleure qualité des logements vendus dans le centre. Le modèle hédonique a pour objectif d'**isoler l'effet de la localisation sur les prix.** Le principe de cette méthode est que le prix d'un bien est une combinaison des prix de ses différents attributs.

$$y_i = \beta_0 + \sum_{k}^{p} \beta_k x_{ik} + \varepsilon_i \tag{9.2}$$

avec x_{ik} la caractéristique k du bien i, β_k le coefficient associé à cette caractéristique et p le nombre de variables explicatives.

Les hypothèses sous-jacentes au modèle hédonique sont que les vendeurs et les acheteurs sont des agents individuels, sans pouvoir de marché, et qu'il s'agit d'une situation de concurrence parfaite. Le coefficient de la régression hédonique associé à une caractéristique informe sur la valeur que les acheteurs **à l'équilibre à un instant donné** accorderaient à **une augmentation de la quantité de cette caractéristique.**

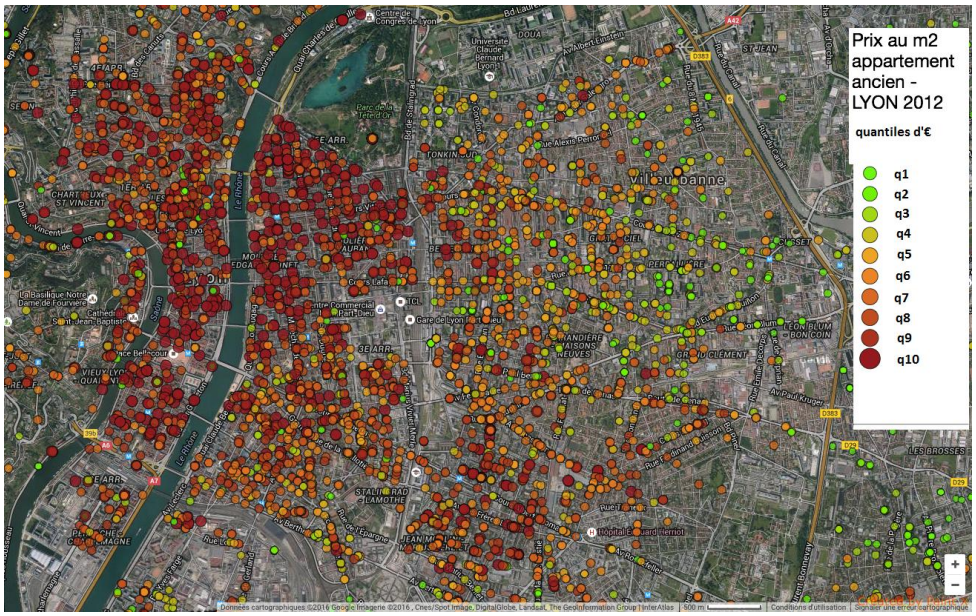

FIGURE 9.1 – Prix de vente au m² d'un appartement ancien - 2012
Source : *base PERVAL*
Champ : agglomération lyonnaise

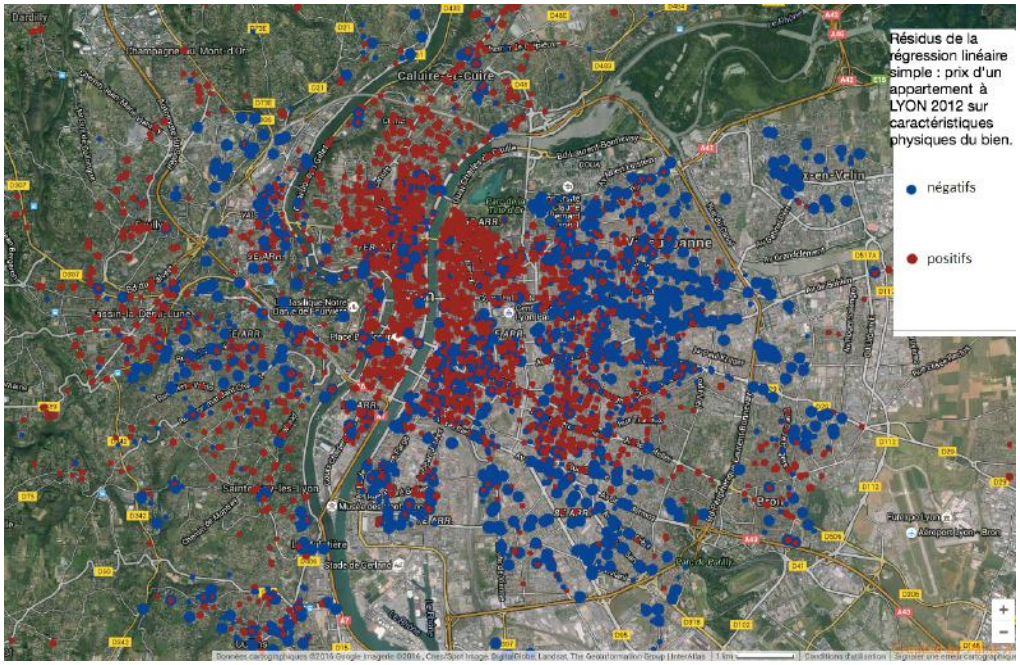

FIGURE 9.2 – Résidus de la régression hédonique du prix sur les caractéristiques du bien
Source : *base PERVAL*
Champ : agglomération lyonnaise

La figure 9.2 représente les résidus de la régression hédonique des prix des appartements sur leurs caractéristiques physiques. Ces résidus ne sont pas distribués aléatoirement dans l'espace (l'hypothèse nulle du test de Moran est rejetée). Le I de Moran de la distribution des résidus est positif, signe d'une corrélation spatiale positive des résidus. **L'hypothèse de stationnarité spatiale de la relation entre prix et caractéristique du bien n'est donc pas valide**, ce qui indique l'existence d'un phénomène d'**hétérogénéité spatiale**.
Comme cela a été vu précédemment, pour tenir compte de la variation des paramètres du modèle avec la localisation, une méthode couramment utilisée consiste à introduire des indicatrices de localisation comme paramètres explicatifs. Considérons à titre d'exemple les prix au m² des appartements à Lyon en 2012 et étudions l'impact sur ces prix de la période de construction de ces appartements. Pour cela nous introduisons la variable indicatrice qui indique pour chaque appartement si celui-ci a été construit entre 1992 et 2000 (indicatrice = 1) plutôt qu'entre 1948 et 1969 (indicatrice = 0). La table 9.1 indique la valeur des coefficients associés à cette indicatrice pour chaque arrondissement lyonnais.

Arrondissement	Coefficient de la régression	Significativité
1er	1,1511	.
2ème	1,1499	.
3ème	**1,1481**	***
4ème	1,1860	**
5ème	1,4909	***
6ème	**1,3085**	***
7ème	1,1897	***
8ème	1,1487	***
9ème	1,1981	***

TABLE 9.1 – Significativité des coefficients de régression associés à une époque de construction récente
*, **, *** indiquent la significativité aux seuils de 10, 5 et 1% **Source :** *base PERVAL*
Champ : agglomération lyonnaise

La valeur et la significativité des coefficients varient en fonction des arrondissements (table 9.1). Les acheteurs valorisent donc différemment l'époque de construction d'un logement suivant sa localisation. Cependant, pourquoi les frontières qui définissent les changements de modèle correspondraient-elles nécessairement à des contours administratifs ? **La régression géographiquement pondérée permet de répondre à cette question et d'étudier un phénomène qui varie continûment dans l'espace.**

■

9.2 La régression géographiquement pondérée

9.2.1 Un modèle à coefficients variables

La régression géographiquement pondérée appartient à la catégorie des modèles à coefficients variables. Les coefficients de la régression ne sont pas fixes : ils dépendent des coordonnées géographiques des observations. Dit autrement : **les coefficients des paramètres explicatifs forment des surfaces continues qu'on estime en certains points de l'espace**

$$y_i = \beta_0(u_i, v_i) + \sum_k^p \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{9.3}$$

avec (u_i,v_i) les coordonnées géographiques.

9.2.2 Comment estimer le modèle ?

Pour estimer le modèle, on utilise l'hypothèse suivante : **plus deux observations sont proches dans l'espace, plus l'influence des variables explicatives sur la variable dépendante est proche ; c'est à dire, plus les coefficients des paramètres explicatifs de la régression sont proches.** Par conséquent, pour estimer le modèle à coefficients variables au point i, on souhaite utiliser le modèle à coefficients fixes en incluant dans la régression uniquement les observations proches de i. Or plus on inclut de points dans l'échantillon, plus la variance est faible, mais plus le biais est élevé. La solution consiste donc à **diminuer l'importance des observations les plus éloignées en accordant à chaque observation un poids décroissant avec la distance au point d'intérêt.**

Le modèle à estimer est le suivant :

$$\mathbf{Y} = (\beta \otimes \mathbf{X})\mathbf{1} + \varepsilon \tag{9.4}$$

$\mathbf{Y}$: vecteur $n \times 1$ de la variable dépendante.
$\mathbf{X}$: matrice $n \times (p+1)$ des p variables explicatives + la constante.
$\mathbf{1}$: vecteur $(p+1) \times 1$ de 1

Les coefficients β du modèle peuvent être exprimés sous forme matricielle :

$$\beta = \begin{bmatrix} \beta_0(u_1,v_1) & \dots & \beta_p(u_1,v_1) \\ \beta_0(u_j,v_j) & \dots & \beta_p(u_j,v_j) \\ \beta_0(u_n,v_n) & \dots & \beta_p(u_n,v_n) \end{bmatrix} \tag{9.5}$$

L'opérateur $\otimes$ multiplie chaque élément de la matrice des coefficients β par l'élément correspondant de la matrice X des caractéristiques des observations.

