JMS Journées de méthodologie statistique de l'Insee 2018

L'estimation des taux d'artificialisation et d'imperméabilisation en France

Patrick SILLARD(*)[1]

(*)INSEE/DMCSI

patrick.sillard@insee.fr

Mots-clés. artificialisation, imperméabilisation, géostatistique, autocorrélation, processus spatial, simulation.
JEL Codes : R1 ; C1.

Version provisoire

Résumé

Cette étude vise à revisiter les incohérences apparentes portées par les différents chiffrages de taux de surfaces imperméabilisées en France métropolitaine, issus en particulier de l'enquête Teruti-Lucas d'une part et, d'autre part, de la grille raster CORINE Land Cover à haute résolution établie par interprétation automatique d'image satellitaires. La première permet de chiffrer ce taux à 4,6% du territoire métropolitian, tandis que le second permet de l'estimer à 2,8%. Ces écarts semblent, à première vue, trop importants pour ne pas découler de biais liés à des erreurs systématiques de mesure ou à des différences de concepts mesurés.

La notion de variance de mesure est discutée, en lien avec l'adoption (ou non) d'un modèle pour le processus d'imperméabilisation. En particulier, grâce à l'identification du processus stochastique générateur de l'imperméabilisation, on établit que le phénomène d'imperméabilisation est rare et caractérisé par une très forte autocorrélation spatiale à mémoire longue (i.e. persistante). La précision des taux calculés et l'amplitude des intervalles de confiance s'en trouvent affectées. Ne pas en tenir compte conduit à des incohérences dans les taux estimés à partir des différentes sources, incohérences susceptibles de disparaître lorsqu'on fait l'hypothèse que l'artificialisation découle de processus générateur identifié.

On montre en particulier que, selon l'approche retenue de modéliser ou pas l'imperméabilisation comme découlant d'un processus générateur, les taux d'imperméabilisation issus de l'enquête Teruti et de la couche imperméabilisation de CORINE peuvent apparaître comme compatibles statistiquement et beaucoup plus incertaines que ce qui figure dans la littérature sur le sujet.

1. Insee/DMCSI. Cette étude a débuté tandis que Patrick Sillard faisait partie du service statistique du ministère de l'écologie et du développement durable (Service de la donnée et des études statistiques). L'auteur remercie les participants au séminaire d'études du Sdes d'avril 2017, en particulier Vincent Loonis qui y a discuté une première version de ce texte et Marlène Kraszewski pour sa relecture attentive.

Table des matières

1 Introduction

L'artificialisation et l'imperméabilisation des sols sont des phénomènes abondamment étudiés car l'une comme l'autre modifie l'espace et les habitats des espèces naturelles et sont reconnues comme préjudiciables à l'environnement (voir par exemple Chakir & Madignier (2006) et Pageaud & Carré (2009)). L'artificialisation caractérise le passage, causé par l'homme, d'un état des sols naturel ou agricole à un état non naturel et non agricole. L'imperméabilisation correspond à une artificialisation qui aboutit à des sols artificiels imperméables. Cette dernière est donc une forme particulière d'artificialisation.

L'artificialisation comme l'imperméabilisation sont le résultat des pressions anthropiques qui, à mesure que la démographie et l'économie se développent, poussent à consommer davantage d'espaces naturels et agricoles au profit des implantations humaines. Dans les pays comme la France, les taux d'artificialisation ou d'imperméabilisation sont mesurés par différentes sources et à différentes échelles. On peut citer en particulier, s'agissant de couvertures exhaustives, les bases géographiques de l'IGN (Occupation du sol à grande échelle), du référentiel parcellaire graphique du ministère de l'agriculture, de l'administration en charge du cadastre (fichiers MAJIC) et la base pan-européenne sur l'occupation du sol (CORINE Land Cover). Ces bases ont la particularité d'être constituées de couches géographiques décrivant les contours des objets géographiques qu'elles restituent. Elles sont élaborées avec des observations dont la résolution est variable ce qui aboutit à une forme d'échelle de restitution des détails, même si par construction les bases géographiques ne comportent pas d'échelle. Ainsi, la base CORINE Land Cover est élaborée à partir de fonds d'images satellitaires dont la résolution vise à élaborer des cartes à moyenne échelle. Dans ces conditions, la précision des détails est beaucoup moins fine que lorsque les observations s'appuient sur des relevés de terrain à grande échelle, comme dans la cas des bases cadastrales ou celles de l'IGN.

Plus récemment, des productions nouvelles permettent de compléter ce panorama des sources exhaustives, notamment à l'aide d'images satellitaires interprétées par des algorithmes de reconnaissance automatique, fondés sur un mécanisme d'apprentissage. Ces modes de traitements de l'information aboutissent à des images pixelisées du terrain, chaque pixel (de taille régulière) étant affecté d'une variable caractérisant un niveau d'imperméabilisation. Une telle couche d'information est aujourd'hui disponible en accompagnement des produits CORINE Land Cover. Par rapport à une source comme CORINE Land Cover - base géographique, la couche d'imperméabilisation à haute résolution (notée ci-après CLC-HR), comme son nom l'indique, offre un niveau de résolution des détails, intermédiaire entre les bases de données géographiques à grande échelle et les bases à moyenne échelle (1/100 000$^{\text{ème}}$). Ces informations sont également appréciées pour leur faible coût de production, par rapport aux bases géographiques qui nécessitent, à un moment ou à un autre du processus de production, qu'un opérateur dessine les contours d'objets.

A côté des sources géographiques qui restituent une information exhaustive (i.e. une information qui couvre l'ensemble du territoire), se sont développées depuis de nombreuses années, aux niveaux français et européens, des enquêtes sur l'occupation du sol (Teruti en France, Lucas en Europe). Ces enquêtes s'appuient sur un sondage de points géographiques, généralement selon un plan de sondage systématique (points spatialement équirépartis). Ces points font l'objet d'une observation *in situ* et la résolution des détails relevés par l'enquête est de l'ordre de la précision des bases de données géographiques les plus précises. Donc en termes de résolution des détails restitués, les enquêtes sur l'occupation des sols et les bases géographiques à grande échelle sont comparables.

Comme souvent, des indicateurs d'artificialisation et d'imperméabilisation sont fondés sur ces sources. Même si ces indicateurs ont vocation à traiter de statistiques sur des phénomènes de même nature (i.e. selon des nomenclatures concordantes), en pratique les moyennes obtenues diffèrent bien au-delà des écarts de précision attendues sous l'hypothèse d'absence de biais des estimations produites. Par exemple, le taux d'artificialisation en France mesuré par CORINE Land Cover est de 5,8% en 2012 (Janvier, Nirascou & Sillard 2016), tandis qu'il s'élève à 9,3% lorsque sa mesure s'appuie sur la source Teruti (Fontes-Rousseau & Jean 2015). De même [2], le taux d'imperméabilisation obtenu à l'aide de l'enquête Teruti est de 4,6%, tandis qu'il est estimé à 2,8% avec les couches CLC-HR.

Différents facteurs peuvent expliquer ces écarts : la nomenclature de restitution de l'occupation en est un ; la résolution spatiale des observations élémentaires en est un autre. Cependant, à notre connaissance, il n'existe pas de modèle explicatif de l'ordre de grandeur de tels écarts. En particulier, la compréhension fine impose de décomposer le processus de passage entre le phénomène étudié (artificialisation ou imperméabilisation) tel qu'il se manifeste sur le terrain et l'indicateur qui en découle. Un modèle statistique cohérent est donc requis. Ce document a pour but de proposer un modèle statistique pour les phénomènes d'imperméabilisation permettant de réconcilier ces différentes mesures. La démarche peut se généraliser à l'artificialisation qui est globalement de même

2. Voir annexe A pour le détail du calcul

nature que l'imperméabilisation. Une quantification des écarts d'estimation des taux basée sur ce modèle est proposée.

Dans une première partie, nous établissons quelques résultats empiriques sur les données dont nous disposons. Puis, nous tentons de caractériser les processus statistiques d'imperméabilisation. Ceci afin d'être en mesure de simuler ces processus pour examiner plus en détail la nature des statistiques de taux que nous pouvons établir à partir de ces données. Enfin, dans une dernière partie, nous revenons sur le niveau de cohérence des statistiques que l'on peut fonder sur les sources Teruti-Lucas et CLC-HR, à la lumière des développements précédents.

2 Les données utilisées et quelques statististiques descriptives qui en découlent

L'imperméabilisation [3] peut être modélisée comme un processus stochastique à support spatial. Ce processus est intrinsèquement continu sur son support, c'est-à-dire qu'à tout point [4] $\mathbf{M}$ de l'espace géographique bidimensionnel (coordonnées planes), on peut associer une variable aléatoire $Z(\mathbf{M})$ caractérisant l'artificialisation en ce point. On conviendra qu'en ce point élémentaire, $Z(\mathbf{M})$ vaut 1 si l'espace est totalement imperméable du fait d'une construction humaine et 0, s'il est totalement exempt de construction humaine imperméable. Pour permettre une description plus fine qui rende compte de l'échelle d'analyse, on supposera que Z peut prendre toutes les valeurs de l'intervalle $[0, 1]$ selon qu'au point considéré, le sol est plus ou moins imperméable du fait des éventuelles constructions humaines. Une façon de justifier cette représentation est de considérer que toute valeur de $Z(\mathbf{M})$ est, *in fine*, une moyenne d'une variable élémentaire binaire z à une échelle de petitesse à laquelle n'est pas sensible le système de mesure utilisé. Ainsi [5]

$$Z(\mathbf{M}) = \int_{\mathbf{m}\in\mathscr{V}(\mathbf{M})} z(\mathbf{m})d\mathbf{m} \Big/ \int_{\mathbf{m}\in\mathscr{V}(\mathbf{M})} d\mathbf{m}$$

où $\mathscr{V}(\mathbf{M})$ est un voisinage (dans un sens à préciser ultérieurement) du point $\mathbf{M}$. Avec cette représentation, $Z(\mathbf{M})$ apparaît comme une moyenne arithmétique. Sous certaines hypothèses que nous n'explicitons pas, il est possible de se placer dans les conditions, pour Z, d'application des théorèmes de convergence usuels (type loi des grands nombres). On considèrera alors que Z est une variable continue, c'est-à-dire qu'en tout point $\mathbf{M}$, la variable $Z(\mathbf{M} + d\mathbf{m})$ a pour limite [6] $Z(\mathbf{M})$ quand $d\mathbf{m}$ tend vers $\vec{0}$. Quoiqu'il en soit, nous considèrerons que nous ne travaillons pas sur la variable binaire z mais sur une variable continue, Z, à valeurs dans $[0, 1]$.

Avec ce modèle simple en tête, il est instructif d'examiner quelques statistiques descriptives.

La couche CLC-HR couvre le territoire métropolitain en 3,1 milliards de pixels, correspondant sur le terrain à des carrés de 20m de côté. Chaque pixel est affecté d'une valeur comprise dans l'intervalle $[0, 1]$: 0 indique un terrain non-imperméable et 1 indique un terrain totalement imperméable. Les pixels peuvent aussi être affectés d'une valeur différente de 0 ou de 1, indiquant un terrain partiellement imperméable avec un niveau d'imperméabilité plus ou moins prononcé. Une carte de l'imperméabilisation est proposée à la figure 1. Cette carte décrit donc la répartition spatiale d'une variable, notée $Y(\mathbf{M})$, d'imperméabilisation moyenne sur des pixels de 20m de côté. Cette variable est donc du type de la variable Z dont les propriétés sont exposées au début de ce paragraphe.
A une plus grande échelle, on observe que la restitution des zones imperméabilisées par la couche CLC-HR est relativement détaillée, ainsi que le montre l'image du Stade de France (figure 2) dont la pelouse apparaît bien comme étant non imperméabilisée tandis que les tribunes sont identifiées comme imperméabilisées.
Le phénomène d'imperméabilisation restitué par la couche CLC-HR est fondamentalement peu fréquent : la moyenne de la variable Y sur l'échantillon étudié est de 2,8%. Son écart-type est

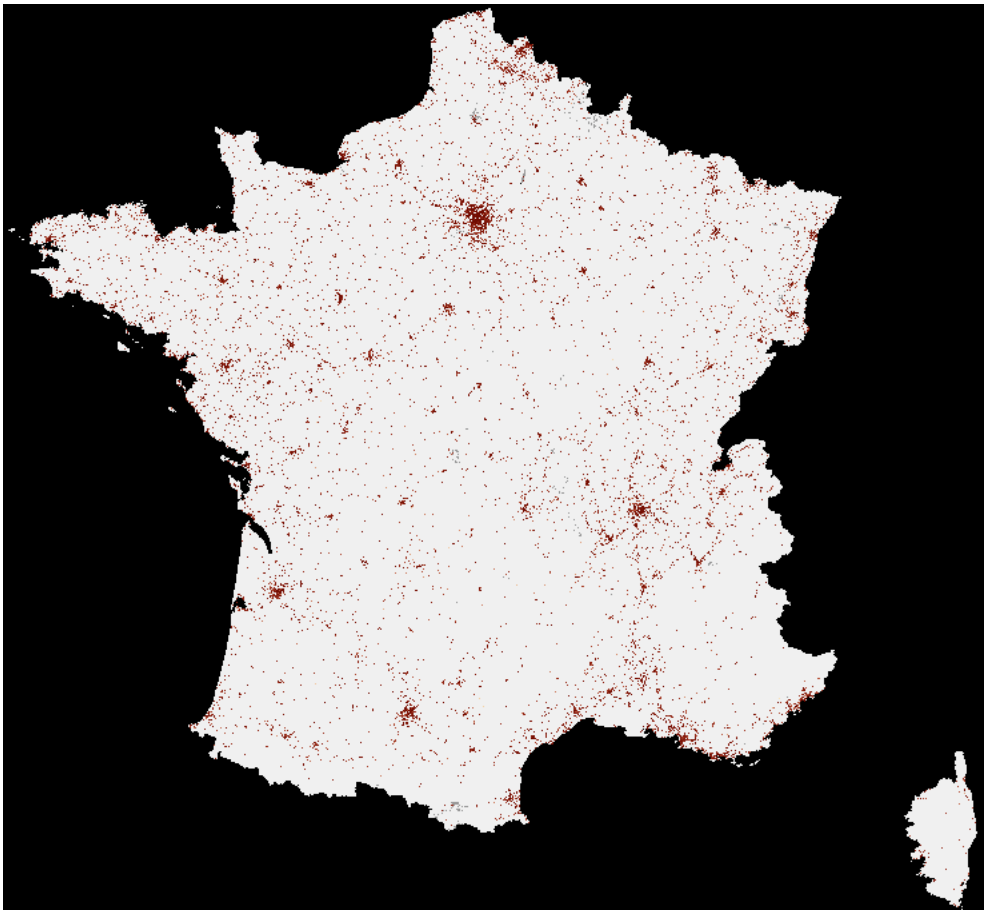
3. Les applications empiriques de ce papier se concentrent sur la notion d'imperméabilisation pour laquelle il est possible d'ajuster un modèle cohérent et complet à partir des données disponibles. Les développements théoriques proposés, en revanche, s'appliquent indifféremment aux notions d'artificialisation ou d'imperméabilisation.

