Caractérisation d'un parcellaire agricole : comparaison des sacs de noeuds obtenus par un chemin de Hilbert adaptatif et un graphe de voisinage

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

Caractérisation d'un parcellaire agricole : comparaison des sacs de nœuds obtenus par un chemin de Hilbert adaptatif et un graphe de voisinage

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 porte sur la comparaison d'approches pour la fouille de motifs spatiaux. Il concerne des données de parcellaires agricoles, qui sont explorées de deux façons, par une méthode de linéarisation de l'espace, qui permet d'obtenir une séquence, et par la construction d'un graphe de voisinage. Ces représentations sont ensuite exploitées par des algorithmes d'énumération de « sacs de nœuds », c'est-à-dire un vecteur de présence / absence d'un type de nœuds dans les sous-structures de la représentation de l'espace (séquence ou graphe). Les sacs de nœuds sont ensuite comparés pour mettre en évidence la capacité qu'ont les deux méthodes proposées à caractériser le parcellaire agricole. Les résultats tendent à montrer que la linéarisation de l'espace capte la majorité de l'information – à l'exception des éléments rares – sur l'organisation du parcellaire agricole et que des outils de fouille de données utilisant ces représentations linéaires peuvent être utilisés à la place d'outils moins efficaces tels que les méthodes de fouille de graphes spatiaux.

1 Introduction

Les évolutions actuelles de l'agriculture, liées entre autres aux contraintes environnementales grandissantes et à la mise en concurrence des terres pour la production alimentaire et la production énergétique, mettent en évidence le besoin de suivre et de comprendre les différents usages des sols. Pour cela différentes sources de données sont disponibles, images satellitaires ou enquêtes de terrain, et différentes méthodes d'analyse peuvent être mises en œuvre, analyses statistiques, méthodes markoviennes, ou recherche de motifs fréquents (par ex. Castellazzi et al. (2007); Le Ber et al. (2006); Lazrak et al. (2010)). Ces études conduisent ensuite à dessiner des scénarios d'évolution ou d'optimisation, qui peuvent s'appuyer sur différentes techniques (par ex. Castellazzi et al. (2010); Sorel et al. (2010)). Ces études se basent sur des caractérisations spatiales et temporelles réalistes de l'organisation des usages des sols à l'échelle d'un territoire. Dans cet article, nous exploitons des données d'enquête sur la Zone Atelier Armorique (ZAA) [1], représentant un parcellaire agricole continu constitué de plus de

1. Une Zone Atelier est un dispositif scientifique qui implique des observations et des travaux sur plusieurs années.

7000 parcelles. Ce secteur d'étude est particulièrement intéressant pour les écologues et les agronomes car il est structuré en zones aux caractéristiques différentes allant des grandes cultures de plaine aux prairies du bocage. L'objectif de notre travail est d'étudier et de développer des méthodes de fouille de données pour caractériser les assolements (*i.e.* les allocations d'occupations du sol à des parcelles agricoles une année donnée) au travers des motifs spatiaux fréquents qui peuvent en être extraits.

La recherche de motifs spatiaux et plus généralement de motifs spatio-temporels est un domaine de recherche actif et qui trouve de nombreuses applications. On trouvera un panorama du domaine, concernant plus particulièrement les applications agro-environnementales par Selmaoui-Folcher et al. (2013). Toutes ces méthodes de fouille de données spatiales s'appuient sur une représentation de l'espace dans des structures de données dont les propriétés formelles peuvent être utilisées pour proposer des algorithmes efficaces. Il s'agit classiquement de graphes, d'arbres ou de chemins. Pour chacune de ces représentations, plusieurs constructions alternatives sont possibles. Les graphes peuvent être construits à partir des informations d'adjacence ou de proximité entre parcelles, les arbres peuvent être construits par exemple à partir des graphes par la recherche du *minimum spanning tree*, les chemins peuvent être obtenus par des marches aléatoires dans les graphes ou en parcourant l'espace par des courbes fractales, telles que la courbe de Hilbert-Peano (voir Peano (1890)).

