Modélisation géostatistique

David Causeur

Laboratoire de Mathématiques Appliquées
Pôle d'Enseignement Supérieur et de Recherche Agronomique de Rennes
65, rue de St-Brieuc - CS 84215
35042 Rennes Cedex
email : david.causeur@agrocampus-rennes.fr

Table des matières

1 Introduction

1.1 Objectifs pédagogiques

Ce cours d'introduction à la modélisation géostatistique s'inscrit dans le cadre de formations de 3ème cycle à vocation professionnalisante (type 2ème année de master professionnel). Les compétences requises sont :

- d'une part une bonne connaissance des démarches de modélisation et des modèles de régression,
- d'autre part un savoir-faire minimal dans le domaine de l'application des méthodes statistiques.

La présentation des modèles de statistiques spatiales et de leurs propriétés passe nécessairement par une formalisation mathématique. L'objectif de ce cours est de limiter cette formalisation pour mettre en relief les concepts importants et privilégier la confrontation avec des problèmes concrets.

1.2 Statistiques spatiales

La définition des contours d'un ensemble de méthodes statistiques dédiées à l'étude de la variabilité spatiale peut être considérée comme récente à l'échelle de l'histoire de la statistique puisqu'attribuable à Cressie (1993). Auparavant, les concepts sous-jacents aux différentes méthodes qui composent aujourd'hui les **statistiques spatiales** apparaissent essentiellement dans les sciences de l'environnement sous des formes plus orientées vers des applications spécifiques : évaluation des ressources minières, gestion des ressources forestières ou halieutiques, météorologie, ...

Récemment, le développement des technologies de positionnement et simultanément des ressources informatiques a permis l'explosion des Systèmes d'Information Géographique. Ces nouveaux outils informatiques facilitent la gestion d'une information d'ordre géographique et favorisent donc naturellement le recours aux méthodes de statistiques spatiales. L'intrusion de ces outils a suscité le développement de méthodes d'analyse spatiale dans de nouveaux domaines d'application où la géographie joue un rôle prépondérant : l'économie, l'aménagement du territoire, la gestion du paysage, ...

Indépendamment de son contexte d'application, l'unité des méthodes de statistiques spatiales se traduit par le fait que l'objet d'étude peut toujours être représenté de la façon suivante : $Z(S)$, où Z est une variable aléatoire mesurée au site géographique S. Pour fixer les idées, dans un domaine d'application très classique de ce type de méthodes, Z peut être une caractéristique physique ou chimique d'un sol et S l'ensemble des coordonnées géographiques, en l'occurrence abscisse et ordonnée, du site d'observation de Z.

En pratique, il faut distinguer plusieurs types de situations appelant des traitements statistiques spécifiques :

- le cas où la variable de localisation S est elle-même l'objet de l'étude, par exemple lors de l'étude de la répartition spatiale d'espèces végétales ou animales. Les sites de localisation de ces espèces se répartissent en effet selon un mécanisme aléatoire qui constitue le cœur de la problématique.

- le cas où les sites de localisation ne sont pas répartis de manière aléatoire mais agencés de manière régulière. Par exemple, certaines techniques spatiales d'analyse d'image reposent sur la représentation suivante : les pixels de l'image constituent une grille régulière de sites de mesures d'une quantité d'intérêt Z, à savoir le niveau de gris. Les variations de Z en fonction du pixel sont ici au centre de l'étude.

- le cas où une quantité d'intérêt est mesurée en des sites expérimentaux choisis. L'exemple emblématique de ce type de situation est, dans le domaine de la prospection minière, celui où Z est la hauteur d'un filon mesurée par carottage en différents sites. Ici, la géographie des sites joue un rôle central dans la loi de variation de Z. Cette propriété définit le cadre de travail de la **géostatistique**.

1.3 Un exemple d'application de la géostatistique : la variabilité intra-parcellaire d'un rendement cultural

La caractérisation de la variabilité intra-parcellaire est un enjeu important de l'agriculture moderne, qui appelle le développement de nouvelles technologies et par conséquent de nouvelles méthodologies regroupées sous l'appellation « agriculture de précision ». L'objectif de ces méthodes est d'évaluer la variabilité des rendements culturaux et des caractéristiques physico-chimiques du sol à l'échelle d'une parcelle. Dans le cas de grandes parcelles, la bonne connaissance de cette variabilité permet par exemple une répartition intelligente des engrais chimiques et par conséquent une économie dans le traitement des parcelles.

On dispose ainsi de mesures de rendements culturaux sur une parcelle expérimentale parcourue par un engin agricole équipé d'un système de positionnement. La figure 1 représente les positions des sites de mesure sur une parcelle.

Dans un premier temps, par souci de simplicité, on ne s'intéresse à la variabilité d'une mesure de rendement que le long d'un trajet rectiligne de l'engin agricole signalé par une flèche sur la figure 1. La figure 2 représente les mesures du rendement en fonction de la distance parcourue par l'engin agricole, si l'on suppose qu'il part de la flèche de la figure 1.