Pour accorder un poids décroissant aux observations en fonction de leur distance au point d'intérêt, on effectue une estimation par moindres carrés pondérés, la pondération étant régie par la matrice de poids $W_{(u_i,v_i)}$. Les paramètres régissant la construction de cette matrice sont détaillés dans la section 9.2.3.

Conformément au principe des moindres carrés pondérés, les coefficients $\hat{\beta}(u_i,v_i)$ au point de coordonnées géographiques (u_i,v_i) minimisent la somme 9.6 :

$$\sum_{j=1}^{n} w_j(i)(y_j - \beta_0(u_i,v_i) - \beta_1(u_i,v_i)x_{j1} - \dots - \beta_p(u_i,v_i)x_{jp})^2 \tag{9.6}$$

$$\hat{\beta}(u_i,v_i) = (\mathbf{X^T W_{(u_i,v_i)} X})^{-1}\mathbf{X^T W_{(u_i,v_i)} Y} \tag{9.7}$$

On peut écrire $\hat{\mathbf{Y}} = \mathbf{SY}$ avec S dénommée la "matrice chapeau" et définie par l'équation 9.8 : En notant $\mathbf{x_i^T} = (1 \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{ip})$ la $i$ème colonne de $\mathbf{X}$, la matrice des variables explicatives, on a alors

$$\mathbf{S} = \begin{bmatrix} (\mathbf{x_1^T X^T W_{(u_1,v_1)} X})^{-1}\mathbf{X^T W_{(u_1,v_1)}} \\ \vdots \\ \mathbf{x_n^T X^T W_{(u_n,v_n)} X})^{-1}\mathbf{X^T W_{(u_n,v_n)}} \end{bmatrix} \tag{9.8}$$

Rappel : estimation par Moindres Carrés Ordinaires

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{9.9}$$

$$\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y} \tag{9.10}$$

Y : vecteur nx1 de la variable dépendante.
X : matrice nx(p+1) des p variables explicatives + la constante.

9.2.3 Choisir les paramètres d'estimation

La matrice $W_{(u_i,v_i)}$ contient le poids de chaque observation en fonction de sa distance au point i de coordonnées (u_i, v_i) (figure 9.3). On suppose que les observations proches du point i exercent plus d'influence sur les paramètres estimés au lieu i que les observations plus lointaines. Le poids des observations est donc décroissant avec la distance au point i. Il existe plusieurs manières de spécifier cette décroissance. Nous présentons ici les principaux paramètres de décroissance.

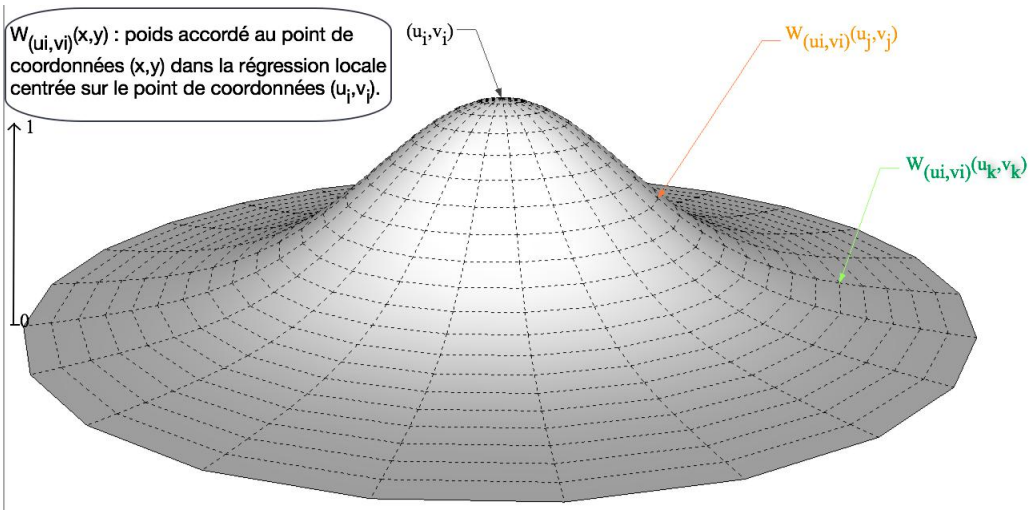

FIGURE 9.3 – Représentation graphique de la matrice W

La décroissance du poids de chaque observation avec la distance au point d'origine est déterminée par une **fonction de noyau**. Les paramètres clés de la fonction de noyau sont :

— la forme du noyau ;
— le noyau fixe ou adaptatif ;
— la taille de la bande passante.

La forme du noyau

On peut distinguer les noyaux continus qui accordent un poids à toutes les observations (figure 9.4 ; table 9.2) des noyaux à support compact (figure 9.5 ; table 9.3) pour lesquels le poids des observations est nul au delà d'une certaine distance. Cependant, **la forme du noyau ne modifie que légèrement les résultats** (BRUNSDON et al. 1998).

— Choisir un noyau uniforme revient à effectuer une régression par moindres carrés ordinaires en chaque point.

Noyau uniforme	$w(d_{ij}) = 1$
Noyau gaussien	$w(d_{ij}) = exp(-\frac{1}{2}(\frac{d_{ij}}{h})^2)$
Noyau exponentiel	$w(d_{ij}) = exp(-\frac{1}{2}(\frac{\lvert d_{ij}\rvert}{h}))$

TABLE 9.2 – Expression fonctionnelle de noyaux continus

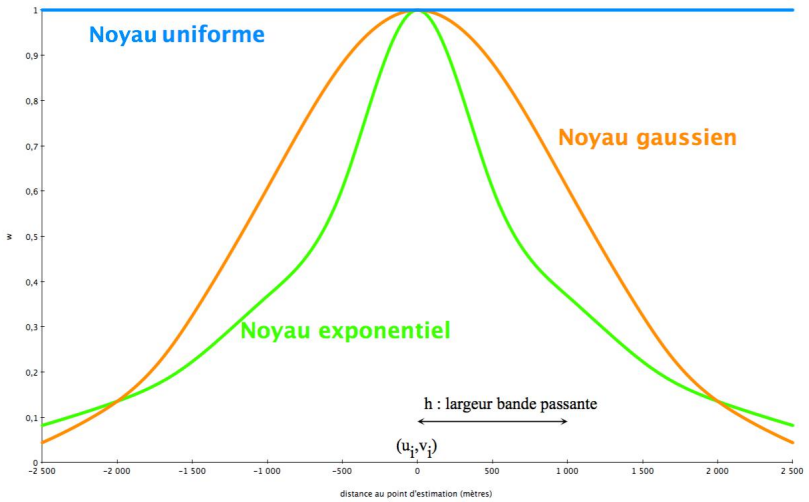

FIGURE 9.4 – Représentation graphique de noyaux continus

Noyau Box-Car	$w(d_{ij}) = 1$ si $\lvert d_{ij}\rvert < h$, 0 sinon
Noyau Bi-square	$w(d_{ij}) = (1-(\frac{d_{ij}}{h})^2)^2$ si $\lvert d_{ij}\rvert < h$, 0 sinon
Noyau Tri-cube	$w(d_{ij}) = (1-(\frac{\lvert d_{ij}\rvert}{h})^3)^3$ si $\lvert d_{ij}\rvert < h$, 0 sinon

TABLE 9.3 – Expression fonctionnelle de noyaux à support compact

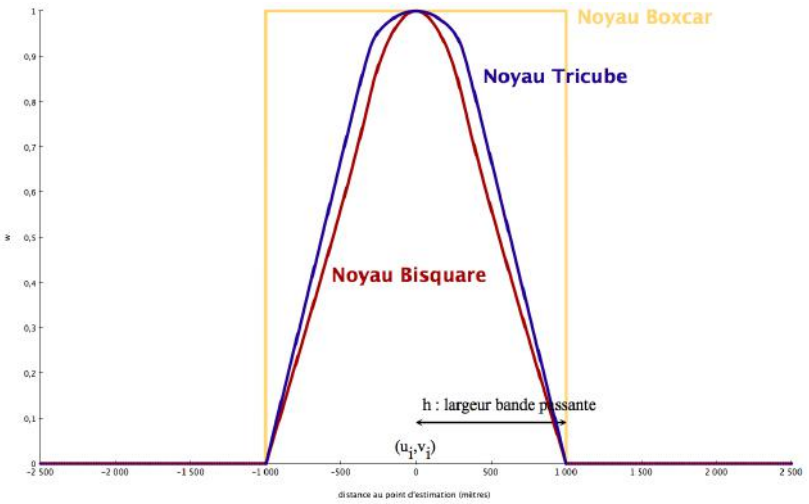

FIGURE 9.5 – Représentation graphique de noyaux à support compact

— Le noyau Box-Car traite un phénomène continu de façon discontinue.
— Les noyaux gaussiens et exponentiels pondèrent toutes les observations, avec un poids qui tend vers zéro avec la distance au point estimé.
— Les noyaux Bisquare et Tricube accordent également aux observations un poids décroissant avec la distance, mais ce poids est nul au delà d'une certaine distance h appelée *bande passante*.

