Fouille de données spatiales et modélisation de linéaires de paysages agricoles

Sébastien da Silva

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Fouille de données spatiales et modélisation de linéaires de paysages agricoles

THÈSE

présentée et soutenue publiquement le 11 Septembre 2014

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Sébastien Da Silva

Composition du jury

Rapporteurs :	Marie-Odile Cordier	Professeur, Université de Rennes
	Didier Josselin	Directeur de Recherche CNRS, Avignon
Examinateurs :	Katarzyna Adamczyk	Ingénieur de Recherche, INRA Jouy-en-Josas
	Isabelle Debled-Rennesson	Professeur, Université de Lorraine
	Alexandre Joannon	Chargé de Recherche INRA, Rennes
	Amedeo Napoli	Directeur de Recherche CNRS, Nancy
Directrices :	Florence Le Ber	Directrice de la Recherche, ENGEES Strasbourg
	Claire Lavigne	Directrice de Recherche, INRA-PACA Avignon

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

INRA SCIENCE & IMPACT

Loria Laboratoire lorrain de recherche en informatique et ses applications

Inria INVENTEURS DU MONDE NUMÉRIQUE

Mis en page avec la classe thesul.

Remerciements

Il est courant d'entendre au sujet des thésards qu'ils doivent manger «thèse», dormir «thèse», respirer «thèse» durant toute la durée de leur recherche. Je dois dire que ce fut effectivement mon cas, et ce jusqu'à l'apothéose du 11 septembre 2014, date de la soutenance. Mais pour en arriver là, j'ai dû et pu m'appuyer sur de très nombreuses personnes rencontrées depuis mes débuts. Je vais essayer, ici, d'évoquer, afin de remercier, tous ceux qui ont rendu cela possible, en remontant le cours de mon histoire.

Tout d'abord, je tiens à remercier les membres de mon jury d'avoir accepté d'y participer. Leurs conseils et leurs remarques, ainsi que leurs encouragements m'ont aidé à apprécier tout le chemin parcouru depuis le mois d'octobre 2010. Je remercie également, et pour les mêmes raisons, les membres de mon comité de thèse, Mme ANGEVIN Frédérique, Mme MIGNOLET Catherine, Mr FAIVRE Robert, et Mr MONESTIEZ Pascal. Leur œil d'expert sur mon travail m'a permis de toujours retrouver mon chemin dans la recherche et d'obtenir les résultats présents dans ce manuscrit.

Les derniers mois de recherche furent financés par mon poste d'ATER à l'université de Lorraine. Grâce à cela, j'ai pu enseigner au sein de l'école d'ingénieur TELECOM Nancy. Je remercie Mr FESTOR Olivier, Mme CHRISTMENT Isabelle et Mme COLLIN Suzanne de m'avoir accordé leur confiance dont j'espère avoir été digne. Je les remercie également de cette opportunité qu'ils m'ont donnée d'enseigner de nouveau. La transmission du savoir est une chose importante, et d'y prendre part à nouveau m'a permis de me rappeler à quel point j'aimais cela.

Les trois premières années de cette thèse furent financées avec un Contrat Jeune Scientifique, partenariat entre l'INRA (Institut National de la Recherche Agronomique) et Inria. Ce fut une vrai chance, cela m'a permis de travailler dans deux équipes de recherche aux thématiques très différentes. Ma vie au quotidien s'est écrite entre ces deux instituts.

Tout d'abord, l'Équipe de la Production Intégrée (EPI) dirigée par Mme LESCOURET Françoise puis par Mme LAVIGNE Claire, dans l'unité Plantes et Systèmes Horticoles (PSH), dirigée par Mr GÉNARD Michel, à l'INRA d'Avignon. Je fus bien accueilli par les membres de l'unité et rapidement considéré comme un doctorant de l'équipe, même si mon travail de recherche dépendait de l'école doctorale d'informatique à Nancy. Je n'ai été effectivement qu'un an dans les murs de l'équipe mais j'ai pu rencontrer des personnes méritant d'être connues. Je garderai en mémoire autant les réunions de travail que les sorties entre thésards, avec une mention spéciale pour le «Goûter de Noël des Thésards», tradition se perpétuant encore aujourd'hui... Je

tiens à remercier l'ensemble des personnes de l'unité que j'ai côtoyé, durant une simple discussion ou plus longtemps. Plus particulièrement, je souhaite mettre en avant trois personnes grâce à qui les moments de doute et de déprime furent beaucoup plus courts. Tout d'abord, Mariline, thésarde de son état, particulièrement investie par sa mission de recherche. Elle m'a permis de prendre du recul sur mon sujet et de mettre en perspective mes tracas de jeune chercheur, nos longues heures de discussion ont été pour chacun, je l'espère, des moments de calme et de réconfort avant de se donner corps et âme à la thèse. Ensuite, Jean-François, également thésard, mais surtout grand supporter de l'OM (personne n'est parfait).Toujours de bonne humeur, j'ai eu la chance de pouvoir compter sur lui lors de mes baisses de régime et profiter de son expérience dans l'unité afin de toujours trouver ce dont j'avais besoin. Finalement, je tiens à rendre hommage à Eva, ingénieur de l'unité, qui a eu le mérite de partager mon bureau durant mes séjours dans l'équipe, et je n'étais pas forcément tous les jours facile à vivre. Mais ses facéties incessantes et son énergie débordante m'ont toujours été d'un grand secours. Elle fut un soleil de substitution quand, dans notre petit bureau du sous-sol, les jours paraissaient bien gris et bien froid... et puis, jamais je n'oublierai son interprétation de «Hello Goodbye» des Beatles.

Ensuite, l'équipe dans laquelle j'ai passé le plus clair de mon temps, à savoir l'équipe ORPAILLEUR dirigé par Mr NAPOLI Amédéo au LORIA de Nancy. Je dois dire qu'au départ, le changement de climat fut très difficile entre mon Sud natal et cet Est Sibérien, mais rendu supportable grâce aux membres de l'équipe. Il est très difficile donc de ne pas remercier personnellement chacun d'entre eux pour leur aide, leur soutien, leur présence, enfin bref, d'avoir été là quand il le fallait... alors dans le désordre, je remercie Zainab, Mehdi, Chedy, Adrien, Sébastien, Jean-Hyacinthe alias «In the Kitchen» , Aleksey, Julien, Mehwish, Jérémie, Mathieu, Victor, My-Thao, Laura, Luis-Felipe, Thomas, Amédéo, Inaki, Léo, Maxime, Sylvain et Elias devenu un ami précieux. Dans cette équipe où aucune distinction de niveau n'est pratiquée, j'ai vécu les meilleurs moments de ces quatre années de recherche, dans et en dehors du laboratoire, et pour cela, je vous en remercie encore.

Je remercie également mes deux directrices de thèse, Mme LAVIGNE Claire et Mme LE BER Florence de m'avoir permis de travailler avec elles, de m'avoir donner l'opportunité de travailler dans la recherche. Je les remercie également pour leur soutien et leur disponibilité dans mon quotidien, il n'est pas toujours aisé, en tant que thésard, de trouver son chemin en ayant la tête coller à ses résultats.

Enfin, je souhaite remercier certaines personnes ne faisant pas parti de ce monde professionnel et me connaissant depuis plus longtemps que le début de la thèse. Tout

d'abord Stéphanie, ma sœur, qui m'a aidé durant toutes mes années d'études, toujours présente afin de me faire avancer et réussir avec sa bienveillance et son soutien indéfectible. Et ensuite, mes parents qui, en plus de m'avoir donné la vie, m'ont permis d'arriver où je suis aujourd'hui, mais plus encore, comme je suis aujourd'hui, sans eux, tout ceci ne serait pas de moi. Plus particulièrement, merci maman, tu es ce que tu es et c'est pour cela que je suis devenu ce que je suis. Et si tu ne l'avais pas été, je n'en serai pas là, comme quoi, les "coups de pieds aux fesses" ont parfois du bon...