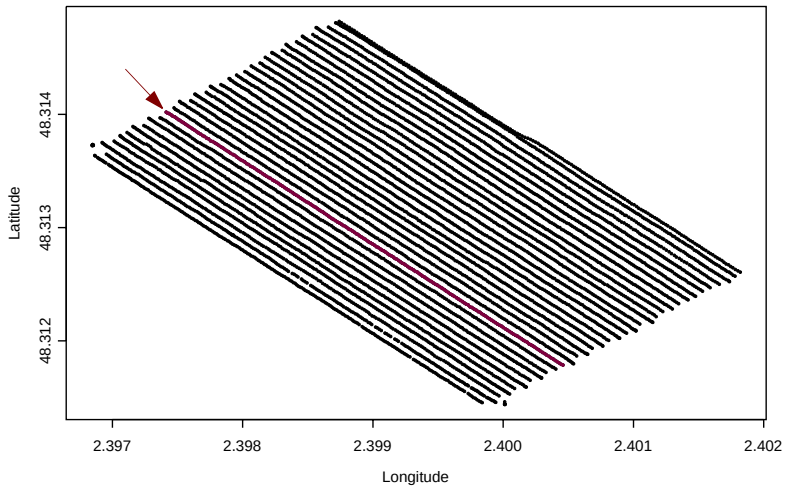


FIG. 1: *Sites de mesures du rendement sur une parcelle*

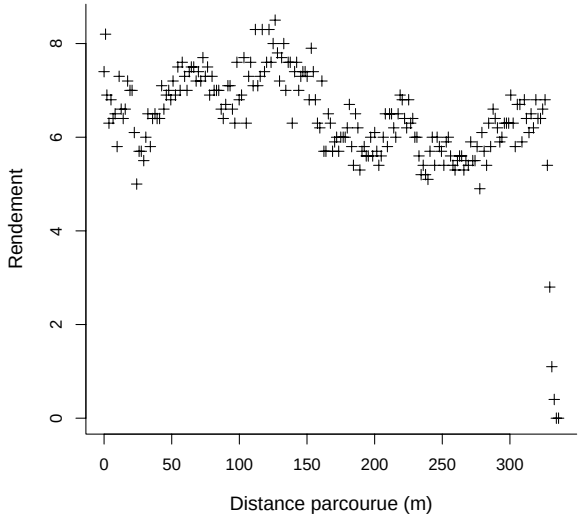


FIG. 2: *Rendement le long d'un trajet rectiligne de l'engin agricole*

1.4 Plan du cours

Dans la deuxième partie de ce cours, on propose une approche descriptive de la variabilité spatiale. L'essentiel des techniques qui y sont présentées ne relèvent pas spécifiquement des statistiques spatiales. Cette première approche est le plus souvent un outil d'aide à la modélisation. Dans la troisième partie, on s'intéresse à la modélisation géostatistique et aux propriétés des modèles géostatistiques. Enfin, la quatrième partie est consacrée à l'utilisation de modèles géostatistiques dans un objectif de cartographie.

2 Description de données géostatistiques

L'analyse exploratoire des données est toujours une étape préalable utile, voire nécessaire, à une démarche de modélisation. Elle permet une première prise de connaissance des plages de variations des données, des liens éventuels entre les caractéristiques que l'on étudie, d'une éventuelle structuration des données ... et oriente donc naturellement le choix du modèle qui formalisera l'ensemble de ces caractéristiques.

2.1 Données géostatistiques

Soit $Z(s)$ la mesure de la caractéristique que l'on étudie au site s et $\{s_1, s_2, \ldots, s_n\}$, $n \geq 1$, l'ensemble des sites de mesure. Les données géostatistiques sont constituées par le vecteur $[Z(s_1),\ Z(s_2), \ldots,\ Z(s_n)]$.

En pratique, l'essentiel des outils logiciels permettant la description de données géostatistiques se restreignent aux cas de sites de mesure à support dans $\mathbb{R}$ ou $\mathbb{R}^2$. Dans le cas de l'exemple 1.3 concernant la modélisation de la variabilité d'un rendement cultural le long d'un trajet de l'engin agricole, le support des mesures peut être assimilé à un segment de $\mathbb{R}$. On peut alors représenter de manière simple les variations du rendement sur ce trajet : la figure 2 est d'ailleurs un exemple de représentation naïve des données observées du rendement le long du segment.

L'objectif de la suite de cette partie est de présenter des outils de représentation des données géostatistiques lorsque le support des sites de mesure est dans $\mathbb{R}^2$.

2.2 Représentations graphiques de données géostatistiques

La représentation de données géostatistiques dont le support des sites de mesure est dans $\mathbb{R}^2$ soulève des problèmes de lisibilité graphique liés au caractère nécessairement tridimensionnel de cette représentation. Dans la suite, on explore trois possibilités de remédier à ces problèmes de lisibilité :

- la représentation perspective,
- la représentation topographique
- et la représentation picturale.

2.2.1 Représentation perspective

La représentation perspective présentée ici est une extension de la représentation naïve de la figure 2 dans laquelle les différents points du nuage seraient reliés entre eux en respectant l'ordre du support. Lorsque ce support est une partie de $\mathbb{R}$, cette nouvelle représentation est

de réalisation simple et donne, dans le cas des données de la figure 2, le graphique de la figure 3.

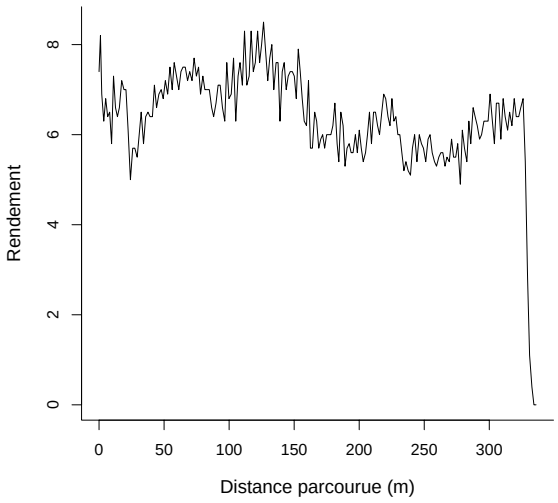


FIG. 3: *Rendement le long d'un trajet rectiligne de l'engin agricole*

L'extension de ce type de graphique au cas d'un support dans $\mathbb{R}^2$ suppose l'existence d'un ordre sur l'ensemble des sites de mesure. Cet ordre n'est naturel que si les sites de mesure sont disposés sur une grille définissant sans ambiguïté les successeurs et les antécédents de chaque site de mesure. Cependant, ce cas particulier de disposition des sites de mesure est marginal dans le cadre des applications de la géostatistique. Par exemple, la figure 4 montre la localisation de sites de mesures de la teneur en phosphates sur la même parcelle que celle mentionnée dans la partie 1.3.

Comme le montre la figure 5, on peut alors superposer une grille rectangulaire à mailles régulières à la disposition des sites de mesure. Il reste alors à calculer une « évaluation » de la teneur en phosphates en chaque nœud de la grille. Chacun de ces nœuds appartient nécessairement à un triangle de sites de mesure : on évalue alors la teneur en phosphates en un nœud par interpolation linéaire à partir des valeurs observées sur les sites de mesure formant les sommets du plus petit triangle entourant le nœud. Dans le cas présent, on obtient la représentation perspective de la figure 6.

Sauf dans le cas d'une évolution spatiale très progressive et mettant en évidence une structuration simple de la variabilité spatiale, la représentation perspective est le plus souvent plus spectaculaire que réellement efficace.