⇒ **Le noyau Bisquare est à privilégier pour optimiser le temps de calcul.**

Noyau fixe ou adaptatif

Définition 9.2.1 — Noyau fixe. L'étendue du noyau est déterminée par la **distance** au point d'intérêt. Le noyau est identique en tout point de l'espace (figure 9.6).

Définition 9.2.2 — Noyau adaptatif. L'étendue du noyau est déterminée par le **nombre de voisins** du point d'intérêt. Plus la densité des observations est faible, moins le noyau est étendu (figure 9.7).

— Un noyau fixe est adapté à une répartition des données uniforme dans l'espace mais peu efficace dans le cas d'une répartition inhomogène. Son rayon doit être au moins égal à la distance entre le point le plus isolé et son premier voisin ce qui peut conduire à un nombre variable de points inclus dans la régression.
— Dans les zones peu denses, un noyau fixe trop petit inclura trop peu de points dans la régression. La variance sera plus élevée.
— Dans les zones très denses, un noyau fixe trop grand négligera les variations à une échelle fine. Le biais sera plus élevé.

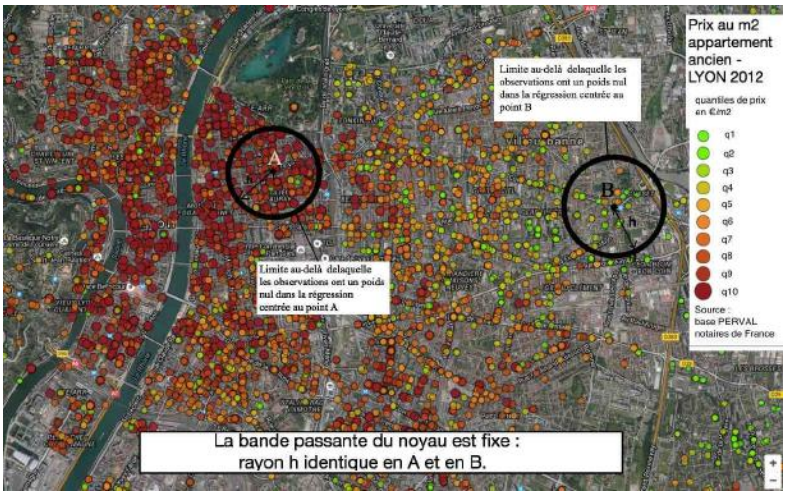

FIGURE 9.6 – Noyau fixe
Source : *base PERVAL*

Définition et choix de la bande passante

La bande passante est une distance au-delà de laquelle le poids des observations est considéré comme nul. **La valeur de la bande passante h est le paramètre dont le choix a la plus forte influence sur les résultats.** Plus la valeur de la bande passante est élevée, plus le nombre d'observations auxquelles le noyau accorde un poids non nul est élevé. La régression locale inclura alors davantage d'observations et les résultats seront plus lissés qu'avec une faible bande passante.

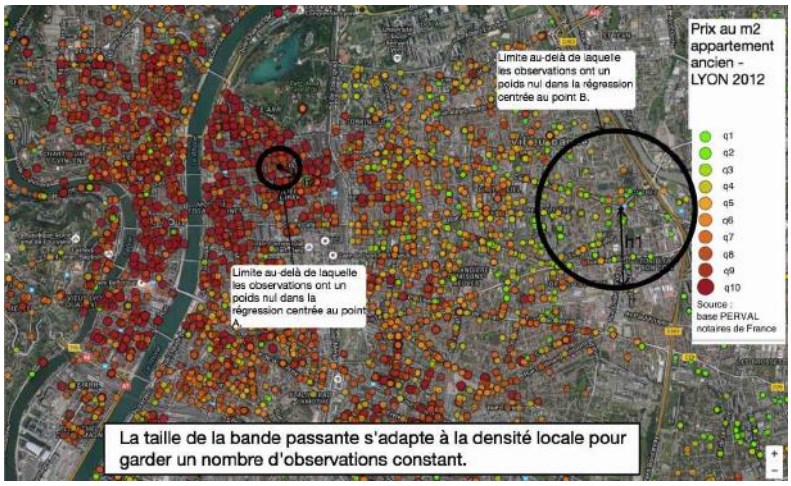

FIGURE 9.7 – Noyau adaptatif
Source : *base PERVAL*

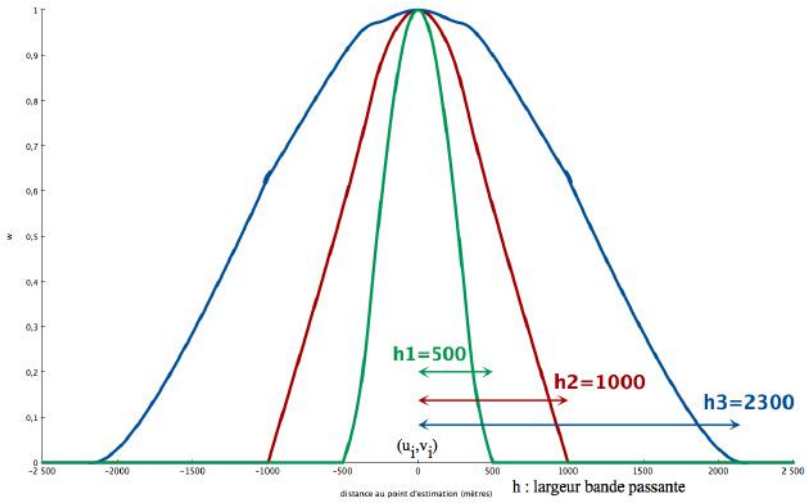

FIGURE 9.8 – Influence du choix de la bande passante sur le noyau

Lorsque la bande passante tend vers l'infini, les résultats de la régression locale s'approchent de ceux d'une régression par moindres carrés ordinaires.

Le choix de la bande passante n'est pas lié au modèle lui même, mais à la stratégie de calibration. Si le noyau inclut des points trop éloignés, la variance sera faible mais le biais élevé. Si le noyau inclut uniquement les points trop proches, le biais sera faible mais la variance élevée. Plusieurs critères statistiques peuvent aider à choisir la bande passante la plus adaptée. Le package R *GW.model* permet d'obtenir la bande passante minimisant l'un ou l'autre des deux critères : le critère de validation croisée et le critère d'Akaike corrigé (voir encadrés 9.2.1 et 9.2.1).

La valeur de la bande passante minimisant ces critères est aussi une indication précieuse quant à la pertinence d'une modélisation par une régression géographiquement pondérée. Si la bande passante tend vers le maximum possible (toute la taille de la zone d'étude, ou tous les points), alors l'hétérogénéité locale n'est probablement pas significative et la RGP n'est pas nécessaire. Inversement, une bande passante extrêmement faible doit alerter sur le risque que le processus sous-jacent soit aléatoire (GOLLINI et al. 2013). Il faut également garder en tête que la bande passante qui minimise les critères statistiques s'appuie sur la prédiction de la variable dépendante, et sur celle des coefficients de la régression (qui sont pourtant ceux utilisés ensuite pour tester la validité de l'hypothèse de non-stationarité).

Encadré 9.2.1 — Critère de validation croisée.

$$CV = \sum_{i=1}^{n} \left[y_i - \hat{y}_{\neq i}(h) \right]^2$$

$\hat{y}_{\neq i}(h)$ est la valeur de y au point i prédite lorsqu'on calibre le modèle avec toutes les observations sauf y_i. En effet, si on estimait le modèle avec l'intégralité des observations, la bande passante optimale serait 0 puisque, lorsque $h = 0$, on n'inclut aucun autre point que y_i dans la régression ; on a donc $\hat{y}_i = y_i$ ce qui est l'optimum atteignable.
La bande passante h qui minimise le score de validation croisée CV **maximise le pouvoir prédictif du modèle**.

Encadré 9.2.2 — Critère d'Akaike corrigé.

$$AIC_c(h) = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left\{ \frac{n + tr(S)}{n - 2 - tr(S)} \right\}$$

n est la taille de l'échantillon ; $\hat{\sigma}$ est l'estimation de la déviation standard du terme d'erreur ; $tr(S)$ est la trace de la matrice de projection (matrice chapeau) de la variable observée y sur la variable estimée $\hat{y}$.
Le critère AIC favorise un **compromis entre le pouvoir prédictif du modèle et sa complexité.** Plus la bande passante est faible, plus le modèle global est complexe. Le critère AIC favorise généralement des bandes passantes plus larges que le critère CV.