L'utilisation de méthodes de fouilles de sous-graphes dans le graphe de voisinage présente *a priori* un intérêt supérieur à d'autres méthodes de fouille de données plus simples, notamment sur des chemins ou séquences. Mais relativement aux temps de calcul importants qu'impliquent les méthodes de fouille de graphe, il pourrait tout de même être plus intéressant de travailler sur des représentations plus simples de l'espace.

Outre les aspects calculatoires, Il faut également considérer les apports des différentes représentations en terme de résultats. En effet, les possibilités de caractérisation de l'espace par des méthodes de fouille de motifs sont dépendantes, d'une part, de l'algorithme de fouille de motifs et, d'autre part, du choix de la représentation de l'espace. Les propriétés sur les algorithmes des fouilles de motifs sont bien connues, mais peu d'intérêt a été porté au choix de la représentation de l'espace dans le cadre de la mise en place d'une extraction de motifs fréquents. En première approche, il semble que le choix de la représentation influe fortement sur les caractérisations qui peuvent être extraites des données, mais ceci mérite d'être exploré plus avant. C'est pourquoi, nous proposons une méthode pour comparer des représentations de l'espace, indépendamment de l'algorithme de fouille de données qui sera utilisé. Nous comparons ici plus particulièrement deux représentations de l'espace : un chemin de Hilbert adaptatif et un graphe spatial.

Pour comparer les caractérisations obtenues par les graphes et les chemins, l'approche classique consiste à établir des corrélations entre les localisations des motifs dans les chemins et les motifs dans les graphes. Deux motifs seront d'autant plus corrélés qu'il se situent fréquemment et exclusivement co-localisés. Nous sommes alors confrontés à la difficulté de comparer (par exemple, par une matrice de confusion) des caractérisations de natures différentes, sous-graphes et sous-chemins.

Dans le cadre de ce travail, nous introduisons les « sacs de nœuds » comme outil de caractérisation de l'organisation spatiale. Cette représentation reprend l'idée des *bags of word* initialement introduits dans le contexte de l'analyse de texte (Salton et al. (1975)) puis appliqués en analyse d'images (Weber et al. (2000)) ou de séries temporelles (Megalooikonomou

et al. (2005)). Ce modèle représente une information complexe tel qu'un texte ou une image au moyen d'une liste de termes référencés dans un dictionnaire. On dispose alors d'une représentation vectorielle d'un document qui facilite sa manipulation (apprentissage supervisé, classification, etc.). Le succès de la méthode tient dans la simplicité de cette représentation et dans son efficacité pour répondre aux tâches de classification et recherche d'information.

L'article est organisé comme suit : dans une deuxième section, nous recensons quelques méthodes de comparaison de caractéristiques d'un document ou d'un jeu de données. Dans la section 3, nous présentons les données et les deux méthodes de représentation sur lesquelles nous nous focalisons. Nous introduisons également les « sacs de nœuds » comme outil de caractérisation de l'organisation spatiale qui peut être construit à partir des chemins et des graphes. La section 4 s'attache à la comparaison des résultats obtenus. La dernière section offre une discussion suivie d'une conclusion.

2 Travaux connexes

Peu de travaux se sont portés sur la comparaison de différentes représentations structurelles de l'espace. Dans la plupart des travaux, cette comparaison s'appuie sur des résultats de performance dans une tâche de classification supervisée. Par exemple, des résultats de classification de SVM ont été utilisés pour comparer des descriptions de l'information spatiale encodées dans des noyaux SVM, pour représenter des molécules (Wale et al. (2008)), ou des données multimédia (Gosselin (2011)). Pour une tâche d'indexation de graphes, Zhao et al. (2007) comparent l'utilisation de chemins, d'arbres et de sous-graphes sur les performances et l'efficacité de l'indexation. Les approches d'évaluation en cascade (Candillier et al. (2006)) utilisent des schémas d'apprentissage supervisé pour évaluer l'apport d'une caractérisation en ajoutant cette caractérisation aux données d'entrée des données d'un classifieur. Dans le cas de la caractérisation des parcellaires agricoles, nous ne disposons pas d'étiquetage permettant d'évaluer les performances en classification. Nous restons donc dans un schéma d'évaluation fondé sur la comparaison de caractéristiques extraites, dans un contexte non-supervisé.