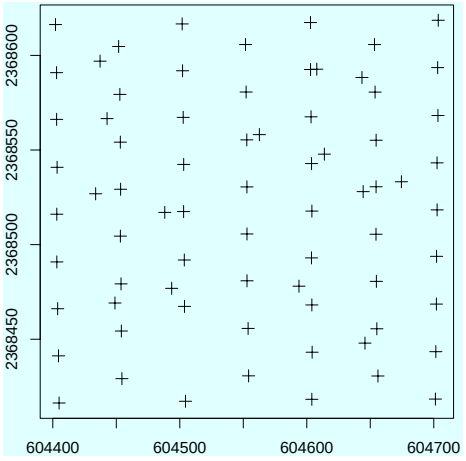


FIG. 4: *Disposition des sites de mesures de la teneur en phosphate sur la parcelle*

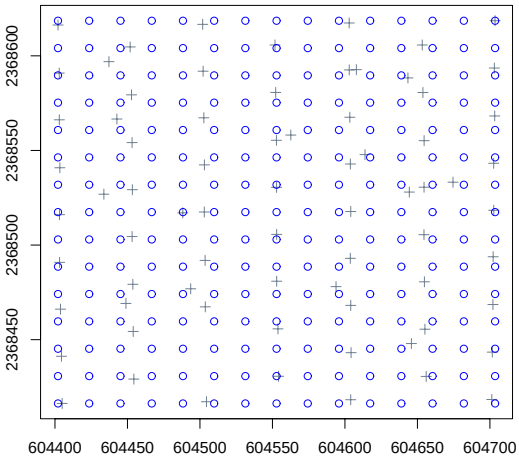


FIG. 5: *Grille régulière servant du support à la représentation perspective*

2.2.2 Représentation topographique

La représentation topographique est un exemple de conversion en deux dimensions de l'information graphique contenue dans la représentation perspective. Le principe en est le

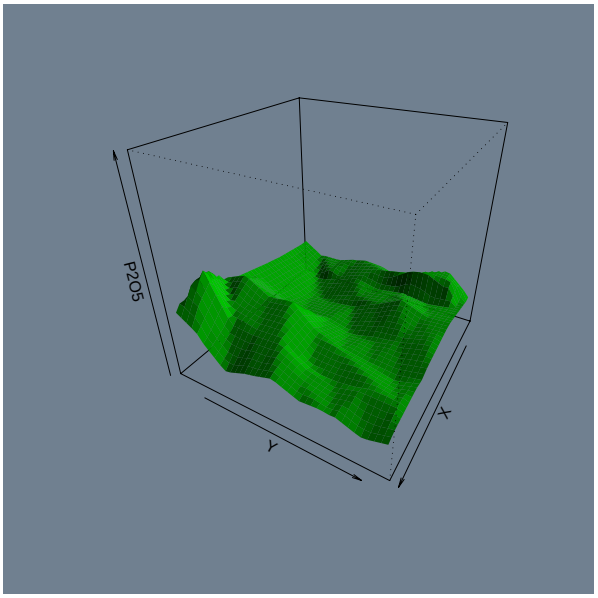


FIG. 6: *Représentation perspective de la variabilité intra-parcellaire des teneurs en phosphates*

suivant : par intersection entre la surface obtenue par la représentation perspective et un plan parallèle à celui contenant les sites de mesure, on obtient des courbes le long desquelles Z garde une valeur constante. La projection de ces courbes sur le plan des sites de mesure donne les courbes de niveaux. La représentation topographique consiste alors en un tracé de courbes de niveaux sur le plan contenant les sites de mesure.

Par exemple, la figure 7 donne une représentation topographique des variations spatiales de teneurs en phosphates sur la parcelle.

2.2.3 Représentation picturale

Comme la représentation perspective, la représentation picturale s'appuie sur une grille à mailles régulières pour laquelle on dispose d'une évaluation par interpolation de la valeur de Z en chaque nœud. La représentation picturale consiste alors à associer à chaque nœud de la grille un carré dont le niveau de gris est proportionnel à la valeur de Z en ce nœud.

Dans le cas de la représentation des variations spatiales de teneurs en phosphates, on obtient par exemple le graphique de la figure 8.

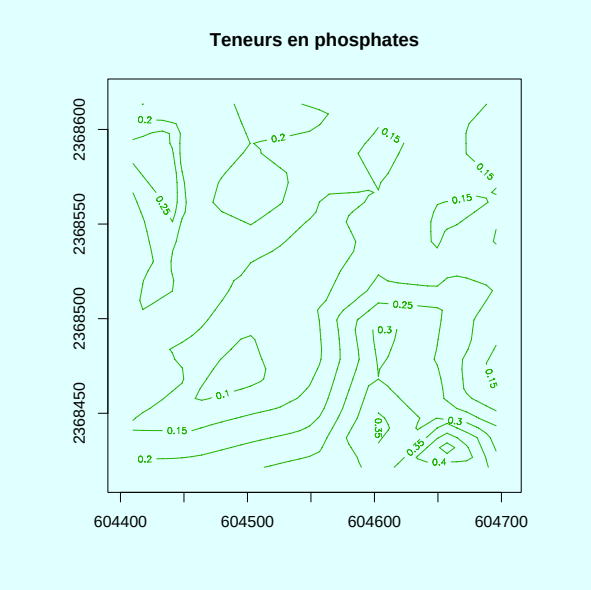


FIG. 7: *Représentation topographique de la variabilité intra-parcellaire des teneurs en phosphates*

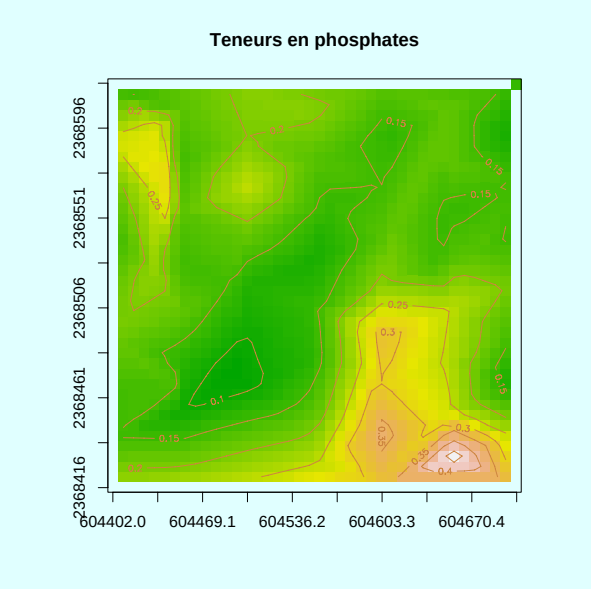


FIG. 8: *Représentation picturale de la variabilité intra-parcellaire des teneurs en phosphates*

3 Modélisation géostatistique

3.1 Introduction

A l'origine, l'introduction d'une approche mathématique dans le contexte de la prospection minière vise essentiellement à évaluer le volume d'un gisement à partir de mesures ponctuelles de la hauteur de ce gisement. Plus généralement, la modélisation géostatistique intervient aujourd'hui dans tous les problèmes de cartographie.