4. On notera un point de l'espace géographique par une lettre grasse.

5. On notera simplement les intégrales de surface par un unique signe somme. Il n'y a généralement pas d'ambiguïté avec les intégrales 1-D puisque les intégrales 2-D opèrent sur des vecteurs notés en gras. Les cas potentiellement ambigus sont néanmoins signalés.

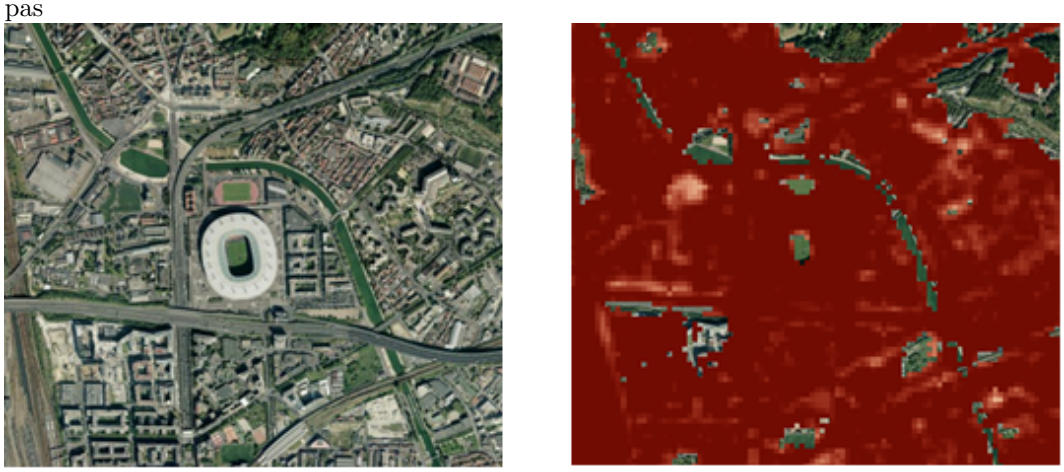
6. en probabilité ; par exemple, $\forall\alpha > 0$, $\lim_{\|d\mathbf{m}\|\to 0} \mathbb{P}\left(|Z(\mathbf{M} + d\mathbf{m}) - Z(\mathbf{M})| \geqslant \alpha\right) = 0$.

FIGURE 1 – Carte des pixels d'imperméabilisation à partir de la couche CLC-HR correspondante



Note : 96,2% des pixels comportent un taux d'imperméabilisation nul (représentés en blanc). Les pixels imperméabilisés sont représentés en niveaux de rouge qui, du fait de leur concentration, apparaissent comme étant saturés (voir aussi histogramme des valeurs, figure 3).

FIGURE 2 – Le Stade de France : des tribunes imperméabilisées et une pelouse qui ne l'est pas



Note : à gauche, orthophoto du stade de France (IGN) ; à droite superposition des pixels de la couche CLC-HR (Rouhaud 2016)

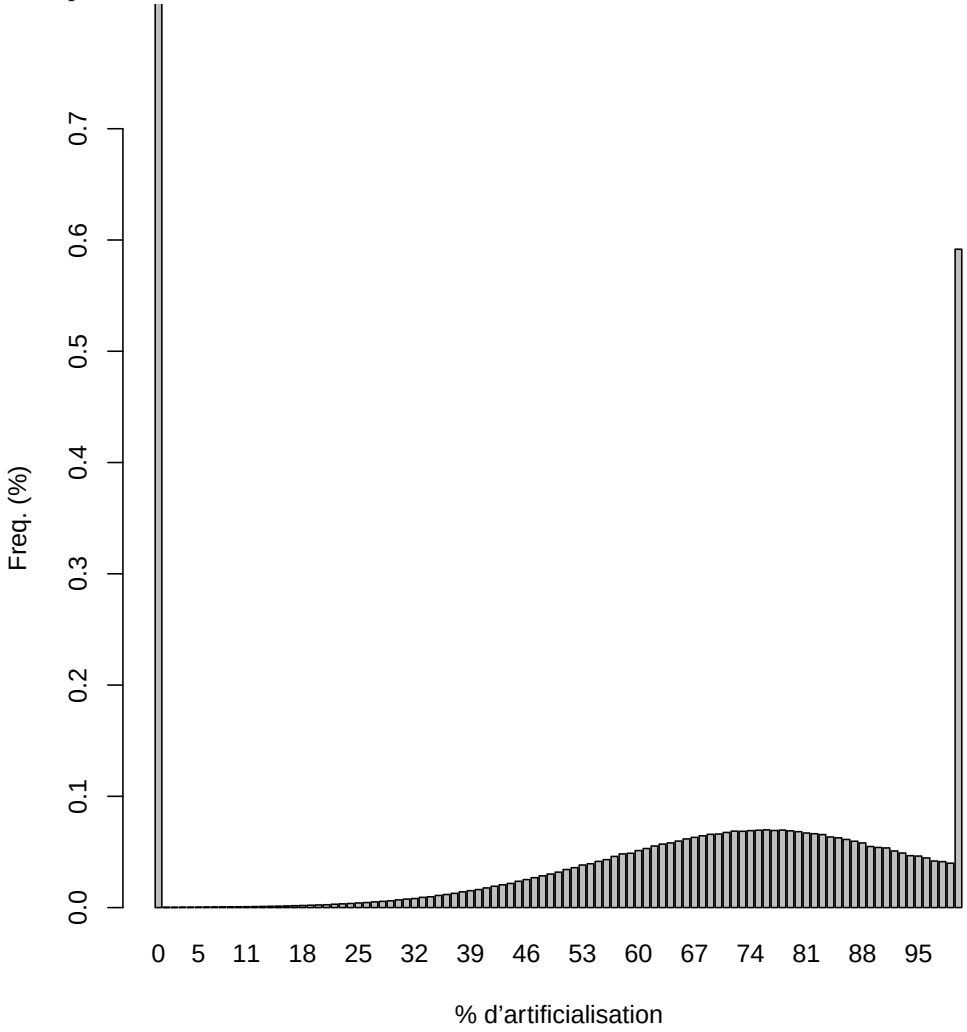
de 14,8%. La figure 3 montre le tracé de l'histogramme en fréquences des valeurs de la variable Y d'imperméabilisation. La distribution des fréquences de la variable Y comporte deux points-masses : en 0 où 96,2% des pixels se concentrent et en 1 où 0,6% des pixels se concentrent. Les autres valeurs de la distribution empirique sont réparties de manière quasi-continue sur $]0, 1[$.

L'enquête Teruti-Lucas est réalisée chaque année par le service statistique du ministère de l'agriculture. Elle est destinée à suivre l'évolution de l'occupation [7] et de l'utilisation [8] des sols. Elle est fondée sur un échantillonnage systématique du territoire à raison d'un point par km^2. Les enquêteurs stationnent les points d'enquête (dont le coordonnées sont connues au mètre près) et relèvent la nature de la couverture dans un cercle de rayon 1,5m, selon une double nomenclature [9] européenne, propre à l'occupation du sol (122 postes) d'une part et propre à l'usage du sol (38 postes) d'autre part. Il est possible, à l'aide de cette enquête, d'isoler les sols artificialisés par lecture directe de la nomenclature d'occupation (poste) et, les soles imperméabilisés, par sélection fondé sur la double nomenclature telle que décrite en annexe A. Pour le territoire métropolitain, le taux d'artificialisation est estimé à 9,3% et le taux d'imperméabilisation, à 4,6%.
Compte tenu de la nature de l'observation, réalisée *in situ* sur un périmètre beaucoup moins étendu que les pixels de la couche CLC-HR, le processus observé dans le cadre de l'enquête Teruti-Lucas est de nature un peu différente du processus Y évoqué à l'endroit de CLC-HR. Nous anticipons dès à présent que la moyenne des deux processus devrait néanmoins être proche. Pour distinguer [10] le processus associé à Teruti-Lucas, nous le notons X et nous examinons plus précisément son lien avec le processus Y dans la partie 4.

Ainsi, les sources CLC-HR et Teruti-Lucas conduisent à des taux d'imperméabilisation différents, de 2,8% pour la première à 4,6% pour la seconde. La comparaison de ces deux chiffres se heurte à la difficulté d'estimer leur précision. Pour la couche d'informations CLC-HR, le passage du processus ponctuel à une moyenne sur des carreaux de 20m de côté génère forcément un certain niveau d'incertitude. Les statistiques fondées sur l'enquête Teruti-Lucas sont affectées d'une imprécision liée, d'une part à l'échantillonnage caractéristique du plan de sondage et, d'autre part, à la variable d'intérêt elle-même. En effet, si on note $(X_i)_{i\in\{1,\dots,n\}}$ les différentes observations réalisées par l'enquête du processus d'imperméabilisation, l'estimateur de la moyenne empirique, $\hat{\mu} = \frac{1}{n}\mathbb{1}_n^T\mathbf{X}$, sans

7. C'est-à-dire la couverture physique du sol.
8. C'est-à-dire l'usage socio-économique du sol.
9. La double nomenclature permet, par combinaison, de mieux identifier certaines catégories d'occupations. Par exemple, les chemins utilisés pour le déplacement de matériels agricoles entre différents champs s'obtiennent par la combinaison de l'occupation A22 : structures linéaires non construites et de l'utilisation U11 : agriculture.
10. Un troisième processus, générique de l'imperméabilisation, sera noté Z (voir parties 3 et 4). X et Y sont donc des processus mesurés, associés au processus parent Z.

FIGURE 3 – Histogramme des pixels d'imperméabilisation à partir de la couche CLC-HR correspondante



Lecture : 96,2% des pixels comportent un taux d'imperméabilisation nul. La barre de l'histogramme représentée dans la figure en 0 est tronquée pour permettre la représentation des autres modalités d'imperméabilisation qui autrement seraient quasiment confondues avec l'axe des abscisses.

biais pour un sondage équiprobable, a pour variance :

$$\mathrm{var}(\hat{\mu}) = \frac{1}{n^2} \mathbb{1}_n^T \sqrt{\Delta}\ R\ \sqrt{\Delta} \mathbb{1}_n$$

où $\sqrt{\Delta}$ est la matrice diagonale des écart-types et R la matrice de corrélation des (X_i). Si on note $\hat{\mu}_0$ la variance de $\hat{\mu}$ en l'absence de corrélation (i.e. R est l'identité), on observe que :

$$\mathrm{var}(\hat{\mu}) = \underbrace{\left(1 + \frac{1}{n} \sum_{i \neq k} R_{i,k}\right)}_{\text{facteur d'inflation de la variance}} \quad \mathrm{var}(\hat{\mu})_0 \tag{1}$$

Il découle de cette expression que la connaissance de la structure – en particulier de la corrélation – des processus X et Y est cruciale pour examiner la cohérence des estimateurs obtenus. Nous nous concentrons désormais sur la connaissance de la structure du processus Y, associé à CLC-HR. Nous allons montrer que d'une part, il peut être connu intégralement et que d'autre part, la connaissance de sa structure nous renseigne sur X et *in fine* nous permet, sous certaines hypothèses que nous expliciterons, de comparer les taux d'imperméabilisation associés aux deux sources de données.

3 La structure du phénomène d'imperméabilisation

Sur un plan technique, la modélisation du processus repose d'une part sur l'identification de sa loi marginale et d'autre part, sur l'identification de sa structure d'autocorrélation. Le préalable à l'identification de ces deux composantes est d'établir la stationnarité du processus. Nous revenons sur ces différents points dans les sous-parties de ce paragraphe. Puis dans une dernière sous-partie, nous proposons une simulation du processus modélisé.

D'un point de vue pratique, l'analyse du processus suppose de disposer de mesures qui permettent d'identifier ses composantes caractéristiques. Un candidat pourrait être les données de l'enquête Teruti-Lucas dont on a vu précédemment qu'elles correspondent à l'observation d'un phénomène dont la résolution spatiale est celle qui nous intéresse. Cependant, le sondage systématique sur lequel elle repose (approximativement un point tous les kilomètres selon une grille régulière) ne permet pas d'identifier une structure d'autorcorrélation continues, en particulier pour les échelle de distances comprises entre 0 et 1 kilomètre, puis entre 1 et 2 kilomètres, etc. Or la connaissance du processus à ces échelles est très importantes pour en identifier le comportement. Un alternative crédible est la base de données CLC-HR qui offre une information continue sur le territoire métropolitain, discrétisée à un pas uniforme de 20m. Comme l'information est moyennée sur cette distance, le système d'observation est *de facto* sensible au phénomène de résolution métrique que l'on cherche à observer. Sous certaines hypothèses de continuité, il est possible de déduire de la connaissance du processus observé sur un pas de 20m (processus Y) le processus parent métrique Z le premier apparaissant une simple moyenne sur un intervalle de 20m du second. Nous choisissons donc cette approche et étudions dans un premier temps le processus Y.

3.1 Modélisation du processus $Y(\mathbf{M})$

La répartition de la variable Y est tracée à la figure 3. Il peut être utile de modéliser la loi correspondante, notamment afin d'être en mesure de simuler le processus. Pour modliser la densité nous observons que la densité estimée peut s'exprimer sous la forme d'une composition de trois lois β. On suppose que la densité peut se modéliser sous la forme :

$$f_Y(x) = p_1 \beta_{a_1,b_1}(x) + (1 - p_1 - p_3)\beta_{a_2,b_2}(x) + p_3 \beta_{a_3,b_3}(x) \tag{2}$$

où $\beta_{a,b}$ est la densité de la loi β de paramètres a et b ; soit une répartition donnée par :

$$F_Y(x) = p_1 \mathscr{B}_{a_1,b_1}(x) + (1 - p_1 - p_3)\mathscr{B}_{a_2,b_2}(x) + p_3 \mathscr{B}_{a_3,b_3}(x) \tag{3}$$

où $\mathscr{B}_{a,b}$ est la fonction de répartition de la loi β de paramètres a et b. Les valeurs estimées des paramètres sont indiqués au tableau 1.