Parmi les approches non-supervisées, le système PANDA (*PAtterns for Next-generation DAtabase systems*) propose un cadre pour la comparaison de motifs simples et complexes (cf. Bartolini et al. (2009, 2004)). La comparaison de motifs s'appuie sur une définition unifiée des motifs par un ensemble de composants (définis en fonction des types de motifs). Cette comparaison est réalisée en recherchant les appariements possibles entre les ensembles de composants. Les exemples proposés par les auteurs ne concernent que la comparaison de motifs de même nature (par ex. *itemsets* avec *itemsets*). Or, dans notre travail, nous cherchons à mettre en regard d'une part des structures linéaires, d'autre part des structures prenant la forme de graphes. Dans les tentatives d'unification des méthodes de fouille de données structurées (cf. Nourine et Petit (2012); Négrevergne et al. (2013)), ces types de motifs n'ont actuellement pas été pris en compte.

3 Matériel et méthodes

3.1 Présentation des données

Les données sur lesquelles nous travaillons ont été collectées dans le cadre de la Zone Atelier Armorique, autour de la commune de Pleine-Fougères (35) [2]. Ces données sont sous une forme vectorielle : chaque parcelle agricole est définie par une forme géométrique (un polygone) associée 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. La figure 1 illustre la répartition des types d'occupation du sol, en nombre de parcelles, pour la zone qui nous intéresse. On remarque un très fort déséquilibre entre les classes, les prairies semées, puis les bois, maïs et céréales étant largement majoritaires.

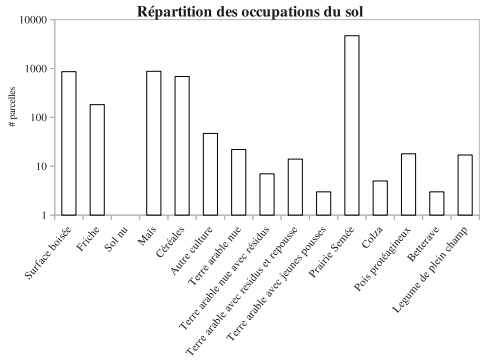

FIG. 1 – *Répartition des types d'occupation du sol en nombre de parcelles dans la zone d'étude (échelle logarithmique)*

Pour mieux mettre en avant des caractéristiques agricoles dans les motifs, nous nous limitons à la matrice agricole : seules les parcelles cultivées sont conservées. Toutes les parcelles de bâti, de route et de chemin ont été supprimées pour éviter d'extraire des motifs qui les impliqueraient. La matrice agricole comporte 7421 parcelles (cf. figure 2).

La matrice agricole a ensuite été décomposée en 14 sous-zones, numérotées selon leur position dans la grille de découpage. Ces différentes sous-zones n'ont pas les mêmes caractéristiques, en particulier en nombre de parcelles (de 189 à 1231). Les zones les plus au nord (par ex. zones 42 et 52) sont caractérisées par des grands champs de céréales, tandis que les zones au sud (par ex. zones 11 et 21) correspondent à un secteur bocager constitué de petites parcelles de prairies.

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

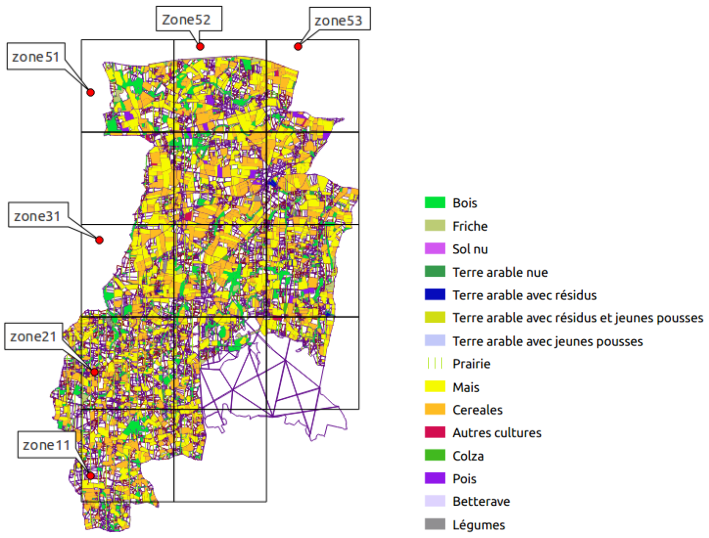

FIG. 2 – *Matrice agricole et découpage du parcellaire en sous-zones : la couleur de chaque polygone correspond à son occupation du sol*