En pratique, il est naturel de distinguer 2 types de variabilités dans les mesures d'un phénomène : d'une part, des variations spatiales à l'échelle de la plage d'observation et d'autre part des variations locales autour de la tendance spatiale. Dans le cas de la modélisation de la variabilité intra-parcellaire présenté dans la partie 1.3, on peut matérialiser les variations à grande échelle du rendement en fonction de la distance au bord de la parcelle par la courbe de la figure 9. La figure 10 représente les écarts entre les valeurs observées du rendement et la tendance, telle qu'elle est représentée dans la figure 9 : ces écarts représentent donc les variations locales du rendement autour de la tendance spatiale.

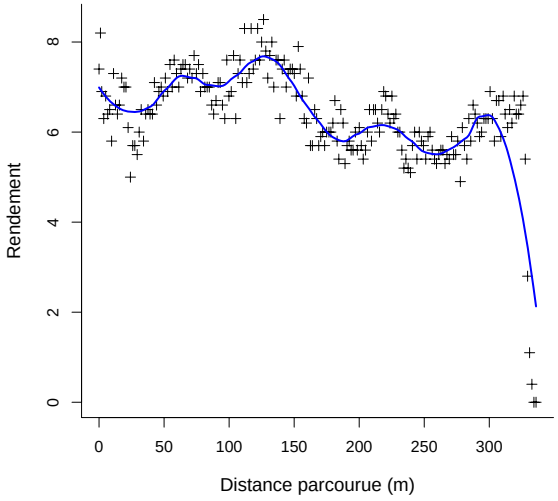


FIG. 9: *Variations à grande échelle du rendement*

La partie 3.2 est consacrée à la décomposition de la variabilité dans le modèle géostatistique. La modélisation des variations locales et de la tendance spatiale font l'objet d'une troisième partie et d'une quatrième partie respectivement.

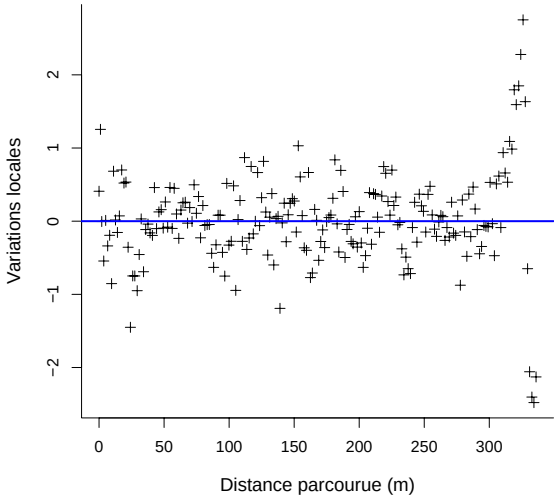


FIG. 10: *Variations locales du rendement*

3.2 Décomposition d'un processus

Soit $Z(s)$ la variable aléatoire modélisant la mesure au site s, on décompose alors $Z(s)$ de la manière suivante :

$$Z(s) \quad = \quad \mu(s) + \varepsilon(s) \tag{1}$$

où

- $\mu(.)$ est la tendance spatiale,
- $\varepsilon(s)$ est une variable aléatoire centrée et telle que :

$$C\left[\varepsilon(s), \varepsilon(s')\right] \quad = \quad c(s, s'),$$

et $c(.,.)$ est appelée dans la suite, fonction de **covariance** du processus spatial Z.

Différentes hypothèses sur la tendance spatiale et sur ε permettent de spécifier le modèle dans l'optique d'une interprétation plus facile. En pratique, ces hypothèses viseront essentiellement à supposer une certaine régularité dans les variations de $Z(s)$.

3.3 Modélisation des variations locales

3.3.1 Stationnarité d'un processus spatial

Dans un premier temps, il est intéressant de se demander si la variabilité locale du processus obéit à la même loi dans toute la plage d'observation du processus. Dans le cas contraire, on ne peut pas considérer que la variabilité locale peut faire l'objet d'un traitement global sur l'ensemble de la plage d'observation du processus. Cette propriété d'invariance est appelée stationnarité du processus.

Definition 3.1 Stationnarité d'un processus spatial

Soit ε un processus spatial, on dit que ε est stationnaire si, pour tout $n \geq 1$, pour tout $n-$uplet $\{s_1, s_2, \ldots, s_n\}$ de sites de mesures, pour tout h,

$$\mathcal{L}\left[\varepsilon(s_1), \varepsilon(s_2), \ldots, \varepsilon(s_n)\right] \quad = \quad \mathcal{L}\left[\varepsilon(s_1+h), \varepsilon(s_2+h), \ldots, \varepsilon(s_n+h)\right]$$

où $\mathcal{L}[.]$ désigne la loi d'un vecteur aléatoire.

En d'autres termes, la loi d'un $n-$uplet de mesures d'un processus spatial stationnaire ne dépend que des positions relatives des n sites en lesquels ces mesures sont effectuées et non de la localisation de ces n sites sur la plage d'observation du processus.

Si ε est stationnaire alors il est facile de montrer que, pour tout couple (s, s') de sites de mesure,

$$\begin{aligned} \mathbb{E}\left[\varepsilon(s)\right] &= \mathbb{E}\left[\varepsilon(s')\right], \\ \mathrm{Var}\left[\varepsilon(s)\right] &= \mathrm{Var}\left[\varepsilon(s')\right], \end{aligned}$$

et il existe une fonction C telle que,

$$\mathrm{Cov}\left[\varepsilon(s), \varepsilon(s')\right] \quad = \quad C(s'-s).$$

La stationnarité d'un processus spatial induit donc une forme de stationnarité plus faible dont la définition ne porte que sur les moments d'ordre 1 et 2 du processus.

Definition 3.2 Stationnarité à l'ordre 2 d'un processus spatial

On dit que ε est stationnaire à l'ordre 2 si l'espérance de $\varepsilon(s)$ ne dépend pas de s et s'il existe une fonction C telle que, pour tout (s, s'),

$$\mathrm{Cov}\left[\varepsilon(s), \varepsilon(s')\right] \quad = \quad C(s'-s).$$

La fonction symétrique C est appelée **autocovariance** du processus ε.

3.3.2 Régularité des variations locales

Dans le domaine de l'analyse des fonctions, la régularité est une notion générale que l'on peut assimiler à l'ordre de dérivabilité. Ainsi, l'absence de continuité d'une fonction traduit la situation de plus forte irrégularité. Au contraire, plus une fonction est continûment dérivable, plus son tracé est lisse et donc régulier. Lorsqu'il s'agit de processus aléatoire, la notion de régularité s'inspire fortement de celle utilisée en analyse fonctionnelle.