TABLE 1 – Valeurs des paramètres de l'Eq. (3) retenus

Paramètre	Valeur
p_1	96.2%
p_3	0.6%
a_1	0.5
b_1	2000
a_2	3.94 (0.09)
b_2	1.69 (0.03)
a_3	200
b_3	0.5

Note : Les paramètres p_1 et p_3 sont directement issus des valeurs en 0 et 1 de l'histogramme de la figure 3. (a_1, b_1) et (a_3, b_3) sont conventionnels et symétriques de sorte que les valeurs de la loi β associées soient respectivement concentrées en 0 et 1. Les paramètres a_2 et b_2 sont estimés par moindres carrés sur la densité empirique correspondant à l'histogramme de la figure 3 sur $]0, 1[$ (entre parenthèses les écarts-types d'estimation).

Avec les paramètres du tableau 1, la moyenne de la loi modélisée résultante est [11] $2,9\%$ et l'écart-type est de $14,8\%$. Ces valeurs sont à comparer aux estimations empiriques fondées sur la couche CLC-HR, respectivement $2,8\%$ et $14,8\%$ (voir paragraphe 2). Le modèle, calé sur la combinaison de lois β, est donc très proche de l'empirique, ce qui est évidemment l'objectif visé.

Le tracé de la densité modélisée est proposé à la figure 4. La répartition obtenue est tracée à la figure 5.

3.2 Modélisation de l'autocorrélation de $Y(\mathbf{M})$

L'imperméabilisation est aussi caractérisée par une très forte corrélation spatiale ce qui, du point de vue physique, se traduit par une très grande concentration spatiale. On l'observe immédiatement sur la figure 1 dans le cas de l'imperméabilisation : le phénomène, bien que très peu fréquent, se trouve, de manière évidente, concentré en certains lieux.
La quantification de ce phénomène de corrélation spatiale passe par la détermination d'une fonction d'autocovariance ou, de manière équivalente pour les processus stationnaires, par l'estimation du semi-variogramme (plus simplement appelé ci-après variogramme). Ce dernier étant de portée plus générale, on utilise cette fonction. Avec les notations précédentes et sous l'hypothèse que la corrélation spatiale soit isotrope, c'est-à-dire qu'elle ne dépende pas (en moyenne) de la direction considérée, le variogramme est, pour un processus spatial Z, défini par :

$$\gamma_Z(h) = \frac{1}{2}\mathbb{E}_{\mathbf{M},\vec{u}}\left[Z(\mathbf{M} + h.\vec{u}) - Z(\mathbf{M})\right]^2 \tag{4}$$

où $\vec{u}$ est un vecteur unitaire quelconque. $\mathbb{E}_{\mathbf{M},\vec{u}}$ indique que l'espérance est indépendante de $\mathbf{M}$ et $\vec{u}$ ou qu'elle s'entend quels que soient $\mathbf{M}$ et $\vec{u}$. Cette notation est cohérente avec l'hypothèse de stationnarité du processus.
Si on définit la fonction d'autocovariance (isotrope) du processus Z par :

$$C_Z(h) = \mathbb{E}_{\mathbf{M},\vec{u}}\left[(Z(\mathbf{M} + h.\vec{u}) - \mathbb{E}Z)\,(Z(\mathbf{M}) - \mathbb{E}Z)\right]$$

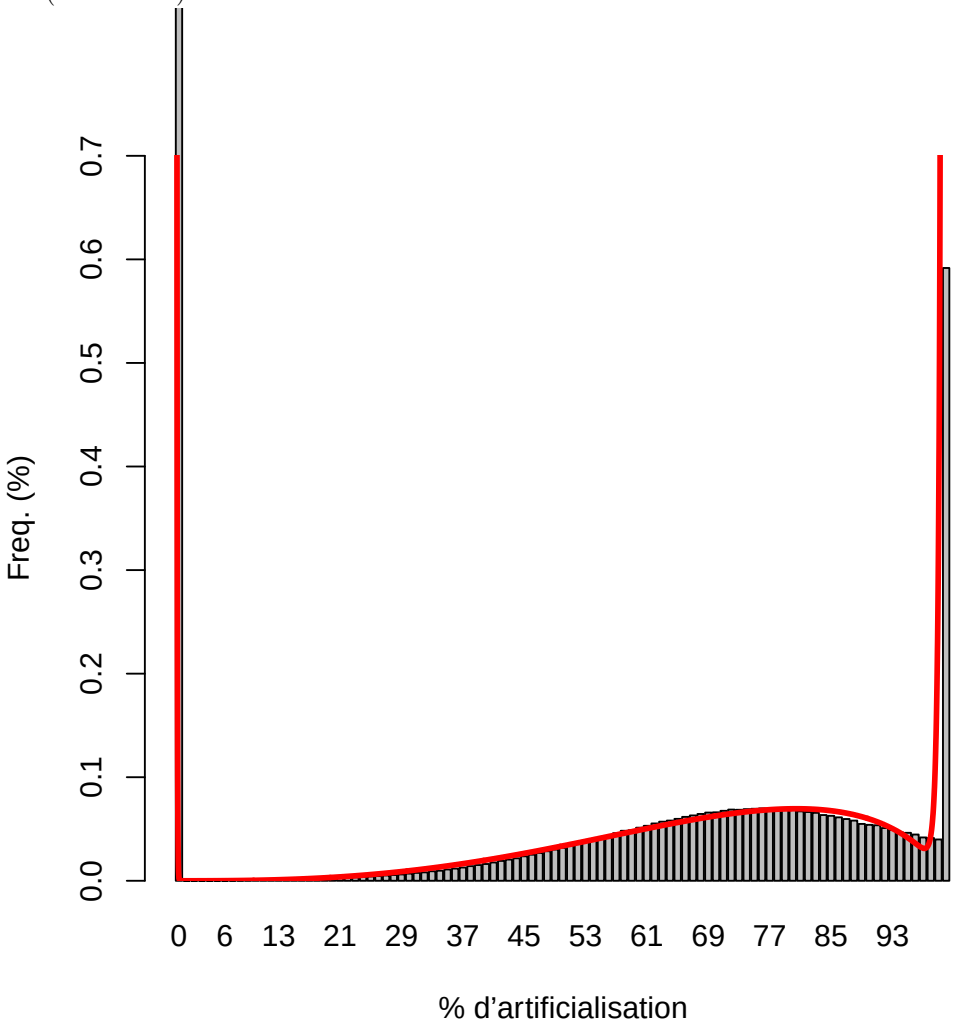
où $\mathbb{E}Z$ désigne l'espérance [12] de la variable Z, on vérifie facilement que

$$\gamma_Z(h) = \frac{1}{2}\left(2C_Z(0) - 2C_Z(h)\right) = C_Z(0) - C_Z(h) \tag{5}$$

11. La loi $\beta(p, q)$ (dite « de première espèce » – voir Tassi (1992)) a pour espérance $p/(p+q)$ et pour variance $pq/(p+q)^2(p+q+1)$. A l'aide de l'expression de la densité donnée à la relation (2), on calcule aisément l'espérance et la variance de la variable modélisée. Avec les expressions de la relation (2), l'espérance s'écrit $\mu = \sum_{i=1}^{3} p_i \frac{a_i}{a_i+b_i}$ avec $p_2 = 1 - p_1 - p_3$. La variance s'écrit : $\sigma^2 = -\mu^2 + \sum_{i=1}^{3} p_i \left[\frac{a_i b_i}{(a_i+b_i)^2(a_i+b_i+1)} + \left(\frac{a_i}{a_i+b_i}\right)^2\right]$

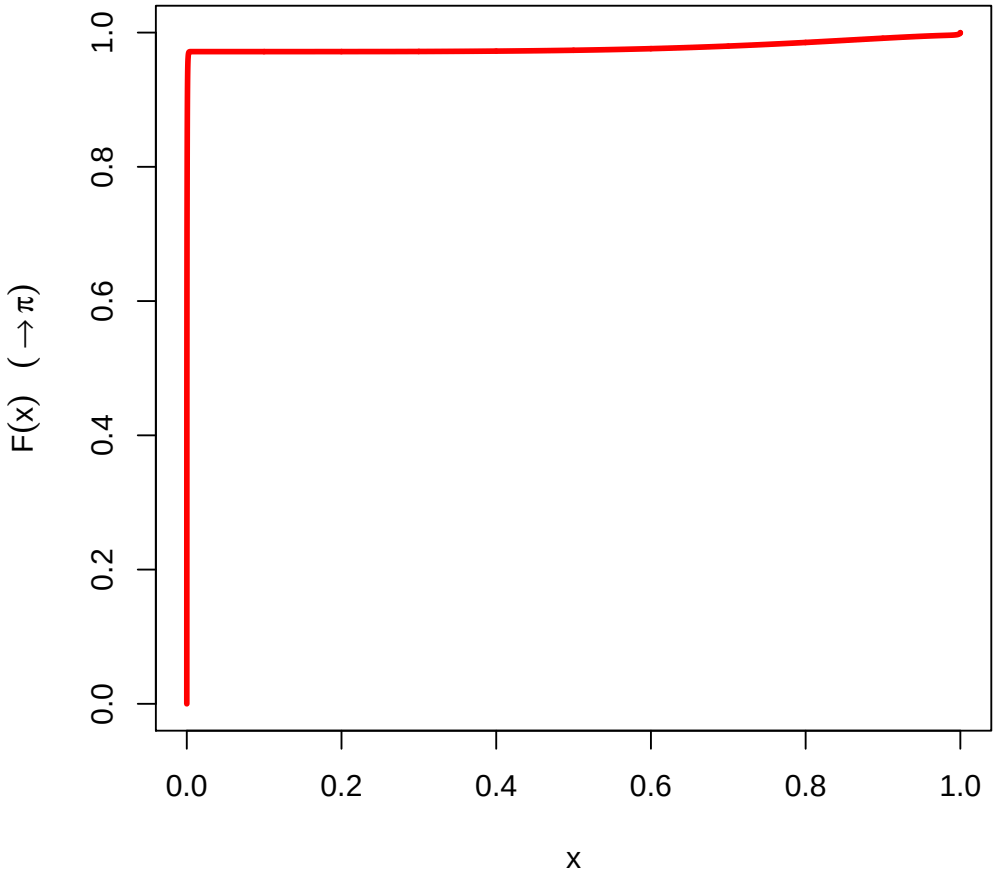
12. Avec les mêmes notations, $\mathbb{E}Z = \mathbb{E}_{\mathbf{M}}\left(Z(\mathbf{M})\right)$.

FIGURE 4 – Modélisation de la loi marginale du taux d'imperméabilisation des pixels CLC-HR (variable Y)



Note : histogramme identique à celui de la figure 3. La courbe en rouge est la densité correspondant à la relation (2).

FIGURE 5 – Répartition de la loi marginale du taux d'imperméabilisation des pixels modélisée



Note : tracé de la fonction de répartition correspondant à la relation (3). La densité associée est présentée à la figure 4.

On utilisera aussi la fonction d'autocorrélation, $R_Z(h)$ définie à l'aide de l'autocovariance par :

$$R_Z(h) = C_Z(h)/C_Z(0) \tag{6}$$

Par définition, cette fonction caractérise la corrélation existant entre deux points séparés d'une distance h.

Il existe dans la littérature différents estimateurs de $\gamma(h)$ (voir Matheron (1970) ou Cressie (1993)). Et beaucoup d'algorithmes de sélection des couples de points séparés d'une distance donnée ont été développés. Nous utilisons celui mis en œuvre par Naimi, Skidmore, Groen & Hamm (2011), conçu pour s'adapter au cas d'une grille régulière de points (raster) ainsi que se présente la couche imperméabilisation de CLC-HR.

Le variogramme de la variable Y caractérisant l'imperméabilisation telle que vue par CLC-HR et dont la distribution est examinée à la figure 4 est tracé à la figure 6. Les détails pratiques de son estimation sont donnés en annexe B.

L'autocorrélation associée est présentée à la figure 7. Elle témoigne à la fois d'une atténuation de la corrélation spatiale en fonction de la distance assez rapide sur les premières centaines de mètres, puis d'une dépendance à plus longue distance qui s'atténue très lentement. Le niveau de corrélation à 4km atteint encore 11,5%. En ce sens, le processus d'imperméabilisation est à mémoire longue. Ceci ne remet toutefois pas en cause l'hypothèse de stationnarité du processus.
Pour compléter le modèle, il convient de modéliser la fonction d'autocovariance de l'imperméabilisation (variable Y). Après différents tests de fonctions, il s'avère que la forme en :

$$\hat{C}_Y(h) = A\left(1 + \frac{|h|}{b}\right)^{-\alpha} \tag{7}$$

est bien adaptée dans le cas présent (figure 6). A, b et α sont des paramètres positifs à déterminer. La relation équivalente sous forme de variogramme s'écrit :

$$\hat{\gamma}_Y(h) = A\left[1 - \left(1 + \frac{|h|}{b}\right)^{-\alpha}\right] \tag{8}$$

L'estimation est réalisée sous cette deuxième forme par moindres carrés non linéaires. Les résultats sont donnés à la table 2. Le tracé du modèle obtenu est proposé à la figure 8.

TABLE 2 – Valeurs des paramètres de l'Eq. (8) retenus

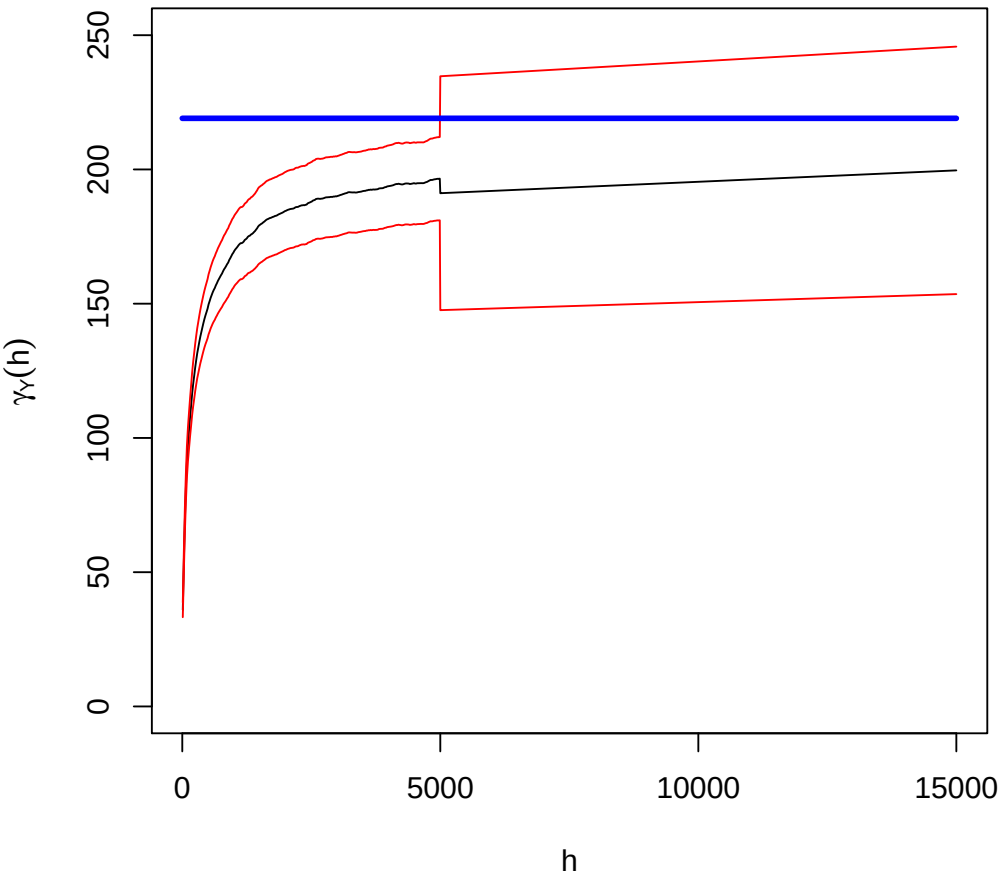
Paramètre	Valeur
A	227.9
	(1.5)
b	38.1
	(1.6)
α	0.41
	(0.01)

Note : Les paramètres sont estimés par moindres carrés non linéaires sur le variogramme empirique tracé à la figure 6 (entre parenthèses les écarts-types d'estimation). Tous les paramètres sont significatifs.

