Comparaison des chemins de Hilbert adaptatif et des graphes de voisinage pour la caractérisation d'un parcellaire agricole

Thomas Guyet*, Sébastien da Silva**,***, Claire Lavigne***, Florence Le Ber****

* Agrocampus Ouest, Rennes
** LORIA – INRIA Grand Est, Villers-lès-Nancy
*** INRA PSH, Avignon
**** ICUBE, Université de Strasbourg/ENGEES - CNRS, Strasbourg

Résumé. Cet article compare deux représentations de données spatiales, les graphes de voisinages et les chemins de Hilbert-Peano, utilisées par des algorithmes de fouille. Cette comparaison s'appuie sur la mise en œuvre d'une méthode d'énumération de « sacs de nœuds », qui permet d'obtenir des caractérisations homogènes à partir des deux représentations. La méthode est appliquée à la caractérisation de parcellaires agricoles et les résultats tendent à montrer que la linéarisation de l'espace capte la majorité de l'information, à l'exception des éléments rares, sur cet exemple particulier.

1 Introduction

Les méthodes de recherche de motifs spatiaux sont utilisées couramment pour construire des caractérisations de données spatiales (voir Selmaoui-Folcher et al. (2013)). Ces approches s'appuient sur des représentations de l'espace tels que des graphes de voisinage ou des chemins construits sur des courbes fractales (par ex. chemins de Hibert-Peano, voir Mari et Le Ber (2006)). Les graphes de voisinages contiennent une information spatiale riche, mais ils sont plus complexes à fouiller, tandis que les chemins sont faciles à fouiller, mais ils réduisent l'information spatiale disponible.

Dans ce travail, nous confrontons ces deux représentations vis-à-vis de la caractérisation d'un parcellaire agricole. En particulier, nous cherchons à savoir si l'approximation par un chemin fractal permet de conserver une bonne caractérisation de l'organisation spatiale des parcelles agricoles en vue de l'application d'une méthode de recherche de motifs. Cette question soulève deux difficultés. La première est l'absence d'étiquetage des données qui permettrait d'évaluer une représentation sur une tâche de classification. Nous nous plaçons donc dans un contexte non-supervisé. La deuxième difficulté porte sur la comparaison des caractérisations obtenues, qui sont de natures différentes. Nous ne pouvons pas nous appuyer comme classiquement sur des calculs de corrélation (par ex. par une matrice de confusion) entre les localisations des motifs dans les chemins et dans les graphes.

Pour résoudre ces difficultés, nous utilisons des « sacs de nœuds » , inspirés des sacs de mots introduits dans le contexte de l'analyse de texte (Salton et al. (1975)). Ce modèle permet de construire des caractérisations homogènes d'un même espace pour les graphes et les chemins.

2 Représentations de l'espace et caractérisation

Le parcellaire agricole a été collecté dans le cadre de la Zone Atelier Armorique, autour de la commune de Pleine-Fougères (35) [1]. Ces données sont sous une forme vectorielle : chaque parcelle agricole est définie par un polygone associé à un attribut catégoriel qui donne l'occupation du sol en été (céréale, prairie, etc.). On note $\mathcal{O} \subset \mathbb{N}$ l'ensemble des identifiants des types d'occupation du sol.

Un graphe $S_G = \langle V, E_G, \sigma \rangle$ est construit à partir des données selon la méthode décrite par Guyet (2010). Un nœud $v \in V$ est construit pour chaque parcelle, représentée par son barycentre. Chaque nœud $v \in V$ est associé à une occupation du sol. Un arc $e \in E_G \subset V \times V$ lie deux parcelles voisines, c'est-à-dire connexes ou séparées par un faible espace (séparation par une route ou imprécision géométrique des données). Les arcs ne sont pas étiquetés.

Une méthode de calcul du chemin de Hibert-Peano Adaptatif (CHA) a été décrit par Quinqueton et Berthod (1981). Cette méthode a été utilisée par Da Silva (2013) pour extraire les structures spatiales de linéaires agricoles. Elle utilise un ensemble de points spatialement distribués (ici les barycentres des parcelles), qui sont parcourus de manière déterministe. Le chemin est ensuite simplifié pour se ramener à une succession de parcelles, qui est transcrite sous la forme d'une séquence d'occupations du sol. Pour l'unification des notations, un CHA peut être décrit de la même façon qu'un graphe, par $S_{CHA} = \langle V, E_{CHA}, \sigma \rangle$, avec un nœud par *item* de la séquence et un arc pour deux *items* successifs. On peut noter que la séquence S_{CHA} obtenue sur les mêmes données n'est pas un sous-graphe de S_G.