3.2 Représentations de l'espace

3.2.1 Représentation par un graphe spatial

Un graphe $S_G = \langle V, E_G, \sigma \rangle$ est construit à partir des données au format vectoriel selon la méthode décrite par Guyet (2010). Un nœud $v \in V$ est construit pour chaque parcelle, représentée par son barycentre (cf. figure 3). Chaque nœud $v \in V$ est associé à une occupation du sol $\sigma(v) \in \mathcal{O}$ (o, j, v pour orange, jaune et vert sur la figure). 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.

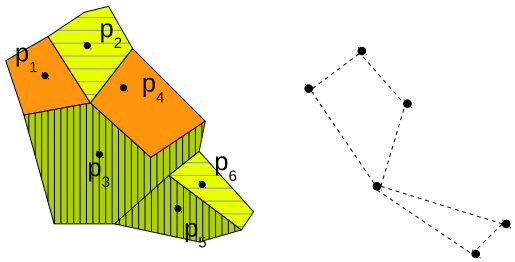

FIG. 3 – *À gauche, exemple de parcellaire avec les barycentres de parcelles et les occupations du sol (vert : hachures verticales, orange : sans hachures, jaune : hachures horizontales) ; à droite, graphe de voisinage obtenu pour l'exemple*

3.2.2 Représentation par un chemin de Hilbert adaptatif

Les courbes remplissant l'espace sont utilisées pour linéariser une information spatiale et sont couramment employées en traitement d'images mais aussi pour l'étude de phénomènes spatiaux. On utilise en particulier la courbe de Hilbert-Peano qui s'appuie sur un motif simple. Cependant, si le phénomène spatial étudié présente des irrégularités, des zones très denses et des zones vides d'information, la courbe classique, régulière, est peu appropriée. Dans ce cas, il faut soit choisir une échelle moyenne, qui délaisse une partie de l'information, comme cela a été fait par Lazrak et al. (2010), soit utiliser une méthode qui s'adapte localement à l'information : c'est ce que fait la courbe adaptative de Hilbert (CHA), dont l'algorithme est décrit par Quinqueton et Berthod (1981). Cette dernière méthode a été utilisée par exemple pour extraire les structures spatiales de linéaires agricoles (Da Silva (2013)).

La méthode de construction d'un CHA utilise un ensemble de points spatialement distribués, qui sont parcourus de manière déterministe. Sur la figure 4, les parcelles de la figure 3 ont été représentées par leurs barycentres. Le point de départ de l'algorithme est la case englobant tous les barycentres ainsi qu'une direction principale (haut-droite-bas sur l'illustration). À une étape de l'algorithme, on considère le nombre de barycentres dans la case courante : si ce nombre est égal à 1 ou 0, on passe à la case suivante ; sinon on découpe la case en quatre sous-cases.

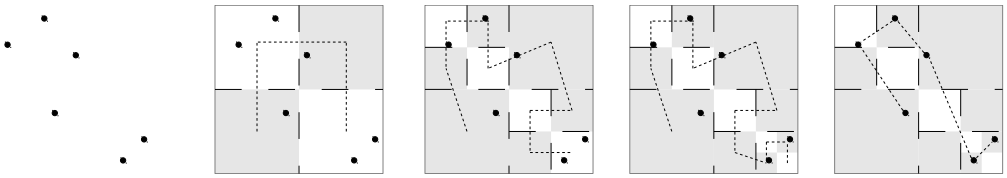

FIG. 4 – *Illustration de la construction du chemin de Hilbert Adaptatitif (CHA)*

Le chemin est ensuite simplifié par suppression des nœuds qui ne correspondent à aucun barycentre. Par cette opération, on perd donc la structure originelle du CHA et on se ramène à une simple succession de parcelles, qui est ensuite transcrite sous la forme d'une séquence d'occupations du sol. Dans l'exemple de la figure 4, on obtient la séquence suivante : $v > o > j > o > v > j$.