Definition 3.3 Continuité en moyenne quadratique

Soit ε un processus spatial, on dit que ε est **continu en moyenne quadratique** *si, pour tout s,*

$$\lim_{h\to 0} \mathbb{E}\left[(\varepsilon(s+h)-\varepsilon(s))^2\right] \quad = \quad 0.$$

Notons que $\mathbb{E}\left[(\varepsilon(s+h)-\varepsilon(s))^2\right] = \mathrm{Var}\left[\varepsilon(s+h)-\varepsilon(s)\right]$. Cette dernière quantité mesure donc la dispersion de l'écart entre $\varepsilon(s)$ et $\varepsilon(s+h)$. Dire que ε est continu en moyenne quadratique revient donc à dire que la dispersion de l'écart entre $\varepsilon(s)$ et $\varepsilon(s+h)$ est d'autant plus proche de 0 que l'écart h entre les sites est faible.

Dans le cas d'un processus ε stationnaire à l'ordre 2 :

$$\begin{aligned}\mathrm{Var}\left[\varepsilon(s+h)-\varepsilon(s)\right] &= \mathrm{Var}\left[\varepsilon(s+h)\right] + \mathrm{Var}\left[\varepsilon(s)\right] - 2\,\mathrm{Cov}\left[\varepsilon(s+h),\varepsilon(s)\right], \\ &= 2\left[C(0)-C(h)\right],\end{aligned}$$

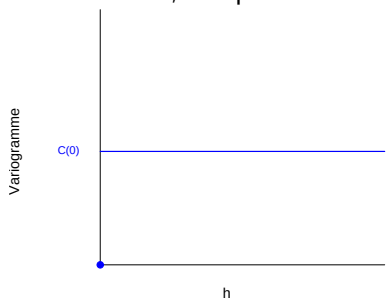
où C est la fonction d'auto-covariance du processus ε.

Dans la suite, la fonction $\gamma(h) = [C(0)-C(h)]$ est appelée **variogramme** du processus. Par définition, γ est une fonction symétrique, positive et $\gamma(0)=0$.

La situation de plus forte irrégularité des variations locales correspond à une absence de dépendances entre des mesures effectuées en des sites différents, ce qui se traduit par :

Pour tout $h \neq 0$,

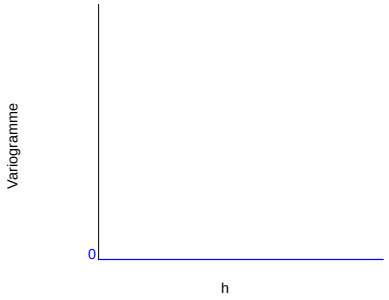
$$\begin{aligned}C(h) &= 0, \\ \gamma(h) &= C(0).\end{aligned}$$



Inversement, la situation théorique de plus grande régularité correspond à des dépendances locales maximales :

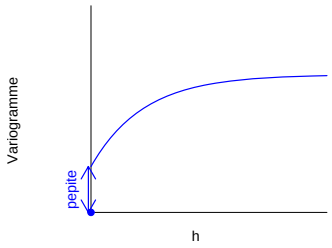
Pour tout $h \neq 0$,

$$\begin{aligned} C(h) &= C(0), \\ \gamma(h) &= 0. \end{aligned}$$



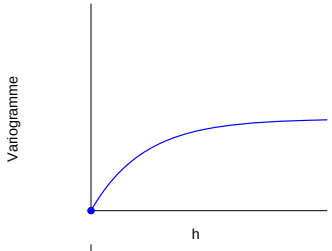
Entre ces deux situations extrêmes, la régularité du processus des variations locales se mesure à la régularité fonctionnelle du variogramme de ce processus au voisinage de 0. Ainsi, comme on l'a vu, ε est continu en moyenne quadratique si, et seulement si, le variogramme est continu en 0. Lorsque ce n'est pas le cas, on parle d'**effet pépite** et on appelle $\gamma_0 = \lim_{h \to 0} \gamma(h)$ la **pépite** du variogramme.

ε n'est pas continu en moyenne quadratique

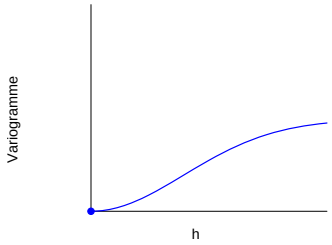


0n pourrait également montrer que la dérivabilité en moyenne quadratique de ε est équivalente à la dérivabilité de γ en 0 et $\gamma'(0) = 0$. Ainsi, les graphiques suivants de variogrammes illustrent différents niveaux de régularité des variations locales :

ε est continu en moyenne quadratique et non-dérivable



ε est continu et dérivable en moyenne quadratique



3.3.3 Modèles théoriques de variogramme

On peut schématiser l'ensemble des modèles théoriques de variogramme par le graphique de la figure 11. Ce graphique met en avant 3 paramètres fondamentaux. En premier lieu, on

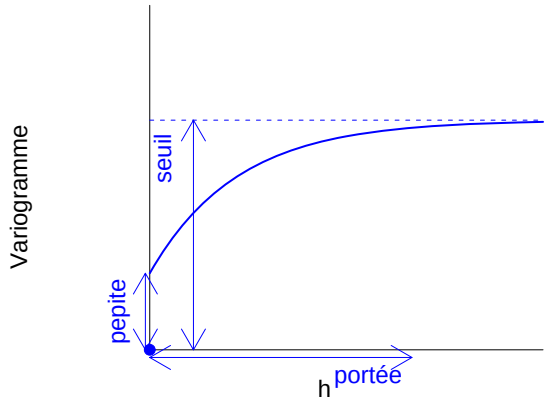
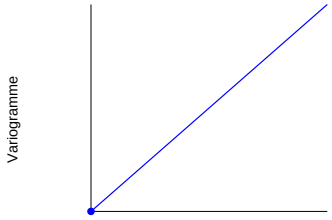


FIG. 11: *Schéma type d'un variogramme*

appelle **seuil** du variogramme sa valeur limite pour de grandes valeurs de h. Ce paramètre suscite un intérêt très important dans l'analyse de la régularité des variations locales. En effet, si le seuil d'un variogramme est infini, alors le processus des variations locales n'est pas stationnaire. Les deux modèles de variogrammes suivants sont parmi les plus utilisés pour des variations non-stationnaires.