Pour atténuer l'influence des paramètres de construction des CHA, nous construisons plusieurs chemins 1) en faisant varier aléatoirement les limites de la cellule initiale autour des limites géographiques qui définissent une zone (3 cadrages aléatoires initiaux) et 2) en générant les chemins pour les 4 directions principales.

La caractérisation des régions par SdN débute par l'énumération de toutes les sous-structures d'une représentation de l'espace. Soit $S = \langle V, E, \sigma \rangle$ une représentation de l'espace, chemin de Hilbert adaptatif ou graphe. Une sous-structure de S est un triplet $\langle V', E', \sigma \rangle$ où $V' \subset V$, $E' \subset E$ tel que $\forall u, v \in V'$, il existe $e \in E'$ tq e soit un arc entre v et u. Dans le cas des CHA, on s'intéresse à des sous-séquences de taille fixe w_s. Dans le cas des graphes de voisinage, on s'intéresse à des sous-graphes contenant exactement w_g arcs. L'énumération de toutes les structures est possible en temps output-polynomial (Bonzini et Pozzi (2007)), sans seuil de fréquence à fixer. Pour les sous-graphes, nous utilisons l'outil TGE de Uno (2005).

Le « sac de nœuds » (SdN) d'une sous-structure s_S, noté $SdN(s_S) \in 2^{\mathcal{O}}$, est un vecteur de présence / absence des types d'occupations du sol dans la sous-structure s_S. Nous avons préféré ne conserver que l'information de présence / absence à la place d'un dénombrement pour éviter la multiplication combinatoire des sacs de nœuds.

3 Comparaison des représentations sur les parcellaires

Le parcellaire a été décomposé en 14 sous-zones aux caractéristiques variées : les zones les plus au nord sont caractérisées par des grands champs de céréales, tandis que les zones au sud correspondent à un secteur bocager constitué de petites parcelles de prairies. La décomposition

1. Les données ont été acquises par le laboratoire COSTEL et l'unité SAD-Paysage de l'INRA.

réduit le nombre de parcelles à prendre en compte pour obtenir des résultats plus rapidement et permet d'étudier le caractère spécifique des motifs d'une zone par rapport à une autre ou le comportement des deux méthodes en relation avec les caractéristiques des zones.

Le nombre total de sacs (de 1 à 4 occupations du sol) distincts s'élève à 475 pour les graphes (toutes zones comprises), 206 pour les chemins : le rapport moyen du nombre de sacs différents trouvés par les graphes et par les chemins dans les différentes zones s'élève à 1,9. Le nombre moyen d'éléments dans les sacs issus des graphes s'élève à 879 (toutes zones confondues), et à 28 pour les sacs issus des chemins.

En considérant ensemble tous les sacs obtenus sur les 14 zones par le CHA d'une part et par le graphe d'autre part, on obtient un indice de corrélation (Spearman) $I_r(graphes, chemins) = 0,712$, ce qui indique une forte corrélation ; pour le test du χ^2 (avec distribution sous H_0 simulée à cause des faibles valeurs), on obtient $I_\chi(graphes, chemins) = 30094$, valeur de $p < 0,0005$, soit des distributions très différentes. On observe les mêmes résultats sur les zones prises séparément. Les proportions des sacs estimées par les deux méthodes sont donc différentes mais corrélées.

Pour expliciter ces distributions différentes, on s'intéresse maintenant aux sacs oubliés par les CHA. On peut calculer le nombre d'éléments par sac à partir duquel un sac présent dans le graphe disparaît dans le chemin (moyenne sur les 14 zones $1215,6 \pm 855,9$) et le rapporter au nombre maximum d'éléments dans les sacs issus des graphes ($81307,0 \pm 64722,3$) ou au nombre moyen ($5891 \pm 6322,74$). Le premier taux s'élève à 2,1%, le second à 44,6%. Ces chiffres dépendent fortement du nombre de parcelles dans chaque zone et de la diversité des occupations représentées.