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. L'ensemble de nœuds V est le même pour les représentations avec un graphe ou avec un chemin. Toutefois, on peut remarquer que la séquence S_{CHA} obtenue sur les mêmes données n'est pas un sous-graphe de S_G. En effet la construction de S_{CHA} ne s'appuie pas sur les informations de voisinage, mais simplement sur les positions des barycentres dans l'espace des parcelles. Certains arcs ne sont donc pas présents (par ex. entre p_3 et p_6) tandis que d'autres s'ajoutent entre les parcelles proches mais non connexes (par ex. entre p_4 et p_5).

Par ailleurs la méthode de construction d'un CHA est sensible aux deux choix initiaux de la construction du chemin ; la définition des limites de la case initiale et le choix de la direction principale (quatre possibilités). Les chemins qui sont construits peuvent varier en fonction de ces choix initiaux :

– si la direction principale est modifiée, l'ordre de décomposition des zones change et les chemins sont modifiés ;
– si les limites de la case initiale sont déplacées, la médiane de chaque case l'est également et deux parcelles – spatialement proches – dont les barycentres se situaient de part et d'autre de la médiane (horizontale ou verticale) d'une case peuvent se retrouver dans la même case. La décomposition se trouve ainsi modifiée.

Pour atténuer l'influence de ces paramètres de construction des CHA, nous construisons plusieurs chemins pour une zone d'étude en faisant varier les limites de la case initiale aléatoirement autour des limites géographiques qui définissent une zone (+- 10%) et en générant les chemins pour les 4 directions principales. Concrètement, dans l'expérience décrite ici, douze chemins ont été construits pour chacune des 14 sous-zones (3 cadrages aléatoires initiaux parcourus selon les 4 directions principales).

3.2.3 Exemple

La figure 5 illustre, à droite, un chemin de Hilbert adaptatif généré sur une sous-zone, et réduit aux barycentres des parcelles. Il comporte autant de points que de parcelles, soit 674 nœuds. La figure de gauche illustre le graphe sur la même zone, il comporte 1693 arcs.

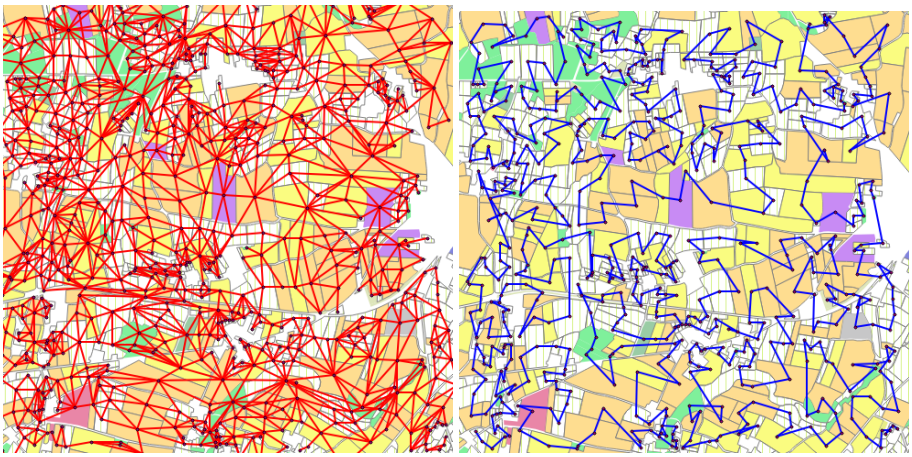

FIG. 5 – *À gauche : graphe de connexité d'une sous-zone ; à droite : chemin de Hilbert Adaptatif sur la même sous-zone*

3.3 Caractérisation d'une région par un « sac de nœuds » (SdN)

La caractérisation des régions par SdN débute par l'énumération de toutes les sous-structures d'une représentation de l'espace. Les sous-structures organisent le jeu de données initial au travers d'une vue spécifique des méthodes de fouille de motifs. Ces méthodes s'intéressent à dénombrer des motifs parmis ces sous-structures.

