12. Petits domaines et corrélation spatiale

PASCAL ARDILLY
Insee

PAUL BOUCHE
Ensai - Sciences Po

WENCAN ZHU
Ensai

Résumé

Lorsque l'on veut diffuser les résultats d'une enquête sur des petites populations, en particulier s'il s'agit de zones géographiques restreintes, les faibles effectifs de l'échantillon recoupant ces populations peuvent conduire à des estimations trop imprécises. La théorie classique des sondages n'apporte pas de solution satisfaisante à ce problème et il faut donc faire appel à des techniques d'estimation spécifiques, fondées sur l'utilisation d'informations auxiliaires et sur des modèles plus ou moins complexes. Tous ces modèles ont en commun de formaliser la liaison entre la variable d'intérêt et les variables auxiliaires. La liaison linéaire est la forme la plus simple mais on trouve d'autres modèles, de nature non linéaire (modèle de Poisson, modèle logistique). La plupart des modèles isolent des effets locaux, propres aux domaines. On peut introduire des corrélations entre ces effets, d'autant plus fortes que les domaines sont géographiquement proches. Cette corrélation spatiale est alors de nature à améliorer la qualité des estimations localisées.

Ce chapitre est consacré à la présentation générale de la problématique appelée "estimation sur petits domaines" en portant une attention plus particulière à la prise en compte de la corrélation spatiale dans les modèles.

12.1 Mise en place du modèle

12.1.1 Contexte et objectifs

Les statisticiens d'enquête portent un intérêt particulier à l'estimation de paramètres θ inconnus, définis dans une population finie et généralement de grande taille. La plupart des paramètres sont des totaux ou des dérivés immédiats de totaux tels que des moyennes ou des proportions. Plus rarement, on trouve des fonctions non linéaires mais que l'on peut tout de même exprimer comme des fonctions de totaux (ratios, variances dans la population, coefficients de corrélation ou de régression). Selon le sujet de l'enquête, on peut aussi vouloir estimer des paramètres fortement non linéaires, comme des quantiles ou des indicateurs d'inégalités, lesquels ne s'écrivent pas comme des fonctions de totaux.

Les paramètres sont définis à partir d'une (ou plusieurs) variables(s) d'intérêt et se formalisent par des expressions impliquant en toute généralité l'ensemble des individus de la population U. On notera Y la variable d'intérêt, que l'on considérera par la suite comme unique. Les individus de U étant identifiés par l'indice i, si le paramètre θ est un total T, alors $T = \sum_{i \in U} Y_i$.

Lorsqu'on ne dispose pas de la valeur individuelle Y_i pour chaque individu i de U, on doit procéder à une estimation de T par sondage, c'est-à-dire à partir d'informations Y_i obtenues sur un échantillon répondant, noté s, inclus dans U. Le tirage de l'échantillon relève généralement d'un plan d'échantillonnage complexe, associant par exemple une stratification, des tirages à probabilités inégales et des tirages à plusieurs degrés. Certains paramètres d'intérêt ne sont pas définis sur la population U toute entière mais sur une sous-population, notée d. Une telle sous-population s'appelle un *domaine*, et on a alors affaire à une estimation sur domaine. Dans ce cas, le paramètre d'intérêt θ peut être le total sur le domaine, soit $T_d = \sum_{i \in d} Y_i$, qu'il faut estimer à partir des données collectées. La théorie classique des sondages attribue à chaque unité échantillonnée i un poids de sondage w_i, coefficient réel positif qui dépend de la méthode d'échantillonnage et de traitement de la non-réponse et qui vient "dilater" la valeur de la variable d'intérêt Y_i. Pour estimer un total T défini sur la population complète U, l'estimateur prend une forme linéaire $\hat{T} = \sum_{i \in s} w_i Y_i$. Pour estimer un total sur un domaine d, on se contente de restreindre la somme aux éléments de d sans toucher à leur pondération, soit $\hat{T}_d = \sum_{i \in s \cap d} w_i Y_i$. Si le paramètre θ est une moyenne sur d, notée désormais $\bar{Y}_d$ (ce qui inclut les proportions, qui sont des moyennes de variables booléennes), on estime la taille N_d du domaine par $\hat{N}_d = \sum_{i \in s \cap d} w_i$ (une taille est un total de valeurs individuelles constantes égales à 1) et on forme le ratio $\hat{\bar{Y}}_d = \hat{T}_d / \hat{N}_d$. Mais si on connaît N_d, on peut aussi utiliser l'estimation alternative $\hat{\bar{Y}}_d = \hat{T}_d / N_d$.

Dans tous les cas de figure, l'échantillonnage entraîne une erreur spécifique des estimateurs $\hat{T}_d$ et $\hat{\bar{Y}}_d$ que l'on résume au moyen de deux indicateurs appelés respectivement *biais* et *variance d'échantillonnage*. Considérons le cas de $\hat{\bar{Y}}_d$. Le biais désigne la différence entre l'espérance de $\hat{\bar{Y}}_d$, c'est-à-dire l'estimation attendue "en moyenne" compte tenu de l'aléa qui conduit à la constitution de s, et le paramètre $\bar{Y}_d$, tandis que la variance d'échantillonnage mesure la sensibilité de l'estimation $\hat{\bar{Y}}_d$ à l'échantillon répondant s. Un plan de sondage précis se traduit par un faible biais et une faible variance. Les estimateurs issus de la théorie des sondages et utilisés par les statisticiens d'enquête sont généralement sans biais ou ont un biais négligeable. La variance d'échantillonnage est une fonction décroissante de n_d, où n_d est la taille de l'échantillon répondant recoupant le domaine d, c'est-à-dire la taille de $s \cap d$. Lorsque n_d est suffisamment petit pour que les objectifs de qualité de l'estimation $\hat{\bar{Y}}_d$ ne soient pas atteints, on est face à un problème d'estimation sur *petit domaine*.

Pour traiter cette difficulté, lorsqu'il n'est plus possible d'augmenter la valeur de la taille de s, notée n, il faut créer un contexte théorique nouveau qui permette de rendre l'estimation finale du paramètre θ (total T_d ou moyenne $\bar{Y}_d$) moins sensible à l'échantillon répondant s (ou $s \cap d$, ce qui est équivalent). C'est une technique de *modélisation* qui va le permettre. Il s'agit de se placer dans un cadre hypothétique simplificateur de la réalité (c'est la définition générale d'un modèle). L'approche habituelle consiste à considérer que Y_i s'explique essentiellement par un ensemble de variables individuelles X_i connues pour chaque unité i de la population tout en impliquant quelques grandeurs δ *a priori* inconnues - les paramètres du modèle. Il suffira d'estimer ces grandeurs δ pour pouvoir en déduire n'importe quelle valeur inconnue Y_i (correspondant aux cas $i \notin s$), et donc *in fine* la valeur du paramètre θ.

La mise en œuvre de la modélisation passe fondamentalement par la disponibilité d'informations auxiliaires. Naturellement, on pense à des variables connues au niveau individuel sur l'intégralité de la population U. Supposons que l'information auxiliaire relative à l'individu i soit composée de p variables individuelles, notées $X_{i,1}, X_{i,2}, ..., X_{i,p}$, et partons du principe qu'il existe une liaison "suffisamment fiable" entre ces valeurs et la variable d'intérêt Y_i. Cette liaison est par construction considérée comme valable lorsqu'on l'applique à l'intégralité de la population U, sans autre connaissance que $X_{i,1}, X_{i,2}, ..., X_{i,p}$. Il est essentiel qu'elle reste valable si on se limite à l'échantillon répondant s, ce qui signifie que l'information qu'apporte l'appartenance à l'échantillon répondant ne doit pas amener le statisticien à modifier l'expression formelle de cette relation (plan de sondage dit "non informatif"). Dans un monde idéal où tout serait simple, il existerait une certaine fonction f telle que pour tout individu i de U on a $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}; \delta)$ où δ est un paramètre vectoriel inconnu à ce stade, dit paramètre du modèle. Dans ce contexte parfait, la forme fonctionnelle de la fonction f est parfaitement connue mais elle est néanmoins paramétrée par δ. Si on parvient, grâce à l'information collectée à l'occasion de l'enquête, à estimer de façon satisfaisante le paramètre δ, on pourra prédire les valeurs Y_i de tous les individus i non échantillonnés (ou échantillonnés mais non-répondants) et donc prédire θ.

Le cadre traditionnel de la statistique d'enquête qu'est la théorie des sondages ne s'appuie sur aucune modélisation et considère que la variable d'intérêt Y n'est pas aléatoire (elle est donc déterministe). C'est la procédure de sélection de l'échantillon et le mécanisme de non-réponse qui introduisent un aléa et cet aléa permet de considérer tout estimateur, comme par exemple l'estimateur de la moyenne $\bar{Y}_d$, comme une variable aléatoire. Or, en présence d'une modélisation de Y, puisque la réalité n'est pas celle d'un monde idéal et simple, il ne serait pas raisonnable de supposer qu'il y a égalité entre la valeur Y_i et une quelconque valeur de type $f(X_{i,1}, X_{i,2}, ..., X_{i,p}; \delta)$, car la relation entre Y_i et $X_{i,1}, X_{i,2}, ..., X_{i,p}$ serait trop contrainte et donc non crédible. C'est pourquoi il faut considérer que la fonction f comprend une composante aléatoire U_i, dont la première caractéristique est d'être guidée par le hasard. On doit désormais abandonner l'environnement traditionnel de la théorie des sondages et considérer que *les variables Y sont des variables aléatoires*, telles que $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}, U_i; \delta)$.

12.1.2 Le modèle linéaire individuel standard

Le formalisme du modèle utilise par conséquent des aléas explicites qui lui sont propres et qui n'ont aucune relation avec l'aléa d'échantillonnage. Dans certaines circonstances, on a coutume d'introduire une variable aléatoire individuelle U_i, qui est en moyenne nulle et que l'on relie ainsi à Y_i, pour tout i de U (équation 12.1).

