Stationnarité d'un modèle géostatistique

Pour réaliser notre étude de la *canopy_cover* et du *rh98*, il nous faut considérer que la valeur de la variable en un point (x, y) n'est pas indépendante de la valeur de cette même variable en $(x \pm \epsilon, y \pm \epsilon')$.

Notre première idée est donc de sous-échantilloner notre riche jeu de données (10^6 données en ordre de grandeurs pour chacune des 3 écorégions sélectionnées) de sorte à sélectionner des points à des distances où les valeurs prises par les variables sont considérées ne pas avoir d'influence l'une sur l'autre, distance qu'on estime à environ 5 km d'après les connaissances écologiques actuelles. Dans un second temps, nous nous apprêtons à déterminer cette valeur avec un variogramme.

Dans un troisième enfin, nous entendons prendre en compte la corrélation spatiale à travers un coefficient dans le modèle lui-même. Après étude bibliographique, les modèles géostatistiques nous sont apparus plus pertinents pour notre travail que les modèles spatiaux régionaux, notamment de type autoregressifs comme SAR ou CAR, car nous avons beaucoup de données, et que leurs positions recouvrent bien des espaces pourtant très larges (du Sénégal jusqu'à l'Érythrée concernant Sahelian Acacia).

Pour déterminer le rayon de la "zone d'influence", nous utilisons un variogramme. Afin de m'approprier l'outil, je consulte de la bibliographie, et je découvre notamment qu'il y a (eu ?) une école française de géostatistique de Fontainebleau (liée à l'école des Mines) assez ambitieuse dans son approche mathématique et philosophique de la discipline. Pierre me conseille de regarder ce qui avait été fait dans l'article de 2015[1] avec *gstat*. Je découvre en fait que le package utilisé pour le variogramme dans cet article et *geoR* et je remonte dans la bibliographie jusqu'au livre de Diggle[2]. Diggle ayant coréalisé le package *geoR*, son ouvrage semble bien correspondre à ce je recherche, sans entrer dans les détails mathématiques profonds comme dans la documentation de l'école de Fontainebleau, notamment le livre de Chauvet 1989[3]. Toutefois je lis dans son ouvrage chapitre 4 que le processus stochastique géoréférencé doit être stationnaire, cette notion étant présentée comme le fait que la loi de la variable à expliquer doit être la même en tout point de l'espace, ce qui n'est pas ce qu'on observe, notamment pour la zone Sahelian Acacia. Je retrouve cette notion de stationnarité dans de nombreux articles se voulant une introduction à la géostatique, certains auteurs allant même à écrire que cette hypothèse empêche de considérer des données présentant un gradient spatial.

Aubry présente tout d'abord la notion de variable observée $z()$ en tout point x de l'espace $D \subset \mathbb{R}^2$. Aucune hypothèse n'est faite sur z en terme de régularité. z peut même ne pas être une fonction analytique. Elle est définie comme une *variable régionalisée* (VR).

Chaque valeur $z(x)$ est ainsi une réalisation d'une variable aléatoire (VA) $Z(x)$, pour tout $x \in D$. Les variables aléatoires $\{Z(x), x \in D\}$ ne sont pas indépendantes les unes des autres mais sont liées par une structure de corrélation. Ces deux hypothèses consituent un modèle dit probabiliste ou topo-probabiliste par l'école Matheron-Chauvet. $\{Z(x), x \in D\}$ est une *fonction aléatoire* (FA). Le modèle de fonction aléatoire est à la fois défini dans un espace topologique et dans un espace probabiliste.

Il y a deux abus de langage principaux (qu'on retrouve donc dans les ouvrages et articles que j'avais consultés). Le premier est l'indentification de $z(x)$ et $Z(x, \omega)$ (où ω appartient à la tribu de l'espace probabilisé), de probabilité $\mathbb{P}(\omega)$. Cette première identification est globalement assez simple à légitimiter. En revanche il y a souvent une identification de $z()$ et $\{Z(x, .), x \in D\}$, qui est plus problématique pour la bonne compréhension de la géostatistique. Aubry note bien dans sa thèse : "ce qui n'a aucun sens". Nous voici rassurés.