Au vu de l'expression de l'autocovariance (7), il apparaît que le processus est stationnaire à mémoire longue. En effet, le processus est stationnaire puisque son espérance est finie et son autocovariance ne dépend que de la distance qui sépare deux points de réalisation du processus.
Et par ailleurs, son autocovariance vérifie les propriétés $\int C_Y = \infty$ et $C_Y \overset{h\to\infty}{\sim} Kh^{-\alpha}$ où $\alpha \approx 0.41$ et K un réel positif. Par conséquent, elle est de la forme Kh^{2d-1} avec $d \approx 0.29 \in [0, \frac{1}{2}]$. Par définition (Beran, Feng, Ghosh & Kulik 2016), le processus est à mémoire longue.

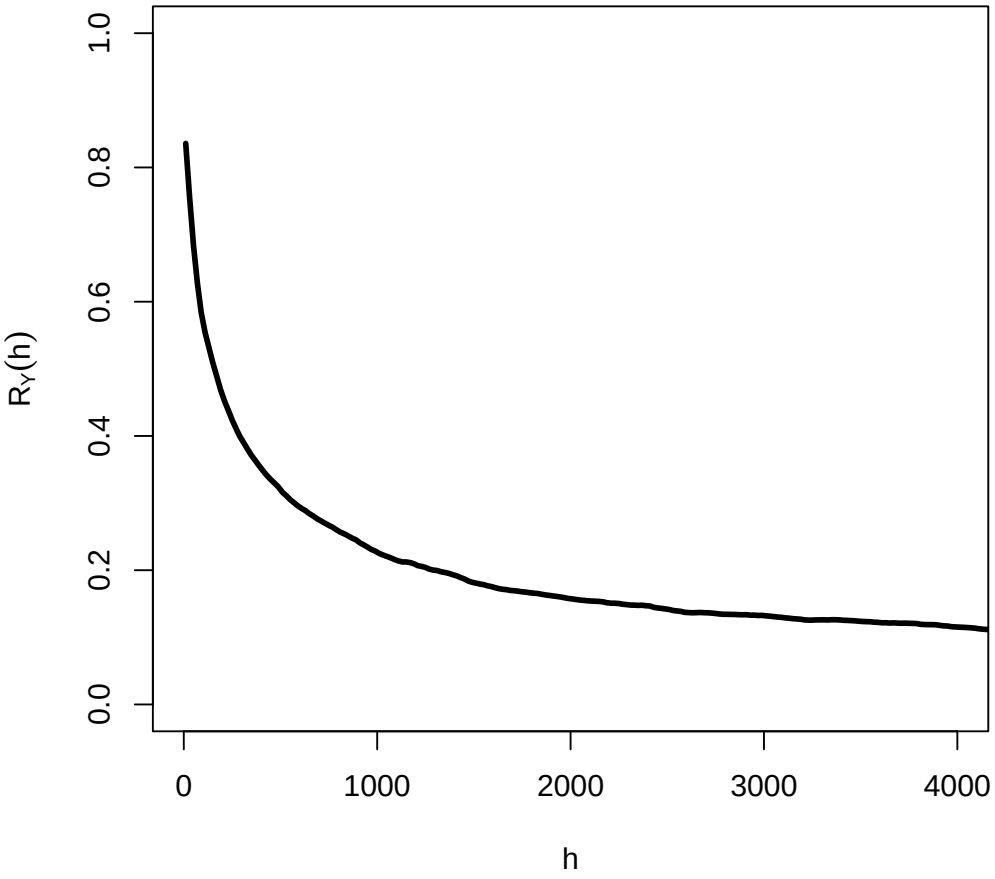
Ce point est d'importance car il traduit la persistance du processus dans l'espace et le caractère non négligeable de cette persistance dans les estimateurs de variance. Par exemple, si on reprend l'exemple (1), la matrice R qui apparaît dans cette expression peut être quasiment pleine. Dans ces conditions, la somme de ses termes non diagonaux peut tendre vers l'infini lorsque $n \to \infty$ et la

FIGURE 6 – Estimation du variogramme de l'imperméabilisation à partir de la couche CLC-HR du thème imperméabilisation



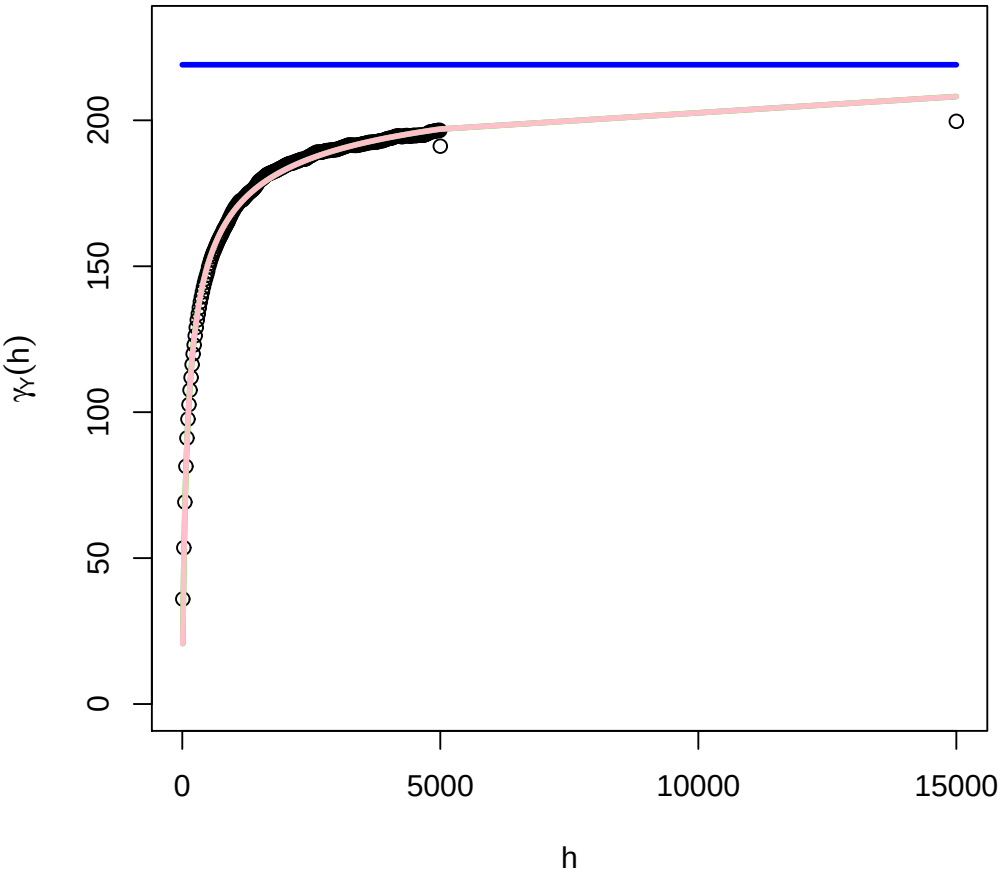
Note : Variogramme estimé empiriquement sur l'ensemble de la France métropolitaine (échelle des abscisses en m). Les courbes en rouges indiquent les intervalles de confiance à deux écart-types calculés point par point, sans exploiter d'hypothèse de continuité du variogramme. Si cette dernière hypothèse était utilisée, par exemple avec une technique d'estimation par méthode de régression locale, alors l'amplitude des intervalles serait beaucoup moins grande. Les intervalles proposés sont donc certainement très pessimistes par rapport à la réalité. La droite horizontale en bleu indique la variance du processus Y. Cette droite est par construction l'asymptote de $\gamma_Y(h)$ quand $h \to \infty$, sous les hypothèses formulées dans le texte pour le processus Y.

FIGURE 7 – Autocorrélation du processus imperméabilisation telle que vue par CLC-HR (Y) déterminée à partir de la couche CLC-HR



Note : Tracé issu de l'estimation du variogramme (cf. figure 6) et des relations (5) et (6). Echelle des abscisses en m

FIGURE 8 – Modélisation du variogramme du taux d'imperméabilisation des pixels CLC-HR (variable Y)



Note : variogramme identique à celui de la figure 3. La courbe en rose est le modèle estimé correspondant la relation (8). En bleu figure le niveau de variance empirique de la variable d'artificialisation CLC-HR dont l'estimation est donnée au §2 (écart-type de 14,9%). Comme évoqué plus haut, ce niveau est une asymptote pour le variogramme (quand $h \to \infty$).

convergence de la variance empirique peut s'en trouver profondément modifiée, voire non satisfaite. Pour fixer un ordre de grandeurs et à partir des valeurs estimées (table 2), il est de coutume en géostatistique de définir la *portée pratique* (Allard 2012) comme étant la distance à partir de laquelle l'autocovariance est inférieure [13] à 5% de sa valeur en 0. La portée pratique prend ici, pour le processus Y, la valeur de 56,6km. Autrement, pour toute distance inférieure à 56km, la corrélation entre les points est supérieure à 5%.

3.3 Simulation du processus Y

Formellement, la simulation d'un champ de variables dont les lois marginales sont du type de celle présentée à la figure 3 et dont la structure de corrélation est donnée à la figure 7 nécessite de disposer de la loi jointe de ces variables.
Notons plus généralement un n-uplet de variables $(Z_1, \ldots, Z_n)$ de répartitions marginales F_i et F leur répartition jointe. Formellement, la simulation du vecteur aléatoire $\mathbf{Z} = (Z_1, \ldots, Z_n)$ suppose de connaître F. Si F est continue, de même que les F_i (cas des variables aléatoires continues), alors il existe une unique fonction C, appelée copule, qui permet de passer des lois marginales à la loi jointe :

$$\exists C : [0,1]^n \to [0,1] \ \mid \forall (z_1, \ldots, z_n) \ , \ F(z_1, \ldots, z_n) = C\left(F_1(z_1), \ldots, F_n(z_n)\right)$$

Ce résultat est connu dans la littérature sous le nom de théorème de Sklar (1973).
Le point important, dans ce théorème, est que la copule est porteuse de la relation qui existe entre les composantes du vecteur aléatoire, hors les lois marginales proprement dites. On montre que la copule est invariante par transformation strictement croissante du vecteur aléatoire $\mathbf{Z}$ (Gobet 2013). Pour illustrer cette propriété, prenons l'exemple de la simulation d'un vecteur aléatoire $\mathbf{W}$ de lois marginales continues strictement monotones F_i et de variance fixée Σ. Pour simuler $\mathbf{W}$, on procède de la manière suivante :

1. On génère le vecteur aléatoire $\mathbf{Z}$ de loi marginales normales centrées réduites et de variance Σ. Ceci s'effectue en simulant un vecteur $\mathbf{V}$ dont les composantes sont normales centrés réduites indépendantes. Puis on définit $\mathbf{Z} = L^T\mathbf{V}$ où L est une factorisée de Σ ($\Sigma = L^T L$). Ainsi défini, $\mathbf{Z}$ est de lois marginales normales centrées réduites et de variance Σ.
2. On note Φ la fonction de répartition de la loi normale centrée réduite. Le vecteur aléatoire $\mathbf{W}$ de composantes $W_i = F_i^{-1} \circ \Phi(Z_i)$ a pour lois marginales les F_i. La copule est conservée par transformation strictement croissante de $\mathbf{Z}$, ce qui est ici le cas, donc $\mathbf{W}$ a pour variance Σ.

Ce principe est appliqué pour une matrice de cellules dont les lois marginales correspondent à la loi modélisée telle que présentée à la figure 4 et dont la corrélation entre cellules est fixée conformément au modèle de corrélation donnée à la relation (7). La figure 5 donne le tracé de la fonction de répartition F_i et la figure 9 celui de F_i^{-1}. Différents résultats de simulations sont présentés à la figure 10 sur des cellules de 20m de côté regroupées dans des blocs de terrain de 1 500m de côté.

On observe que ces quelques éléments de modélisation permettent très convenablement de restituer les différentes situation observées sur le terrain, depuis l'absence de sols imperméabilisés jusqu'à de très fortes concentrations (figure 10).