À vous tous, MERCI!!!
Sébastien

Je dédie cette thèse
à mes anges gardiens

Sommaire

Partie III Simulation et évaluation

Partie IV Conclusion

Partie V Bibliographie

Table des figures

Chapitre 1

Introduction

Multiplication des données

Actuellement, le volume de données numériques produit quotidiennement augmente en permanence. Il ne s'agit pas toujours d'informations précises mais d'un véritable déluge difficile à appréhender et à traiter car les données peuvent être hétérogènes et non structurées. Les professionnels travaillant autour du secteur de l'exploration de données, qu'il s'agisse de chercheurs, de marketeurs, ou de décideurs publics, doivent faire face ce flot de données, issu de la multiplicité des canaux de communications et d'outils numériques.

Historiquement, les premiers producteurs de données étaient des scientifiques, en quête de compréhension de phénomènes physiques naturels et qui utilisaient des systèmes d'acquisition de données. Aujourd'hui, tout un chacun produit, sans forcément en être conscient, des données plus ou moins accessibles. Il existait déjà les données de connexion et de navigation Internet, les informations sur les habitudes de consommation sur les sites marchands, sur les publicités vues ou cliquées, sur le temps de consultation. Les mobiles ont ajouté les données de géolocalisation. Facebook, Twitter, les médias sociaux ou collaboratifs en général, ont ouvert les vannes à une production libre, en temps réel et sans fin, mais qui reste stratégique pour ceux qui savent en extraire l'essentiel. A titre d'exemple, 30 milliards de documents sont ajoutés chaque mois sur Facebook, 140 millions de tweets sont écrits par jour, 20 millions de SMS ont été échangés par minute en 2013 et 35 heures de vidéos sont chargées chaque minute sur YouTube. Ce phénomène a pris une telle ampleur que la totalité des données produites dans le monde en 2013 dépasse de plus de 35% les capacités de stockage existantes.

Parmi ce flot de données, un type particulier nous intéresse, les données spatiales. C'est-à-dire celles incluant une information géolocalisée. Les données les plus communes

actuellement sont l'œuvre des systèmes GPS (*Global Positionning System*), qu'ils soient dans une voiture ou dans un smartphone. Les données ainsi produites sont spatiales (et temporelles), et l'industrie se veut une grande consommatrice de celles-ci. Ainsi, elle peut connaître le point d'entrée ou de sortie d'une zone spécifique, le temps passé dans un endroit ou les "centres d'intérêt" des utilisateurs de ce type d'appareillage. Ces informations se révèlent digne d'intérêt pour adapter, par exemple, les messages publicitaires affichés à proximité d'un utilisateur identifié.

D'autres mesures spatiales peuvent être le résultat de l'utilisation de satellites d'observation de la terre, de relevés topographiques, de la numérisation du cadastre ou de déclarations des agriculteurs, dans le cadre de la PAC (politique agricole commune).

Dans le contexte actuel, les données spatiales sont particulièrement intéressantes pour les agronomes et les pouvoirs publics car l'un comme l'autre sont à la recherche de solutions pour la préservation de l'environnement et des ressources naturelles dans l'optique du développement durable. Les pouvoirs publics cherchent, par exemple, à extraire les connaissances des données spatiales afin d'être mieux préparés et faire face aux risques climatiques. Ils rejoignent les agronomes dans leurs travaux pour repousser les limites actuelles des systèmes productivistes ou implanter de nouveaux systèmes de culture. Le dernier exemple en date, en France, correspond aux tempêtes qui ont sévi en Bretagne durant l'hiver 2014. D'un coté, l'opinion publique pointait du doigt les agriculteurs estimant que la destruction du bocage historique pour l'implantation d'une agriculture intensive avait renforcé de façon considérable les conséquences néfastes de cet épisode climatique. De l'autre, certains agriculteurs expliquaient devoir répondre à la consommation de masse en se pliant aux exigences du marché dictées par ceux-là même qui composent l'opinion publique. Dans ce cas, les pouvoirs publics et les scientifiques interviennent pour trouver un équilibre entre la protection des citoyens et le maintien d'une agriculture moderne.

Une telle quantité de données avec une telle diversité de nature, de forme et de stockage nécessite un travail considérable pour en retirer les informations utiles aux chercheurs ou aux décideurs publics. Depuis une vingtaine d'années, tout un champ de recherche s'est développé afin d'extraire de cette masse de données ces informations utiles, il s'agit de l'exploration de données également nommée "Extraction de Connaissances à partir de Bases de Données" (ECBD), en anglais "Knowledge Discovery in Databases" (KDD). Dans cette optique, des outils logiciels sont développés pour optimiser le stockage, l'étiquetage des données mais également pour analyser les contenus, y compris au niveau sémantique.

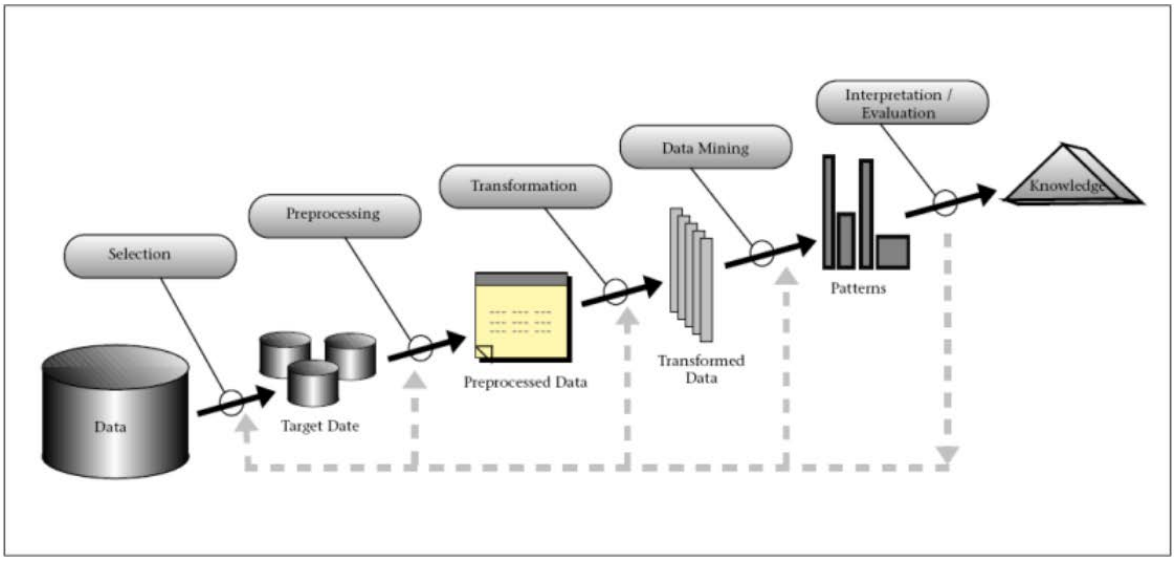


FIGURE 1.1 – Un aperçu des étapes qui composent le processus d'Exploration de Données [30]

L'exploration de données ou Extraction de Connaissances à partir de Bases de Données

Il existe deux façons d'envisager l'exploration de données. La première, une fouille à l'aveugle en procédant, par exemple, avec une segmentation non supervisée ou une analyse de données. Il s'agit alors de travailler à partir d'une base de données sans connaître à l'avance les informations recherchées. La deuxième, une fouille ciblée qui consiste à trouver une information particulière dans une base de données, le premier défi consiste justement à définir cette information. Dans les deux cas, la démarche de l'extraction de connaissances mise en place suit un cheminement identique pour tous les types de données (Figure 1.1).