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + U_i \tag{12.1}$$

Les variables auxiliaires $X_{i,1}, X_{i,2}, ..., X_{i,p}$, parfaitement déterministes, sont dites "effets fixes". L'aléa du modèle portant sur les valeurs Y_i ne doit pas être confondu avec l'aléa de sondage qui

détermine la composition de l'échantillon s. C'est à ce stade que le contexte de l'estimation sur petits domaines apporte sa spécificité. La population U étant partitionnée en D domaines, on considère que si i appartient au domaine d, l'aléa U_i - nul *en moyenne* - est composé d'un effet (aléatoire) propre au domaine d, noté τ_d, et d'un résidu (aléatoire) individuel noté e_i. On a donc :

$$U_i = \tau_d + e_i. \tag{12.2}$$

Dans l'approche la plus simple, les deux composantes τ_d et e_i sont supposées indépendantes, les τ_d sont deux à deux indépendants, de même les e_i sont deux à deux indépendants. L'espérance et la variance associées à l'aléa du modèle seront notés respectivement ε et ν, si bien que les hypothèses les plus simples accompagnant ce modèle sont :

— pour les espérances : $\varepsilon(\tau_d) = 0$ et $\varepsilon(e_i) = 0$;
— pour les variances : $\nu(\tau_d) = \sigma_\tau^2$ et $\nu(e_i) = \sigma_e^2$.

Par ailleurs, toutes les covariances envisageables impliquant ces composantes élémentaires sont nulles. Ainsi, globalement $\varepsilon(U_i) = 0$ et $\nu(U_i) = \sigma_\tau^2 + \sigma_e^2$. Le formalisme de ce modèle permet de créer une corrélation entre les variables d'intérêt associées à des unités d'un même domaine puisque $\forall i \in U$, $\forall j \in U$, $j \neq i$: si $i \in d$ et $j \notin d$ alors $cov(Y_i, Y_j) = cov(U_i, U_j) = 0$ et si $i \in d$ et $j \in d$ alors $cov(Y_i, Y_j) = cov(U_i, U_j) = \sigma_\tau^2$. Ainsi, la matrice de variances-covariances du vecteur des Y_i, où i parcourt U, a la forme d'une matrice diagonale par blocs, chaque bloc étant associé à un domaine et pouvant être décrit par une diagonale comprenant partout $\sigma_\tau^2 + \sigma_e^2$ alors que tous les autres éléments du bloc prennent la valeur constante σ_τ^2.

Du fait des hypothèses portant sur les moments des aléas, un tel modèle ne peut s'appliquer, en toute rigueur, qu'à des variables Y quantitatives et continues - ce qui exclut en particulier toute variable d'intérêt de nature qualitative (et donc les paramètres définis comme des proportions). L'effet aléatoire τ_d est un effet local qui s'interprète comme étant la composante de la variable d'intérêt expliquée par l'appartenance au domaine au-delà de l'information contenue dans les variables individuelles $X_{i,1}, X_{i,2}, ..., X_{i,p}$. Souvent, les domaines sont des zones géographiques et τ_d prétend traduire la part d'explication purement due à la localisation géographique de l'unité. Apprécier la vraie part explicative de la localisation sur telle ou telle zone, et même définir ce qu'est un effet géographique, constitue une question un peu philosophique. En effet, parce que c'est une explication facile et bien pratique, on peut toujours considérer comme effet géographique un effet résiduel significatif qui serait dû à une insuffisante prise en compte des variables auxiliaires individuelles réellement explicatives. Autrement dit, s'il y a des éléments géographiques qui expliquent Y, l'idéal consiste à les traduire d'une façon ou d'une autre dans le vecteur d'effets fixes $X_{i,1}, X_{i,2}, ..., X_{i,p}$. Il faut donc concevoir *a priori* l'effet local τ_d comme un effet "parasite" et chercher à en diminuer au maximum l'importance : plus le paramètre σ_τ^2 sera petit, c'est-à-dire plus les valeurs τ_d seront numériquement faibles, plus le caractère explicatif reposera sur les effets fixes $X_{i,1}, X_{i,2}, ..., X_{i,p}$, et donc meilleur sera le modèle. La structure de covariance ayant une certaine complexité, on dit que le modèle appartient à la famille des modèles linéaires généraux.

Avec un tel modèle, l'espérance de la variable aléatoire Y_i est une fonction linéaire des paramètres β. La composante explicative de Y_i la plus importante est constituée d'effets non aléatoires $X_{i,j}$ (les effets fixes) mais la composante résiduelle τ_d attribuée exclusivement au domaine est en revanche de nature aléatoire (l'effet aléatoire). Pour ces raisons, on parle de *modèle linéaire mixte*.

Si on reprend les notations de la partie 12.1.1, on vérifie bien que $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}, U_i; \delta)$, où le paramètre vectoriel δ rassemble toutes les grandeurs inconnues apparaissant dans le modèle, à savoir $\delta = (\beta_0, \beta_1, \beta_2, ..., \beta_p, \sigma_\tau^2, \sigma_e^2)$. Il a une dimension $p+3$, distinguant $p+1$ paramètres réels associés aux effets fixes explicatifs et deux paramètres réels associés à la structure de variances-covariances attachée au modèle.

12.1.3 Le modèle linéaire individuel avec corrélation spatiale

Le modèle linéaire mixte standard formule l'hypothèse d'une corrélation nulle entre les aléas U_i associés à des individus appartenant à deux domaines distincts. Cette situation n'est pas nécessairement crédible parce que les limites des zonages géographiques constituant les domaines n'ont aucune raison de constituer une barrière stoppant brutalement toute propagation des phénomènes mesurés. En général, il y a une forme de continuité spatiale naturelle des comportements des individus localisés et deux individus géographiquement proches sur le terrain ont plus de chance d'afficher des valeurs de Y voisines que deux individus éloignés. De ce point de vue, une relation entre les effets géographiques caractérisant des domaines proches apparaît assez naturelle.

Sur le plan technique, on peut chercher à traduire cette situation en introduisant une corrélation qui ne tienne plus compte que de la distance entre les domaines. La forme analytique de la corrélation est libre, pourvu qu'elle diminue quand la distance augmente. Dans cet esprit, on peut s'appuyer sur un modèle qui conserve exactement les formalisations 12.1 et 12.2 mais qui assure $\forall i \in d, \forall j \in d'$, si $i \neq j$:

$$cov(Y_i, Y_j) = cov(U_i, U_j) = cov(\tau_d, \tau_{d'}) = \sigma_\tau^2 exp\left(-\frac{1}{\rho} dist(d, d')\right) \tag{12.3}$$

où $dist(d, d')$ est une distance définie entre les domaines d et d'. On peut prendre par exemple la distance euclidienne habituelle calculée à partir des coordonnées des centroïdes des deux domaines concernés. Le coefficient ρ est un paramètre d'échelle qui offre l'opportunité d'un meilleur ajustement du modèle : plus la distance sera influente sur la covariance, plus ρ sera proche de zéro. Dans le cas particulier où $d = d'$, et lorsque $i \neq j$, alors $cov(Y_i, Y_j) = cov(\tau_d, \tau_d) = \sigma_\tau^2 exp(0) = \sigma_\tau^2$. Si $i = j$, s'ajoute la variance de l'effet individuel, soit $cov(Y_i, Y_j) = cov(U_i, U_j) = \sigma_\tau^2 + \sigma_e^2$. Cette fois, la matrice de variances-covariances est une matrice pleine, sans zéros. On peut néanmoins considérer, à titre de variante intéressante, que la distance devient infinie lorsqu'elle a dépassé un certain seuil. Cela permet de réintroduire de nombreux zéros dans la matrice, facilitant ainsi les traitements numériques ultérieurs (en particulier en épargnant de la mémoire vive). Dans ces circonstances, les paramètres du modèle sont un peu plus nombreux puisqu'il faut tenir compte du nouveau paramètre ρ, si bien que $\delta = (\beta_0, \beta_1, \beta_2, ..., \beta_p, \rho, \sigma_\tau^2, \sigma_e^2)$.

Une autre approche consiste à introduire une relation simple entre les effets locaux τ_d des différents domaines, en faisant en sorte que cette relation soit d'autant plus forte que les domaines sont plus proches. Ainsi, on peut concevoir que l'effet local associé à un domaine donné soit "presque" une combinaison linéaire des effets locaux des domaines qui l'entourent, avec une intensité de liaison qui diminue au fur et à mesure qu'on s'éloigne du domaine donné. L'intensité de la liaison entre les effets τ_d est traduite par deux éléments. D'une part un système de coefficients $\alpha_{d,d\prime}$ [1] qui règlent l'influence relative que peuvent avoir les différents domaines distingués $d\prime$ sur un domaine donné d, d'autre part un paramètre ρ compris entre -1 et 1 qui règle la valeur absolue de l'intensité de liaison. On impose pour tout i : $\sum_{d\prime=1, d\prime\neq d}^{D} \alpha_{d,d\prime} = 1$. La relation postulée entre les effets aléatoires est $\tau_d \approx \rho \sum_{d\prime=1, d\prime\neq d}^{D} \alpha_{d,d\prime} \tau_{d\prime}$. En écriture matricielle, cela devient :

$$\begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_D \end{pmatrix} = \rho . \begin{pmatrix} 0 & \alpha_{1,2} & \dots & \alpha_{1,D} \\ \alpha_{2,1} & 0 & \dots & \alpha_{2,D} \\ \vdots & \ddots & 0 & \vdots \\ \alpha_{D,1} & \dots & \dots & 0 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_D \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_D \end{pmatrix} \tag{12.4}$$

1. Les paramètres $\alpha_{d,d\prime}$ correspondent aux poids $w_{d,d\prime}$ de la matrice de poids W utilisée dans les chapitres précédents. Dans ce chapitre, w désigne les poids de sondage.

en introduisant un vecteur d'aléas u_d qui suit une loi de Gauss centrée et de variance $\sigma_u^2 I_D$. On désigne ce modèle sous le nom de modèle SAR (*Simultaneous AutoRegressive model*).