4 De l'imperméabilisation sur un raster de pas 20m au processus métrique Z

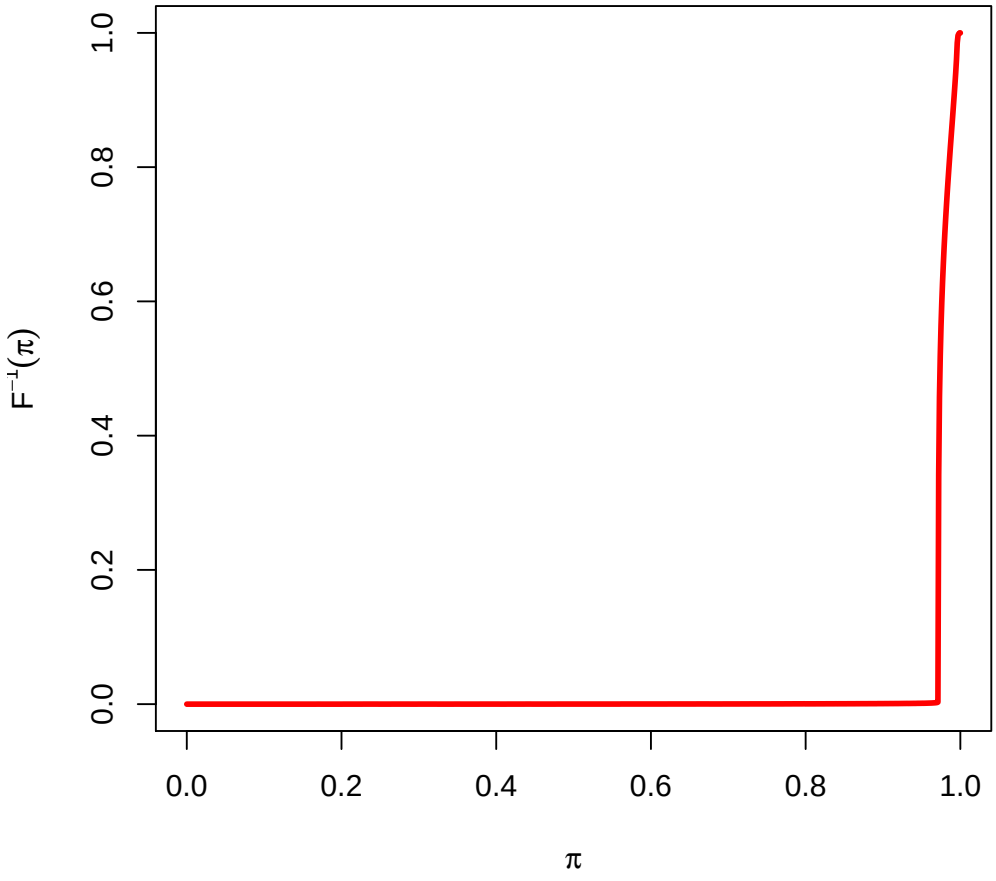
Jusqu'à présent, nous avons travaillé sur une variable d'imperméabilisation issue d'une moyennisation opérée sur des pixels de 20m. Le processus parent, celui qui est observé sur le terrain par les enquêteurs chargés de l'enquête Teruti, est typiquement un processus métrique. Comme exposé au paragraphe 2, le passage du processus métrique, Z, au processus moyenné sur pixels de 20m de côté, Y, peut se représenter comme une transformation opérée par moyenne mobile :

$$Y(\mathbf{M}) = \frac{\displaystyle\int_{\mathbf{m} \in \mathscr{P}(\mathbf{M})} Z(\mathbf{m}) d\mathbf{m}}{\displaystyle\int_{\mathbf{m} \in \mathscr{P}(\mathbf{M})} d\mathbf{m}} \tag{9}$$

où $\mathscr{P}(\mathbf{M})$ désigne le pixel de 20m de côté centré sur le point $\mathbf{M}$.

13. Cette valeur correspond donc à $C_Y^{-1}(0.05 C_Y(0))$.

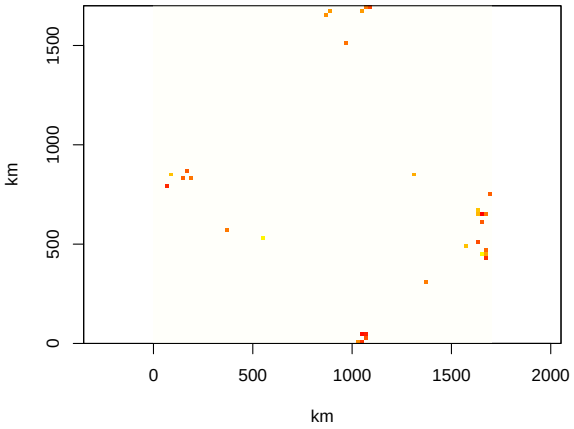
FIGURE 9 – Inverse de la fonction de répartition du taux d'imperméabilisation des pixels modélisée



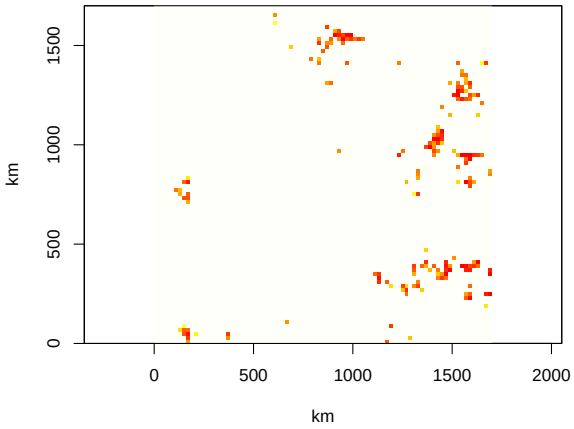
Note : inverse de la fonction de répartition dont le tracé est donné à la figure 5.

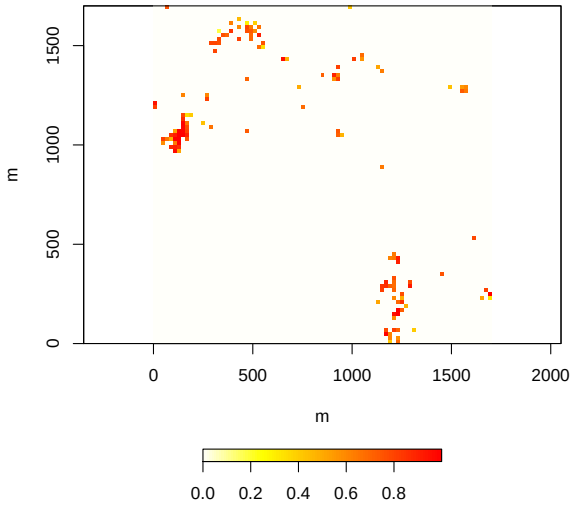
FIGURE 10 – Simulations de la variable d'imperméabilisation (variable Y) modélisée

(a)

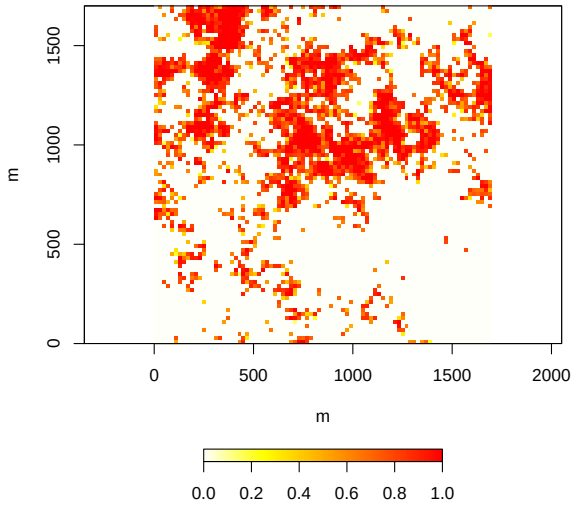


(b)





(c)

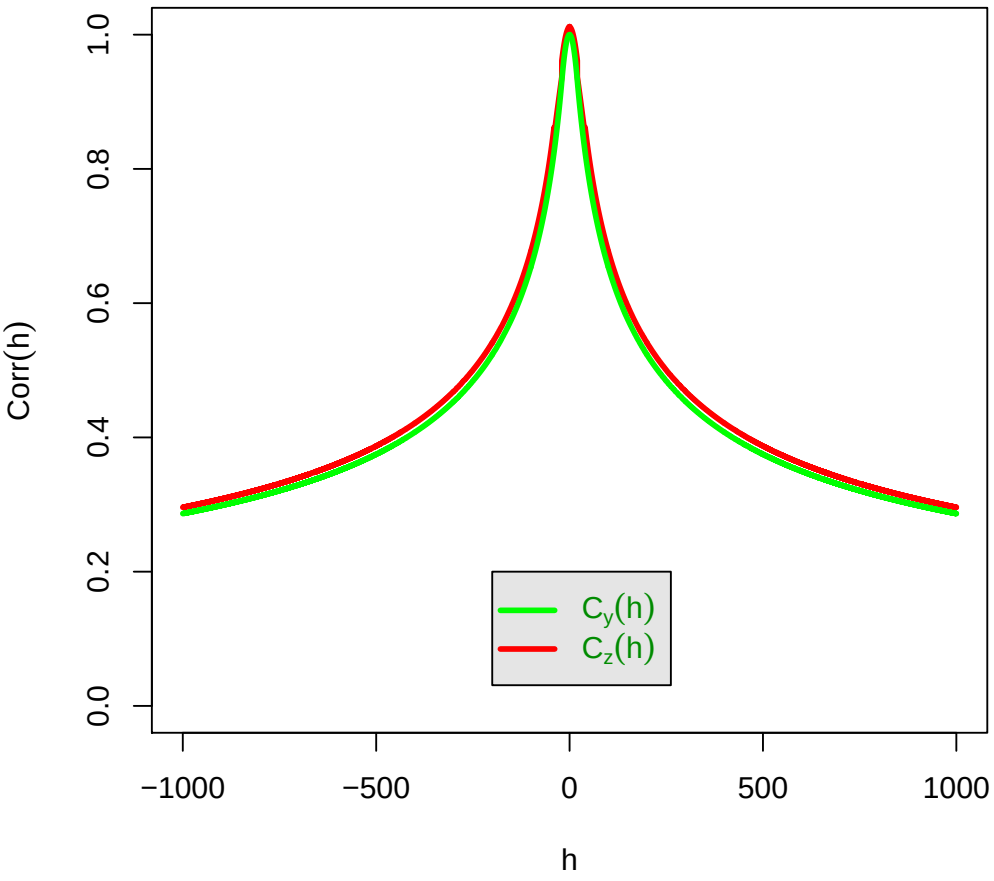


(d)

Remarque : Différentes simulations sont proposées ici pour illustrer la variété des situations dont rend compte le modèle proposé. On note les extrêmes que représentent les exemples (a) où très peu de pixels sont imperméabilisés et (d) où au contraire, beaucoup le sont. Rappelons que le processus générateur correspond à un taux moyen d'imperméabilisation de 2,8%.

Sous certaines hypothèses de régularité (stationnarité et double dérivabilité) du processus Z, il est possible de déterminer l'autocorrélation du processus Z, connaissant celle du processus Y. Les hypothèses et le calcul sont précisés en annexe C. Si ces conditions sont respectées et avec les valeurs numériques mesurées empiriquement, l'autocorrélation du processus Z est extrêmement proche de celle du processus Y. Retenons néanmoins que le comportement de l'autocorrélation de Z au voisinage de 0 est, par construction, mal connu. En effet, le lissage (9) gomme la variabilité de court horizon qui peut être présente dans le comportement de la variable Z, à laquelle rappelons-le, nous n'avons pas directement accès. Un tracé des courbes d'autocorrélations obtenues pour Y et, pour Z, déduite de celle de Y, est donné à la figure 11. On en retient, pour la suite de l'article, que l'autocorrélation du processus métrique Z est convenablement approximée par celle du processus Y, dont les éléments de définition sont donnés à la relation (7) et les valeurs numériques associées, à la table 2.

FIGURE 11 – Autcorrélation des processus Y et Z



Lecture : L'autocorrélation du processus Y correspond à la courbe déjà présentée à la figure 7, à ceci près qu'en 0, l'autocorrélation prend la forme régularisée d'une fonction dérivable dont la dérivée seconde est négative. Les raisons de cette modification sont expliquées en annexe C. On note la très grande proximité des autocorrélations de Y et Z.

Sur cette base, il est possible de simuler la variable Z en procédant de la même manière que pour Y avec la corrélation ainsi calculée, et une densité correspondant à deux masses en 0 et 1 ($p_0 = 4,6\%$) approximée à l'aide de deux lois β de paramètres $(0.5, 2000)$ et $(2000, 0.5)$ respectivement. Cette

approximation permet de fonder la densité modélisée sur un modèle continu et strictement monotone et donc de réaliser une simulation à l'aide de la méthode des copules, comme pour la variable Y. Quelques exemples de simulations produites sur des blocs de terrain de 300m de côté sont proposées à la figure 12.

5 La précision des estimateurs de moyenne

On s'intéresse désormais à la variable aléatoire moyenne de Z sur un ensemble (en dimension 2) V définie par [14] :

$$\bar{Z}(V) = \frac{1}{|V|}\int_V Z(\mathbf{x})d\mathbf{x} \tag{10}$$

Par définition, si la variable Z est stationnaire :

$$\begin{aligned} \mathrm{var}(\bar{Z}(V)) &= \frac{1}{|V|^2}\mathbb{E}\int_V\int_V \tilde{Z}(\mathbf{u})\tilde{Z}(\mathbf{v})d\mathbf{u}d\mathbf{v} \\ &= \frac{1}{|V|^2}\int_V\int_V C(\mathbf{u}-\mathbf{v})d\mathbf{u}d\mathbf{v} \end{aligned} \tag{11}$$

où $\tilde{Z}$ désigne le processus centré associé à Z et C est la fonction d'autocorrélation du processus Z. A ce stade, aucune hypothèse n'est réalisée sur la fonction d'autocorrélation hormis que nous postulons la stationnarité du processus Z. On peut montrer, en toute généralité sous ces hypothèses que (voir annexe D) :

$$\mathrm{var}(\bar{Z}(V)) = \frac{1}{|V|^2}\int_{\mathbf{h}\in\mathbb{R}^2} C(\mathbf{h})f(\mathbf{h})d\mathbf{h} \tag{12}$$

avec

$$f(\mathbf{h}) = \int_{\mathbf{x}\in\mathbb{R}^2} \mathbb{1}_V(\mathbf{x}+\mathbf{h})\mathbb{1}_V(\mathbf{x})d\mathbf{x}$$

Dans le cas où l'ensemble V peut s'approximer comme un disque de rayon T, on montre (annexe D) que $\mathrm{var}(\bar{Z}(V)) = \Phi_C(T)$ où Φ_C correspond à une intégrale simple dont le noyau est la fonction d'autocorrélation isotrope C du processus Z. On peut calculer numériquement cette intégrale pour la fonction $\hat{C}_Y$ estimée pour le processus Y (Eq. (7) et table 2), laquelle correspond approximativement à celle du processus métrique Z, en vertu des développements du paragraphe 4. Dans ce calcul, le territoire métropolitain est approximé à un disque de rayon $T = 417km$ dont la surface correspond à la surface du territoire métropolitain (540 000km^2). La variance ainsi estimée est celle du taux d'imperméabilisation moyen du territoire métropolitain. Numériquement, on obtient

$$\sqrt{\hat{\Phi}_{\hat{C}}(T = 417km)} = 2,40 \tag{13}$$

Ainsi, la précision théorique d'un taux d'imperméabilisation calculée sur le territoire métropolitain dans son intégralité a un écart-type, compte tenu de son autocorrélation et de la variance de sa loi marginale, de $2,40$ points de pourcentage.