Tout d'abord, une première étape permet de sélectionner, parmi les données disponibles, celles qui seront traitées, et ainsi réduire le volume de données à traiter. Ensuite, une étape de prétraitement est utilisée afin d'épurer les données sélectionnées et obtenir des données adaptées à l'analyse. Nous pouvons voir le prétraitement comme une phase de "nettoyage" des données (par exemple, gestion des données manquantes). L'étape suivante permet de transformer les données. Il s'agit là d'écrire l'information contenue dans les données d'origine sous une forme en adéquation avec une méthode de fouille. La transformation peut consister en une discrétisation ou agrégation des données, en une linéarisation de données bidimensionnelles ou à la construction de graphes [41]. L'étape suivante consistera à appliquer aux données transformées des méthodes

de fouille de données. La dernière étape permet de visualiser les informations extraites et ainsi facilite leurs interprétations par les experts ou les analystes afin, par exemple, d'aider à la décision des pouvoirs publics, d'instruire l'opinion publique. Cette dernière étape est un champ de recherche à part entière.

L'ECBD peut s'appliquer à tous les types de données, notamment les données spatiales, et leur traitement emploie indifféremment des méthodes mathématiques et informatiques. La fouille de données spatiales est un domaine actif, qu'il s'agisse de travailler sur des photos satellitaires [66], des trajectoires de soins [28], des habitudes de consommation ou de transport. Les agronomes conjointement avec les informaticiens et les mathématiciens déploient également beaucoup d'énergie pour développer des méthodes de fouille de données applicables à leur domaine.

Démarche et Contribution

Notre travail s'inscrit dans le champ de l'Extraction de Connaissances à partir de Bases de Données sur des données spatialisées. La question initiale de ce projet résulte de la problématique agronomique exposée et développée dans le chapitre 2. Il s'agit d'analyser et de générer les structures des paysages agricoles afin que les agronomes puissent comprendre et prévoir - en vue d'une meilleure maîtrise - leurs impacts sur les processus agro-écologiques (dynamique et régulation des ravageurs, pollinisation...). Pour cela il est question d'une part, de décrire les structures du paysage pour caractériser les relations existantes entre celles-ci et les processus agro-écologiques. D'autre part, nous devons analyser les structures de paysages pour mettre en évidence les règles sous-jacentes à leurs implantations et permettre de considérer ensemble des paysages ayant des caractéristiques proches. Et enfin, nous devons générer des structures de paysages.

Les méthodes de fouille de données, développées dans ce travail, ont donc pour but d'analyser les structures de haies dans un paysage, mais doivent servir également, par la suite, à générer des structures de haies dans des paysages virtuels.

Notre parti pris est de générer des structures de haies dans un paysage virtuel, pour servir de bases à l'étude des processus agro-écologiques supportés par le paysage réel, sans utiliser d'information quant aux processus agro-écologiques qui ont structuré ce paysage. Par exemple, nous ne considérons pas la dispersion des graines d'arbres, ni l'action de l'agriculteur qui implante des haies. Nous avons fait le choix de générer des modèles s'appuyant sur les caractéristiques géométriques des structures car celles-ci sont importantes pour les processus agro-écologiques étudiés. A l'heure actuelle, les

méthodes d'analyse et de génération des paysages ne portent que sur le parcellaire mais pas sur les linéaires. Notre objectif est donc de développer une méthode d'apprentissage sur l'implantation des haies dans les paysages agricoles permettant ensuite de les générer.

Ce problème est aussi un problème d'exploration des données puisqu'il s'agit d'extraire des informations d'une grande quantité de données concernant l'implantation des haies dans les paysages agricoles mais aussi l'implantation des réseaux routiers et hydriques. La démarche adoptée dans ce travail est donc inspirée de la démarche générale de l'exploration des données.

Les données de départ, qui concernent deux zones aux caractéristiques bien différentes, ont été fournies sous forme de polylignes, au format des Systèmes d'Information Géographique (SIG). Elles regroupaient les informations d'implantation des haies et des réseaux routier et hydrique. Dans la phase de prétraitement, nous avons choisi d'extraire de ces données tous les segments de haies, de routes et de canaux et de les sauvegarder dans un format texte (Flèche 1 sur la figure 1.2). Ce format de données est plus facilement manipulable par les méthodes de fouille de données. Nous avons aussi pu définir les variables caractéristiques attachées aux segments (Flèche 2 sur la figure 1.2).

Nous n'avons pas travaillé l'ensemble des données comme un tout car les zones d'étude sont hétérogènes et nous ne voulions pas risquer de lisser l'information sur l'ensemble. Nous avons donc découpé les zones en cellules carrées de tailles identiques (Flèche 3 sur la figure 1.2). Ce découpage nous a permis également de disposer de plusieurs jeux de données, et donc d'opérer des répétitions.

Ensuite, nous avons réalisé une classification hiérarchique sur les cellules carrées selon le nombre de segments qu'elles renfermaient (Flèche 4 sur la figure 1.2) et une classification des segments de haies selon leur valeur pour les variables caractéristiques (Flèche 5 sur la figure 1.2). À l'issue de ces segmentations, nous disposons de groupes de cellules et de groupes de segments similaires. Ainsi, nous avons construit des données avec deux niveaux d'échelle permettant des analyses variées. Nous avons, par là même, construit des répétitions de jeux de données avec des caractéristiques similaires ce qui permet de vérifier la stabilité des méthodes.

Après ces premières étapes, nous avons engagé deux approches distinctes dans les phases de transformation et de fouille.

Avec la première approche, nous voulions caractériser l'implantation des structures en prenant totalement en compte l'aspect multidirectionnel des relations de voisinage (en opposition avec la démarche suivante qui ne garde qu'une partie de l'information

spatiale). Nous avons donc introduit une nouvelle variable construite à partir de la distance entre voisins (Flèche 6 sur la figure 1.2). La fouille de données mise en place sur cette variable a permis d'extraire des motifs de voisinage sur les classes de cellules (Flèche 7 sur la figure 1.2). Nous avons ainsi pu mettre en évidence des principes d'implantation des segments de haies par rapport aux autres segments (Flèche 8 sur la figure 1.2) (ex : les haies orientées Nord-Sud sont parallèles aux routes à faible distance).

La seconde approche comporte une transformation fondée sur les méthodes de linéarisation du plan par les courbes de Hilbert-Peano (Flèche 9 sur la figure 1.2). Les méthodes de Markov ont été utilisées pour fouiller cette information linéaire (Flèche 10 sur la figure 1.2) et cet apprentissage a permis la création d'un modèle de génération pour chaque cellule. Ces deux méthodes sont très utilisées dans certains traitements de l'information spatiale. Lorsqu'elles sont couplées, elles présentent l'avantage de simplifier, de façon raisonnable, les données et leurs analyses. Nous avons privilégié une approche markovienne car nous voulions disposer d'un modèle capable de générer des données. De plus, les modèles appris ont été classés, grâce à une distance entre les matrices de transition (Flèche 11 sur la figure 1.2). Cela a permis de confronter les résultats des deux méthodes de classement des cellules pour évaluer la cohérence des deux approches (Flèche 12 sur la figure 1.2). L'objectif est aussi de disposer de classes de modèles reliés à des types de paysages, classes dans lesquelles nous puisons pour la génération.