L'arbitrage entre cette méthode et la précédente n'a rien d'évident *a priori*, c'est pourquoi le seul conseil à prodiguer à ce stade consiste à tester les deux méthodes puis à utiliser les outils d'appréciation de la qualité dont on dispose, en particulier ceux mentionnés dans la partie 12.3.

L'introduction d'une corrélation spatiale dans le modèle linéaire mixte de base ne change rien aux conditions restrictives d'usage : un tel modèle ne peut être utilisé que pour l'estimation de paramètres θ construits à partir d'une variable d'intérêt quantitative et continue. En outre, il perd une grande partie de son intérêt si les domaines sont géographiquement de grande taille car la distance considérée se mesure entre les centroïdes des domaines.

En pratique, pour limiter le nombre de coefficients non nuls dans la matrice de variances-covariances des effets locaux (et ainsi accélérer les calculs et/ou éviter des problèmes de mémoire insuffisante), on neutralise complètement l'influence $\alpha_{d,d\prime}$ des domaines $d\prime$ situés au-delà d'une certaine distance de d, ou même éventuellement qui ne sont pas dans un voisinage immédiat du domaine de référence d. Néanmoins, il est difficile d'éviter les problèmes posés par les "effets de bord" qui surviennent lorsqu'un domaine se trouve en périphérie d'un territoire plus vaste, parce qu'on ne peut pas prendre en compte tous ses voisins. C'est par exemple presque systématiquement le cas pour les territoires frontière des États.

12.1.4 Traitement des variables qualitatives par un modèle individuel linéaire mixte généralisé

Le modèle logistique

Les paramètres de dénombrement d'une sous-population quelconque s'appuient sur des variables individuelles de nature qualitative. Supposons que l'on cherche à estimer le nombre total d'individus θ vérifiant une propriété donnée Γ - comme par exemple "être une femme" ou "être un agriculteur de moins de 50 ans". Si on définit la variable individuelle $Y_i = 1$ lorsque i vérifie Γ et $Y_i = 0$ dans le cas contraire, il est facile de vérifier que $\theta = \sum_{i \in U} Y_i$. La variable aléatoire Y ainsi définie est une variable dite "indicatrice" qui quantifie une information individuelle initialement qualitative. En divisant θ par la taille de U, on obtient la proportion d'individus de la population qui vérifient la propriété Γ. Malheureusement, le modèle 12.1 n'est pas du tout adapté à ce type de variable. On va contourner la difficulté en optant pour une modélisation parfaitement compatible avec les variables indicatrices : la loi de Bernoulli traduira la distribution des Y_i. Il s'agit d'une distribution qui charge la valeur 1 avec une probabilité P_i et la valeur 0 avec une probabilité $1 - P_i$. On peut donc considérer que pour tout individu i de la population globale U, la variable Y_i est une variable aléatoire qui suit une loi de Bernoulli $\mathscr{B}(1, P_i)$. Le cœur de la modélisation suit : on va relier le paramètre P_i aux caractéristiques individuelles de i résumées par les variables auxiliaires $X_{i,1}, X_{i,2}, ..., X_{i,p}$ et on va introduire un effet aléatoire local τ_d. La forme fonctionnelle qui relie P_i aux $X_{i,1}, X_{i,2}, ..., X_{i,p}$ et à τ_d doit être compatible avec la contrainte $P_i \in [0,1]$. Différentes options existent, mais la plus commune consiste à poser, pour tout i dans d :

$$log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \tau_d. \tag{12.5}$$

On parle de *modèle logistique*. L'espérance de la variable aléatoire Y_i est P_i, qui n'est manifestement pas une fonction linéaire des paramètres β (contrairement au cas de 12.1). Pour cette raison, on dit que le modèle représenté par l'équation 12.5 est un *modèle linéaire mixte généralisé*. La classe des modèles de type 12.5 distingue les modèles où les effets locaux τ_d sont deux à deux indépendants, comme dans 12.2, et les modèles avec corrélation spatiale, comme dans 12.3 ou 12.4.

Le modèle de Poisson

Il arrive que l'information qualitative se présente de manière agrégée lorsqu'on traite les unités statistiques. Si on reprend l'exemple précédent, dans le cas où les unités sont des ménages et non plus des individus physiques, on dispose pour chaque ménage i du nombre total d'individus Y_i vérifiant la propriété Γ (le nombre de femmes dans le ménage, ou le nombre d'agriculteurs de moins de 50 ans dans le ménage). Cette variable n'est plus une variable indicatrice mais une variable qui peut prendre n'importe quelle valeur dans $\mathbb{N}$, ensemble des entiers naturels (en théorie, puisqu'en pratique elle est toujours bornée supérieurement). Dans ces conditions, la loi de Poisson est une loi naturelle assez simple que l'on peut associer à Y_i. Elle possède un unique paramètre λ_i réel (strictement positif), que l'on va faire dépendre de l'unité i au travers de caractéristiques individuelles $X_{i,1}, X_{i,2}, ..., X_{i,p}$ et d'un effet aléatoire local τ_d. Le paramètre λ_i est souvent transformé par une fonction simple avant d'être relié aux facteurs explicatifs. En pratique, on utilise essentiellement la fonction logarithme, ce qui fait que le modèle complet - linéaire mixte généralisé - se formalise ainsi :

$$\begin{aligned} Y_i &\sim Poisson(\lambda_i) \\ log(\lambda_i) &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \tau_d. \end{aligned} \tag{12.6}$$

Une fois encore, les effets aléatoires locaux τ_d peuvent être considérés comme deux à deux indépendants, comme dans 12.2, ou corrélés spatialement, comme dans 12.3 ou 12.4.

12.1.5 Extension aux modèles définis au niveau du domaine

Le modèle de Fay et Herriot

En tenant compte de l'échantillonnage, on peut produire des estimateurs de n'importe quel paramètre, en particulier les totaux T_d (ou les moyennes $\bar{Y}_d$) définis au niveau du domaine d. Ces estimateurs sont construits avec les poids de sondage individuels w_i (eux-mêmes fonction de la méthode d'échantillonnage utilisée). Ils n'utilisent que l'information relative au domaine d, c'est pourquoi on les appelle *estimateurs directs*. Il est possible de construire une modélisation qui s'appuie sur ces estimateurs, notés $\hat{T}_d$ pour les totaux et $\hat{\bar{Y}}_d$ pour les moyennes. L'unité statistique modélisée n'est plus alors l'individu mais le domaine. L'objectif est de relier l'information disponible $\hat{T}_d$ ou $\hat{\bar{Y}}_d$ à un ensemble de variables explicatives, ces dernières étant adaptées au niveau traité : il faut naturellement qu'elles caractérisent les domaines et non plus les individus. Les effets locaux τ_d conservent leur nature et leur interprétation, exactement comme dans l'équation 12.2.

Un célèbre modèle est le modèle dit de *Fay et Herriot*, qui appartient à la famille des modèles linéaires mixtes. Si on note $X_{d,1}, X_{d,2}, ..., X_{d,p}$ les variables explicatives retenues au niveau domaine, la version la plus élémentaire de la modélisation s'écrit :

$$\bar{Y}_d = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d. \tag{12.7}$$

La variable expliquée est ici la vraie moyenne dans le domaine d. Puisque cette valeur est inconnue, il faut ajouter une étape pour lui substituer une estimation. À ce stade, l'estimation $\hat{\bar{Y}}_d$ issue de l'enquête n'est certes pas de bonne qualité puisque l'échantillon $s \cap d$ est de petite taille, néanmoins elle existe et peut être reliée à la vraie valeur en introduisant un terme d'erreur err_d selon :

$$\hat{\bar{Y}}_d = \bar{Y}_d + err_d. \tag{12.8}$$

La variable err_d est l'erreur d'échantillonnage. Cette dernière équation n'a rien à voir avec un modèle, il s'agit simplement de la définition de l'erreur d'échantillonnage. Généralement

l'estimateur $\hat{\bar{Y}}_d$ est pondéré de façon à être sans biais ou de biais négligeable (s'il y a eu redressement par exemple, et si toutefois on considère que la non-réponse a été correctement traitée) si bien que l'erreur d'échantillonnage a une espérance nulle lorsqu'on prend en compte l'aléa de sondage, soit $\mathbb{E}(err_d) = 0$. La variance de l'erreur dépend de l'échantillonnage mais on sait qu'elle varie comme l'inverse de n_d. On notera désormais ψ_d cette variance. La combinaison des deux équations précédentes conduit à la formule opérationnelle :

$$\hat{\bar{Y}}_d = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d + err_d. \tag{12.9}$$

À l'image de ce que l'on a vu avec les modèles individuels, on peut formuler une hypothèse d'indépendance entre les effets locaux τ_d, ou au contraire postuler une corrélation spatiale, structurée comme dans les équations 12.3 ou 12.4.

Les hypothèses sur l'espérance et la variance des effets τ_d n'étant en toute rigueur compatibles qu'avec des vraies moyennes $\bar{Y}_d$ qui ont des distributions continues, autant dire que la variable d'intérêt individuelle Y_i collectée au niveau des individus devrait être quantitative et continue. Cela étant, si la variable d'intérêt Y_i est par nature qualitative et si le domaine d a une taille N_d suffisamment grande, on peut considérer - avec un peu d'audace parfois ! - que la vraie moyenne $\bar{Y}_d$ peut prendre *a priori* un nombre suffisamment grand de valeurs pour que cet ensemble puisse être considéré comme continu, c'est-à-dire sans "trou". C'est bien la taille N_d qui est le paramètre essentiel. Considérons par exemple Y_i la variable indicatrice caractérisant la modalité "femme". La moyenne $\bar{Y}_d$ est alors la proportion de femmes dans la population du domaine. Si $N_d = 10$, cette moyenne peut prendre les valeurs $k/10$, où k est un entier compris entre 0 et 10, ce qui est très loin d'occasionner une situation "continue". Si $N_d = 10\,000$, la moyenne peut prendre les valeurs $k/10\,000$, où k est un entier compris entre 0 et 10 000, ce qui rend beaucoup plus plausible l'hypothèse de continuité. C'est pourquoi on peut conclure que la modélisation 12.9 est acceptable pour l'estimation de proportions (variables d'intérêt qualitatives) dès lors que les domaines d ne sont pas trop petits.