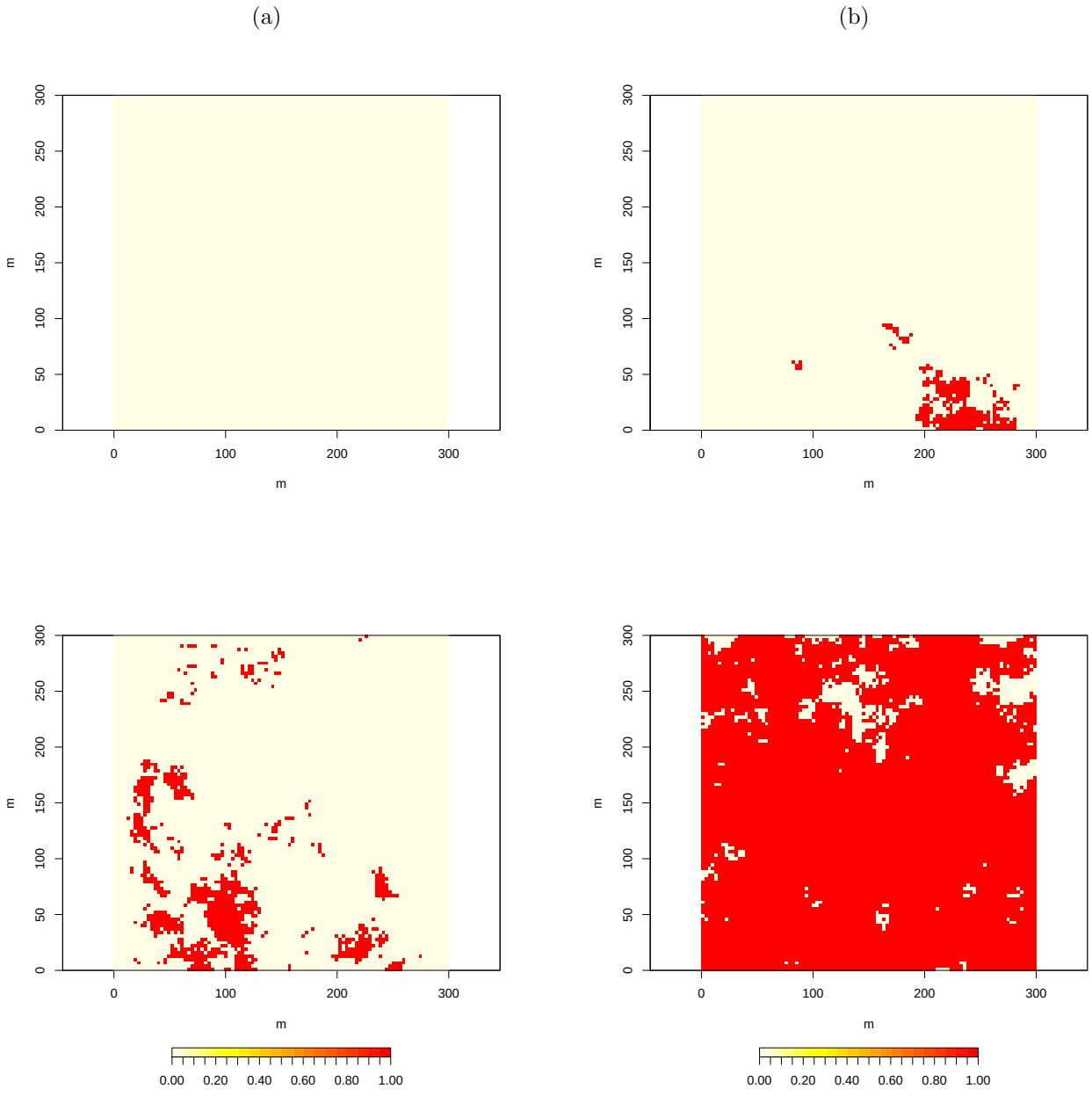
Cet écart-type est très élevé, pour un taux d'imperméabilisation estimé à 3 ou 4 points, selon la source considérée. Elle correspond à la variabilité de la moyenne (i.e. du taux d'imperméabilisation) que l'on obtiendrait si on simulait un grand nombre de fois un processus stochastique correspondant à celui qu'on a identifié pour l'imperméabilisation. Par exemple, si on veut comparer le niveau d'imperméabilisation de deux territoires différents et déterminer si ces niveaux correspondent à des processus générateurs significativement différents (eu égard à leur paramètre de moyenne en l'occurrence), alors l'intervalle de confiance de la différence requis pour cette inférence requiert le calcul d'une variance du type de celle présentée ici. C'est, en d'autres termes, une *variance d'estimation* du paramètre d'espérance du processus spatial sous-jacent à l'imperméabilisation, au sens de Diggle & Ribeiro (2007). Il est intéressant de noter d'ailleurs que cet écart-type élevé est la conséquence directe de l'autocorrélation persistante du processus. Ainsi, la situation polaire dans laquelle le processus serait structuré de manière totalement décorrélée conduirait une variance [15] de $\bar{Z}(V)$ égale à $A^2/|V|$ (avec les notations de la table 2), soit un écart-type de $2,3 \times 10^{-5}$ points de pourcentage [16] !

14. $|V|$ désigne la mesure de la surface de V.

15. En reprenant l'expression 11 avec $C(h) = A\delta(h)$, $\mathrm{var}(\bar{Z}(V)) = \frac{1}{|V|^2}\int\int A^2\mathbb{1}_V(\mathbf{u})\mathbb{1}_V(\mathbf{v})\delta(\mathbf{u}-\mathbf{v})d\mathbf{u}d\mathbf{v}$, soit $\mathrm{var}(\bar{Z}(V)) = \frac{1}{|V|^2}\int A^2\mathbb{1}_V(\mathbf{v})d\mathbf{v} = A^2/|V|$.

16. toujours dans le modèle d'un territoire circulaire de rayon 417km

FIGURE 12 – Simulations de la variable d'imperméabilisation métrique (variable Z) modélisée



Remarque : Différentes simulations sont proposées ici pour illustrer la variété des situations dont rend compte le modèle proposé pour la variable binaire (métrique) Z. Les simulations sont réalisées sur des blocs de terrain carrés de 300m de côté. On note les extrêmes que représentent les exemples (a) où aucun pixel n'est imperméabilisé et (d) où au contraire, beaucoup le sont. Rappelons que le processus générateur correspond à un taux moyen d'imperméabilisation de 4,6% estimé à l'aide de l'enquête Teruti.

S'agissant de la cohérence d'estimations de taux d'imperméabilisation pour un même territoire, la situation est un peu différente. En effet, lorsqu'on compare les taux d'imperméabilisation issus de deux procédés de mesure différents sur un même territoire, comme c'est le cas avec CLC-HR d'une part et Teruti-Lucas d'autre part, on compare les résultats obtenus sur une même réalisation du processus spatial Z. Pour modéliser cette situation, on peut procéder de la manière suivante.

Il est d'usage d'examiner la variance d'échantillonnage pour désigner la variance de l'écart entre l'estimateur et la variable d'intérêt, telle qu'elle est donnée à la relation (10). Plusieurs modèles synthétiques sont possibles pour caractériser les situations respectives de CLC-HR et de Teruti-Lucas. Par exemple, pour modéliser le cas de l'enquête Teruti-Lucas, on peut considérer qu'on estime la variable $\bar{Z}(V)$ en observant ses valeurs sur n points distincts de l'ensemble V, soit :

$$\bar{z}^{TL} = \frac{1}{n}\sum_{i=1}^{n} Z(\mathbf{x}_i)$$

Dans le même ordre d'idées, dans le cas de CLC-HR, pour estimer $\bar{Z}(V)$, on construit un pavage de V, noté $(v_i)_{,i\in\{1,\dots,N\}}$ sur lequel on mesure les variables $z(v_i)$ et l'estimateur est la moyenne empirique de ces variables :

$$\begin{aligned} z(v_i) &= \frac{1}{|v_i|}\int_{v_i} Z(\mathbf{x})d\mathbf{x} \\ \bar{z}^{CL} &= \frac{1}{N}\sum_{i=1}^{N} z(v_i) \end{aligned}$$

Pour apprécier la cohérence des estimateurs $\bar{z}^{TL}$ et $\bar{z}^{CL}$, on peut s'intéresser à la variable caractérisant leur écart à la variable estimée $\bar{Z}(V)$ et en considérer la variance $\mathrm{var}(\bar{z}^{(.)} - \bar{Z}(V))$. Dans cette vision, la variance est une *variance de prédiction* au sens de Diggle & Ribeiro (2007). Cette double dimension de la variance de prédiction et d'estimation n'est pas sans lien avec la question, en théorie des sondages, de l'hypothèse portant sur l'existence d'une superpopulation dans laquelle serait tirée la population d'intérêt faisant, elle-même, l'objet du sondage (Tillé 2011).

Différentes approximations de ces termes de variances existent dans la littérature géostatistique, connues sous le terme de variance d'extension ou d'estimation (Matheron 1966, Matheron 1970, Chilès & Delfiner 1999, Chauvet 2008). On trouvera en annexe D le détail des calculs et approximations réalisés. On montre que lorsque l'on peut associer les points d'observation $\mathbf{x}_i$ à un pavage régulier de V, et pour un modèle de territoire métropolitain circulaire et un pavage composé lui-même de $N = 309000$ cercles (le nombre de points de l'enquête Teruti-Lucas), alors la variance se simplifie en :

$$\mathrm{var}(\bar{z}^{(.)} - \bar{Z}(V)) = \frac{1}{N}\left[\Phi_C(t) - \Phi_C(T)\right]$$

où T est le rayon de V et t celui de v. En l'occurrence, $t = 750m$. L'écart-type d'estimation s'élève à 0,015 point de pourcentage. Cette valeur est d'un ordre de grandeur inférieure à celle de 0,18 point annoncée dans la documentation méthodologique de l'enquête Teruti-Lucas (SSP 2014). Compte tenu des approximations réalisées [17], on peut considérer que ces résultats sont compatibles.

6 Conclusion

Au terme de ce document, nous avons établi l'importance pour l'analyse de tenir compte des corrélation spatiales. Nous nous sommes dotés d'outils qui permettent de simuler les observations corrélées d'artificialisation et d'imperméabilisation. Cette capacité à simuler permet de reproduire les différentes étapes de production d'une source statistique et d'examiner dans quelle mesure ces étapes modifient l'estimation des statistiques que l'on peut fonder sur ces sources. Nous avons établi que les résultats de calcul de taux d'imperméabilisation peuvent apparaître comme cohérents entre les différentes sources examinées dans ce document (CLC-HR et l'enquête Teruti), si on les adosse à un processus générateur de l'imperméabilisation. En revanche, les écarts de taux mesurés résistent à la modélisation d'erreurs de mesures centrées. En d'autre termes, il existe manifestement des biais de mesure associés à Teruti-Lucas et CLC-HR qui font que les taux d'imperméabilisation ne sont pas les mêmes, conditionnellement à la réalisation unique du processus d'imperméabilisation observée sur le territoire français.

17. territoire métropolitain considéré comme circulaire, plan de sondage systématique appliqué ici tandis qu'il est plus compliqué que cela dans Teruti-Lucas, coefficient de variation Teruti-Lucas publié pour les terrains artificialisés appliqué aux sols imperméabilisés (non publié).

Références

Allard, D. (2012). Statistiques spatiales : introduction à la géostatistique, *Lecture notes*, University of Montpellier.

Beran, J., Feng, Y., Ghosh, S. & Kulik, R. (2016). *Long-Memory Processes.*, Springer.

Chakir, R. & Madignier, A.-C. (2006). Analyse des changements d'occupation des sols en France entre 1992 et 2003, *Économie rurale* **296**.

Chauvet, P. (2008). *Aide-mémoire de géostatistique linéaire*, Presses des MINES.

Chilès, J.-P. & Delfiner, P. (1999). *Geostatistics : modeling spatial uncertainty*, John Wiley & Sons Inc., New York.

Cressie, N. (1993). *Statistics for Spatial Data*, Wiley.

Diggle, P. J. & Ribeiro, P. J. (2007). *Model-based Geostatistics*, Springer series in statistics, Springer.

Fontes-Rousseau, C. & Jean, R. (2015). Utilisation du territoire. L'artificialisation des terres de 2006 à 2014 : pour deux tiers sur des espaces agricoles, *Agreste Primeur* **326**.

Gobet, E. (2013). *Méthodes de Monte-Carlo et processus stochastiques : du linéaire au non-linéaire*, Editions de l'Ecole Polytechnique.

Janvier, F., Nirascou, F. & Sillard, P. (2016). L'occupation des sols en France : progression plus modérée de l'artificialisation entre 2006 et 2012, *Le Point sur* **219**.

Matheron, G. (1966). Présentation des variables régionalisées, *Journal de la société française de statistique* **107** : 263–275.

Matheron, G. (1970). *La théorie des variables régionalisées et ses applications*, École Nationale Supérieure des Mines de Paris.

Naimi, B., Skidmore, A. K., Groen, T. A. & Hamm, N. A. S. (2011). Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling, *Journal of biogeography* **38** : 1497–1509.

Pageaud, D. & Carré, C. (2009). La France vue par Corine Land Cover, outil européen de suivi de l'occupation des sols, *Le Point sur* **10**.

Papoulis, A. & Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*, McGraw-Hill.

Rouhaud, M. (2016). La géomatique dans le domaine environnemental : contribution à la modélisation de l'occupation des sols en France métropolitaine, *Rapport de master I Environnement*, Université Paris 1 Panthéon-Sorbonne.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas, *Kybernetika* **9**(6) : 449–460.

SSP (2014). L'utilisation du territoire en 2014 : Teruti-Lucas, *Agreste Chiffres et données Agricuture* **229**.

Tassi, P. (1992). *Méthodes statistiques*, Economica.

Tillé, Y. (2011). *Théorie des sondages : échantillonnage et estimation en population finie*, Dunod.