Définition 1 (Sous-structure)**.** *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 du CHA de taille fixe w_s. Dans le cas des graphes de voisinage, on s'intéresse à des sous-graphes contenant exactement w_g arcs. En reprenant l'exemple de la figure 3, pour $w_s = 3, w_g = 2$, les sous-structures construites sont les suivantes :

- pour le CHA : $\{p_3, p_1, p_2\}$, $\{p_1, p_2, p_4\}$, $\{p_2, p_4, p_5\}$, $\{p_4, p_5, p_6\}$
- pour le graphe : $\{p_1, p_2, p_3\}$, $\{p_4, p_2, p_3\}$, $\{p_1, p_4, p_3\}$, $\{p_4, p_6, p_3\}$, $\{p_4, p_5, p_3\}$, $\{p_1, p_6, p_3\}$, $\{p_1, p_5, p_3\}$, $\{p_6, p_5, p_3\}$.

L'énumération de toutes les structures est possible en temps output-polynomial (Bonzini et Pozzi (2007)). Contrairement à des approches de fouille de sous-séquences ou de fouille de graphes, il n'y a pas de seuil de fréquence à fixer. Pour l'énumération des sous-graphes, nous utilisons l'outil TGE de Uno (2005).

Définition 2 (« Sac de nœuds » (SdN))**.** *Soit S une représentation de l'espace, s_S une sous-structure. On appelle sac de nœuds, noté $SdN(s_S) \in 2^{\mathcal{O}}$, un vecteur de présence / absence des types d'occupations du sol dans la sous-structure s_S.*

Il faut noter qu'un sac de nœuds ne conserve que l'information de présence / absence. Cette solution a été préférée à celle consistant à compter le nombre de nœuds pour chaque type d'occupation du sol. Ce choix évite la multiplication combinatoire des sacs de nœuds, par ex. trois nœuds voisins, représentant deux parcelles vertes et une orange, seront regroupés dans le mêmes sac que trois nœuds représentant deux parcelles oranges et une verte. Sur l'exemple, les SdN construits sont donnés ci-dessous. On liste dans chaque ensemble uniquement les éléments de $\mathcal{O}$ qui sont présents ; selon des occupations répétées, les ensembles n'ont pas nécessairement tous la même cardinalité :

- pour le CHA : $\{v, o, j\}$, $\{o, j\}$, $\{j, o, v\}$, $\{o, v, j\}$
- pour le graphe : $\{o, j, v\}$, $\{o, j, v\}$, $\{o, v\}$, $\{o, j, v\}$, $\{o, v\}$, $\{o, j, v\}$, $\{o, v\}$, $\{j, v\}$.

Finalement, on dénombre trois occurrences du SdN $\{v, o, j\}$ et une occurrence de $\{o, j\}$ pour le CHA. Pour le graphe, on dénombre quatre occurrences du SdN $\{v, o, j\}$, trois du SdN $\{v, o\}$ et une du SdN $\{j, v\}$. On constate sur cet exemple que les deux méthodes extraient presque également le SdN le plus fréquent mais différemment les SdN moins fréquents : ainsi les SdN $\{v, o\}$ et $\{j, v\}$ n'ont pas été captés par le CHA ; inversement le CHA a capté un sac de nœuds $\{o, j\}$ inexistant pour le graphe. Notons qu'ici un seul chemin a été construit.

4 Comparaison des méthodes sur les données complètes

4.1 Principes

La comparaison s'effectue selon deux étapes. Lors de la première étape, on évalue la corrélation globale entre les sacs de nœuds obtenus par l'approche graphe et par l'approche chemin sur un ensemble de zones issues de la région étudiée. On utilise pour cela la mesure de corrélation de Spearman, suivie d'une mesure du χ^2 (Saporta (1990)). Dans la seconde étape, on examine les différences entre les sacs de nœuds obtenus par le graphe et par le chemin, en

particulier les sacs oubliés par la méthode CHA et la part qu'ils représentent dans l'ensemble des sacs obtenus par la recherche par graphe.