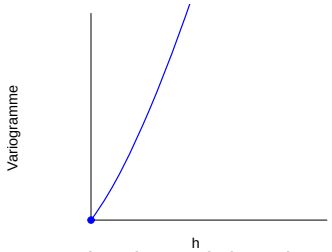
Le variogramme linéaire

$$\gamma(h) \quad = \quad \gamma_0 + \beta h.$$



Le variogramme puissance

$$\gamma(h) \quad = \quad \gamma_0 + \beta h^a.$$

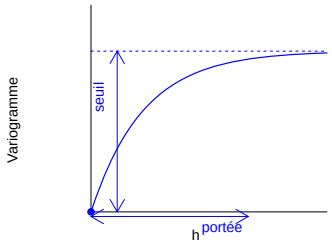


Lorsqu'au contraire le variogramme atteint une limite, on s'intéresse à la valeur de h pour laquelle cette limite est atteinte. En effet, cette valeur, qu'on appelle la **portée** des variations

locales, caractérise l'écart entre sites de mesures au-delà duquel les dépendances entre mesures du processus sont nulles. Les deux modèles suivants constituent deux exemples très classiques de variogrammes pour processus stationnaires se différenciant par leur régularité au voisinage de 0.

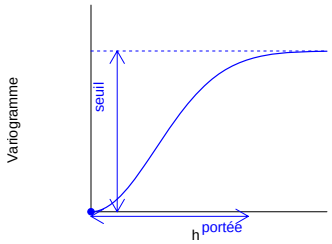
Le variogramme exponentiel

$$\gamma(h) \quad = \quad \gamma_0 + \sigma^2 \left[1 - \exp(-\frac{h}{a})\right].$$



Le variogramme gaussien

$$\gamma(h) \quad = \quad \gamma_0 + \sigma^2 \left[1 - \exp(-\frac{h^2}{a^2})\right].$$



Enfin, comme on l'a vu ci-dessus, lorsque la pépite est non-nulle, alors le processus des variations locales n'est pas continu en moyenne quadratique, ce qui traduit une forte irrégularité.

3.3.4 Variogramme empirique

L'analyse de la régularité du processus des variations locales repose sur le variogramme :

$$\begin{aligned} 2\gamma(h) &= \mathrm{Var}\left[\varepsilon(s+h) - \varepsilon(s)\right], \\ &= \mathrm{Var}\left[Z(s+h) - Z(s)\right]. \end{aligned}$$

Pour une valeur de h donnée, on obtient une estimation empirique de $\gamma(h)$ de la manière suivante : soit $N(h)$ l'ensemble des couples (s_i, s_j) de sites de mesure tels que $s_i - s_j = h$, alors

$$\hat{\gamma}(h) \quad = \quad \frac{1}{2|N(h)|} \sum_{(s_i,s_j)\in N(h)} \left[Z(s_i) - Z(s_j)\right]^2.$$

Dans un premier temps, examinons le cas des données uni-dimensionnelles de la figure 2. Ici, comme dans la plupart des situations d'application de la géostatistique, les sites de mesure se répartissent de manière irrégulière le long de la plage d'observation du processus. Le plus souvent, pour une valeur de h donnée, l'ensemble $N(h)$ est donc soit vide, soit constitué d'un seul couple de sites de mesures. Afin de pallier ce problème, il est souhaitable d'introduire

dans la définition du variogramme empirique un paramètre $\eta > 0$ de tolérance : $N(h)$ est alors l'ensemble des couples (s_i, s_j) tels que $h - \eta \leq s_i - s_j \leq h + \eta$.

Le graphique de la figure 12 montre les valeurs du variogramme empirique ainsi calculé à partir des données de la figure 2.

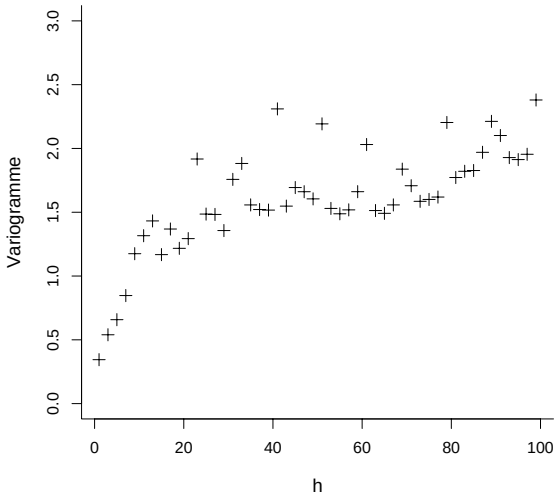


FIG. 12: *Variogramme empirique des données de la figure 2*

Dans l'exemple bi-dimensionnel décrit par la figure 4, chaque site est caractérisé par deux coordonnées. Par conséquent, fixer une valeur de h ne revient plus seulement à se donner une distance entre sites mais aussi une direction dans le plan. En théorie, à chaque direction dans la plage d'observation du processus on peut donc associer un variogramme analogue à celui calculé dans le cas uni-dimensionnel. Toutefois en pratique, deux situations sont possibles :

- soit, pour des raisons d'ordre géologique, physique ou encore climatique, on a de bonnes raisons de penser que les variations locales ne sont pas les mêmes dans un petit nombre de directions privilégiées. On parle alors de **variations locales anisotropiques**. La figure 13 montre deux variogrammes des données de teneurs en phosphates selon les 2 directions définies par les bords de la parcelle. Même si des différences existent entre ces deux variogrammes, on va considérer qu'elles ne sont pas assez évidentes pour justifier une hypothèse d'anisotropie.

- soit on postule que la dépendance entre mesures voisines ne dépend que de la distance entre les sites de mesures. On parle alors de **variations locales isotropiques**. La figure

14 montre le variogramme empirique des teneurs en phosphates.

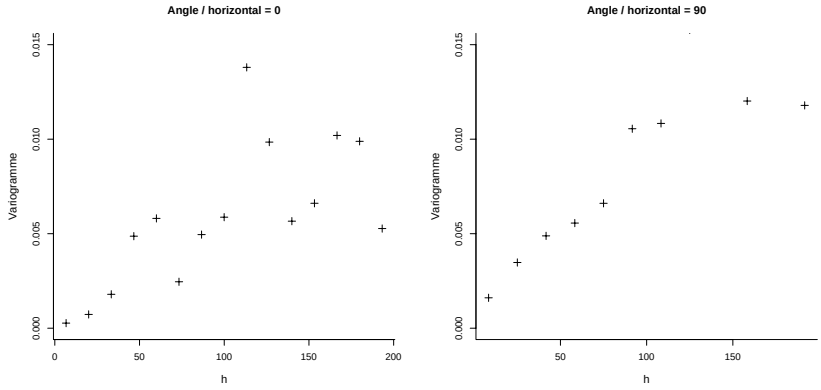


FIG. 13: *Variogrammes empiriques des teneurs en phosphates selon deux directions*

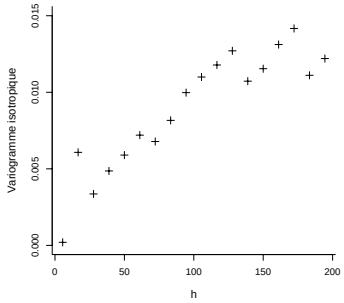


FIG. 14: *Variogramme isotropique des teneurs en phosphates*

Les variogrammes empiriques servent dans la suite de support au choix d'un modèle théorique de variogramme décrivant de manière satisfaisante la régularité des variations locales.