Le choix de la FA fait, la géostatistique se propose d'opérer dans l'espace probabilisé. Il est nécessaire d'invoquer la stationnarité d'une FA car on ne peut pas inférer la loi spatiale à partir d'une seule réalisation. En quelque sorte "l'hypothèse de la stationnarité revient à compenser l'absence de plusieurs réalisations de la FA par une forme de redondance de l'information au sein d'une seule réalisation (i.e., la VR)". Il convient toutefois de définir avec quelle stationnarité on souhaite travailler.

Tout d'abord on ne distinguera deux VA $Z(x_1)$ et $Z(x_2)$ que sur la base de leurs moments d'ordre 1

et 2 (espérance et variance/covariance). Pour revenir sur les notations, même Aubry utilise le premier abus de notation puisque $Z(x_1)$ et $Z(x_2)$ ne sont pas deux réalisations d'une même variable Z, elles sont à interpréter comme $\omega \mapsto Z(x_1, \omega)$ et $\omega \mapsto Z(x_2, \omega)$ deux variables aléatoires distinctes car prises en deux points $x_2 \neq x_2$ différents.

Ensuite on peut définir une stationnarité stricte qui est que pour tout n fini, et pour tout vecteur inter-suppport h, la fonction de répartition conjointe de $\{Z(x_i), i = 1, .., n\}$ est la même que celle de $\{Z(x_i+h), i = 1, .., n\}$. C'est toutefois une hypothèse "irréaliste parce que beaucoup trop forte vis-à-vis de l'homogénéité spatiale de la VR".

Il convient donc d'utiliser une autre hypothèse en introduisant la notion de stationnarité d'ordre 2. Une fonction aléatoire est dite stationnaire d'ordre 2 si la covariance $cov\,(Z(x), Z(x+h))$ existe et ne dépend que du vecteur inter-support h (ce qui implique que l'espérance et la variance de Z ne dépendent pas de x). Il n'y a aucune implication dans un sens ou dans l'autre entre la stationnarité stricte et la stationnarité d'ordre 2.

Les VR ne présentent toutefois pas toujours de variation spatiale bornée au sein d'un domaine d'étude D, on affaiblit encore l'hypothèse jusqu'à ce qu'on appelle *l'hypothèse intrinsèque d'ordre 0*, qui est que les espérances et variances de $Z(x+h) - Z(x)$ existent et ne dépendent pas de x :

$$\mathbb{E}[Z(x+h) - Z(x)] = m \tag{1}$$

$$\mathbb{V}[Z(x+h) - Z(x)] = 2\gamma(h) \tag{2}$$

Cette hypothèse permet d'introduire la définition du variogramme :

$$\gamma(h) = \frac{1}{2}\mathbb{E}[(Z(x+h) - Z(x))^2] \tag{3}$$

Il n'existe donc pas d'hypothèse sur l'espérance ni les variances des VA elles-même dans ce cadre, uniquement sur leurs accroissements.

"L'ergodicité est une seconde hypothèse introduite dans le cadre de la modélisation probabiliste. Une FA est dite ergodique si ses paramètres peuvent être inférés à partir d'une seule réalisation, autrement dit, si les espérances peuvent être estimées par des moyennes spatiales (Cressie 1988a, Chauvet 1993). L'ergodicité établit le passage entre la loi de probabilité de la FA et sa structure spatiale qui seule sera "observable" à travers ce qui est considéré dans le modèle comme une de ses réalisations possibles, i.e. la VR (Chauvet 1994)." Cela revient à faire l'hypothèse que :

$$\frac{1}{[D]}\int_D Z(x)dx \underset{[D]\to\infty}{\longrightarrow} \mathbb{E}[Z(x)] \tag{4}$$

Références

[1] Stéphane Guitet et al. "Spatial Structure of Above-Ground Biomass Limits Accuracy of Carbon Mapping in Rainforest but Large Scale Forest Inventories Can Help to Overcome". en. In : *PLOS ONE* 10.9 (sept. 2015). Sous la dir. de Krishna Prasad Vadrevu, e0138456. issn : 1932-6203. doi : 10.1371/journal.pone.0138456. url : https://dx.plos.org/10.1371/journal.pone.0138456 (visité le 26/02/2024).

[2] Peter J Diggle et Paulo Justiniano Ribeiro Jr. "Model-based Geostatistics". en. In : ().

[3] Chauvet. *Aide Mémoire de Géostatistique Linéaire*. 1989.