La corrélation globale est évaluée en comparant les décomptes des différents SdN pour le graphe et pour le chemin sur la zone étudiée. Le tableau 4.1 ci-dessous illustre les décomptes obtenus pour l'exemple de la figure 3. Outre le calcul de corrélation, ce tableau permet de mettre en évidence les SdN trouvés conjointement par les deux méthodes et ceux qui sont oubliés par l'une ou l'autre.

	graphe	chemin
{o, j, v}	4	3
{o, j}	0	1
{j, v}	1	0
{o, v}	3	0

TAB. 1 – *Décompte des SdN pour le graphe et le chemin de l'exemple de la figure 3*

Pour une grande zone telle que celle que nous voulons étudier (cf. figure 2), le parcellaire est découpé en sous-zones analysées séparément pour réduire le nombre de parcelles à prendre en compte et donc obtenir des résultats plus rapidement. Cela permet également d'étudier le caractère spécifique des motifs d'une zone par rapport à une autre et le comportement des deux méthodes en relation avec les caractéristiques des zones.

4.2 Résultats

L'analyse des zones par les deux méthodes a permis d'extraire pour chacune un certain nombre de sacs de nœuds distincts dont on a dénombré les répétitions.

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. Toutefois, on observe deux zones remarquables, l'une où le nombre de sacs chemins est presque équivalent au nombre de sacs graphes (zone 11, rapport 1,1), l'autre au contraire où le rapport est fortement plus élevé (zone 33, rapport 2,9). Soulignons que la zone 11 est celle qui compte le moins de sacs (29 sacs extraits des graphes), c'est en effet une zone très homogène (bocage avec une large majorité de prairies semées), tandis que la zone 33, très hétérogène, est celle qui en compte le plus (301 sacs distincts extraits des graphes). Si on exclut ces deux zones de l'analyse, le rapport ne bouge pas.

Pour les graphes, le nombre d'éléments dans chaque sac varie entre 1 et 237630 (zone 32, sac {bois, prairie semée}) sur les différentes zones. Pour les sacs issus des chemins, le nombre maximum d'éléments s'élève à 3006 (zone 21, sac {prairie semée}). 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 (soit un rapport de 1/31). 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_χ(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%. On remarque que ces chiffres sont très variables, ils 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 (voir tableau 2). On retrouve les comportements atypiques des zones 11 et 33. Au delà on peut séparer des zones plutôt homogènes – peu de sacs perdus et peu remplis – (zones 22, 12, 31, 21, 41) et des zones plutôt hétérogènes – sacs plus nombreux et plus remplis (zones 51, 42, 23, 52, 43). 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 types de voisinages relativement fréquents même si non majoritaires.

nombre	numéro de zone													
d'éléments	22	51	53	11	33	12	31	32	42	23	21	52	43	41
1-10	8	19	16	0	76	4	11	14	28	16	12	13	36	12
11-100	5	35	19	2	79	1	7	15	35	23	8	18	36	7
101-500	2	12	8	0	32	5	5	7	9	15	4	24	15	2
> 500	0	4	3	1	11	1	2	3	5	1	4	6	1	0

TAB. 2 – *Répartition des sacs perdus dans les CHA selon le nombre d'éléments dans les sacs correspondants des graphes sur les différentes zones*

5 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 (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 acceptable et pertinente en termes de rapidité et d'efficacité.

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 les plus rares mis en évidence par la méthode appuyée sur le graphe de voisinage. En effet, du fait de sa structure le CHA perd certains voisinages : un motif fréquent dans l'espace ne l'est pas forcément sur le CHA alors que l'inverse est vrai. La méthode est donc incomplète. De plus il est nécessaire de répéter les chemins sur une zone selon différents

cadrages afin d'obtenir une certaine « robustesse » des motifs extraits. Cependant, nous avons montré que cet inconvénient est variable selon l'hétérogénéité des zones étudiées.