9.3 Régression géographiquement robuste

Tout comme la régression linéaire classique, la régression géographiquement pondérée est sensible aux points aberrants. Ces points distordent la surface des paramètres estimés. Puisque la régression géographiquement pondérée estime un modèle différent en chaque point de l'espace, il suffit qu'un point soit aberrant **par rapport au contexte local** pour que l'estimation soit faussée. Or il y a plus de chances pour qu'un point soit aberrant par rapport au contexte local que par rapport

au contexte global. Rechercher les points aberrants au niveau global risque donc de laisser passer des points qui sont aberrants localement mais pas globalement. Deux méthodes ont été développées pour pallier ce problème.

Méthode 1 : filtrer en fonction des résidus standardisés

L'objectif de la méthode 1 est de détecter les observations dont les résidus sont très élevés et de les exclure de la régression.

Soit $e_i = y_i - \hat{y}_i$ le résidu de l'estimation au point i. Si y_i est un point aberrant, e_i devrait avoir une valeur très élevée. Cependant, les résidus n'ont pas tous la même variance. Il faut donc les standardiser afin de pouvoir les comparer et juger de ceux qu'il est nécessaire d'éliminer de la régression.

Notons $\hat{y} = \mathbf{Sy}$ où S est la matrice chapeau définie plus haut. On a $\mathbf{e} = \mathbf{y} - \mathbf{Sy} = (\mathbf{I} - \mathbf{S})\mathbf{y}$ avec $\mathbf{e}$ le vecteur des résidus et $var(\mathbf{e}) = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{\mathbf{T}} var(\mathbf{y}) = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{\mathbf{T}}\sigma^2$ avec σ la déviation standard de y. Les variances des e_i sont donc les éléments de la diagonale de la matrice $(\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{\mathbf{T}}\sigma^2$, *a priori* différents.

Soit $\mathbf{Q} = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{\mathbf{T}}$ et q_{ii} le $i$ème élément de la diagonale de $\mathbf{Q}$.
$r_i = \frac{e_i}{\hat{\sigma}\sqrt{q_{ii}}}$ est appelé *résidu standardisé intérieur*.
Si le point i est aberrant, l'inclure dans l'estimation de $\hat{\sigma}^2$ risque de produire un biais. On estime donc la valeur de σ en excluant l'observation i potentiellement aberrante : σ_{-i}
$r_i^* = \frac{e_i}{\hat{\sigma_{-i}}\sqrt{q_{ii}}}$ est appelé *résidu standardisé extérieur*.
Avec la méthode 1, les observations pour lesquelles $|\mathbf{r_i^*}| > 3$ sont filtrées (le seuil de 3 est proposé par CHATFIELD 2006).

Inconvénient de la méthode : $\mathbf{Q}$ est une matrice $n * n$ dont **le temps de calcul est à ce jour rédhibitoire pour de grosses bases de données**, avec une machine dotée d'une puissance de calcul usuelle. Par exemple, BRUNSDON et al. 1996 jugent qu'au-delà de 10 000 observations l'emploi de cette méthode n'est pas envisageable.

Méthode 2 : diminuer les poids des observations aux résidus élevés

L'objectif de la méthode 2 est de diminuer le poids des observations ayant des résidus élevés (HUBER 1981). Après une première estimation du modèle, on accorde un poids $w_r(e_i)$ supplémentaire à chaque observation i. Ce poids doit être multiplié avec le poids qui varie en fonction la distance au point i. On a donc une nouvelle matrice W qui est le produit terme à terme entre l'ancienne matrice W et une matrice W_r des poids des résidus, ainsi définis :

$$w_r(e_i) = \left\{ \begin{array}{cc} 1 & si \quad |e_i| \leq 2\hat{\sigma} \\ \left[1 - (|e_i| - 2)^2\right]^2 & si \quad 2\hat{\sigma} < |e_i| < 3\hat{\sigma} \\ 0 & sinon \end{array} \right\} \tag{9.11}$$

Si aucun des résidus de la première régression n'est plus élevé que deux déviations standard, le deuxième modèle est identique au premier. Les observations dont les résidus sont compris entre deux et trois déviations standard voient leur poids diminué dans la deuxième régression, tandis que les observations dont les résidus sont supérieurs à trois déviations standard sont carrément exclues du modèle.

Discussion

La méthode 2 est beaucoup plus rapide à calculer que la méthode 1 puisque chaque cycle demande seulement le calcul des n résidus et non celui d'une matrice $n \times n$. Cependant, elle ne prend pas en compte les différences de variance entre les résidus et élimine davantage de points que la méthode 1.

Application avec R

Le package *GWmodel* permet de mettre en œuvre la régression géographiquement pondérée. La première étape consiste à calculer les distances entre toutes les observations grâce à la fonction `gw.dist`. Ensuite, la fonction `bw.gwr` permet de calculer de manière optimale, au sens d'un critère statistique donné, la bande passante de la fonction de noyau, . Enfin, les coefficients locaux de la régression géographiquement pondérée sont obtenus grâce à la fonction `gwr.robust`. Les résultats sont contenus dans un objet de classe `gwrm`, contenant en particulier un objet de type `SpatialPointsDataFrame`, dont le contenu est détaillé ci-après.

Options de la fonction `gw.dist`

— `dp.locat` : coordonnées des observations ;
— `rp.locat` : coordonnées des points sur lesquels sera calibré le modèle (par exemple : les points d'une grille régulière) ;
— `p` : régit le choix de la distance (p=1 : Manhattan ; p=2 : euclidienne) ;
— `theta` : angle avec lequel on effectue une rotation du système de coordonnées (utile pour distance Manhattan).

Options de la fonction `bw.gwr`

— `formula` : le modèle $y \sim x_1 + x_2 + ... + x_p$;
— `approach` : méthode de calcul de la bande passante optimale : CV (Validation Croisée) ou AIC (Critère d'Information d'Akaike) ;
— `kernel` : type de noyau : "gaussian", "exponential", "bisquare", "tricube", "boxcar" ;
— `adaptive` : si TRUE, la bande passante est un nombre de voisins, le noyau est adaptatif et si FALSE, la bande passante est une distance, le noyau est fixe ;
— `dMat` : la matrice de distances pré-calculée.

Options de la fonction `gwr.robust`

— `regression points` : les coordonnées géographiques des points à partir desquels le modèle sera évalué ;
— `bw` : la taille de la bande passante ;
— `filtered` : si TRUE, filtre les observations en fonction de la valeur des résidus standardisés (Méthode 1 de régression robuste). Si FALSE, estime le modèle une deuxième fois en pondérant les observations en fonction de la valeur de leurs résidus (Méthode 2 de régression robuste) ;
— `F123.test` : calcule la statistique de Fischer (défaut FALSE) ;
— `maxiter` : nombre maximum d'itérations de l'approche automatique (Méthode 2) : vaut 20 par défaut ;
— `cut1` : $\sigma cut1$ est le seuil de valeur des résidus au-delà duquel les observations ont un poids < 1 (vaut 2 par défaut) ;
— `cut2` : $\sigma cut2$ est le seuil de valeur des résidus au-delà duquel les observations ont un poids nul (vaut 3 par défaut) ;
— `delta` : seuil de tolérance de l'algorithme itératif (vaut 10^{-5} par défaut).

Interprétation des résultats : le contenu du fichier $SDF

— Le fichier $SDF est de nature "SpatialPointsDataFrame", il contient des attributs associés à des coordonnées géographiques.

— *c_x* : estimation du coefficient associé à la caractéristique x en chaque point.
— *yhat* : valeur de y prédite.
— *residual*, *Stud_residual* : résidu et résidu standardisé
— *CV_score* : score de validation croisée
— *x_SE* : erreur standard de l'estimation du coefficient devant la caractéristique x.
— *x_TV* : t-value de l'estimation du coefficient devant la caractéristique x.
— *E_weight* : poids des observations dans la régression robuste (à multiplier au poids lié à la fonction de noyau).

■ **Exemple 9.2 — Application à l'étude des prix de l'immobilier lyonnais.** La régression géographiquement pondérée permet d'étudier l'influence de la localisation d'un bien immobilier sur son prix, tout en prenant en compte l'hétérogénéité spatiale c'est-à-dire le fait que l'influence des caractéristiques d'un bien immobilier sur son prix dépende de sa localisation. Le coefficient associé à la constante de la régression géographiquement pondérée est le prix d'un appartement de référence : le prix d'un appartement, une fois prise en compte l'influence de ses caractéristiques physiques.