Nous avons ensuite développé un algorithme pour générer des structures virtuelles de segments (Flèche 13 sur la figure 1.2) en nous appuyant sur les modèles de Markov précédents. Ces structures devront, par la suite, être post-traitées avec les résultats obtenus lors de la première phase de fouille (Flèche 14 sur la figure 1.2). De plus, un algorithme plus précis devra permettre de générer des structures de paysages sur une zone d'étude complète (Flèche 15 sur la figure 1.2). Dans une étape future, la comparaison, avant et après post-traitement, des paysages réels et virtuels permettra d'évaluer la méthode de génération.

Enfin, durant le travail de recherche, une attention particulière a été apportée à l'automatisation de la démarche afin de la rendre accessible aux non spécialistes de l'exploration de données. La création d'un logiciel en est le résultat.

Finalement, nous avons implanté une méthode pour la fouille de données capable d'effectuer l'apprentissage et l'analyse mais aussi de générer à son tour des données. Nous pouvons ainsi produire des données utilisables pour l'étude des processus agro-écologiques présents dans le paysage. De plus, notre approche spatiale s'appliquant à

des segments, nous avons développé une méthode de fouille de données adaptée à ces objets unidimensionnels localisés dans un espace bidimensionnel et prenant en compte une vraie relation de proximité spatiale, c'est-à-dire un point de vue géographique, dont les modèles d'aujourd'hui sont peu pourvus.

Plan

Nous présenterons le résultat de nos recherches en suivant la démarche scientifique utilisée pendant cette thèse. Tout d'abord, nous présenterons plus avant la problématique agronomique dans le chapitre 2 et la relation entre les haies et les processus agro-écologiques dans le paysage agricole. Nous parlerons ensuite des travaux déjà réalisés dans le domaine de l'exploration des données et proches de notre problématique (chapitre 3).

Le reste du manuscrit suit le même déroulement que celui mis en œuvre pour l'exploration de données. Dans la suite de ce plan, le numéro entre parenthèse fera référence à la flèche correspondante dans la version graphique de ce plan, sur la figure 1.2.

Le chapitre 4 présentera les données et les pré-traitements (1, 2 et 3) qui y ont été appliqués, à l'échelle des segments et des cellules, ainsi qu'une première phase des transformations (4 et 5).

Dans le chapitre 5, nous présenterons la première méthode de transformation (6) et de fouille de données (7) pour la caractérisation des structures spatiales. Cette méthode se fera par l'étude des voisinages. Nous présenterons l'interprétation des résultats (8) de cette approche pour les segments de haies dans le paysage.

Le chapitre 6 présentera la deuxième méthode de transformation (9) et de fouille de données (10) permettant de caractériser les structures spatiales pour les générer. La caractérisation (classification) des données par cette méthode (11) sera comparée à celle obtenue dans le chapitre 4 (12).

Finalement, dans le chapitre 7, nous détaillerons les algorithmes développés à partir des résultats précédents pour la génération de structures de segments dans une cellule (13). Nous présenterons les résultats obtenus ainsi qu'une première analyse de ceux-ci. Nous présenterons également une nouvelle version de l'algorithme précédent (15) mais ajusté à partir des connaissances du domaine obtenus dans le chapitre 5 (14).

Le dernier chapitre est la conclusion de ce travail de recherche synthétisant les résultats obtenus, et présentant les perspectives pouvant y faire suite.

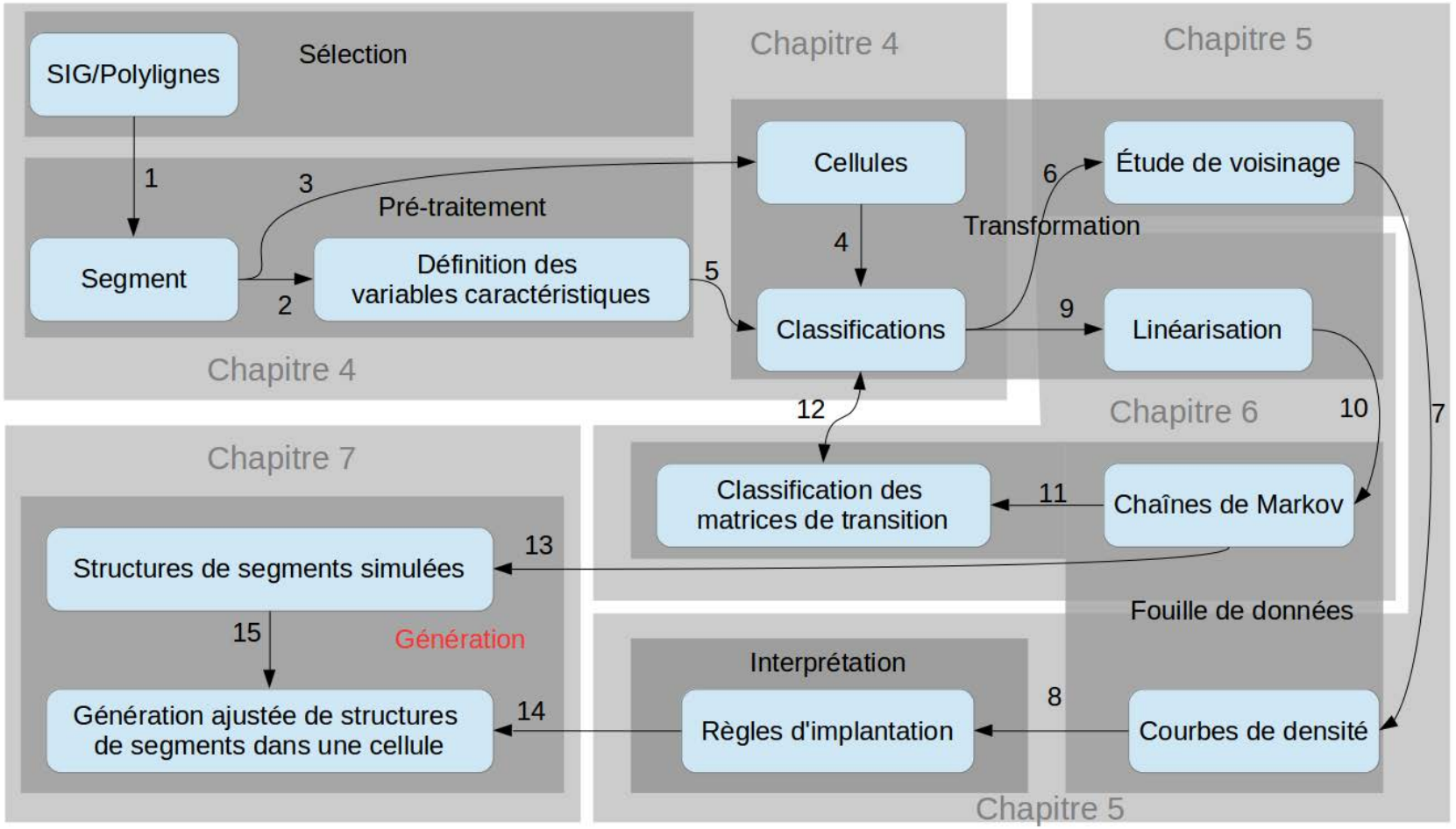


FIGURE 1.2 – Un aperçu des étapes qui composent le cheminement du travail de recherche durant la thèse

Première partie

Contexte

Chapitre 2

Problématique

Concrètement, un paysage agricole est une surface continue composée d'éléments de nature diverse aux fonctions différentes [33], des terres arables, des chemins, des canaux d'irrigation ou des haies par exemple. Ces éléments ne sont pas disposés de façon similaire sur l'ensemble des paysages agricoles, ni de façon homogène sur un seul paysage agricole. Plus généralement, le paysage est un niveau d'organisation des systèmes écologiques, supérieur à l'écosystème ; il se caractérise par son hétérogénéité et par sa dynamique gouvernée pour partie par les activités humaines [16]. Cette diversité d'implantation signifie que les éléments créent une structure différente pour chaque paysage. Cette structure peut être caractérisée par la composition du paysage, à savoir la surface relative de chaque élément ou leur nombre, et par la configuration du paysage, à savoir la distribution spatiale de ces éléments. Or, cette structure a une influence sur de nombreux processus qui se déroulent dans les paysages agricoles. Avec la prise de conscience des problèmes liés à l'utilisation intensive de ces paysages, de plus en plus d'acteurs de la vie publique cherchent à obtenir des réponses quant aux problèmes de la préservation des qualités sanitaires, écologiques ou économiques [12, 82] des paysages agricoles. En particulier, comprendre comment la structure du paysage affecte la dynamique des populations animales ou végétales est devenu une question clef dans les problèmes de conservation des espèces [10] mais aussi dans les études agro-écologiques [37].