Le modèle de Poisson

Bien que le modèle de Fay et Herriot s'accommode bien des variables qualitatives, c'est-à-dire des paramètres qui se définissent comme des proportions par domaine d'individus vérifiant une propriété Γ (assimilable à une sous-population Γ) ou comme les effectifs par domaine de ces mêmes individus, on peut lui préférer dans certaines circonstances un modèle plus spécifiquement adapté aux dénombrements. Notons $N_{\Gamma,d}$ le nombre total d'individus du domaine d appartenant à la sous-population Γ. L'échantillon permet de former l'estimateur sans biais (ou presque) $\hat{N}_{\Gamma,d} = \sum_{i \in s \cap d \cap \Gamma} w_i$. Cet estimateur n'utilise que l'information liée au domaine, c'est donc un estimateur direct, et il est de qualité médiocre puisque l'échantillon $s \cap d$ est de petite taille. Néanmoins, il s'agit d'une variable aléatoire calculable dont on peut modéliser la distribution par une loi de Poisson. Cette loi, dépendant d'un unique paramètre réel λ_d fonction du domaine, est particulièrement adaptée aux dénombrements. On peut montrer que λ_d est l'espérance mathématique de $\hat{N}_{\Gamma,d}$: il doit donc être numériquement assez proche de cette estimation. Il s'agit bien à ce stade d'une première hypothèse et non d'une propriété qui découlerait de la théorie des sondages. Néanmoins, le risque pris reste modeste parce que le comportement asymptotique des estimateurs directs est proche d'une loi de Gauss, dont la loi de Poisson est elle-même proche si son paramètre est suffisamment grand.

Le cœur de la modélisation relève de la suite : on considère généralement que le logarithme du paramètre λ_d s'écrit ainsi :

$$log(\lambda_d) = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d \tag{12.10}$$

en reprenant les notations des parties précédentes. La variable aléatoire τ_d conserve la même interprétation : il s'agit de distinguer l'effet de la localisation des unités statistiques au-delà de ce que les effets fixes $X_{d,1}, X_{d,2}, ..., X_{d,p}$ sont capables de traduire. Les hypothèses portant sur les corrélations entre les effets locaux τ_d sont identiques à celles des modèles déjà présentés : ou bien on considère que ces effets sont deux à deux indépendants, ce qui est plus simple mais peut-être parfois en décalage avec la réalité du terrain, ou bien on introduit des corrélations spatiales, en reprenant par exemple les formulations 12.3 ou 12.4. Dans les deux cas, le modèle est un modèle linéaire mixte généralisé.

12.2 Formation de l'estimateur "petits domaines"

Définir le modèle sur lequel on va s'appuyer ne constitue qu'une première étape du processus. À ce stade, on ne perçoit encore qu'assez qualitativement l'intérêt du modèle, qui consiste à réduire la dimension du problème en simplifiant considérablement la réalité. Il est en effet beaucoup plus facile de procéder à des estimations dans un univers où toute l'information d'intérêt est supposée s'expliquer par quelques variables bien connues et par quelques paramètres plutôt qu'à évoluer dans un système non cadré qui de fait dépendrait d'une infinité de composantes non maîtrisées... comme le suppose au demeurant la théorie classique des sondages !

L'étape suivante est celle du choix de la stratégie d'estimation - on devrait d'ailleurs désormais parler de prédiction puisque le paramètre d'intérêt est devenu une variable aléatoire à la suite de la modélisation.

12.2.1 Stratégie d'estimation BLUP : cas du modèle individuel standard

Dans cette partie, on considère uniquement des modèles linéaires. Dans ce cadre, plusieurs stratégies d'estimation/prédiction du paramètre d'intérêt peuvent être mises en œuvre mais nous présentons maintenant celle qui est probablement la plus commune, la stratégie *Best Linear Unbiased Predictor* (BLUP). Considérons le cas où le paramètre est la moyenne $\bar{Y}_d$. Son prédicteur est en toute généralité une fonction des données collectées, c'est-à-dire des Y_i où i décrit l'échantillon global répondant s. Le statisticien cherche avant tout un prédicteur qui soit linéaire, du type $\sum_{i\in s} a_i Y_i$ où les a_i sont des coefficients réels, et sans biais, c'est-à-dire que son espérance soit égale à celle de $\bar{Y}_d$. Enfin, il cherche à minimiser l'erreur quadratique moyenne (*Mean Square Error* ou MSE) qui est l'espérance du carré de l'écart entre le prédicteur et la valeur $\bar{Y}_d$ qu'il doit prédire. La solution de ce problème mathématique est l'estimateur (ou prédicteur) BLUP, dit aussi dans la littérature estimateur de Henderson. On le notera $\tilde{\bar{Y}}_d^H$.

Dans le cas spécifique du modèle linéaire mixte individuel standard (voir partie 12.1.2), lorsque la fraction de sondage est négligeable, on vérifie que l'estimateur BLUP s'écrit :

$$\tilde{\bar{Y}}_d^H = \gamma_d[\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T \tilde{\beta}] + (1-\gamma_d)\bar{X}_d^T \tilde{\beta} \tag{12.11}$$

Tous les vecteurs sont des vecteurs colonnes, le vecteur transposé étant repéré par l'exposant T. En notant D le nombre total de domaines d'intérêt, en notant $X_i = (X_{i,1}, X_{i,2}, ..., X_{i,p})^T$ le vecteur des variables auxiliaires, $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ le vecteur des paramètres de modèle associés à ces variables, $\bar{x}_d = \frac{1}{n_d}\sum_{i\in s\cap d} X_i$, $\bar{y}_d = \frac{1}{n_d}\sum_{i\in s\cap d} Y_i$ et $\bar{X}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} X_i$, on a :

$$\gamma_d = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \frac{\sigma_e^2}{n_d}} \tag{12.12}$$

$$\tilde{\beta} = \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i X_i^T - \gamma_d n_d \bar{x}_d \bar{x}_d^T \right) \right)^{-1} . \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i Y_i - \gamma_d n_d \bar{x}_d \bar{y}_d \right) \right). \quad (12.13)$$

Le vecteur de coefficients $\tilde{\beta}$ n'a pas ici une expression familière, mais on peut vérifier qu'il s'agit de l'estimateur classique et bien connu dit des "moindres carrés généralisés", rencontré fréquemment dans la théorie des modèles de régression linéaire. Il estime de manière optimale le vecteur de paramètres inconnus β du modèle.

Il est fondamental de noter qu'on a besoin de connaître les vraies moyennes par domaine $\bar{X}_d$. En pratique, cela signifie que les variables individuelles X_i sont disponibles dans un certain fichier exhaustif couvrant le champ de l'enquête (ce qui ne signifie pas que ces valeurs individuelles soient accessibles au statisticien en charge de l'estimation, qui ne dispose peut-être que des $\bar{X}_d$). Toutefois, il est possible que ce fichier ne soit pas la base de sondage et que l'information X_i mobilisée pour le calcul de $\tilde{\beta}$ provienne du fichier de collecte de l'enquête, exactement au même titre que Y_i. Dans ce cas, qui est courant en pratique, il convient de s'assurer que la variable X relève bien des mêmes concepts dans les deux sources (fichier exhaustif et fichier de collecte). Par exemple, calculer $\tilde{\beta}$ à partir d'une enquête Emploi où X représente le statut d'activité collecté dans l'enquête et former $\tilde{\bar{Y}}_d^H$ en utilisant des $\bar{X}_d$ qui représentent les statuts d'activité déclarés dans le recensement, s'avérerait très périlleux.

Formellement, l'estimateur de Henderson est constitué de deux éléments qui sont combinés grâce au coefficient réel γ_d. Le premier élément - situé entre les crochets de l'équation 12.11 - est un estimateur de circonstance dont l'interprétation est un peu compliquée mais qui a les mêmes performances statistiques que $\bar{y}_d$, estimateur construit à partir du sous-échantillon $s \cap d$: il a une variance d'échantillonnage fonction décroissante de n_d, donc *a priori* grande. Parce que cette caractéristique est associée aux estimateurs directs, mais qu'en même temps la présence du coefficient $\tilde{\beta}$, formé à partir de l'échantillon complet, ne permet pas de le qualifier rigoureusement d'estimateur direct, on parlera d'*estimateur pseudo direct*. Le second élément est un estimateur construit en multipliant le coefficient de régression $\tilde{\beta}$ par la vraie moyenne de la variable auxiliaire $\bar{X}_d$, ce qui intuitivement devrait donner une valeur proche de la vraie moyenne de la variable d'intérêt si le modèle est pertinent. Cet estimateur $\bar{X}_d^T \tilde{\beta}$ s'appelle *estimateur synthétique*. Ses propriétés statistiques sont totalement dépendantes de celles de $\tilde{\beta}$ puisque la moyenne $\bar{X}_d$ n'a aucun caractère aléatoire. Or on constate *de visu* que $\tilde{\beta}$ est constitué de termes impliquant l'intégralité de l'échantillon répondant s et non pas seulement la partie $s \cap d$. Cela lui confère par nature une grande stabilité, autrement dit une faible dépendance à l'échantillon répondant s. Si on ne considère que l'aléa de sondage, on peut donc dire que la composante synthétique offre une faible variance d'échantillonnage. La contrepartie de cette stabilité est l'existence d'un biais d'échantillonnage, qui peut être numériquement fort si le modèle est inadapté.