Si on regarde plus précisément la répartition des sacs perdus, on observe qu'il s'agit la plupart du temps de « petits » sacs, comptant moins de 100 éléments. Au delà on peut séparer des zones plutôt homogènes – peu de sacs perdus et peu remplis – et des zones plutôt hétérogènes – sacs plus nombreux et plus remplis. Les premières rassemblent des zones de petites parcelles de bocage (prairies très majoritaires) et des zones de grands parcelles cultivées (céréales et maïs majoritaires) : les sacs perdus par la méthode CHA représentent des occupations et des voisinages très minoritaires. Les deuxièmes sont plus diversifiées en taille de parcelles et occupations du sol : les sacs perdus peuvent correspondre à des voisinages relativement fréquents même si non majoritaires.

4 Discussion et conclusion

La comparaison d'une analyse fondée sur les données extraites par un chemin et d'une analyse sur les données complètes a été réalisée dans le cadre des modèles de Markov en analyse d'images par Benmiloud et Pieczynski (1995). Pour ces modèles, l'analyse fondée sur un chemin, bien que moins ajustée à la réalité des données, s'est montrée pertinente et acceptable en termes de rapidité.

Pour l'étude que nous avons menée, nous aboutissons à une conclusion similaire tout en mettant en évidence certains manquements de la méthode fondée sur le CHA, qui conduit à oublier les motifs rares mis en évidence par la méthode appuyée sur le graphe de voisinage.

Finalement, si on s'intéresse à des voisinages fréquents et à des zones relativement homogènes, la recherche de motifs par linéarisation de l'espace s'avère pertinente et efficace. Le fait que les différences se trouvent principalement sur les motifs rares laisse également espé-

rer des résultats de caractérisation des parcellaires intéressants par des méthodes de fouille de données. Ces méthodes ne s'intéressant qu'aux motifs fréquents, on peut s'attendre à ce que l'information extraite sur les séquences soit très similaire à celle obtenue sur des graphes tout en réduisant considérablement les temps de calcul.

Dans le futur et pour l'exemple traité, il reste à approfondir l'étude en regardant de plus près à quels ensembles d'occupation correspondent les sacs trouvés et oubliés par la méthode CHA. Une extension du travail portera aussi sur la recherche de motifs ordonnés afin de mieux spécifier les voisinages, en lien avec des problématiques agro-écologiques.

Références

Benmiloud, B. et W. Pieczynski (1995). Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images. *Traitement du signal 12*(5), 433–454.

Bonzini, P. et L. Pozzi (2007). Polynomial-time subgraph enumeration for automated instruction set extension. In *Design, Automation Test in Europe Conference Exhibition*, pp. 1–6.

Da Silva, S. (2013). Fouille de données spatiales et modélisation de paysages. Rapport interne, INRA – INRIA Nancy Grand Est.

Guyet, T. (2010). Fouille de données spatiales pour la caractérisation spatiale de paysages en lien avec des fonctionnalités agro-écologiques. In *Spatial Analysis and GEOmatics (SAGEO'10)*, pp. 3.

Mari, J.-F. et F. Le Ber (2006). Temporal and Spatial Data Mining with Second-Order Hidden Markov Models. *Soft Computing – A Fusion of Foundations, Methodologies and Applications 10*(5), 406–414.

Quinqueton, J. et M. Berthod (1981). A Locally Adaptive Peano Scanning Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-3*(4).

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Communication ACM 18*(11), 613–620.

Selmaoui-Folcher, N., F. Flouvat, H. Alatrista Salas, et S. Bringay (2013). Motifs spatio-temporels – enjeux et applications à l'environnement. *Revue d'Intelligence Artificielle 2013*, 619–648.

Uno, T. (2005). TGE : subtree/subgraph/connected components enumeration algorithm. URL : http ://research.nii.ac.jp/ uno/code/tge.html.

Summary

This article focuses on the comparison of approaches for spatial patterns mining. It deals with agricultural fields, which are mined in two ways, 1) by a fractal linearization method of space which provides a sequence of fields and 2) by the construction of a neighborhood graph. These representations are then used by enumeration algorithms to extract "bags of nodes" (BoN). The results suggest that the linearization of space captures most of the information – except some rare elements – about the organization of agricultural fields.