Des études ont porté sur les déplacements des insectes sur le paysage agricole [96] et, une attention particulière a été apportée à la compréhension du lien entre distribution spatiale des éléments semi-naturels et abondance et répartition de certaines espèces [20, 70, 14, 101]. De ce fait, de nombreux indices ont été développés pour caractériser les éléments du paysage agricole et leur répartition spatiale. Ces indices peuvent porter sur les éléments eux-mêmes (superficie, forme), leur connectivité ou l'hétérogénéité

du paysage à différentes échelles [86]. Ces indices ont également été utilisés dans les approches de modélisation qui visent à mettre en évidence comment les interactions entre la structure du paysage et la dynamique de population influent sur l'abondance des espèces et leur structure génétique [103, 23]. Parmi les études empiriques portant sur les effets de la composition du paysage agricole, peu traitent des éléments linéaires (i.e. des éléments qui peuvent être représentés par des lignes, par exemple, canaux d'irrigation [4] ou haies [32]) et du rôle qu'ils peuvent jouer. Cependant, les haies sont des éléments majeurs dans la caractérisation des paysages agricoles et jouent plusieurs rôles pour les espèces qui évoluent sur ce paysage. Elles peuvent être des habitats pour des espèces végétales (arbres, plantes) spécifiques, elles forment des couloirs de circulation pour les individus entre les parcelles de forêt ou au contraire, un obstacle à la dispersion des espèces spécialisées dans les zones ouvertes [15, 24, 77]. Elles offrent un habitat aux espèces vivant sur les paysages agricoles [2, 44]. En outre, les effets de brise-vent et l'ombre peuvent produire des modifications locales du micro-climat et des turbulences dans les flux d'air [42], qui peuvent affecter la survie des espèces ou leur reproduction [96]. Les haies peuvent également être utiles pour maintenir la qualité des eaux et protéger les sols [63]. Et, dans une certaine mesure, elles fournissent un complément de revenus à l'agriculteur qui exploite son bois [72].

Le travail présenté ici ne tient pas compte des processus écologiques qui sont affectés par les haies. De plus, nous nous concentrons sur la caractérisation de la distribution spatiale des haies, contrairement à d'autres études qui définissent des méthodes pour caractériser le type et la composition des haies [78, 56]. Plusieurs indices ont déjà été proposés au niveau du paysage par Groot *et al.* [39] et appliqués aux haies dans une zone agro-écologique aux Pays Bas. Une étude récente caractérise la densité des lignes vertes (haies et bandes enherbées) sur les paysages européens [97]. Toutefois, les interactions spatiales locales entre les éléments du paysage n'ont pas été prises en compte dans ces études. L'emplacement des linéaires, dont les haies, n'est pas aléatoire et répond à des contraintes extérieures [71]. A l'échelle européenne, Van der Zanden *et al.* [97] ont montré que les méthodes basées sur l'auto-corrélation spatiale ont échoué parce que l'implantation des lignes vertes dépendait de la présence d'autres utilisations des terres telles des cultures de rente ou des élevages aussi bien que de la vitesse du vent. En conséquence, nous pouvons nous attendre à ce que la distribution spatiale des haies sur les paysages présente des caractéristiques spécifiques basées sur la fonction de ces haies (par exemple, clôture, brise-vent, bois, habitat d'auxiliaire,...) et la répartition spatiale des autres éléments plus pérennes tels que les routes et les voies d'eaux (ou fossés). Nous nous intéressons à ces deux hypothèses dans la première partie de notre

étude en nous attardant sur le deuxième point.

De plus, nous avons voulu construire des modèles neutres [95] qui simulent des haies dans un paysage agricole et servent de support à l'étude de processus agro-écologiques. Le propre d'un modèle neutre est de générer le comportement espéré d'un système en l'absence de tout processus spécifique qui pourrait affecter le système (d'après Caswell [19] cité dans [35]). Les modèles neutres ont été développés en écologie du paysage et popularisés par l'agroforesterie ; ils sont utilisés dans l'analyse des paysages et permettent de tester des hypothèses sur les phénomènes constituant un paysage. Gardner *et al.* [35] utilisent des modèles neutres définis par morceaux (*patches*) afin d'analyser les motifs de paysage à grande échelle. Les modèles neutres de paysage peuvent aussi être utilisés pour étudier l'effet de la variation des types de paysage sur les processus écologiques [71] ou liés à l'agriculture [57]. Nous cherchons alors à maîtriser les variations des paysages pour les relier aux variations des processus. En ce sens, nous pouvons parler de paysages virtuels, construits à partir des caractéristiques de paysages réels (Castellazzi *et al.*, [18]). Dans notre étude, nous nous inspirons des approches géométriques récentes telles que celles proposées par Gaucherel *et al.* [36] et Le Ber *et al.* [59] pour simuler des parcellaires de paysages agricoles. Toutefois, ces approches ne prennent pas en compte les éléments linéaires auxquels nous nous intéressons ici.

Chapitre 3

État de l'art

Introduction

L'information spatiale se présente sous différentes formes : données ponctuelles, polylignes, polygones... Pour la traiter, nous disposons d'un ensemble de méthodes présentées, par exemple, dans *Statistics for Spatial Data* [22] ou, de manière plus appliquée dans *Geospatial analysis : a Comprehensive Guide to Principles, Techniques and Software Tools* [25]. Les plus répandues se basent sur les modèles de Markov ou sur les processus ponctuels, présentés dans l'ouvrage *Modélisation et statistique spatiales* [34], ou sur le couplage de ces deux méthodes dont une présentation est faite par Van Lieshout dans *Markov point processes and their applications* [98].

La génération de données spatiales virtuelles, c'est-à-dire la reconstitution de données, peut s'appuyer, quant à elle, sur des outils statistiques généraux présentés dans *Geostatistical simulation : models and algorithms* [55], sur des méthodes géométriques stochastiques, présentées par Stoyan dans *Stochastic geometry and its applications* [91] ou géométriques discrètes dont un état de l'art se trouve dans *Géométrie discrète et images numériques* [21].

Le paysage réel est une source d'information spatiale qui peut se présenter sous différentes formes et provenir de différentes sources de données. Plusieurs méthodes s'appliquant à l'information spatiale sont particulièrement adaptées à l'étude des paysages et seront présentées ci-après. Il s'agit des méthodes statistiques pour l'étude du voisinage (section 3.1), mais aussi des méthodes de pavage (Voronoï, STIT, T-tesselations) (sous-section 3.1.3) ou encore des méthodes permettant la linéarisation de l'espace (section 3.2) auxquelles se couplent les modèles de Markov (section 3.3).