Le coefficient γ_d, qui est toujours compris entre 0 et 1, est un coefficient remarquable parce qu'il pondère de manière optimale (on rappelle qu'il minimise la MSE) les deux composantes distinguées, lesquelles ont des comportements tout à fait opposés en matière à la fois de biais et de variance d'échantillonnage. En cela, on dit que $\tilde{\bar{Y}}_d^H$ est un *estimateur composite* (ou *mixte*). La stratégie BLUP conduit donc à une expression de γ_d qui donne priorité à celle des deux composantes qui est la plus efficace. Prenons le cas où σ_τ^2 est petit, ce qui correspond à des effets locaux τ_d petits, autrement dit à un modèle performant puisqu'il fait porter le véritable caractère explicatif sur les variables auxiliaires maîtrisées X_i et non sur le terme résiduel "attrape-tout" τ_d. Dans de telles circonstances, on a tendance à faire confiance au modèle et à construire l'estimateur final en s'appuyant au maximum sur le modèle, c'est-à-dire sur l'estimateur synthétique. C'est effectivement ce qu'il advient puisque γ_d est petit. Prenons maintenant le cas où la taille d'échantillon répondant

n_d est grande. Un tel contexte donne confiance dans l'estimateur pseudo direct, qui n'utilise pas (ou très peu) le modèle, et donc par construction qui ne risque pas d'être déprécié par un manque de pertinence du modèle (l'estimateur pseudo direct a un biais faible, et ici une faible variance puisque n_d est grand) : c'est bien ce à quoi on aboutit puisque γ_d est grand, proche de 1.

Ajoutons qu'avec cette théorie, on est en mesure de prédire facilement chaque effet local τ_d. Après des calculs simples mais néanmoins fastidieux, on obtient :

$$\tilde{\tau}_d = \gamma_d \left(\bar{y}_d - \bar{x}_d \tilde{\beta} \right) \tag{12.14}$$

ce qui permet d'écrire l'estimateur de Henderson sous une forme plus intuitive

$$\tilde{\bar{Y}}_d^H = \bar{X}_d^T \tilde{\beta} + \tilde{\tau}_d. \tag{12.15}$$

Il reste encore une étape à franchir pour atteindre le stade opérationnel. En effet, l'estimateur BLUP $\tilde{\bar{Y}}_d^H$ a une expression complexe qui dépend à ce stade de certaines composantes du vecteur des paramètres du modèle δ introduit au 12.1.2. L'application de la stratégie BLUP a permis de produire des estimateurs $\tilde{\beta}$ de β qui ont réduit la dimension du problème : le vecteur de paramètres initiaux δ s'avère désormais limité aux composantes de variance, c'est-à-dire aux deux valeurs réelles σ_τ^2 et σ_e^2. On les résumera par le vecteur $\Sigma = (\sigma_\tau^2, \sigma_e^2)$. De fait, ce que l'on appelle - communément mais abusivement - l'estimateur $\tilde{\bar{Y}}_d^H$ n'en est pas un puisque cette expression n'est pas calculable et on devrait donc en toute rigueur le noter $\tilde{\bar{Y}}_d^H(\Sigma)$ et parler de "pseudo estimateur". Les composantes de Σ étant inconnues, il va falloir les estimer au moyen des données collectées. Une fois le paramètre Σ estimé par $\hat{\Sigma}$, on substituera $\hat{\Sigma}$ à Σ dans $\tilde{\bar{Y}}_d^H(\Sigma)$ pour aboutir à une nouvelle expression, soit $\tilde{\bar{Y}}_d^H(\hat{\Sigma})$, qui cette fois mérite bien le nom d'estimateur puisqu'elle est calculable. On donne à l'estimateur/prédicteur ainsi obtenu le nom de *Empirical Best Linear Unbiased Predictor* (EBLUP).

On estime fréquemment Σ par la méthode du maximum de vraisemblance. On dispose aussi d'une variante appelée maximum de vraisemblance restreint, qui est recommandable car elle réduit les biais des estimateurs lorsque les tailles d'échantillon sont modestes. Cette approche impose néanmoins une hypothèse supplémentaire sur la loi des variables aléatoires τ_d et e_i, que l'on considère presque systématiquement comme des variables suivant une loi de Gauss. Il n'existe pas d'expressions analytiques donnant $\hat{\sigma}_\tau^2$ et $\hat{\sigma}_e^2$, mais des algorithmes d'analyse numérique sont capables de produire des estimations conformes à la théorie. Partant de ces estimations, on obtient $\hat{\gamma}_d = \dfrac{\hat{\sigma}_\tau^2}{\hat{\sigma}_\tau^2 + \frac{\hat{\sigma}_e^2}{n_d}}$, puis :

$$\hat{\beta} = \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i X_i^T - \hat{\gamma}_d n_d \bar{x}_d \bar{x}_d^T \right) \right)^{-1} \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i Y_i - \hat{\gamma}_d n_d \bar{x}_d \bar{y}_d \right) \right) \tag{12.16}$$

et finalement l'estimateur EBLUP :

$$\hat{\bar{Y}}_d^H = \hat{\gamma}_d [\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T \hat{\beta}] + (1 - \hat{\gamma}_d) \bar{X}_d^T \hat{\beta}. \tag{12.17}$$

Noter qu'on peut éviter toute hypothèse portant sur la loi de Y_i en utilisant une méthode de type "méthode des moments", mais en contrepartie elle s'avère théoriquement moins efficace si la distribution des variables aléatoires τ_d et e_i est effectivement gaussienne.

12.2.2 Application au modèle linéaire individuel avec corrélation spatiale

La stratégie BLUP, avec son prolongement naturel qu'est l'EBLUP, s'applique exactement de la même façon dès lors que l'on introduit des corrélations spatiales entre les effets locaux. La différence avec le modèle linéaire standard réside uniquement dans les expressions mathématiques des différents estimateurs, qui sont évidemment beaucoup plus compliquées, mais les principes ne changent pas. Détailler l'expression formelle de l'estimateur de Henderson en présence de corrélations spatiales ne peut raisonnablement se faire qu'en utilisant des notations matricielles, qui sont très lourdes et sans valeur ajoutée didactique.

L'estimateur BLUP (ou EBLUP) reste une combinaison d'un estimateur direct et d'un estimateur synthétique, avec une pondération optimale calculée en tenant compte du contexte, selon la confiance que l'on peut accorder au modèle et selon la taille de l'échantillon répondant n_d. Le coefficient σ_τ^2 introduit dans l'équation 12.3 conserve un rôle essentiel, mais les calculs doivent désormais se faire en prenant également en compte le coefficient supplémentaire ρ, lequel règle l'intensité de la corrélation spatiale. Le paramètre de modèle à estimer est donc $\Sigma = (\rho, \sigma_\tau^2, \sigma_e^2)$.

Les algorithmes de calcul du maximum de vraisemblance (restreint le cas échéant) s'accommodent de l'introduction d'un paramètre supplémentaire, et ils produisent une estimation de ρ, de σ_τ^2 et de σ_e^2. La complexité de la structure de variances-covariances ne semble pas autoriser d'autres méthodes d'estimation de Σ que celle du maximum de vraisemblance ou du maximum de vraisemblance restreint.

12.2.3 Application au modèle de Fay et Herriot

Le modèle de Fay et Herriot revêt une grande importance car il est très utilisé en pratique. Dans de nombreuses circonstances, il s'ajuste bien et produit des estimations satisfaisantes, préférables aux estimations directes. Bien que l'on se place à un degré d'agrégation plus élevé que dans les modèles précédents, la stratégie BLUP se décline aussi dans le cadre de ce modèle. Dans l'expression de l'estimateur optimum de Henderson avec le modèle standard, les σ_e^2 ont évidemment disparu mais on trouve en revanche les valeurs des vraies variances d'échantillonnage par domaine ψ_d. Il est important de noter que dans la théorie standard, les vraies variances d'échantillonnage sont supposées connues. Ce n'est évidemment pas le cas en réalité, et il faut *in fine* remplacer les expressions théoriques ψ_d par les estimateurs $\hat{\psi}_d$ que l'on obtient en appliquant les méthodes traditionnelles de calcul de variance d'échantillonnage. À ce stade, il est conseillé de terminer par un lissage des valeurs $\hat{\psi}_d$. Cette opération protège contre la prise en compte d'estimations $\hat{\psi}_d$ anormalement faibles ou anormalement fortes, évitant ainsi un impact fortement dégradant sur la qualité des estimations finales par domaine. On aboutit à :

$$\tilde{\bar{Y}}_d^H = \gamma_d \hat{\bar{Y}}_d + (1-\gamma_d)\bar{X}_d^T \tilde{\beta} \tag{12.18}$$

avec

$$\begin{aligned} \gamma_d &= \frac{\sigma_\tau^2}{\sigma_\tau^2 + \hat{\psi}_d} \\ \tilde{\beta} &= \left[\sum_{d=1}^{D} \frac{\bar{X}_d . \bar{X}_d^T}{\sigma_\tau^2 + \hat{\psi}_d}\right]^{-1} \cdot \left[\sum_{d=1}^{D} \frac{\bar{X}_d . \hat{\bar{Y}}_d}{\sigma_\tau^2 + \hat{\psi}_d}\right]. \end{aligned} \tag{12.19}$$

L'estimateur $\tilde{\bar{Y}}_d^H$ conserve une forme composite et la stratégie BLUP produit la pondération γ_d idéale, partagée entre l'estimation directe $\hat{\bar{Y}}_d$, indépendante du modèle mais instable, et l'estimation synthétique $\bar{X}_d^T \tilde{\beta}$, qui est pour sa part totalement dépendante du modèle, mais en contrepartie peu

sensible à la composition de l'échantillon répondant. Dans de rares circonstances, lorsqu'il s'agit d'estimer une proportion, il peut arriver que l'estimation $\tilde{\bar{Y}}_d^H$ sorte de l'intervalle $[0,1]$. Dans ce cas, il est nécessaire d'adapter le modèle initial.