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 espérer des résultats de caractérisation des parcellaires intéressants par des méthodes de fouille de données. En effet, les méthodes de fouille de données 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. L'analyse peut être développée en considérant des voisinages plus éloignés, au prix d'une complexité calculatoire plus élevée, en particulier pour les graphes. Une extension du travail portera aussi sur la recherche de motifs ordonnés afin de mieux spécifier les voisinages, et d'en garder partiellement la structure spatiale, en lien avec des problématiques agro-écologiques.

Références

Bartolini, I., P. Ciaccia, I. Ntoutsi, M. Patella, et Y. Theodoridis (2004). A unified and flexible framework for comparing simple and complex patterns. In *Knowledge Discovery in Databases, PKDD 2004*, pp. 496–499. Springer.

Bartolini, I., P. Ciaccia, I. Ntoutsi, M. Patella, et Y. Theodoridis (2009). The Panda framework for comparing patterns. *Data Knowl. Eng. 68*(2), 244–260.

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.

Candillier, L., I. Tellier, F. Torre, et O. Bousquet (2006). Cascade evaluation of clustering algorithms. In *Machine Learning : ECML 2006*, pp. 574–581. Springer.

Castellazzi, M., J. Matthews, F. Angevin, C. Sausse, G. Wood, P. Burgess, I. Brown, K. Conrad, et J. Perry (2010). Simulation scenarios of spatio-temporal arrangement of crops at the landscape scale. *Environmental Modelling & Software 25*(12), 1881–1889.

Castellazzi, M., J. Perry, N. Colbach, H. Monod, K. Adamczyk, V. Viaud, et K. Conrad (2007). New measures and tests of temporal and spatial pattern of crops in agricultural landscapes. *Agriculture, Ecosystems & Environment 118*, 339–349.

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

Gosselin, P.-H. (2011). *Apprentissage interactif pour la recherche par le contenu dans les bases multimédias*. Habilitation à diriger des recherches, Université de Cergy Pontoise.

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.

Lazrak, E., J.-F. Mari, et M. Benoît (2010). Landscape regularity modelling for environmental challenges in agriculture. *Landscape Ecology 25*, 169–183.

Le Ber, F., M. Benoît, C. Schott, J.-F. Mari, et C. Mignolet (2006). Studying crop sequences with CARROTAGE, a HMM-based data mining software. *Ecological Modelling 191*(1), 170–185.

Megalooikonomou, V., Q. Wang, G. Li, et C. Faloutsos (2005). A multiresolution symbolic representation of time series. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pp. 668–679. IEEE.

Nourine, L. et J.-M. Petit (2012). Extending set-based dualization : Application to pattern mining. In *ECAI*, Volume 242 of *Frontiers in Artificial Intelligence and Applications*, pp. 630–635. IOS Press.

Négrevergne, B., A. Termier, M.-C. Rousset, et J.-F. Méhaut (2013). Paraminer : a generic pattern mining algorithm for multi-core architectures. *Data Mining and Knowledge Discovery 20*(3), 300–320.

Peano, G. (1890). Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen 36*(1), 157–160.

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.

Saporta, G. (1990). *Probabilités, analyse des données et statistiques*. Édition Technip.

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.

Sorel, L., V. Viaud, P. Durand, et C. Walter (2010). Modeling spatio-temporal crop allocation patterns by a stochastic decision tree method, considering agronomic driving factors. *Agricultural Systems 103*(9), 647 – 655.

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

Wale, N., I. Watson, et G. Karypis (2008). Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems 14*(3), 347–375.

Weber, M., M. Welling, et P. Perona (2000). Unsupervised learning of models for recognition. In *Computer Vision - ECCV 2000*, Volume 1842 of *Lecture Notes in Computer Science*, pp. 18–32. Springer Berlin Heidelberg.

Zhao, P., J. X. Yu, et P. S. Yu (2007). Graph indexing : tree + delta <= graph. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pp. 938–949.

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). A BoN is a vector representing the presence or absence of node types in the sub-structures of sequences or graph. BoNs are compared to highlight the ability of the two proposed methods to characterize the neighborhood of agricultural fields. The results suggest that the linearization of space captures most of the information – except some rare elements – about the organization of agricultural fields. Thus, data mining algorithms using these linear representations can be used instead of less efficient tools such as graphs mining algorithms.