Signification des variables de l'exemple ci-dessous :
f_lgpx : logarithme du prix au mètre carré.
c_epoqueA : indicatrice de construction avant 1850
c_epoqueF : indicatrice de construction entre 1981 et 1991
c_epoqueG : indicatrice de construction entre 1992 et 2000
c_mmut1, 2, 3 : indicatrice d'une mutation ayant eu lieu au mois de janvier, février, mars, etc.
c_sdbn_2 : indicatrice de l'existence de deux salles de bain
c_cave1 : indicatrice de l'existence d'une cave

```
library(GWmodel)
dm.calib <- gw.dist(dp.locat=coordinates(lyon2012))

#Calcule une matrice de distances entre les points
bw0 <- bw.gwr(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+
                c_mmut_3+c_epoqueA+c_epoqueF+c_sdbn_2+c_cave1,
              data=lyon2012, approach="AIC", kernel="bisquare",
              adaptive=TRUE,dMat=dm.calib)

gwr.robust.lyon2012 <- gwr.robust(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+
                c_mmut_3+c_epoqueA+c_epoqueF+c_sdbn_2+c_cave1,
                bw=bw0, kernel="bisquare", filtered=FALSE, adaptive=TRUE,
                                                dMat=dm.calib)

#Extraction de la constante : prix du bien de référence (figure 9.9)

lyon2012.intercept.robust <- gwr.robust.lyon2012$SDF[,c(1)]
# 1 correspond à la position de la constante dans le fichier contenant les
    résultats de la régression.
lyon2012.intercept.robust$Intercept <- exp(lyon2012.intercept.robust$
    Intercept)

#Extraction du coefficient lié au fait d'avoir été construit avant 1850
    plutot qu'entre 1948 et 1969 (époque de référence) - figure 9.10
lyon2012.epoqueA.robust <- gwr.robust.lyon2012$SDF[,c(15)]
```

```
lyon2012.epoqueA.robust$c_epoqueA <- exp(lyon2012.intercept.robust$c_
    epoqueA)

#Estimation du modèle (non robuste) sur un carroyage de 100 mètres de coté
    (figure 9.11)
#Soit "quadrillage" un fichier de type SpatialGridDataFrame recouvrant la
    zone à étudier
dm.calib.quadrillage<- coordinates(quadrillage) gw.dist(dp.locat=
    coordinates(lyon2012),rp.locat=coordinates(quadrillage))
gwr.lyon2012<-gwr.basic(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+c_mmut_3+c_
    epoqueA+c_epoqueF+c_sdbn_2+c_cave1,regression.point=quadrillage,bw=bw0,
     kernel="bisquare", filtered=FALSE, adaptive=TRUE,    dMat=dm.calib.
    quadrillage)
```

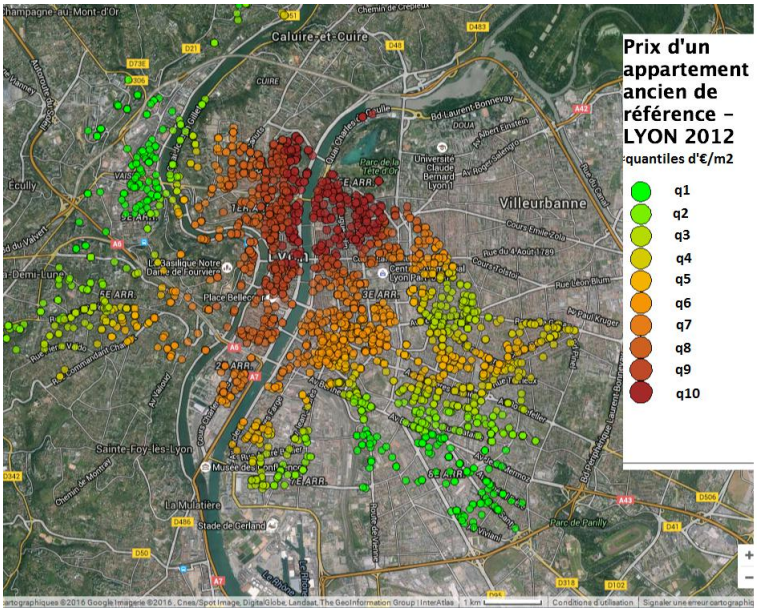

FIGURE 9.9 – Constante locale : prix du bien de référence
Source : *base PERVAL*

Coefficient	Min	1er quartile	Médiane	Moyenne	3ème quartile	Max
constante	1666	2220	2668	2705	3088	4030
époque A	0.6250	0.9533	1.1480	1.1070	1.2470	1.8190

TABLE 9.4 – Statistiques descriptives des coefficients de la RGP des prix immobiliers sur leurs caractéristiques
Source : *base PERVAL*

Les coefficients de la régression hédonique varient dans l'espace (Table 9.4). La régression géographiquement pondérée a permis de mieux appréhender la richesse spatiale de l'évolution des paramètres explicatifs des prix immobiliers puisque les estimations sont indépendantes de la frontière administrative des arrondissements. Sur les Figures 9.9 et 9.10, les points où les coefficients ont été estimés sont ceux où des transactions avaient eu lieu. Cependant, un des intérêts de la RGP est aussi de pouvoir estimer les valeurs des coefficients de façon continue. La Figure 9.11

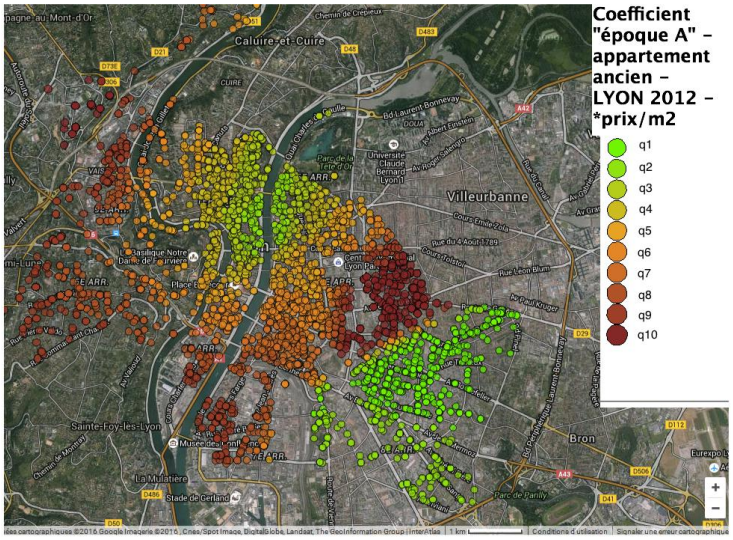

FIGURE 9.10 – Coefficient associé au fait d'avoir été construit avant 1850 plutôt qu'entre 1948 et 1969 (époque de référence)
Source : *base PERVAL*

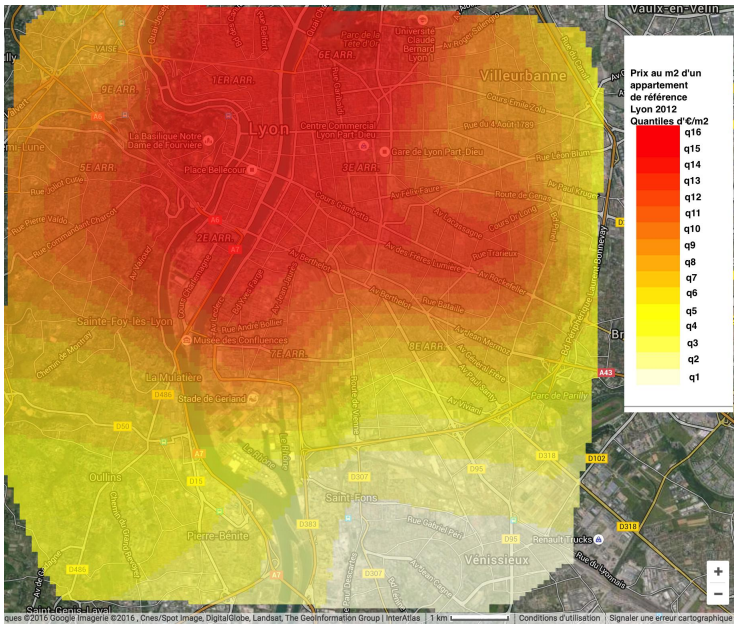

FIGURE 9.11 – Estimation des prix immobiliers sur un carroyage de 100 mètres de côté
Source : *base PERVAL*

présente une estimation des paramètres sur un quadrillage de 100 mètres de côté. La Section 9.4 présente une méthode permettant d'évaluer dans quelle mesure la variation spatiale des paramètres est significative. ■

9.4 Qualité des estimations

9.4.1 Précision de l'estimation des coefficients

Quand on estime une RGP avec un noyau adaptatif dans une zone où les observations sont peu denses, les points servant à calibrer le modèle peuvent tous avoir un poids très faible (ils sont situés à une grande distance du point d'estimation).

Soit $\mathbf{C}$ la matrice telle que :

$$\hat{\beta}(u_i, v_i) = (\mathbf{X^T W_{(u_i,v_i)} X})^{-1} \mathbf{X^T W_{(u_i,v_i)} y} = \mathbf{CY} \tag{9.12}$$

La variance du paramètre estimé est :

$$Var\left[\hat{\beta}(u_i, v_i)\right] = \mathbf{CC^T}\sigma^2 \tag{9.13}$$

Avec σ^2 la somme des résidus normalisés de la régression locale :

$$\sigma^2 = \sum_i (y_i - \hat{y}_i)/(n - 2\nu_1 + \nu_2) \tag{9.14}$$

$$\nu_1 = tr(\mathbf{S}) \tag{9.15}$$

$$\nu_2 = tr(\mathbf{S^T S}) \tag{9.16}$$

$$\hat{Y} = \mathbf{SY} \tag{9.17}$$

Une fois que la variance de chaque paramètre a été estimée, les erreurs standard sont obtenues avec l'équation 9.18

$$SE(\hat{\beta}(u_i, v_i)) = \sqrt{Var}\left[\hat{\beta}(u_i, v_i)\right] \tag{9.18}$$

On peut ainsi calculer des intervalles de confiance pour les coefficients.