Si on introduit des corrélations spatiales, les expressions ci-dessus évoluent en conséquence - en se compliquant considérablement - mais aucun des grands principes n'est modifié. Dans tous les cas, avec ou sans corrélations spatiales, le logiciel sait produire l'estimateur σ_τ^2 par maximum de vraisemblance (éventuellement restreint), dont on déduit immédiatement $\hat{\gamma}_d$ et $\hat{\beta}$, puis l'estimateur final EBLUP $\hat{\bar{Y}}_d^H$. Noter qu'en l'absence de corrélation spatiale, il existe d'autres méthodes d'estimation du paramètre σ_τ^2 que le maximum de vraisemblance.

12.2.4 Stratégie pour les modèles non linéaires

Le monde des modèles non linéaires est techniquement beaucoup plus compliqué que celui des modèles linéaires. En particulier, la stratégie BLUP n'est pas directement adaptée à ce contexte parce qu'elle ne trouve pas de solution mathématique satisfaisante. Elle reste néanmoins une technique de base et c'est pourquoi l'une des façons de traiter les modèles non linéaires, comme le modèle logistique ou le modèle de Poisson par exemple, consiste à les remplacer par des modèles approchés ayant une structure linéaire. Ce qu'est un modèle approché renvoie à une théorie compliquée mais néanmoins opérationnelle. C'est en partant du modèle linéaire approché que l'on applique la stratégie BLUP.

Le modèle d'origine utilise ou non des corrélations spatiales. Au modèle linéaire approché, on applique alors les développements présentés dans les parties qui précèdent.

Cela étant, l'approche la plus convaincante consiste à utiliser une stratégie mieux adaptée à ce contexte non linéaire, comme la stratégie *Empirical Bayes* qui produit des estimations optimales, ou la stratégie *Hierarchical Bayes* qui correspond à l'approche bayésienne classique.

12.3 La qualité des estimateurs

L'approche par modèle aura pour conséquence bien évidente de faire dépendre l'estimation du choix du modèle et se posera donc la question de la pertinence du modèle retenu. En effet, la simplification a un coût en termes de qualité et on peut se demander jusqu'à quel point ce modèle est correctement représentatif de la réalité.

De quoi parle-t-on ?

En matière d'appréciation de la qualité des estimations sur petits domaines, il est plus que jamais nécessaire de préciser le concept de qualité. En effet, le contexte souffre d'une complication toute particulière due à la coexistence d'aléas de natures différentes : d'une part l'aléa de sondage qui décide de la composition de l'échantillon, d'autre part l'aléa du modèle qui traite la variable d'intérêt comme une variable aléatoire. Or on peut apprécier la qualité en prenant en compte ou non l'aléa de modèle.

Sans aléa de modèle, il s'agit de l'approche classique du statisticien d'enquête placé en population finie et traitant de variables individuelles déterministes. De ce point de vue, la situation est extrêmement simple : tous les estimateurs "petits domaines" présentés jusqu'ici sont biaisés. C'est la conséquence naturelle de l'absence de prise en compte des poids de sondage (lorsque l'échantillonnage n'est pas à probabilités égales en tout cas), ou d'une prise en compte seulement partielle de ces poids. Par exemple, dans le modèle linéaire standard individuel, la pondération reflétant l'échantillonnage est systématiquement absente. Dans le modèle de Fay et Herriot, on la retrouve certes dans la composante directe $\hat{\bar{Y}}_d$ mais aucunement dans la partie synthétique $\bar{X}_d^T \hat{\beta}$. En revanche, le modèle apporte un avantage déterminant en termes de variance d'échantillonnage : en effet, les paramètres β qui sont estimés mobilisent l'intégralité de l'échantillon répondant et

c'est pourquoi ils ont une faible variance d'échantillonnage. L'effet aléatoire local estimé $\hat{\tau}_d$ est pour sa part instable, mais si le modèle est bien adapté, il sera numériquement petit et donc sa variance sera d'influence limitée. L'estimateur de Henderson devrait finalement être de variance d'échantillonnage limitée et *a priori* inférieure à celle de l'estimateur direct si le modèle a un bon pouvoir explicatif.

Lorsqu'on prend en compte l'aléa de modèle, si le modèle est linéaire, par construction l'estimateur BLUP est sans biais. Le passage à l'EBLUP n'occasionne que des biais négligeables. Si le modèle n'est pas linéaire, le contexte est beaucoup plus compliqué mais on s'attend à obtenir des biais modestes.

12.3.1 Un processus itératif

L'appréciation de la qualité peut se concevoir selon un mécanisme cyclique (figure 12.1).

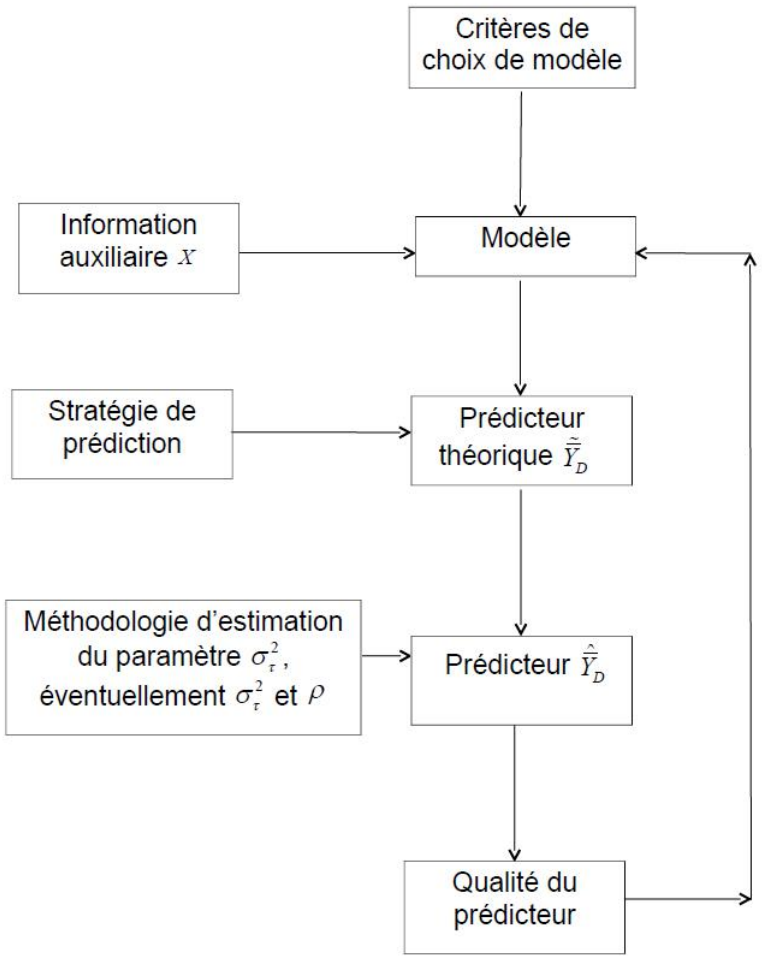

FIGURE 12.1 – Schéma du processus itératif d'appréciation de la qualité des estimateurs

Disposant d'une part de certains critères de sélection de variables explicatives, disposant d'autre part d'un ensemble de variables auxiliaires X potentiellement explicatives de Y, on ajuste un modèle. À ce stade, on dispose d'outils statistiques pour évaluer la qualité de cet ajustement. Ce modèle, associé à une stratégie de prédiction, produit un estimateur théorique $\tilde{\bar{Y}}_D$. Ce dernier est dépendant de paramètres participant à la définition du modèle (au moins le paramètre σ_τ^2, le σ_ε^2 s'il y a lieu et

le ρ s'il y a corrélation spatiale). Ces paramètres sont estimés par une méthode *ad hoc*. À la fin du cycle, on évalue la qualité du prédicteur final (biais, MSE ; voir sections 12.3.3 et 12.3.4). Si elle n'est pas acceptable, on enclenche un nouveau cycle en s'interrogeant de nouveau sur la pertinence du modèle, voire sur celle de la stratégie de prédiction ou encore sur celle de l'estimation des paramètres du modèle. L'appréciation de la qualité passe aussi par une vérification de la pertinence des hypothèses de loi du modèle, s'il y a lieu. C'est pourquoi on vérifiera le caractère gaussien des effets locaux estimés $\hat{\tau}_d$ dès lors qu'une technique de maximum de vraisemblance (restreint ou non) a été utilisée.

12.3.2 Le problème du biais

Les statisticiens d'enquête ont parfois des réticences à utiliser un estimateur dépendant d'un modèle (bien que ce soit incontournable pour traiter la non-réponse). Leur crainte essentielle est celle d'un biais substantiel si on s'en tient à l'aléa d'échantillonnage. Ce risque est inévitable puisque le modèle simplifie, et donc dénature, la réalité. L'important n'est pas d'échapper au biais mais d'obtenir un biais limité qui soit plus que compensé par le gain en termes de variance. Sauf si on travaille sur des populations artificielles, les calculs de biais dûs à l'échantillonnage ne sont pas réalisables, mais on peut utiliser deux outils simples qui permettent d'apprécier la situation, sans toutefois fournir la moindre preuve.

