Impact de l'autocorrélation spatiale sur la qualité des modèles d'apprentissage automatique : étude dans le cadre de la classification d'essences forestières à partir de données satellitaires

Nicolas Karasiak [1], David Sheeren, Jean-François Dejoux, C. Monteil
Dynafor, UMR 1202, – CNRS, INP, ENSAT - Université de Toulouse, [1]

La télédétection, ou l'utilisation d'images prises à partir d'un capteur (satellite, drone, avion…) distant du sujet observé (forêt, ville…), est utilisée depuis des décennies pour prédire la présence de forêts à large échelle mais aussi les grands types de peuplements : feuillus ou résineux. Avec l'arrivée dans les années 70 du premier satellite multispectrale Landsat, des travaux à larges échelles de cartographie des peuplements commencent à voir le jour (Kushwaha, 1990). Si à l'époque les calculs informatiques sont fastidieux et les algorithmes de classification automatique moins performants, il s'avère qu'un biais important est mis en avant (Congalton, 1991) : l'autocorrélation spatiale existant dans les données a pour effet de surestimer la qualité des résultats de classification. L'indépendance des pixels n'étant pas assurée entre pixels voisins, les modèles supervisés construits à partir d'échantillons auto-corrélés spatialement sont biaisés et les qualités prédictives plus optimistes que les qualités réelles constatées dans le produit final. Alors que ce problème est identifié depuis longtemps, plus de 25 ans après cette publication seul un nombre très faible de travaux tiennent compte de cette dépendance spatiale dans les modèles d'apprentissage construits à partir d'images (Fassnacht et al., 2016 ; Griffith, 2016).

L'objectif de ce travail est de comprendre l'impact de l'autocorrélation spatiale dans une série temporelle d'images optiques (Sentinel-2) sur la prédiction des feuillus/résineux en fonction de différentes méthodes de sélection d'échantillons : (i) la méthode la plus répandue qui consiste à séparer en deux le jeu d'échantillons de manière aléatoire avec 50% pour l'apprentissage et 50% pour la validation à l'échelle du pixel (R50-pixel) (ii) une méthode similaire appliquée à l'échelle du peuplement (R50-peuplement) (iii) une méthode de validation croisée leave-one-out pour laquelle l'autocorrélation spatiale est mesurée et prise en compte à l'échelle des pixels de façon à ce que la validation soit réalisée sur un échantillon spatialement indépendant (SLOO-CV-pixel) (iv) une méthode similaire mais à l'échelle des peuplements (SLOO-CV-peuplement). Le protocole a été testé sur un jeu d'échantillons comprenant 4 257 112 pixels de feuillus (soit 2 841 peuplements) et 1 119 079 pixels de conifères (soit 1 972 peuplements) en utilisant l'algorithme Random Forest et en faisant évoluer de manière progressive le nombre de peuplements.

Les premiers résultats montrent qu'un échantillonnage aléatoire qui ne tient pas compte de la dépendance spatiale entre les données surestime de façon importante la performance prédictive du modèle avec toutefois une atténuation de cet effet à partir d'un effectif d'échantillons de très grande dimension (plusieurs centaines de milliers de références). Lorsque la validation est menée en tenant compte de l'autocorrélation spatiale (méthode SLOO-CV-pixel) la performance prédictive (indice kappa) varie de 37 à 73 selon l'effectif comparé à 92 et 93 pour une approche conventionnelle (R50-pixel). Ces résultats confirment le biais optimiste de qualité déjà suspecté depuis longtemps dans la littérature. Ils suggèrent aussi la prise en compte systématique de l'autocorrélation spatiale dans les procédures de classification automatique d'images satellitaires pour fournir une évaluation statistique plus représentative de la carte produite. Afin de faciliter cette transition, une bibliothèque d'algorithmes en Python sera distribuée en libre accès.

Bibliographie

Kushwaha, S. P. S. (1990) «Forest-type mapping and change detection from satellite imagery », ISPRS J. Photogramm. Remote Sens., vol. 45, no 3, p. 175‑181

Congalton, R. G. (1991). «A review of assessing the accuracy of classifications of remotely sensed data», Remote Sens. Environ., vol. 37, no 1, p. 35‑46

Fassnacht, F. E., et al. (2016). «Review of studies on tree species classification from remotely sensed data», Remote Sens. Environ., vol. 186, p. 64‑87

Cánovas-García et al. (2017). « Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery», Computers & Geosciences, 103, pp. 1–11

Griffith, D. A., Chun Spatial, Y. (2016) «Autocorrelation and Uncertainty Associated with Remotely-Sensed Data», Remote Sens,, 8, 535