Application avec R

Le fichier $SDF contenant les résultats de la régression géographiquement pondérée permet d'accéder aux erreurs standard associées aux différents coefficients. Ainsi, dans le cas de l'exemple des prix de l'immobilier lyonnais développé précédemment :

- y : prix de vente ;
- `yhat` : prix de vente estimé ;
- `Intercept_SE` : erreur standard du coefficient associé à la constante ;
- `Intercept_TV` : taux de variation du coefficient associé à la constante.

9.4.2 Test de la non-stationnarité des coefficients

La RGP relâche l'hypothèse que les coefficients sont stationnaires dans une certaine zone géographique. Pour s'assurer de la pertinence du modèle, il est intéressant de tester la non-stationnarité des coefficients. **Les coefficients varient-ils suffisamment dans l'espace pour rejeter l'hypothèse qu'ils sont constants sur toute la surface de l'étude ?**

En termes statistiques, la question équivaut à :
— $H_0 : \forall k, \beta_k(u_1, v_1) = \beta_k(u_2, v_2) = ... = \beta_k(u_n, v_n)$
— $H_1 : \exists k$, tous les $\beta_k(u_i, v_i)$ ne soient pas égaux.
Pour répondre à cette question, on peut utiliser une méthode de simulation du type "simulation de Monte Carlo".

Principe : S'il n'y avait pas de phénomène spatial sous-jacent, on pourrait permuter les coordonnées géographiques des observations aléatoirement dans l'espace et la variance resterait inchangée. Lors d'une simulation de Monte Carlo, on permute n fois les coordonnées géographiques des observations. On obtient donc n estimations de la variance spatiale des coefficients. On peut ensuite estimer la $p - value$ de la variabilité spatiale des coefficients et rejeter - ou non - l'hypothèse nulle selon laquelle ils sont stables dans l'espace.
Rappelons néanmoins que les méthodes simulant une distribution par permutation spatiale des observations dépendent du jeu de données initial. LEUNG et al. 2000 décrivent une méthode plus robuste et moins gourmande en temps de calcul pour tester la non-stationnarité des coefficients.

Application avec R

Fonction `montecarlo.gwr`
— mêmes paramètres que la fonction `gwr.robust` ;
— `nsims` : nombre de simulations ;
— `sortie` : vecteur contenant les p-values de tous les paramètres de la RGP.

9.5 Une application prédictive

La régression géographiquement pondérée a surtout été utilisée pour mettre en évidence l'hétérogénéité spatiale. Comme les autres méthodes de régression, elle peut aussi être utilisée à des fins prédictives, par exemple pour imputer des valeurs à des unités non échantillonnées dans le cadre d'un sondage. Cette partie de l'article s'appuie sur un travail réalisé par E.Lesage et J-M. Floch pour les JMS de 2015 [1], et également présenté aux Journées des Méthodes Avancées pour l'Analyse de Sondages Complexes de 2016. Dans les méthodes d'estimation sur petits domaines, on utilise de plus en plus fréquemment une approche basée sur des modèles qui utilisent des estimateurs BLUP, Best linear unbiased Predictors (CHAMBERS et al. 2012). Les valeurs des unités non échantillonnées sont remplacées par les valeurs prédites à partir d'un modèle dont les paramètres sont estimés à l'aide des valeurs des unités échantillonnées. Une extension de ces méthodes a été proposée par CHANDRA et al. 2012 dans un cadre non stationnaire, en recourant à la régression géographiquement pondérée. L'utilisation de la régression géographiquement pondérée dans les méthodes d'estimation sur petits domaines semble être préférée dans la littérature récente aux méthodes issues de l'économétrie spatiale recourant notamment aux modèles spatiaux autoregressifs (SAR). La régression géographiquement pondérée offre une façon plus flexible de prendre en compte la variabilité spatiale des phénomènes. Cette prise en compte de l'hétérogénéite spatiale doit théoriquement améliorer la précision des estimateurs.

1. Journées de Méthodologie Statistique organisées par l'Insee tous les trois ans environ. Le diaporama de la présentation est accessible à l'adresse `https://maasc2016.sciencesconf.org/data/pages/7.DiapoFlochJeanMichel.pdf`

9.5.1 Présentation du problème

À l'Insee, des travaux empiriques ont utilisé la RGP pour construire des estimateurs à partir de données issues du recensement de la population concernant les quartiers prioritaires. Dans ces quartiers, 40 % des logements sont enquêtés (sur une période de cinq ans), mais le plan de sondage n'est pas optimal, l'appartenance à un quartier prioritaire ne faisant pas partie des variables d'équilibrage. La demande d'une information précise sur ces quartiers étant forte, on a cherché à mobiliser des sources administratives exhaustives ou quasi-exhaustives (Données fiscales, données de l'assurance maladie) pour améliorer la précision des estimateurs. Pour ce faire, on estimait sur les logements de l'échantillon du recensement de la population (RP) un modèle dans lequel la variable d'intérêt était une variable du recensement, les variables auxiliaires des variables tirées des sources administratives, bien corrélées à la variable d'intérêt. Les estimateurs permettaient de prédire une valeur sur les logements non échantillonnés. **L'estimateur du total de la variable d'intérêt correspond à la somme des valeurs observées pour les unités échantillonnées et des valeurs prédites pour les unités non échantillonnées, les poids de sondage n'intervenant plus dans ce calcul**.

Ces travaux empiriques reposaient sur une modélisation utilisant la régression géographiquement pondérée afin de tenir compte de la forte hétérogénéité constatée dans les données urbaines. Mais le gain de précision, par rapport à un modèle non spatial, n'avait pas été étudié. C'est pourquoi on propose une comparaison de trois estimateurs, à partir d'un dispositif expérimental reposant sur des données réelles administratives composées de la source Filosofi, qui permet de calculer la population des ménages à bas revenus, de la source CNAM (Caisse Nationale d'Assurance Maladie) qui fournit le nombre de bénéficiaires et de la CMUC (Couverture Maladie Universelle Complémentaire). La source Filosofi est quasi-exhaustive et permet de disposer des "vraies" valeurs du nombre d'individus ayant des revenus inférieurs au taux de pauvreté.

Les deux sources sont localisées et peuvent être théoriquement appariées à partir de leurs coordonnées géographiques. Pour des raisons de confidentialité, il n'a pas été possible de le faire, et on a calculé le nombre de personnes à bas revenus et le nombre de bénéficiaires de la CMUC sur un maillage formé de carreaux de 100 m de côté, compromis jugé acceptable avec l'utilisation de données individuelles. Ces carreaux de 100 m jouent le rôle d'individus statistiques sur lesquels on va effectuer les mesures.

Le territoire d'intérêt est la commune de Rennes. On tire dans la base de données un échantillon de 40 % des carreaux, comme ce qui est fait dans le recensement de la population. Ces carreaux vont servir de support à l'estimation du nombre de personnes à bas revenus (figure 9.12). On dispose de toutes les informations, mais pour le modèle, les bas revenus ne sont connus que pour les carreaux de l'échantillon, tandis que les bénéficiaires de la CMUC le sont pour l'ensemble des carreaux. On sélectionne un échantillon de taille $n = 856$, qu'on nomme s, par tirage aléatoire simple sans remise (moins complexe que le tirage opéré pour le RP). Le taux de sondage est $n/N = 40$ %. De plus, on note r le complémentaire de l'échantillon s dans U (l'ensemble des carreaux habités sur le territoire de Rennes). Les calculs effectués au niveau du carreau permettent des calculs sur la maille Iris, chaque carreau étant affecté à un Iris (l'Iris est le plus petit découpage administratif français).

Il existe une liaison linéaire forte entre le nombre de personnes à bas revenu et le nombre de bénéficiaires de la CMUC. Les ordonnées à l'origine varient peu d'un carreau à l'autre. La valeur des pentes varient sensiblement de 1.6 à 3.3. Le gradient des situations locales est représenté sur la figure 9.13. Dans l'approche dite "basée sur le modèle", on prédit les valeurs y_i des carreaux non échantillonnés grâce au modèle estimé à partir de l'ensemble des données de l'échantillon et de l'information auxiliaire x disponible pour les carreaux non échantillonnés. On construit trois estimateurs, j désignant l'Iris :

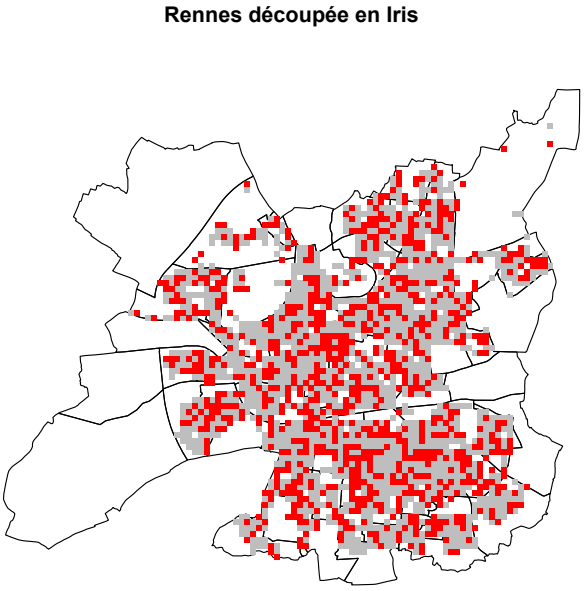

FIGURE 9.12 – Carreaux de 100 m habités à Rennes, échantillonnés (en rouge) ou non (en gris)