3.1 Méthodes statistiques et géométriques pour l'analyse et la génération d'information spatiale

Comment caractériser l'information spatiale, et les segments en particuliers ? Les chercheurs ayant voulu répondre à cette question travaillent surtout dans le domaine de l'analyse d'images. Quelques-uns ont choisi d'analyser les segments mais toujours dans le contexte d'un ensemble connexe donc, d'un réseau. Les méthodes développées dans ces deux cas présentent des axes de réflexion intéressants pour notre étude. La prise en compte de la proximité entre objets spatiaux est un point important. Nous nous intéresserons aussi aux techniques de pavage de l'espace.

3.1.1 Méthodes de caractérisation de données spatiales

Nous abordons dans cette sous-section les approches de type statistiques permettant d'extraire, classer et générer des structures spatiales, que ce soit à partir d'images (satellitaires, photographies) ou de données SIG.

Pour mettre en place le cadre de recherche de la section 5.1, nous avons supposé une liaison forte entre les caractéristiques des éléments du paysage (par exemple : leur taille ou leur orientation) et la configuration paysagère (c'est-à-dire leur distribution spatiale et celles d'autres éléments paysagers). Cette hypothèse fut également posée par d'autres auteurs. Nous pouvons citer Bhattacharya et *al* [11] dont les travaux s'intéressent aux corrélations qui peuvent exister entre les différents types d'objets sur une image satellite. Les auteurs font l'hypothèse que les caractéristiques d'un objet présent sur la plupart des images permettent de classer les images avec une bonne robustesse face au bruit. Ils démontrent que l'utilisation des caractéristiques du réseau routier le permet. En effet, les propriétés du réseau varient considérablement suivant l'environnement géographique. Les auteurs constituent une base de données à partir des images à leur disposition sur les caractéristiques géométriques et topologiques calculées sur le réseau routier dans chaque image. Ainsi, pour toute nouvelle image possédant un réseau routier, il sera possible de calculer les caractéristiques géométriques et topologiques de ce réseau et d'utiliser la base de données créée précédemment pour classer la nouvelle image en fonction des anciennes. Les caractéristiques statistiques considérées sont multiples : le nombre de nœuds, la longueur du réseau, le nombre d'arcs, la taille de l'image...

Dans le travail de recherche que nous avons développé, les données choisies à partir des données initiales furent segmentées et classées de manière automatique, et ceci à

différentes échelles. Cette approche se retrouve dans les travaux de Vannier et Hubert-Moy [100, 99] qui portent également sur des linéaires. Dans ces articles, les auteurs présentent une méthode d'analyse d'image à base d'objets afin d'être capable d'extraire l'information sur la position des haies dans les paysages agricoles. La première étape s'effectue de manière automatique grâce à un algorithme de segmentation multi-échelle et dans ce cas, il existe deux échelles, une au niveau du champ et l'autre au niveau de l'arbre. Dans la seconde étape, les auteurs classent les éléments de la seconde échelle suivant de multiples critères, principalement spatiaux. La dernière étape s'appuie sur des indicateurs résultant d'une analyse statistique et spatiale des arrangements spatiaux, de la morphologie et de la composition d'un petit élément boisé à la frontière des champs.

Nous travaillons sur la caractérisation des éléments du paysage, non pas dans le but de les classer, ou de les extraire d'une image, mais plutôt pour les générer par la suite. Cependant, nous nous sommes inspirés des méthodes d'extraction d'objets qui utilisent des caractéristiques des éléments du paysage pour l'extraction de ces mêmes éléments à partir d'une photo ou d'une image satellite. Nous avons d'ailleurs choisi de créer un indice pour l'étude du voisinage des segments dans le plan. Il sert à caractériser la zone d'étude. Il est également utilisé lors de la phase de post-traitement. Ceci est proche des travaux de Lacoste *et al.* [53]. Dans cet article, les auteurs présentent une méthode pour extraire les réseaux linéaires des images satellites afin de produire des cartes ou de les mettre à jour. Cette méthode est une extension du *Candy* modèle, appelé le modèle *Candy de qualité*, qui utilise des coefficients de qualité en ce qui concerne les interactions entre les segments pour mieux analyser la courbure, les jonctions et les intersections du réseau. Pour calculer ces coefficients, les auteurs construisent un nouvel indice sur les données qui incorpore des propriétés de celles-ci. Ces propriétés sont définies à partir de mesures statistiques. Le calcul des coefficients prend également en compte l'homogénéité locale et le contraste avec le fond à proximité du réseau de lignes. Les auteurs mettent en œuvre une technique d'étalonnage pour choisir les paramètres du modèle afin que le réseau de lignes extrait possède de bonnes propriétés (pas de segments libres, pas de redondance, pas de trous, etc). Dans une autre approche, Lafarge [54] présente un processus ponctuel multi-marqué. Celui-ci est utilisé pour extraire des objets d'une image satellite ou traiter la représentation d'une texture naturelle. Les objets sont extraits à partir d'une librairie géométrique pré-construite, et en utilisant des modèles probabilistes (modèle de Gibbs - Processus Jump-diffusion) pour choisir l'objet optimal.

Descombes et Zhizhina [27] s'intéressent à l'extraction de segments sur une image

de paysage. Ils présentent un aperçu de l'utilisation des champs de Gibbs et de ses variantes dans le traitement de l'image. Plus précisément, ils présentent dans le chapitre 5 de leur publication, les processus ponctuels spatiaux utilisés pour détecter les objets dans les images. Il s'agit, par exemple, des processus ponctuels marqués couplés au saut réversible MCMC ; ou du modèle d'interaction (CANDY modèle) pour l'extraction du réseau routier, comme présenté par Stoica dans [90]. Dans ce dernier cas, les objets considérés sont des segments qui sont déterminés par un point (isobarycentre du segment) et des marques (orientation et longueur). La connaissance a priori est la grande connectivité et la faible courbure du réseau de route. Chaque segment a deux zones attractives (les bouts) et une zone répulsive (le milieu). Si un segment intersecte une zone attractive d'un autre segment, il y a une interaction attractive avec augmentation de l'intensité lorsque la différence d'orientation décroit. Si un segment intersecte une zone répulsive, il y a une interaction répulsive avec une intensité décroissante quand la différence d'orientation tend vers $\frac{\pi}{2}$. De plus, une fonction répulsive existe quand les segments sont connectés plus de deux fois.

Nous remarquerons que le formalisme du processus de Gibbs a déjà été appliqué avec succès comme un processus basé sur des points pour la génération de population d'individus dans la foresterie et en biologie. Il a été appliqué pour organiser des arbres dans les terrains boisés ou pour simuler des groupes d'animaux [93, 38]. Il permet de déterminer, par exemple, une position réaliste pour chaque arbre dans un terrain boisé, en utilisant des expressions de "attraction / répulsion" entre les positions des individus [93, 26, 92]. Gaucherel et *al* [36] adaptent cette méthode pour générer des paysages en remplaçant les individus par les points centraux des unités de paysage (ou *patches*)

Nous pouvons citer également les travaux de Bailly et *al.* [4] et Monestiez et *al.* [73] pour illustrer les travaux existants dans le cadre d'études de réseaux linéaires dans le paysage. Les données, ici, ne sont pas des photos ou des images mais des données sous SIG de réseaux de fossés. Dans ces articles, les auteurs présentent une méthode pour modéliser des réseaux linéaires observés. Ils proposent une approche pour simuler la distribution spatiale des caractéristiques locales le long d'un réseau (de drainage). Cette approche tient compte de la spécificité du réseau considéré comme un arbre et modélise les caractéristiques non seulement sur les nœuds de réseau, mais également à n'importe quel endroit le long de la section linéaire entre les nœuds. Contrairement à des approches markoviennes [40] ou à des modèles de réseau [22] qui mettent l'accent sur ce qui se passe seulement au niveau des nœuds, la géostatistique qui traite de fonctions aléatoires sur un support continu est un cadre théorique plus approprié. Cependant, dans notre cas, les linéaires ne sont pas considérés comme un réseau mais comme un

ensemble de segments non connectés.