Annexes

A Estimation du taux d'imperméabilisation par l'enquête Teruti

L'enquête Teruti consiste à relever, sur un échantillon de points terrain, l'occupation et l'utilisation du sol dans un rayon de 1,5 mètres. Ces deux caractéristiques du sol sont renseignées à l'aide de nomenclatures spécifiques à l'enquête. L'imperméabilisation, c'est-à-dire le caractère imperméable du sol issu d'une artificialisation opérée par l'homme, n'est pas directement relevée. Pour construire une statistique d'imperméabilisation, on s'appuie sur la double nomenclature :

- d'occupation du sol selon la nomenclature NOCC (nouvelle nomenclature d'occupation) codée sur 4 caractères
- d'utilisation du sol LU1N (utilisation principale du sol) codée sur 3 caractères

Dans cette étude, on définit comme imperméabilisés, les points au sol dont les caractéristiques sur la double nomenclature vérifient :

- tous les sols bâtis [18] (NOCC = 11xx)
- les sols artificiels non bâtis de forme linéaire (NOCC = 122x) dont l'utilisation correspond à :
  - LU1N=31x) : Chemins de fer Routes et autoroutes, Transport par eau, Transports aériens, Transport par conduite (gazoduc. . .) et électricité ; Télécommunications ; Stockage, services auxiliaires des transports ; Infrastructure de protection (protection active des biens et des personnes)
  - LU1N=32x) : Fourniture et traitement des eaux ; Traitement des déchets
  - LU1N=33x) : Construction
  - LU1N=34x) : Commerce, finances, services
  - LU1N=35x) : Administrations, collectivités locales, établissements publics, activités associatives, religions
  - LU1N=37x) : Habitat individuel ; Habitat collectif
- les sols artificiels non bâtis de forme aréolaire - nu (NOCC=1212) dont l'utilisation correspond à :
  - LU1N=31x) : Chemins de fer Routes et autoroutes, Transport par eau, Transports aériens, Transport par conduite (gazoduc. . .) et électricité ; Télécommunications ; Stockage, services auxiliaires des transports ; Infrastructure de protection (protection active des biens et des personnes)
  - LU1N=32x) : Fourniture et traitement des eaux ; Traitement des déchets
  - LU1N=33x) : Construction
  - LU1N=34x) : Commerce, finances, services
  - LU1N=35x) : Administrations, collectivités locales, établissements publics, activités associatives, religions
  - LU1N=37x) : Habitat individuel ; Habitat collectif

En synthèse dans cette analyse, en dépit de la pratique (Fontes-Rousseau & Jean 2015) qui veut que soient réputés imperméabilisés tous les sols tels que NOCC=(11,12), sont exclus des sols imperméabilisés : les sols artificiels non bâtis de forme linéaire ou aréolaire-nu correspondant à des activités de sports, camps de vacances, jardins d'agrément et parcs publics, chasse, protection du milieu naturel ou d'autres activités liées à la culture et aux loisirs.

Le taux d'imperméabilisation est ensuite calculé à l'aide de la pondération de l'enquête (variable coef2012). Pour le territoire métropolitain, le taux d'imperméabilisation est estimé à 4,6%. Sous l'hypothèse d'indépendance des points d'observation [19], l'écart-type de cette estimation s'élève à 0,03%.

B Calcul du variogramme de la figure 6

La difficulté du calcul du variogramme repose sur la taille de la matrice raster. Théoriquement, un calcul de variogramme correspond au croisement de la grille par elle-même. Compte tenu de la

18. On note x un caractère quelconque. Il faut donc comprendre ∀x.

19. On sait que cette hypothèse est rejetée en pratique – cf. §3.3 – mais elle constitue une référence numérique utile.

taille importante de la grille CLC-HR (3,1 milliards de cellules), le calcul direct n'est pas possible. Le calcul du variogramme, pour un point donné fait appel d'abord aux points du voisinage. Une méthode consiste donc à limiter le raster sur laquelle le calcul est réalisé (fenêtre), et à reproduire l'estimation sur de nombreuses fenêtres sélectionnées au hasard. C'est la technique que nous avons utilisé dans ce document.
Techniquement, le variogramme proposé à la figure 6 est fondé sur la sélection aléatoire d'environ un millier de fenêtres de 20km de côté (figure 13). Un variogramme est estimé sur chaque fenêtre point par point. Un deuxième lot de fenêtres de 50km de côté est sélectionné pour déterminer les points du variogramme $h = 5km$ et $h = 15km$. Ces derniers calculs sont beaucoup plus longs car la taille des fenêtres devient critique à manipuler (6,2 millions de pixels pour chaque fenêtre simulée). Puis, les valeurs point par point de la courbe obtenues par simulation sont moyennées. Les courbes en rouges indiquent les intervalles de confiance à deux écart-types calculés point par point. Pour le calcul de ces intervalles de confiance, on ne tient pas compte de la continuité du variogramme, qui apparaît de manière évidente au vu du tracé. Si on tenait compte de la continuité du variogramme dans l'estimation (par exemple avec une technique d'estimation par régression locale), alors les intervalles seraient nettement réduits. Les intervalles de confiance matérialisés en rouge sur la figure 6 sont donc certainement très pessimistes.

Le calcul s'appuie sur le package usdm (Naimi et al. 2011) du logiciel R. Le calcul a été réalisé sur un serveur avec parallélisation du code R sur 10 cœurs pour accroître la performance. La durée du calcul des 1000 variogrammes pour les fenêtres de 20km de côté est d'environ 3 jours. C'est également le temps de calcul nécessaire à la détermination des deux points complémentaires de la courbe basée sur environ 100 fenêtres de 50km de côté (sans parallélisation).

C Passage de l'autocovariance d'un processus moyenné par moyenne mobile à celle du processus parent

Comme le processus spatial est isotrope, on est ramené à considérer un processus stochastique unidimensionnel. On notera dans cette partie, y le pendant unidimensionnel du processus spatial Y et z, le pendant unidimensionnel du processus Z. On suppose l'autocovariance de y connue, notée C_y. Compte tenu de de la relation 7, elle vaut :

$$C_y(h) = A\left(1 + \frac{|h|}{b}\right)^{-\alpha} \tag{14}$$

où A, b et α sont des réels positifs (voir tableau 2 pour leurs valeurs numériques estimées) et h est la distance dans le plan géographique. Dans sa version unidimensionnelle, la relation (9) qui lie y à z devient [20] :

$$y(t) = \frac{1}{2T}\int_{-T}^{T} z(t-u)du \tag{15}$$

On connaît C_y et on cherche l'autocovariance C_z du processus z. Etablissons l'équation fonctionnelle qui définit C_z conditionnellement à C_y.

On suppose que y et z sont deux fois dérivables. On dérive la relation (15) par rapport à t :

$$y'(t) = \frac{1}{2T}\int_{-T}^{T} z'(t-u)du$$

Le calcul de l'intégrale de droite conduit à l'égalité :

$$z(t+T) - z(t-T) = 2Ty'(t) \tag{16}$$

On dérive l'égalité (16) par rapport à t :

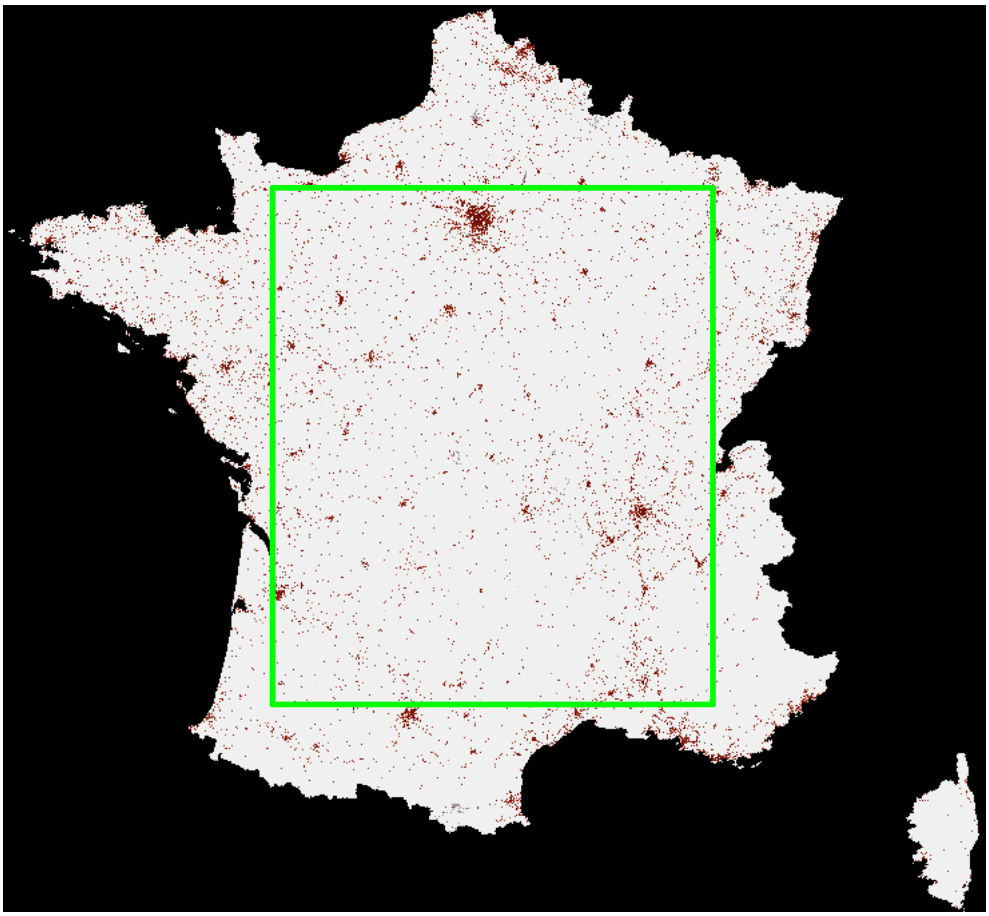
$$z'(t+T) - z'(t-T) = 2Ty''(t)$$

et par rapport à T :

$$z'(t+T) + z'(t-T) = 2y'(t)$$

20. numériquement, en l'espèce, $T = 10m$.

FIGURE 13 – Zone de sélection des centres des fenêtres pour l'estimation du variogramme de la figure 6



Note : le cadre vert précise le lieu des centres de fenêtres de simulation.

Il en découle, par combinaison linéaire des deux relations précédentes, que :

$$\left\{ \begin{array}{rcl} 2z'(t+T) & = & 2Ty''(t) + 2y'(t) \\ 2z'(t-T) & = & 2y'(t) - 2Ty''(t) \end{array} \right.$$

T est un paramètre. On effectue un changement de variables $u = t + T$ dans la première relation et $u = t - T$ dans la seconde. Il vient :

$$\left\{ \begin{array}{rcl} 2z'(u) & = & 2Ty''(u-T) + 2y'(u-T) \\ 2z'(u) & = & 2y'(u+T) - 2Ty''(u+T) \end{array} \right.$$

Finalement,

$$2z'(u) = T\left[y''(u-T) - y''(u+T)\right] + y'(u+T) + y'(u-T)$$

On intègre la relation précédente sur u :

$$2z(u) = K + T\left[y'(u-T) - y'(u+T)\right] + y(u+T) + y(u-T) \tag{17}$$

où K est une constante. On connaît l'autocovariance C_y de y. A partir de l'expression précédente, il est possible de déterminer celle de z :

$$C_z(\tau) = \mathbb{E}_u(\tilde{z}(u)\tilde{z}(u+\tau))$$

où $\tilde{z}$ désigne le processus z centré. En utilisant l'expression de z (relation 17) et en notant [21] C_{wx} la fonction de covariance croisée des processus stationnaires w et x, on a :

$$\begin{array}{rcl} 4C_{zz}(\tau) & = & \mathbb{E}_u\left\{T\left[\tilde{y}'(u+T+\tau) - \tilde{y}'(u+T+\tau)\right] + \tilde{y}(u+T+\tau) + \tilde{y}(u-T+\tau)\right\}. \\ & & \left\{T\left[\tilde{y}'(u+T) - \tilde{y}'(u+T)\right] + \tilde{y}(u+T) + \tilde{y}(u-T)\right\} \end{array}$$

Par construction, pour deux processus stationnaires w et x, $C_{wx}(\tau) = \mathbb{E}_u\left(\tilde{w}(u+\tau)\tilde{x}(u)\right)$. Il découle de cette propriété que l'expression précédente se simplifie en :

$$4C_{zz}(\tau) = \left[C_{yy}(\tau+2T) + 2C_{yy}(\tau) + C_{yy}(\tau-2T)\right] + T^2\left[-C_{y'y'}(\tau+2T) + 2C_{y'y'}(\tau) - C_{y'y'}(\tau-2T)\right]$$

On montre (Papoulis & Pillai (2002) – p.314) que, pour des processus stationnaires dérivables,

$$C_{y'y'}(\tau) = -C''_{yy}(\tau) \tag{18}$$

Il en découle que [22] :

$$4C_z(\tau) = \left[C_y(\tau+2T) + 2C_y(\tau) + C_y(\tau-2T)\right] + T^2\left[C''_y(\tau+2T) - 2C''_y(\tau) + C''_y(\tau-2T)\right] \tag{19}$$

Cette dernière expression fait intervenir l'autocovariance du processus y dont l'expression correspond à la relation (14). Elle fait aussi intervenir la dérivée seconde de cette autocovariance, laquelle correspond, théoriquement (relation 18), à l'opposé de l'autocovariance du processus dérivé y'. Or la dérivée seconde de $C_y(h)$ vaut :

$$C''_y(h) = A\frac{\alpha(1+\alpha)}{b^2}\left(1 + \frac{|h|}{b}\right)^{-\alpha-2}$$

Donc l'autocovariance du processus dérivé y' est négative avec cette expression en tout point, ce qui naturellement pose problème en 0 puisqu'en ce point, une autocovariance est forcément positive (elle peut être négative ailleurs qu'en 0). Il convient donc de régulariser la forme de C_y de sorte que sa dérivée seconde en 0 soit négative. D'un point de vue informationnel, on observe ici que le processus lissé y, s'il apporte une information exploitable sur le processus z pour les valeurs d'autocorrélation correspondant à des pas [23] de « temps » larges (typiquement au-delà de la fenêtre de lissage), il n'apporte pas d'information sur des pas de « temps » inférieurs à la fenêtre de lissage. Ceci est finalement logique.

A ce stade, deux solutions s'offrent à nous pour poursuivre les calculs. Soit on procède à une régularisation arbitraire de C_y en lui substituant une fonction dont :

— la dérivée seconde en 0 est négative ;

21. Avec cette notation $C_y \equiv C_{yy}$.

22. On reprend ici les notations précédentes, $C_y \equiv C_{yy}$.

23. Par abus de langage, on parle ici de pas de temps car tout se passe techniquement comme si nous traitions d'un processus à support temporel (i.e. unidimensionnel), temps et distance étant, dans ce cas, homologues.

— et qui est identique à C_y pour $|h|$ supérieur à un certain seuil a.