Définition 9.5.1 — L'estimateur de Horvitz-Thompson.

$$\widehat{t}_y(j) = \frac{N}{n} \sum_{i \in s_j} y_i \tag{9.19}$$

Définition 9.5.2 — L'estimateur basé sur la régression "classique", sans prise en compte de l'hétérogénéité spatiale.

$$\widehat{t}_{y,reg}(j) = \sum_{i \in s_j} y_i + \sum_{i \in r_l} \tilde{y}_l \tag{9.20}$$

où $\tilde{y}_l = \beta^T x_l$

Définition 9.5.3 — L'estimateur basé sur la régression géographique pondérée.

$$\widehat{t}_{y,RGP}(j) = \sum_{i \in s_j} y_i + \sum_{i \in r_l} \breve{y}_l \tag{9.21}$$

où $\breve{y}_l = \hat{\beta}_l^T x_l$ et $\widehat{\beta}_l$ est le vecteur des coefficients de la régression géographique pondérée pour le carreau l.

9.5.2 Résultats

On répète $K = 1000$ fois ce processus. On obtient pour chaque Iris 1 000 valeurs pour chacun des trois estimateurs. À partir de ces 1 000 valeurs, on construit des estimations Monte Carlo des biais et des erreurs quadratiques moyennes des estimateurs.

Si on note $\widehat{t}_y(j)^{(k)}$ l'estimateur du total de la variable y pour l'Iris j et pour la simulation k, on

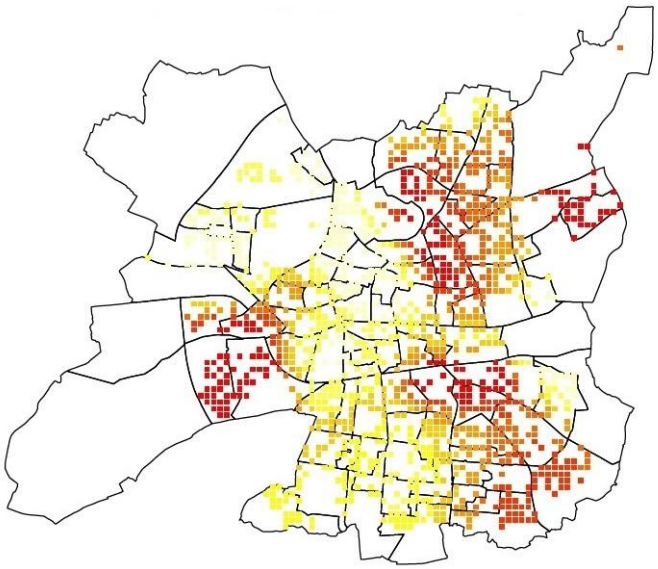

FIGURE 9.13 – Représentation graphique des pentes des régressions géographiques pondérées
Note : L'échelle utilisée est une "échelle de chaleur" qui va de la couleur jaune (valeurs les plus fortes) à la couleur rouge (valeurs les plus faibles).

peut calculer l'erreur quadratique moyenne "Monte Carlo" défini par :

$$EQM(\hat{t}_y(j)) = K^{-1} \sum_{k=1}^{K} (\hat{t}_y(j)^{(k)} - t_y(j))^2 \tag{9.22}$$

puisque l'on connait le total exact $t_y(j)$.

On en déduit l'indicateur qui va servir à comparer les résultats des trois estimations, la racine carrée de l'erreur quadratique moyenne relative :

$$RCEQMR(\hat{t}_y(j)) = \frac{\sqrt{EQM(\hat{t}_y(j))}}{t_y(j)} \tag{9.23}$$

Les Iris de la commune de Rennes sont classés par ordre de taille de population croissante, et on représente sur la figure 9.14 les RCEQMR pour chacun des Iris.

Le premier résultat est l'amélioration de la précision dans les deux approches par les modèles de régression, du fait de la bonne relation linéaire entre la variable y (les personnes à bas revenus) et la variable x (les bénéficiaires de la CMUC). La RCEQMR est de l'ordre de 0.4 pour l'estimateur de Horwitz-Thompson, de l'ordre de 0.12 pour les modèles de régression. La différence entre la régression et la RGP n'est pas très visible sur le graphique de la figure 9.14. Les résultats sont très proches. Les box-plot de la figure 9.15 permettent d'aller un peu plus loin dans la comparaison.

Au vu de la figure 9.14 l'estimateur RGP s'avère néanmoins meilleur que l'estimateur par la régression : pour 75 % des IRIS, la RCEQMR de l'estimateur RGP est inférieure à 0.156, la valeur correspondante pour l'estimateur par la régression étant de 0.178.

9.6 Précautions particulières

9.6.1 Multicolinéarité et corrélation entre les coefficients

Détecter la colinéarité

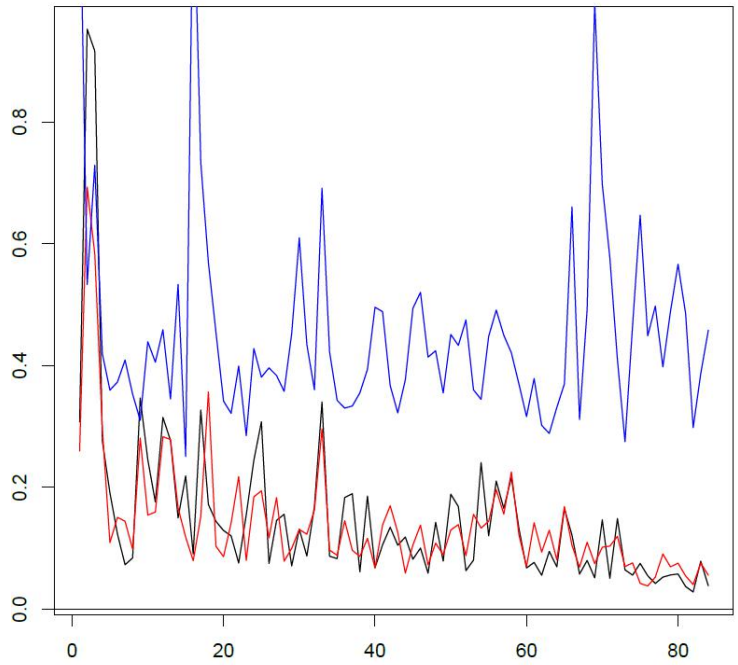

FIGURE 9.14 – RCEQMR (RRMSE sur la figure) de l'estimateur de Horwitz-Thompson (en bleu), de l'estimateur par la régression (en noir) et de l'estimateur par la RGP (en rouge), selon les Iris, classés par taille croissante.

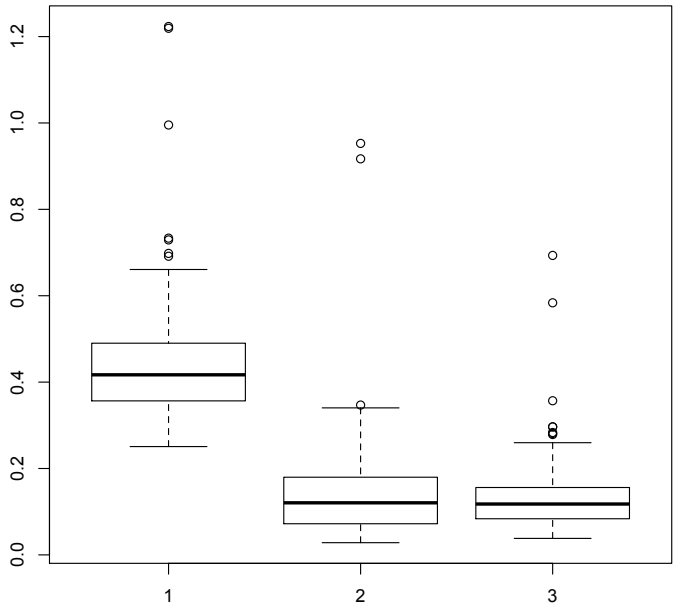

FIGURE 9.15 – Box-plot des RCEQMR de l'estimateur de Horwitz-Thompson (1), de l'estimateur par la régression (2) et de l'estimateur par la RGP (3)

Pour pouvoir estimer les très nombreux coefficients d'une régression géographiquement pondérée, la technique des moindres carrés pondérés impose de nombreuses contraintes sur les paramètres de la régression (LEUNG et al. 2000). Ces contraintes peuvent relier les coefficients de la RGP et créer des problèmes de multicolinéarité.

La multicolinéarité entre les variables peut être responsable d'une grande instabilité dans les coefficients (changement de signe lors de l'ajout d'une nouvelle variable dans la régression) ; du signe contre-intuitif de l'un des coefficients de la régression, ou encore d'erreurs standard des paramètres élevées (WHEELER et al. 2005). Si la structure de corrélation des données est hétérogène dans l'espace, certaines régions peuvent présenter une colinéarité entre leurs variables, tandis que d'autres n'en présenteront pas.