Toutefois, notre travail ne s'attache pas à l'extraction ou à la reconnaissance de formes mais s'inscrit dans la continuité de ceux-ci. Nous analysons les linéaires agricoles afin de les générer par la suite. Les résultats issus de la reconnaissance des linéaires agricoles ne sont que peu utilisés dans cet objectif, et les utilisations qui en sont faites sont récentes et ne concernent pour la plupart que l'aspect agronomique de ces linéaires (cf. chapitre 2).

3.1.2 Prise en compte de la proximité spatiale

Nous étudierons ainsi les différentes approches permettant d'évaluer la proximité entre deux objets spatiaux en commençant par un rappel sur quelques définitions.

Rappel

Soit E un ensemble, une ***distance*** sur E est une application d définie sur le produit $E \times E$ et à valeur dans l'ensemble $\mathbb{R}$,

$$d : E \times E \to \mathbb{R}$$

vérifiant les propriétés suivantes,

— la **symétrie** : $\forall (a, b) \in E \times E, d(a, b) = d(b, a)$

— la **séparation** : $\forall (a, b) \in E \times E, d(a, b) = 0 \Leftrightarrow a = b$

— l'**inégalité triangulaire** : $\forall (a, b, c) \in E \times E \times E, d(a, c) \leqslant d(a, b) + d(b, c)$

Dans le cas de distance sur $\mathbb{R}^n$, la distance la plus répandue et la plus connue est sans nul doute la distance euclidienne, introduite par Euclide dans la Grèce Antique, et qui a la définition suivante :

Définition : Soient $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ et $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, alors la ***distance euclidienne*** d sur $\mathbb{R}^n$ est donnée par

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Cette définition reflète l'idée physique que nous nous faisons d'une distance point à point, mais comme nous pouvons le voir sur le schéma 3.1, ce n'est plus le cas lorsque les objets ne sont plus ponctuels. Pour la distance de Hausdorff, distance usuelle pour l'étude entre des parties compactes d'un espace métrique, nous perdons la notion de proximité géographique avec la propriété de **séparation**.

Définition : Soient (E, δ) un espace métrique et E_H l'ensemble des fermés bornés de E non vides. La ***distance de Hausdorff*** d de E_H est l'application de $E_H \times E_H$ dans $\mathbb{R}$ définie par :

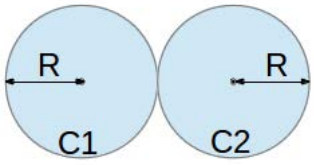
$$\forall (X,Y) \in E_H \times E_H \ , \ d(X,Y) = \max \{ \sup_{y \in Y} \inf_{x \in X} \delta(x,y), \ \sup_{x \in X} \inf_{y \in Y} \delta(x,y) \}$$

Il ne faut pas confondre les distances d et δ car elles ne s'appliquent pas sur les mêmes éléments. Par exemple, la distance entre le vecteur nul et la boule unité fermée $\mathcal{B}$ est égale à 0 pour δ, mais est égale à 1 pour la distance de Hausdorff.

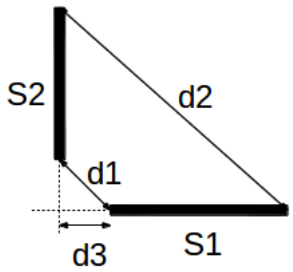
$$\delta(0, \mathcal{B}) = 0 \quad \text{et} \quad d(\{0\}, \mathcal{B}) = 1$$

Si la distance sur E est bornée, la distance de Hausdorff peut même être étendue à l'ensemble des sous-espaces fermés (non nécessairement compacts) de E. Dans le cas contraire, la *distance* ainsi définie peut prendre des valeurs infinies.

Pseudo-distance



(a) Dans le cas de cercles



(b) Dans le cas de segments

FIGURE 3.1 – Exemples pour la comparaison visuelle des différentes distances de $\mathbb{R}^2$

Les distances strictes, comme vues précédemment, présentent l'inconvénient de ne pas traduire exactement la notion de proximité géographique des éléments non-ponctuels. Cependant, lorsque nous étudions des données spatiales, il peut être intéressant de garder cette notion en sacrifiant la propriété de **séparation**. Nous avons donc fait l'hypothèse que la distance euclidienne n'était pas adaptée à l'étude des voisinages, et nous avons choisi de l'adapter afin d'avoir une distance plus cohérente avec la réalité géométrique que nous souhaitions analyser. C'est pourquoi, nous avons défini une pseudo distance permettant d'étudier les relations de voisinage. Cette hypothèse de faiblesse de la distance euclidienne pour l'étude des voisinages géographiques n'est

pas inconnue dans la littérature. En effet, Huang et Kennedy [45], dans leurs travaux, présentent une analyse des relations spatiales de voisinage en utilisant des modèles de Markov cachés et ils avancent que ce type de relation ne peut pas s'étudier avec la distance euclidienne. Dans leur approche, ils découpent le paysage en cellules carrées et utilisent un paramètre connu ou calculable sur l'ensemble du paysage afin d'étudier les relations de voisinage entre cellules adjacentes.

La pseudo-distance entre ensembles en allégeant la propriété de *séparation* peut se définir comme ceci :

Définition : Soient E_1 et E_2 deux parties non vides d'un espace métrique E muni d'une distance d, nous définissons la pseudo-distance entre ces deux ensembles, notée $dist$ comme :

$$dist(E_1, E_2) = \inf\{d(x, y) \mid \forall (x, y) \in E_1 \times E_2\}$$

Cette pseudo-distance est à valeur dans $\mathbb{R}^+$ comme borne inférieure d'un ensemble non vide de réels positifs. Il s'agit bien d'une pseudo-distance car, si la distance entre deux ensembles est nulle, nous ne pouvons pas en déduire que ces ensembles sont égaux.

Sur la figure 3.1, nous présentons deux exemples pour le calcul des différentes distances entre deux objets de $\mathbb{R}^2$. Dans le cas où il s'agit de objets bi-dimensionnels, par exemple des cercles (figure 3.1(a)), dans $\mathbb{R}^2$, nous avons :

$$\text{Dist. de Hausdorff } = 4 \times R \text{ , Dist. Centre à Centre } = 2 \times R \text{ , Pseudo-Distance } = 0$$

Maintenant, s'il s'agit d'objets uni-dimensionnels, par exemple des segments (figure 3.1(b)), dans $\mathbb{R}^2$, nous avons :

$$\text{Distance de Hausdorff } = d2 \text{ , Distance euclidienne } = d3 \text{ , Pseudo-Distance } = d1$$

3.1.3 Méthodes de pavage

L'idée d'utiliser des pavages pour caractériser l'espace se généralise pour générer des objets (paysages, surfaces, textures, structures) en deux dimensions. Weiss, Maier ou Nagel [65] [76] [75] présentent dans leurs travaux l'évolution d'un modèle de tesselations aléatoire stationnaire ; c'est-à-dire dont la distribution est invariante par translation (jusqu'à la forme STIT, c'est-à-dire stable (par distribution) par rapport aux itérations de découpe). Le modèle de tesselations aléatoire se base sur trois processus ponctuels planaires, un processus de nœuds, un processus des centres des arcs, et un processus des