Le premier outil est purement graphique et consiste à construire un nuage de points où chaque point représente un des D domaines traités. Sur l'un des axes, on porte l'estimation directe (donc obtenue sans modèle), sur l'autre axe on porte l'estimation "petits domaines" (donc issue d'un modèle). Si le nuage de points ainsi formé n'est pas symétrique par rapport à la droite $y = x$ (première bissectrice), on soupçonne fortement un biais dû à l'échantillonnage. Néanmoins, il n'y a pas de fatalité (penser à la situation, évidemment idéalisée, d'un modèle traduisant une réalité où toutes les moyennes par domaine sont égales). La situation réciproque est plus convaincante : si le nuage de points est symétrique, il est probable qu'il n'y aura pas de biais significatif dû à l'échantillonnage. Le plus souvent, en pratique on observe un nuage incliné par rapport à la première bissectrice et dont la projection sur l'axe représentant l'estimation "petits domaines" est plus réduite que la projection sur l'axe représentant l'estimation directe. Ce phénomène a reçu le nom de *shrinkage*, et il est donc plutôt annonciateur d'un biais dû à l'échantillonnage. Il traduit une forme de concentration (*a priori* excessive) des estimations. Elle découle mécaniquement du modèle simplificateur, qui a un effet de normalisation et qui a donc plus ou moins tendance à uniformiser les estimations par domaine. Nous insistons sur le fait que cette approche graphique n'offre aucune preuve mais crée seulement des suspicions. En pratique, parce qu'elle ne peut pas traduire fidèlement la réalité, toute modélisation crée fatalement un biais théorique dû à l'échantillonnage et la symétrie éventuelle du nuage de points ne fait qu'indiquer le caractère probablement modeste de ce biais.

La seconde technique est encore plus simple et plus intuitive : il s'agit de sommer les estimations des totaux $\hat{T}_d^H$ obtenues sur les D petits domaines et de comparer le résultat à l'estimation directe du total $\hat{T}$ portant sur la population complète, celle qui résulte de la théorie classique des sondages. En effet, cette dernière est par construction sans biais (l'aléa est ici exclusivement l'aléa de sondage) : s'il existe un biais dû au modèle et que ce biais a un caractère quelque peu systématique, on constatera un écart entre les deux valeurs. En revanche, un biais sans composante systématique n'est pas détectable puisque des compensations peuvent se produire au moment de la sommation.

On a coutume d'exploiter, au profit de la qualité, l'écart dont il est question ci-dessus. En effet, si $\hat{T}_d^H$ est l'estimateur "petits domaines" du vrai total T_d dans le domaine d, si $\hat{T}$ est l'estimateur sans biais direct issu de l'échantillon global répondant s représentant la population complète U, on

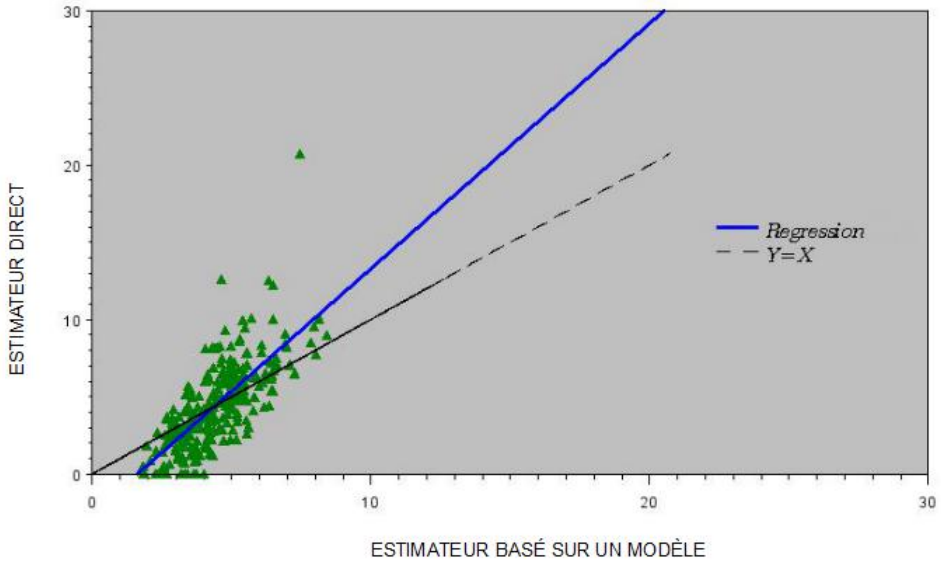

FIGURE 12.2 – Exemple de relation entre l'estimation directe et l'estimation petits domaines

adopte très souvent l'estimation finale suivante :

$$\hat{\hat{T}}_d^H = \hat{T}_d^H \frac{\hat{T}}{\sum_{d=1}^{D} \hat{T}_d^H} \tag{12.20}$$

qui permet de caler sur $\hat{T}$ l'estimation du total dans U. Cette opération reçoit le nom de *benchmarking* et contribue à limiter le biais de $\hat{\hat{T}}_d^H$ tout en assurant une diffusion cohérente.

Par ailleurs, il est toujours intéressant de procéder à une cartographie des estimations de moyenne par domaine $\hat{\hat{Y}}_d^H$, laquelle permet de vérifier visuellement la cohérence du système d'estimation dans son ensemble : normalement, à deux domaines ayant des caractéristiques proches et voisins sur une carte, devraient correspondre deux moyennes estimées $\hat{\hat{Y}}_d^H$ semblables (concrètement, les couleurs représentatives de leurs valeurs respectives devraient se situer dans la même gamme).

12.3.3 L'erreur quadratique moyenne

Dans un environnement où les biais sont possibles, ou probables, ou encore inévitables, le bon concept d'erreur est celui d'erreur quadratique moyenne (ou MSE). Cet indicateur désigne l'espérance du carré de la différence entre l'estimateur et le paramètre. En prenant en compte à la fois l'aléa de sondage et l'aléa de modèle, le cadre théorique offert par le modèle permet d'obtenir l'expression de la MSE et ensuite de l'estimer sans biais ou presque. L'expression de la MSE et son estimation sont très compliquées, même avec le modèle linéaire, et le calcul est donc confié à un logiciel. Néanmoins, en l'absence de corrélation spatiale, on peut vérifier, si le nombre de domaines D est grand, que le terme numériquement le plus important dans l'estimation de la MSE de $\hat{\hat{Y}}_d^H$ est $\hat{\gamma}_d \hat{\psi}_d$ pour le modèle de Fay et Herriot, et $\hat{\gamma}_d \frac{\hat{\sigma}_e^2}{n_d}$ pour le modèle linéaire individuel standard. Concernant tous ces calculs d'erreur, les résultats obtenus supposent fondamentalement que le modèle est spécifié de manière parfaitement conforme à la réalité (le modèle peut être qualifié d'"exact"). Cela n'est certainement pas vrai en toute rigueur ! L'introduction d'une corrélation spatiale crée évidemment une difficulté technique supplémentaire, mais la théorie générale permet

d'aboutir, ce qui ne signifie pas que les outils informatiques actuellement accessibles soient en capacité de la mettre en œuvre. Notez que dans certaines circonstances particulièrement favorables, on peut disposer d'une source externe qui fournit la vraie valeur du paramètre (par exemple à l'issue d'un recensement). Cela permet d'apprécier directement l'erreur d'estimation commise.

12.4 Mise en œuvre avec R

■ **Exemple 12.1 — Diffusion du recensement sur des carreaux.** L'Office statistique de l'Union européenne EUROSTAT souhaite produire en 2021 des statistiques (sexe, tranche d'âge, activité, etc.) portant sur la population complète de chaque pays membre au niveau de carreaux d'un kilomètre de côté. De surcroît, en France, l'Insee a pour ambition de diffuser les données du Recensement de la Population (RP) sur des carreaux de quelques centaines de mètres de côté. Depuis 2004, le recensement de la France est effectué par sondage dans les communes de plus de 10 000 habitants [2]. Dès lors, la superficie ciblée contient trop peu d'observations pour que l'on puisse obtenir de bons estimateurs directs des paramètres d'intérêt. C'est la raison pour laquelle l'estimation *petits domaines* pourrait être une technique statistique appropriée pour l'exploitation de ce type de données.

L'introduction d'une corrélation spatiale dans ce contexte permet de traduire le phénomène de continuité naturelle des caractéristiques sociodémographiques d'individus peuplant des zones géographiquement contiguës. En effet, passant d'un carreau quelconque aux carreaux voisins, on ne peut pas raisonnablement prétendre *a priori* qu'il y ait indépendance entre les comportements des unités statistiques – logements comme individus – qui les composent. ■

Le package *sae* de R permet de calculer les estimations petits domaines aux niveaux "domaine" et "individu", dans le cas de modèles respectivement sans et avec prise en compte de l'autocorrélation spatiale. Implémenté par Molina et Marhuenda, ce package a fait l'objet d'une présentation dans *The R Journal* (MOLINA et al. 2015).

Les principales fonctions qui ont été utilisées pour traiter les données du RP à partir d'un modèle formulé au niveau domaine sont issues du package *sae* : il s'agit de `eblupFH()`, `eblupSFH()`, `mseFH()` et `mseSFH()`.

Pour produire des estimations à partir d'un modèle au niveau individu, on a utilisé les fonctions `eblupBHF()` et `pbmseBHF()` du package *sae*, ainsi que la fonction `corrHLfit()` du package *spaMM*.

Modélisation au niveau domaine : les fonctions de base eblupFH() et mseFH()

Les premières estimations s'appuient sur le modèle de Fay et Herriot sans autocorrélation spatiale. La fonction `eblupFH()` permet d'obtenir en sortie :

i) les estimations de Fay et Herriot pour chaque domaine ;
ii) une estimation de la variance σ_τ^2 de l'effet aléatoire propre aux domaines.

En supplément, la fonction `mseFH()` produit le calcul des erreurs quadratiques moyennes associées à chaque estimation (voir partie 12.3.3).