3.3.5 Ajustement d'un variogramme théorique

Le variogramme de la figure 14 laisse supposer l'existence d'un palier dans la forme du variogramme de la teneur en phosphates : ce constat oriente le choix d'un modèle vers un variogramme pour processus stationnaire. Par ailleurs, la forme du variogramme empirique

en 0 ne permet pas de supposer l'existence d'un effet pépite. On choisit donc d'ajuster au variogramme empirique un variogramme théorique de la forme suivante :

$$\gamma(h) \quad = \quad \sigma^2 \left[1 - \exp(-\frac{h}{a})\right].$$

De manière équivalente, la fonction d'autocovariance associée à ce variogramme est définie par $C(h) = \sigma^2 \exp\left(-\frac{h}{a}\right)$. Le paramètre σ^2 est donc le seuil du variogramme, alors que le paramètre a détermine la portée du processus. L'estimation de ces deux paramètres s'obtient par la méthode des moindres carrés. Le graphique de la figure 15 représente le variogramme ajusté.

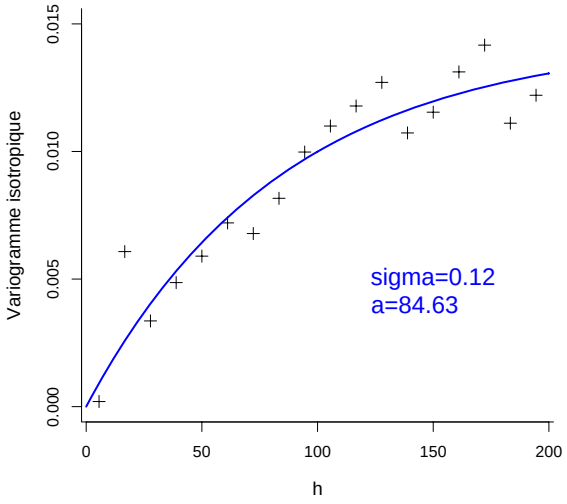


FIG. 15: *Ajustement d'un modèle théorique de variogramme à un variogramme empirique*

3.4 Tendance spatiale

Le choix du modèle de la tendance s'inspire naturellement de la description de la variabilité spatiale présentée dans la partie 2. Dans le cas d'une évolution spatiale très progressive, les modèles polynômiaux permettent le plus souvent un bon compte-rendu de la réalité et une estimation précise de la tendance spatiale. Dans le cas de variations plus changeantes, il est parfois intéressant d'utiliser des approches non-paramétriques, donnant ainsi une plus grande liberté à la forme du modèle. Par souci de simplicité, on se restreint dans le cadre de ce cours à l'évocation de modèles polynômiaux de la tendance spatiale.

Soit $s = (x, y)$ un site de mesure de Z,

$$\exists\ k\ \in\ \mathbb{N},\ k \geq 1,\ \exists\ \beta_i,\ i = 0, \ldots, \frac{k(k+3)}{2},$$

tel que, $\forall (x, y) \in \mathcal{S}$:

$$\mu(x, y) \quad = \quad \beta_0 + \beta_1 x + \beta_2 y + \ldots + \beta_{\frac{k(k+3)}{2}-1} x y^{k-1} + \beta_{\frac{k(k+3)}{2}} y^k.$$

Quoique la surface de réponse μ de ce modèle ne soit, sauf cas particulier d'un polynôme de degré 1, pas linéaire en x et y, le modèle précédent est cependant un modèle de régression linéaire multiple pour lequel les variables prédictrices sont $x,\ y,\ x^2,\ xy,\ y^2, ...,\ xy^{k-1},\ y^k$.

En pratique, le choix de ce type de modèle soulève deux problèmes très liés :

- d'une part le choix du modèle de régression polynômiale ; ce problème relève de la sélection de variables prédictrices d'un modèle de régression linéaire.
- d'autre part l'estimation des paramètres du modèle ; la méthode d'estimation doit tenir compte des dépendances locales modélisées dans la partie 3.3. La méthode des moindres carrés généralisés, qui étend la traditionnelle méthode des moindres carrés au cas de données corrélées, est adaptée à la situation.

Dans l'exemple de la variabilité spatiale de la teneur en phosphates représentée sur la figure 4, le choix du meilleur modèle de régression polynômiale conduit à la sélection d'un polynôme de degré 4. La figure 16 montre l'évolution du critère de sélection du meilleur polynôme en fonction du degré du polynôme et représente la tendance ajustée par le meilleur modèle de régression polynômiale. Le grand nombre de paramètres intervenant dans le meilleur modèle de régression polynômiale est une indication du caractère variant de la tendance spatiale et suggère l'insuffisance d'un modèle de régression polynômiale.