Soit on poursuit les calculs en se rappelant que la fonction C_z qui découle des calculs est inconnue sur un certain intervalle $\tau \in [-a, a]$ centré en 0, *grosso modo* de la taille de la fenêtre de lissage. Dans l'une ou l'autre des deux options précédentes, le comportement précis de la fonction C_z n'est pas connu en 0 (on sait quand même que sa dérivée seconde est négative). Conventionnellement, nous conviendrons que l'intervalle centré sur lequel l'information sur C_z est lacunaire est $[-2T, 2T]$. En effet, d'après l'expression (19), il ressort que l'expression de $C_z(\tau)$ pour $\tau \in]-\infty, -2T[\cup]2T, +\infty[$ ne fait pas intervenir la valeur $C_y''(0)$. Dans l'article nous optons pour la première des deux options en approximant l'autocorrélation C_y par une fonction de la forme $\sigma^2 \exp\left(-h^2/d^2\right)$ avec des paramètres σ^2 et d^2 convenablement choisis pour qu'en 0 la fonction de d'autocovariance corresponde à la variance empirique du processus et que la fonction soit continue au point de raccord $(2T)$ avec la partie identifiée de la fonction C_z telle qu'elle découle de la relation (19).

D Calcul de la variance de processus spatiaux

D.1 Variance de la variable $\bar{Z}(V)$ en coordonnées polaires

L'expression (relation 11)

$$\mathrm{var}(\bar{Z}(V)) = \frac{1}{|V|^2} \int_V \int_V C(\mathbf{u} - \mathbf{v}) d\mathbf{u} d\mathbf{v}$$

peut s'écrire à l'aide de fonctions indicatrices :

$$\mathrm{var}(\bar{Z}(V)) = \frac{1}{|V|^2} \int_{\mathbf{u}\in\mathbb{R}^2} \int_{\mathbf{v}\in\mathbb{R}^2} \mathbb{1}_V(\mathbf{u}) \mathbb{1}_V(\mathbf{v}) C(\mathbf{u} - \mathbf{v}) d\mathbf{u} d\mathbf{v}$$

Puis si on effectue le changement de variables (2D) :

$$\left\{ \begin{array}{rcl} \mathbf{x} & = & \mathbf{u} \\ \mathbf{h} & = & \mathbf{u} - \mathbf{v} \end{array} \right.$$

$$\begin{array}{rcl} \mathrm{var}(\bar{Z}(V)) & = & \frac{1}{|V|^2} \int_{\mathbf{x}} \int_{\mathbf{h}} \mathbb{1}_V(\mathbf{x} + \mathbf{h}) \mathbb{1}_V(\mathbf{x}) C(\mathbf{h}) d\mathbf{x} d\mathbf{h} \\ & = & \frac{1}{|V|^2} \int_{\mathbf{h}} C(\mathbf{h}) \underbrace{\int_{\mathbf{x}} \mathbb{1}_V(\mathbf{x} + \mathbf{h}) \mathbb{1}_V(\mathbf{x}) d\mathbf{x}}_{f(\mathbf{h})} d\mathbf{h} \end{array} \quad (20)$$

Le cas où V est un disque de rayon T et C est isotrope peut être développé analytiquement.

Calculons tout d'abord $f(\mathbf{h})$ en coordonnées polaires. $f(\mathbf{h})$ correspond à la surface de l'intersection de deux disques de rayon T décalés d'une translation $\mathbf{h}$. Du fait de la symétrie du problème (figure 14), f ne dépend que de la norme de $\mathbf{h}$ que l'on notera h. Calculons cette aire en fonction de h et T.

A l'aide de la figure 15, on observe que l'aire recherchée, $f(h)$ vérifie, pour $0 \leqslant h \leqslant 2T$:

$$f(h) = 2 \times \mathscr{A}$$

avec

$$\mathscr{A} = Aire(\widehat{OAB}) - Aire(OAB)$$

où $\widehat{OAB}$ désigne l'arc de cercle et AOB le triangle.

Dans le triangle OHA, $\frac{OH}{OA} = \cos\frac{\theta}{2}$ donc $\frac{\theta}{2} = \arccos\left(\frac{h}{2T}\right)$. Par conséquent,

$$Aire(\widehat{OAB}) = \frac{\theta}{2\pi} \times \pi T^2 = T^2 \arccos\left(\frac{h}{2T}\right)$$

Puis, $Aire(OHA) = \frac{1}{2} AH \times HO = \frac{h}{4} T \sin\frac{\theta}{2}$. Or $\sin(\arccos(x)) = \sqrt{1 - x^2}$ donc

$$Aire(OAB) = 2Aire(OHA) = \frac{hT}{2} \sqrt{1 - \left(\frac{h}{2T}\right)^2}$$

Finalement,

$$f(h) = 2\mathscr{A} = 2 \left(T^2 \arccos\left(\frac{h}{2T}\right) - \frac{hT}{2} \sqrt{1 - \left(\frac{h}{2T}\right)^2} \right) \mathbb{1}_{[0,2T]}(h)$$

FIGURE 14 – Calcul de l'aire d'intersection correspondant à $f(\mathbf{h})$ - I

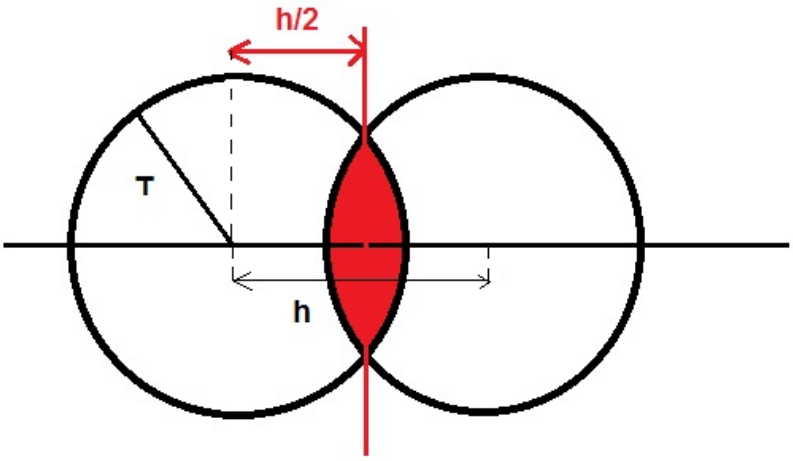
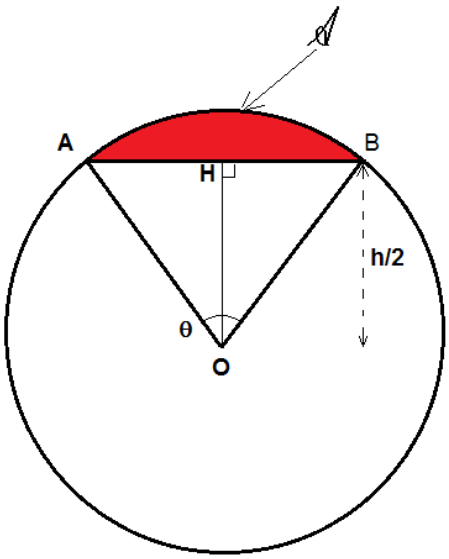


FIGURE 15 – Calcul de l'aire d'intersection correspondant à $f(\mathbf{h})$ - II



Note : $OA = OB = T$

A l'aide de ces éléments, il est possible de calculer $\text{var}(\bar{Z}(V))$ dans le cas où V est un disque de rayon T et C est isotrope (i.e. $C(\mathbf{h}) = C(h)$). Sous ces hypothèses, en effet, la relation (20) s'écrit, en coordonnées polaires [24] ($\mathbf{h} = (r\cos(\theta)\ ;\ r\sin(\theta))'$ et $|V|^2 = (\pi T^2)^2$) :

$$\text{var}(\bar{Z}(V)) = \frac{2}{|V|^2}\int_{r=0}^{2T}\int_{\theta=0}^{2\pi} C(r)\left\{T^2\arccos\frac{r}{2T} - \frac{rT}{2}\sqrt{1-\left(\frac{r}{2T}\right)^2}\right\} r dr d\theta$$

Soit, comme $|V|^2 = (\pi T^2)^2$,

$$\text{var}(\bar{Z}(V)) = \frac{4\pi}{(\pi T^2)^2}\int_{r=0}^{2T} C(r)\left\{T^2\arccos\frac{r}{2T} - \frac{rT}{2}\sqrt{1-\left(\frac{r}{2T}\right)^2}\right\} r dr$$

On effectue un changement de variable en $u = \frac{r}{2T}$ et on conclut que $\text{var}(\bar{Z}(V)) = \Phi_C(T)$ où

$$\Phi_C(T) = \frac{16}{\pi}\int_{u=0}^{1} uC(2uT)\left\{\arccos(u) - u\sqrt{1-u^2}\right\} du \tag{21}$$

D.2 Variance d'estimation

On considère un ensemble V et v un ensemble inclus dans V. On considère les variables

$$\bar{Z}(V) = \frac{1}{|V|}\int_V Z(\mathbf{x})d\mathbf{x}$$

et

$$z(v) = \frac{1}{|v|}\int_v Z(\mathbf{x})d\mathbf{x}$$

On s'intéresse à la variance de la variable d'extension (Chilès & Delfiner 1999) consistant à prédire $\bar{Z}$ par $z(v)$. Nous formons donc [25]

$$\begin{aligned}\text{var}(z(v)-\bar{Z}(V)) &= \mathbb{E}\left[\frac{1}{|V|}\int_V Z(\mathbf{u})d\mathbf{u} - \frac{1}{|v|}\int_v Z(\mathbf{u})d\mathbf{u}\right]\left[\frac{1}{|V|}\int_V Z(\mathbf{v})d\mathbf{v} - \frac{1}{|v|}\int_v Z(\mathbf{v})d\mathbf{v}\right]\\ &= \frac{1}{|V|^2}\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_V(\mathbf{u})\mathbb{1}_V(\mathbf{v})d\mathbf{u}d\mathbf{v} + \frac{1}{|v|^2}\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_v(\mathbf{u})\mathbb{1}_v(\mathbf{v})d\mathbf{u}d\mathbf{v}\\ &\quad -\frac{2}{|v||V|}\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_v(\mathbf{u})\mathbb{1}_V(\mathbf{v})d\mathbf{u}d\mathbf{v}\end{aligned}$$

Considérons à présent un pavage de l'ensemble V à l'aide de briques élémentaires $(v_i)_{i\in\{1,\dots,N\}}$, chaque v_i étant de taille $|v|$. Les v_i sont deux à deux disjoints et leur réunion forme V. Examinons la variance de prédiction de $\bar{Z}(V)$ à l'aide de la variable

$$\bar{z}(v) = \frac{1}{N}\sum_{i=1}^{N} z(v_i)$$

Techniquement, on cherche à déterminer $\text{var}(\bar{z}(v)-\bar{Z}(V)) = \text{var}\left(\frac{1}{N}\sum_{i=1}^{N}(z(v_i)-\bar{Z}(V))\right)$. Comme les v_i sont disjoints,

$$\begin{aligned}\mathbb{E}(z(v_i)z(v_j)) &= \frac{1}{|v|^2}\iint Z(\mathbf{u})\mathbb{1}_{v_i}(\mathbf{u})Z(\mathbf{v})\mathbb{1}_{v_j}(\mathbf{v})d\mathbf{u}d\mathbf{v}\\ &= \frac{1}{|v|^2}\iint C(\mathbf{u}-\mathbf{v})\mathbb{1}_{v_i}(\mathbf{u})\mathbb{1}_{v_j}(\mathbf{v})d\mathbf{u}d\mathbf{v}\end{aligned}$$

On observe que les cas où l'intégrande précédent n'est pas nul correspondent aux situations où $\mathbf{u}-\mathbf{v} \neq 0$. Or C est décroissante par rapport à la norme de son argument. Par conséquent, il est vraisemblable que l'intégrale précédente soit *a priori* petite par rapport aux autres termes de $\text{var}(\bar{z}(v)-\bar{Z}(V))$. On fait donc l'hypothèse, classique en géostatistique (Matheron 1970), qu'il est négligeable.

De ce fait,

$$\begin{aligned}N^2\text{var}(\bar{z}(v)-\bar{Z}(V)) &= \frac{N}{|V|^2}\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_V(\mathbf{u})\mathbb{1}_V(\mathbf{v})d\mathbf{u}d\mathbf{v} + \frac{N}{|v|^2}\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_v(\mathbf{u})\mathbb{1}_v(\mathbf{v})d\mathbf{u}d\mathbf{v}\\ &\quad -\frac{2}{|v||V|}\sum_i\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_{v_i}(\mathbf{u})\mathbb{1}_V(\mathbf{v})d\mathbf{u}d\mathbf{v}\end{aligned}$$

24. intégrales simples
25. On suppose sans perte de généralité que $\mathbb{E}Z(\mathbf{x}) = 0$.

Le dernier terme du membre de droite de l'égalité précédente, en sommant sur les v_i qui forment une partition de V, est égal à $\frac{-21}{|v||V|}\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_V(\mathbf{u})\mathbb{1}_V(\mathbf{v})d\mathbf{u}d\mathbf{v}$. Puis comme $N|v| = |V|$, il vient :

$$N^2\text{var}(\bar{z}(v)-\bar{Z}(V)) = \frac{N}{|v|^2}\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_v(\mathbf{u})\mathbb{1}_v(\mathbf{v})d\mathbf{u}d\mathbf{v} - \frac{N}{|V|^2}\iint C_Y(\mathbf{u}-\mathbf{v})\mathbb{1}_V(\mathbf{u})\mathbb{1}_V(\mathbf{v})d\mathbf{u}d\mathbf{v}$$

Soit :

$$\text{var}(\bar{z}(v)-\bar{Z}(V)) = \frac{1}{N}\left[\frac{1}{|v|^2}\iint_v C_Y(\mathbf{u}-\mathbf{v})d\mathbf{u}d\mathbf{v} - \frac{1}{|V|^2}\iint_V C_Y(\mathbf{u}-\mathbf{v})d\mathbf{u}d\mathbf{v}\right]$$

On se replace dans l'approximation, employée au paragraphe précédent, d'un ensemble V prenant la forme d'un cercle de rayon T. On suppose en outre que v est aussi un ensemble circulaire de rayon t ($t < T$). Moyennant quoi, pour la fonction d'autocorrélation isotrope du processus Y, les résultats du paragraphe précédent s'appliquent et :

$$\text{var}(\bar{z}(v)-\bar{Z}(V)) = \frac{1}{N}\left(\Phi_C(t)-\Phi_C(T)\right) \tag{22}$$

où l'expression de Φ_C est donnée à la relation (21).

Ainsi, il est possible sous certaines hypothèses simplificatrices de calculer la variance d'estimation associé à la prédiction d'une réalisation donnée du champ à l'aide de mesures réalisées sur un ensemble de points. Une pratique courante, notamment pour traiter des contours d'ensemble V particuliers, est de procéder par simulation, d'où l'intérêt d'être en mesure de simuler le processus sous-jacent (cf. §3).