La fonction `gwr.collin.diagno` du package *GWmodel* permet de mettre en œuvre plusieurs types de détection de la colinéarité, notamment les corrélations locales entre les paires de coefficients et les facteurs d'inflation de la variance (VIF) pour chaque coefficient. Ces éléments sont détaillés dans GOLLINI et al. 2013 où des exemples d'application avec R sont présentés.

Encadré 9.6.1 — Facteur d'inflation de la variance : VIF. Soit R_j^2 le coefficient de détermination de la régression de la variable X_j avec les $p-1$ autres variables.

$$VIF_j = \frac{1}{1-R_j^2}$$

Si R_j^2 tend vers 1, VIF_j tend vers $+\infty$ d'où le terme "inflation de la variance". En général, la littérature considère qu'il existe un problème de multicolinéarité lorsqu'un VIF est supérieur à 10, ou que la moyenne des VIF est supérieure à 2 (CHATTERJEE et al. 2015).

Prendre en compte la colinéarité

Une méthode permettant de réduire les problèmes de colinéarité implémentée dans le package *GWModel* est la régression ridge. Le principe est d'augmenter le poids des éléments diagonaux de la matrice de variances-covariances pour diminuer le poids des éléments hors-diagonale (qui contiennent les termes de colinéarité). Dans le cas général, on peut écrire :

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y \tag{9.24}$$

L'inconvénient de cette méthode est que $\hat{\beta}$ est biaisé et que les erreurs standard ne sont plus disponibles.
Dans le cas de la régression géographiquement pondérée, on peut définir une régression ridge locale telle que :

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X + \lambda I(u_i, v_i))^{-1} X^T W(u_i, v_i) Y \tag{9.25}$$

$\lambda I(u_i, v_i)$ est la valeur de λ à la localisation (u_i, v_i). Il est également possible d'utiliser un critère statistique tel que le score de validation croisée pour choisir la bande passante de la régression locale ridge.

9.6.2 Interprétation des paramètres

Problème du test des hypothèses multiples

Lors de l'estimation d'une régression géographiquement pondérée, on obtient en chaque point une évaluation de la significativité de chaque coefficient grâce aux t-values calculées. Pour chaque coefficient, on obtient donc autant de t-values que de points où elles ont été estimées. On se heurte alors au problème du test des hypothèses multiples présenté au chapitre 3 dans le cas des indicateurs

locaux d'autocorrélation spatiale.
Si l'on estime la significativité d'un coefficient en 100 localisations avec un seuil de significativité défini à 95 %, on s'attend à juger le coefficient significatif au moins en 5 localisations, et ceci simplement en raison du principe statistique du test, indépendamment de toute corrélation réelle entre la variable dépendante et les variables explicatives. Pour pallier ce problème, on peut utiliser une méthode d'ajustement de Bonferroni qui augmente la valeur du seuil au-delà duquel le résultat du test local sera jugé non significatif - à niveau de significativité global constant. Cependant les méthodes d'ajustement ont l'inconvénient d'être souvent trop restrictives ce qui peut conduire à juger certains coefficients non significatifs alors qu'ils le sont en réalité.
BRUNSDON et al. 1998 conseillent d'utiliser les t-values produites lors de l'estimation d'une RGP avec précaution. Ils considèrent qu'une surface avec une proportion importante de coefficients très divers localement est un meilleur indicateur de non-stationnarité locale qu'une surface où seule une faible proportion de coefficients dépassent une valeur significative.

Effet du contexte local ou mauvaise spécification

Avant d'interpréter les valeurs des coefficient locaux comme des caractéristiques du contexte local, il est important d'explorer la possibilité d'une mauvaise spécification du modèle. Par exemple, le fait que l'influence d'avoir un garage sur le prix d'un bien immobilier dépende de la localisation peut être dû au fait que la densité de parkings publics varie dans l'espace, ou bien que le modèle hédonique est mal spécifié.

Interprétation de la constante locale

Dans une régression géographiquement pondérée, la constante peut varier localement. Il y a donc un risque qu'elle capture tout le pouvoir explicatif des variables exogènes, en particulier lorsque celles-ci ont une influence nettement plus marquée en certaines localisations (phénomène de clustering spatial). Dans ce cas, les variables explicatives sembleront non significatives. Si l'on soupçonne un tel phénomène, on peut utiliser une régression géographiquement pondérée dite "mixte" dans laquelle la constante ne varie pas.

Conclusion

Proposée en 1998 par BRUNSDON et al. 1998, la RGP a fait l'objet de nombreuses applications pratiques dans les études géographiques et épidémiologiques notamment. Les fondements théoriques ont été considérablement approfondis. Si quelques auteurs ont mis en évidence certaines limites de la méthode, notamment les problèmes de colinéarité (WHEELER et al. 2005, GRIFFITH 2008), elle est désormais partie intégrante des outils d'analyse spatiale. Elle est présentée dans les ouvrages généraux (WALLER et al. 2004, SCHABENBERGER et al. 2017, LLOYD 2010, FISCHER et al. 2009) mais aussi dans les manuels d'économétrie spatiale(ARBIA 2014). Des extensions de la méthode (modèles linéaires généralisés) ont également été proposées.

La RGP peut être utilisée de deux façons différentes. D'un côté elle peut servir de méthode exploratoire pour détecter les zones où apparaissent des phénomènes spatiaux particuliers et les soumettre à une étude approfondie. D'un autre côté, elle peut aider à la construction d'un modèle pertinent : la détection d'une non-stationnarité spatiale est alors symptomatique d'un problème dans la définition du modèle global. BRUNSDON et al. 1998 estiment que la plupart des assertions faites à un niveau global sur la relation spatiale entre les objets mériteraient d'être examinées au niveau local à l'aide de la RGP pour tester leur validité.
La dépendance spatiale entre les termes d'erreur diminue lorsqu'une RGP est utilisée puisque l'autocorrélation spatiale est parfois le résultat d'une instabilité des paramètres non modélisée

(LE GALLO 2004). De plus, la RGP permet de calculer des indicateurs d'autocorrélation spatiale sur une variable, conditionnellement à la distribution spatiale des autres variables, ce qui n'est pas possible avec les indicateurs d'autocorrélation spatiale univariés présentés dans le chapitre 3. Nous encourageons donc à étudier conjointement la dépendance spatiale - avec les indicateurs d'autocorrélation spatiale - et l'hétérogénéité spatiale - avec la régression géographiquement pondérée.

Références - Chapitre 9

ARBIA, Giuseppe (2014). *A primer for spatial econometrics : with applications in R*. Springer.

BRUNSDON, Chris, A Stewart FOTHERINGHAM et Martin E CHARLTON (1996). « Geographically weighted regression : a method for exploring spatial nonstationarity ». *Geographical analysis* 28.4, p. 281–298.

BRUNSDON, Chris, Stewart FOTHERINGHAM et Martin CHARLTON (1998). « Geographically weighted regression ». *Journal of the Royal Statistical Society : Series D (The Statistician)* 47.3, p. 431–443.

CHAMBERS, Ray et Robert CLARK (2012). *An introduction to model-based survey sampling with applications*. T. 37. OUP Oxford.

CHANDRA, Hukum, Ray CHAMBERS et Nicola SALVATI (2012). « Small area estimation of proportions in business surveys ». *Journal of Statistical Computation and Simulation* 82.6, p. 783–795.

CHATFIELD, Chris (2006). « Model uncertainty ». *Encyclopedia of Environmetrics*.

CHATTERJEE, Samprit et Ali S HADI (2015). *Regression analysis by example*. John Wiley & Sons.

FISCHER, Manfred M et Arthur GETIS (2009). *Handbook of applied spatial analysis : software tools, methods and applications*. Springer Science & Business Media.

GOLLINI, Isabella et al. (2013). « GWmodel : an R package for exploring spatial heterogeneity using geographically weighted models ». *arXiv preprint arXiv :1306.0413*.

GRIFFITH, Daniel A (2008). « Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR) ». *Environment and Planning A* 40.11, p. 2751–2769.

HUBER, Peter (1981). « J. 1981. Robust Statistics ». *New York : John Wiley*.

LE GALLO, Julie (2004). « Hétérogénéité spatiale ». *Économie & prévision* 1, p. 151–172.

LEUNG, Yee, Chang-Lin MEI et Wen-Xiu ZHANG (2000). « Statistical tests for spatial nonstationarity based on the geographically weighted regression model ». *Environment and Planning A* 32.1, p. 9–32.

LLOYD, Christopher D (2010). *Local models for spatial analysis*. CRC press.

SCHABENBERGER, Oliver et Carol A GOTWAY (2017). *Statistical methods for spatial data analysis*. CRC press.

WALLER, Lance A et Carol A GOTWAY (2004). *Applied spatial statistics for public health data*. T. 368. John Wiley & Sons.

WHEELER, David et Michael TIEFELSDORF (2005). « Multicollinearity and correlation among local regression coefficients in geographically weighted regression ». *Journal of Geographical Systems* 7.2, p. 161–187.