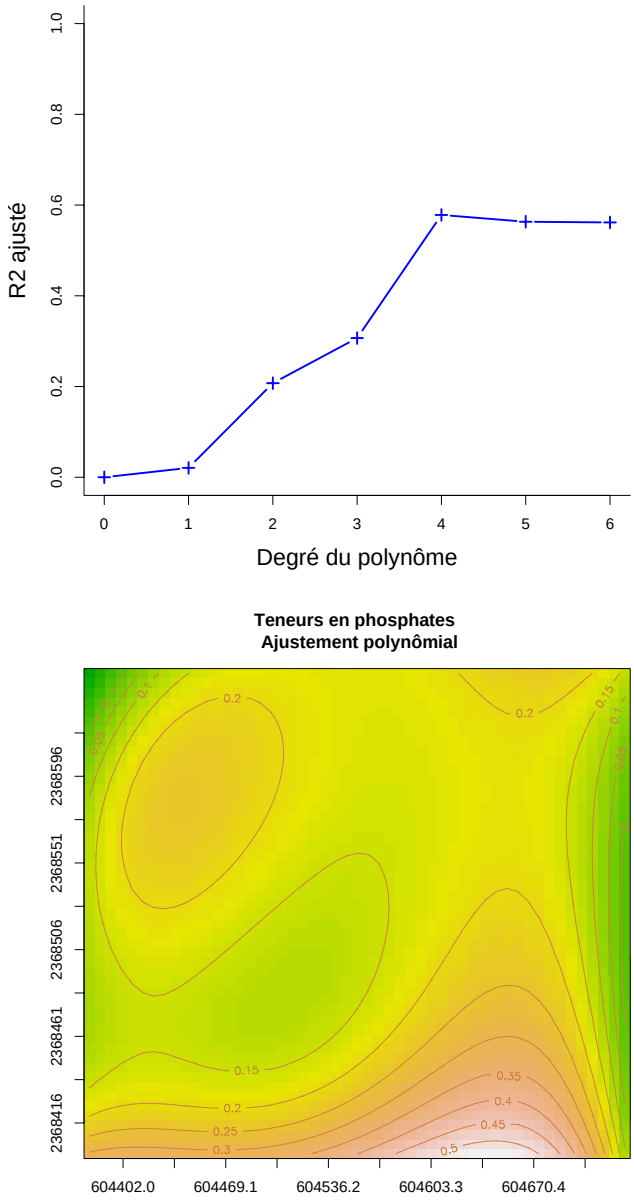


FIG. 16: *Evolution du R^2 ajusté en fonction du degré du polynôme et tendance polynômiale*

4 Interpolation optimale : krigeage

Les développements ci-dessus présentent la modélisation géostatistique de la variabilité spatiale d'un processus. En soi, cette démarche de modélisation est riche d'information sur la variabilité étudiée : elle permet un diagnostic de stationnarité et de régularité de la variabilité spatiale. Le modèle de tendance spatiale permet également une première cartographie de cette variabilité : en l'absence de dépendances locales, cette première approche est d'ailleurs suffisante. L'objectif des méthodes d'interpolation optimale est de proposer une cartographie corrigeant la tendance spatiale par la prise en compte des dépendances locales.

Soit Z un processus spatial et $\{s_1, s_2, \ldots, s_n\}$ l'ensemble des sites de mesures du processus Z. Soit s_0 un site pour lequel on ne dispose pas de la mesure de Z. L'objectif est de construire un prédicteur $\hat{Z}(s_0)$ de $Z(s_0)$ à partir des mesures $Z(s_1), Z(s_2), \ldots, Z(s_n)$. Dans la suite, on se limite aux prédicteurs linéaires :

$$\hat{Z}(s_0) \quad = \quad \lambda_1 Z(s_1) + \lambda_2 Z(s_2) + \ldots + \lambda_n Z(s_n).$$

Definition 4.1 Optimalité de la prédiction
On dit que le prédicteur $\hat{Z}(s_0)$ est optimal si :

- *l'erreur de prédiction $\hat{Z}(s_0) - Z(s_0)$ est d'espérance nulle,*
- *la variance de prédiction de $\hat{Z}(s_0)$ est minimale, parmi tous les prédicteurs linéaires dont l'erreur de prédiction est d'espérance nulle.*

Lorsque le modèle de tendance spatial est linéaire, il est possible de donner une forme analytique explicite au prédicteur optimal (voir par exemple Stein and Corsten, 1991), ainsi qu'à la variance de prédiction. Cette méthode d'interpolation est alors appelée le **krigeage**. Notons que si s_0 est l'un des sites de mesures s_i, alors $\hat{Z}(s_i) = Z(s_i)$. En d'autres termes, la carte construite par krigeage est un interpolateur exacte aux sites d'observation du processus.

La figure 17 permet de mettre en relation la cartographie de la teneur en phosphates par krigeage et celle des écarts-types de prédictions.

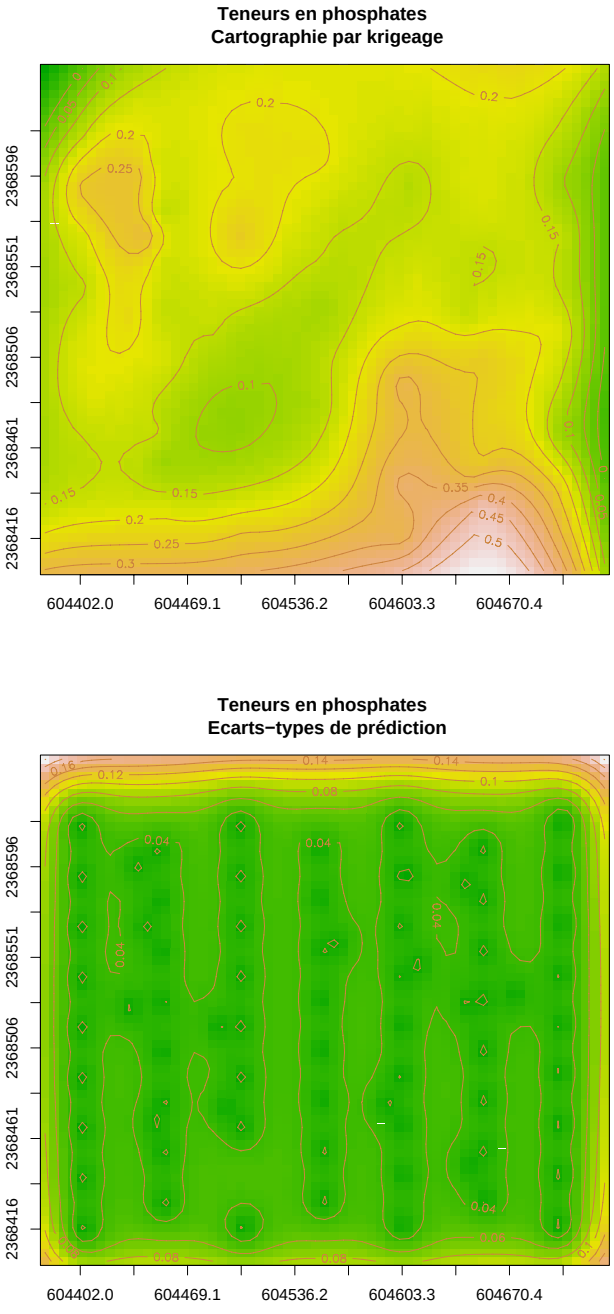


FIG. 17: *Cartographie par krigeage de la teneur en phosphates et écart-types de prédiction*

Références

Cressie, N.A.C. (1993) Statistics for spatial data. *Wiley, New York*
Stein, A. and Corsten, L.C.A. (1991). *Universal Kriging and Cokriging as a Regression Procedure.* Biometrics 47, 575-587.