Les arguments de ces fonctions sont les mêmes. La syntaxe standard est la suivante :

```
mod_FH <- eblupFH(formula = Y ~ X1+...+Xp, vardir = varech, method =  ,
   maxiter = m, precision = e, B = 0, data = )
```

En premier lieu, le paramètre `formula` précise la variable d'intérêt Y ainsi que les variables explicatives retenues $X_1, ..., X_p$. Les valeurs numériques de toutes ces variables doivent être contenues dans un tableau qui associe une ligne à chaque domaine, précisé dans l'argument `data`. Les

2. Il est exhaustif dans les communes de moins de 10 000 habitants

paramètres $\hat{\gamma}_d$ impliqués dans les estimateurs de Henderson sont calculés à l'aide des variances d'échantillonnage (estimées) par domaine $\hat{\psi}_d$, disponibles dans une variable que l'on précise dans l'argument `vardir`.

L'estimation de l'unique paramètre de variance du modèle est obtenue par une technique *ad hoc*. Concrètement, on utilise une méthode itérative qui devrait converger vers la valeur σ_τ^2. Les paramètres `maxiter` et `precision` sont des paramètres techniques (définis soit par l'utilisateur soit par défaut) qui régulent ce processus itératif. À chaque itération, l'algorithme calcule une estimation de σ_τ^2. Le rôle du paramètre `precision` est le suivant : dès que la différence entre deux valeurs obtenues consécutivement est inférieure à celui-ci (e dans notre exemple), l'algorithme s'arrête. Sinon, tant que le nombre maximal d'itérations `maxiter` n'est pas atteint, les itérations se poursuivent. La sortie indique si l'algorithme a convergé ou non. La méthode est également à préciser. On peut choisir parmi trois méthodes, dont la méthode du maximum de vraisemblance et celle du maximum de vraisemblance restreint (respectivement `method = "ML"` et `method = "REML"`). La troisième méthode (`method = "FH"`) est une méthode de type méthode des moments.

On attire l'attention du lecteur sur la nécessité de ne pas rajouter dans les régresseurs de variables indicatrices repérant les domaines. En effet, la constante faisant déjà partie des régresseurs standards, cette pratique conduirait à former une matrice non inversible. La commande suivante conduit donc à un échec :

```
mod_FH <- eblupFH(formula = Y ~ X1+...+Xp+as.factor(Carreau) , vardir =
    varech, method =  , maxiter = m, precision = e, B = 0, data = )
```

Cas d'une corrélation spatiale au niveau domaine : les fonctions eblupSFH() et mseSFH()

Le logiciel estime un modèle SAR (voir partie 12.1.3) du type :

$$\tau = \rho.A.\tau + u. \tag{12.21}$$

Les paramètres de ces fonctions sont les mêmes que ceux des fonctions précédentes, à cela près que l'on doit en plus préciser la matrice de proximité $\mathbf{A}$ (la matrice des coefficients $\alpha_{i,j}$; voir partie 12.1.3). La commande R est de la forme :

```
mod_SFH <- eblupSFH(formula = Y ~ X1+...+Xp, vardir = varech, method =  ,
    maxiter = m, precision = e, proxmat = A, B = 0, data = )
```

La matrice de proximité est décrite par le paramètre `proxmat`. Elle a des lignes standardisées en ce sens où la somme des éléments de chaque ligne vaut toujours 1.

Un processus itératif analogue à celui du cas sans corrélation spatiale permet de calculer :

i) les estimations de Fay et Herriot pour chaque domaine ;
ii) une estimation de la variance de l'effet aléatoire propre aux domaines ;
iii) une estimation du paramètre d'autocorrélation spatiale ρ.

Modélisation au niveau individu, sans corrélation spatiale : les fonctions eblupBHF() et pbmseBHF()

En utilisant une modélisation au niveau individu, la fonction `eblupBHF()` du package *sae* permet de calculer les estimations directes et les estimations "petits domaines" sans corrélation spatiale. La syntaxe est la suivante :

```
mod_BHF <- eblupBHF(formula = Y ~ X1+...+Xp, dom = ,
meanxpop = , popnsize = Popn, data = adr_est)
```

Elle utilise le paramètrage suivant : `formula` pour l'expression formelle du modèle, `dom` pour désigner la variable identifiant les domaines, `popnsize` pour la taille de la population N_d dans chaque domaine, et `meanxpop` pour les moyennes des variables explicatives $\bar{X}_d$ calculées dans la population complète du domaine. Le paramètre `data` désigne la table des données.

La fonction `pbmseBHF()` estime les erreurs (MSE) des estimateurs "petits domaines" par une technique de *bootstrap*. Les paramètres de cette fonction sont les mêmes que pour la fonction précédente, avec en supplément le nombre de ré-échantillonnages du *bootstrap* défini par le paramètre `B` (B=1000 par exemple).

```
mse_BHF <- pbmseBHF(formula = Y ~ X1+...+Xp, dom = ,
meanxpop = , popnsize = , B = 1000, data = )
```

Prise en compte de la corrélation spatiale dans le modèle individuel

Le package *spaMM* peut être utilisé pour prendre en compte la corrélation spatiale. Il peut gérer plusieurs types de modèles, et en particulier le modèle de Poisson (voir partie 12.1.4). La fonction `corrHLfit()` traite le modèle de Poisson au niveau individu avec corrélation spatiale.

```
library(spaMM)
mod_spa <- corrHLfit(formula = Y ~ X1+...+Xp+Matern(1|x+y),
HLmethod = "REML", family = "poisson", ranFix = list(nu=0.5), data = )
```

Concernant le paramètrage de cette fonction, `formula` désigne l'expression formelle du modèle. La composante *Matern(1|x+y)*, propre à la fonction utilisée, permet de prendre en compte les coordonnées x et y des domaines (ici les centres des carreaux), qui doivent donc être présentes dans la table des données, afin de calculer les distances qui interviennent dans la fonction de corrélation spatiale. Par ailleurs, `HLmethod` précise la méthode d'estimation des paramètres de variance et de corrélation spatiale (ici le maximum de vraisemblance restreint), et `family` choisit la distribution de la variable d'intérêt (ici une loi de Poisson). La forme fonctionnelle de la corrélation spatiale peut être choisie parmi une famille paramétrée de fonctions compliquées appelées fonctions de Matérn. Le paramètre `ranFix` précise le paramétrage de cette famille de fonctions. Si on indique `list(nu=0.5)`, on obtient la forme exponentielle de l'équation 12.3, qui est l'expression traditionnellement utilisée - à cela près que le paramètre estimé est $\frac{1}{\rho}$ et non directement ρ. Le paramètre `data` désigne la table des données.

On obtient en sortie, entre autres, les coefficients estimés du modèle, dont le coefficient ρ intervenant dans la corrélation spatiale exponentielle (en fait son inverse si on se réfère à la formulation 12.3), les prédictions optimales $\hat{\tau}_d$ des effets locaux aléatoires, et la variance estimée de l'effet aléatoire $\hat{\sigma}^2_\tau$.

Conclusion

L'estimation sur petits domaines est fondée sur l'utilisation de modèles stochastiques. C'est la contrepartie d'une certaine pauvreté de l'information collectée *via* l'échantillon recoupant le domaine lorsque celui-ci est de petite taille. En effet, pour limiter l'imprécision, il n'y a pas de miracle, il faut bien formuler des hypothèses portant sur l'intégralité de la population et qui compensent le manque d'information obtenue au niveau local. Les modèles font intervenir explicitement des effets géographiques locaux dont l'interprétation est délicate en ce sens qu'on peut toujours considérer qu'il s'agit d'un pis-aller pour camoufler une prise en compte insuffisante d'effets fixes explicatifs du phénomène étudié. Au fond, la première question est bien de savoir jusqu'à quel point l'effet purement géographique existe. Par ailleurs, ces modèles, quels qu'ils soient, créent toujours du biais par rapport à l'aléa de sondage. L'objectif essentiel est d'en limiter l'ampleur, plus que de mesurer la variance d'échantillonnage, qui devient un objectif secondaire pour le statisticien d'enquête. On

dispose certes d'outils statistiques pour apprécier la qualité de l'ajustement d'un modèle, mais cela ne garantit pas l'adéquation du modèle retenu à la situation particulière d'un domaine donné, qui peut être très spécifique sans que le statisticien ne s'en aperçoive. Il n'existe pas d'estimation fiable du biais d'échantillonnage, et on ne dispose actuellement que de quelques outils qualitatifs, plus ou moins convaincants et qui conduisent seulement à l'appréciation d'une situation globale. De façon générale, la théorie des modèles linéaires (*Linear Mixed Models* ou LMM) est beaucoup plus simple que celle des modèles non linéaires traditionnellement mis en œuvre (*Generalized Linear Mixed Models* ou GLMM), qui restent vraiment difficiles d'accès. La présence de corrélation spatiale complique toujours le contexte et se pose alors la question de la disponibilité du code informatique pour procéder aux estimations. Le développement de R est prometteur et on devrait aller à l'avenir vers un élargissement de la gamme des modèles acceptant de la corrélation spatiale.

Références - Chapitre 12

BATTESE, George E, Rachel M HARTER et Wayne A FULLER (1988). « An error-components model for prediction of county crop areas using survey and satellite data ». *Journal of the American Statistical Association* 83.401, p. 28–36.

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.

COELHO, Pedro S et Luıs N PEREIRA (2011). « A spatial unit level model for small area estimation ». *REVSTAT–Statistical Journal* 9.2, p. 155–180.

FAY III, Robert E et Roger A HERRIOT (1979). « Estimates of income for small places : an application of James-Stein procedures to census data ». *Journal of the American Statistical Association* 74.366a, p. 269–277.

MOLINA, Isabel et Yolanda MARHUENDA (2015). « sae : An R package for small area estimation ». *R Journal, in print*.

PRATESI, Monica et Nicola SALVATI (2008). « Small area estimation : the EBLUP estimator based on spatially correlated random area effects ». *Statistical methods and applications* 17.1, p. 113–141.

RAO, John NK (2015). *Small-Area Estimation*. Wiley Online Library.