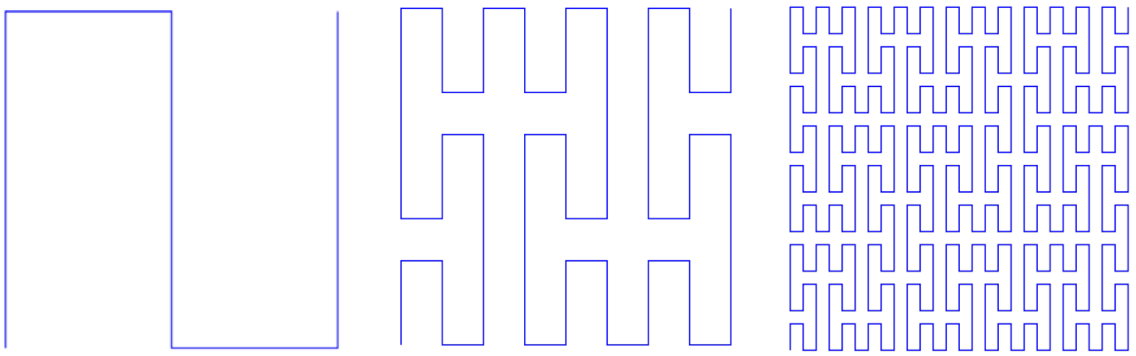
centres de cellules. L'intensité de chaque processus est respectivement le nombre moyen de nœuds, le nombre moyen de centres des arcs et le nombre moyen de centres de cellules par unité d'aire. Ces processus sont dépendants, ils permettent de définir un processus spatial sur les arcs de la tesselation aléatoire, comme un processus aléatoire de segment (un segment étant un arc entre deux nœuds, sans nœuds intérieurs). Ce processus admet deux distributions pour les directions des segments. D'autres méthodes de pavages ont été utilisées dans l'étude du paysage agricole. Ces méthodes sont utilisées, par exemple, dans le logiciel Genexp-LandSiTes [1] [60]. Dans ce contexte, les auteurs utilisent deux formes de pavage pour générer le parcellaire agricole, le pavage de Voronoï et un pavage rectangulaire. Cette approche a été approfondie dans les travaux développés par Kieu et *al.* [51]. Ils ont mis en œuvre des techniques pour construire des pavages qui ont des caractéristiques pouvant être reliées à des caractéristiques de paysages agricoles (par exemple, des angles pas trop aigus, des parcelles de tailles raisonnables,...)

Les techniques de pavage s'appliquent bien à l'analyse et à la génération de parcellaires agricoles mais plus difficilement dans le cas des linéaires qui nous préoccupe.

3.2 Courbe remplissant l'espace

Il existe de nombreuses techniques de fouille de données s'appliquant aux informations séquencées. C'est le cas pour les observations de séries temporelles telles que dans la reconnaissance de la parole ou de l'écriture [85], puisque la séquence est inhérente au processus d'observation. La définition des séquences d'observation avec des données spatiales peut être un défi, en particulier pour les observations d'objets au positionnement aléatoire. Les pratiques courantes sont de constituer un échantillonnage régulier de l'espace et un partitionnement de celui-ci. Par exemple, Zhang et Li [104] définissent les chaînes de Markov en échantillonnant le type de couverture terrestre avec un réseau régulièrement espacé sur une carte vectorielle. Elfeki et Dekking [29] découpent des sections géologiques régulières de 200 km de long et 50 m de profondeur dans le sous-sol et couplent deux chaînes de Markov pour intégrer les probabilités de transition horizontale et verticale. Tjelmeland et Besag [94] proposent un réseau hexagonal pour déterminer le voisinage de premier et de deuxième ordre dans l'étude de champs de Markov. Lovell [64] définit des espaces de recherches discrets circulaires et concentriques dans les images pour la reconnaissance des caractéristiques ; Mari et le Ber [68] présentent une méthode d'échantillonnage plus complexe, la courbe fractale de Hilbert- Peano pour l'échantillonnage des images satellites, dans une étude spatiale

1. `http://engees.unistra.fr/~fleber/Landsites/`



(a) Première itération (b) Deuxième itération (c) Troisième itération

FIGURE 3.2 – Premières itérations pour la construction d'une courbe de Peano

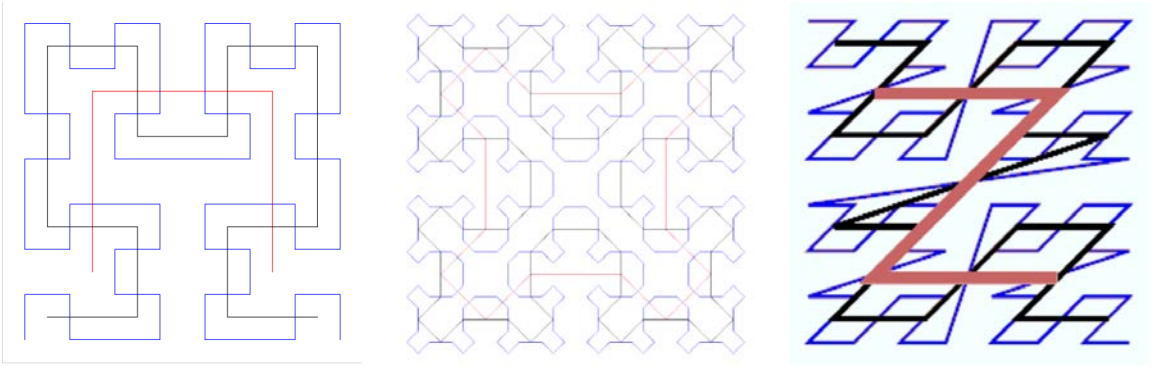
de changement d'utilisation des terres agricoles.

Nous avons utilisé cette dernière approche afin d'ordonner de manière unidimensionnelle des segments (objets d'études) pour obtenir un ordre global sur l'ensemble de ceux-ci dans une zone d'étude. Cet ordonnancement 1D s'apparente donc à une linéarisation des éléments et permet de les représenter le long d'une courbe. De plus, si les segments (objets d'études) proches selon cet ordre 1D sont également proches dans la zone de départ, alors cet ordre respecte, en partie, la localité.

3.2.1 Présentation

L'existence, et donc la création, de courbes remplissant l'espace est directement issue de l'existence d'une bijection entre une droite réelle et une surface. Cette existence est induite par un résultat de Georg Cantor [17], qui établit que l'ensemble des points de l'intervalle unité et celui d'une surface bidimensionnelle finie avaient le même cardinal. Le premier à avoir construit une telle courbe est Giuseppe Peano [81]. Cette construction fut analytique lors de sa première présentation mais, pour notre part, nous l'aborderons de façon géométrique.

Peano proposa de construire une courbe remplissant l'espace en utilisant un motif (figure 3.2(a)) et de le répéter de façon à parcourir le carré unité. Tout d'abord, à l'étape 1, le carré unité est divisé en 9 carrés identiques et le motif présenté à la figure 3.2(a) permet de les parcourir. À l'étape 2, le carré unité est divisé en 81 cases identiques et le motif présenté à la figure 3.2(b) permet de les parcourir. Ainsi, à l'étape $n+1$, le carré unité est divisé en $3^{2*(n+1)}$ carrés identiques et la courbe de Peano, construite à partir de la courbe de Peano de l'étape n, permet de les parcourir. Cette construction se fait en remplaçant chaque segment, issu de la courbe de Peano de l'étape n et inclus



(a) Hilbert (b) Sierpinski (c) Lebesgue

FIGURE 3.3 – Les trois premières itérations pour la construction d'une courbe remplissant l'espace à partir de trois motifs différents

dans une *division*, ou *case*, par un motif de Peano (Figure 3.2(a)) de taille adéquat. Ce dernier peut éventuellement être orienté différemment ou subir une réflexion.

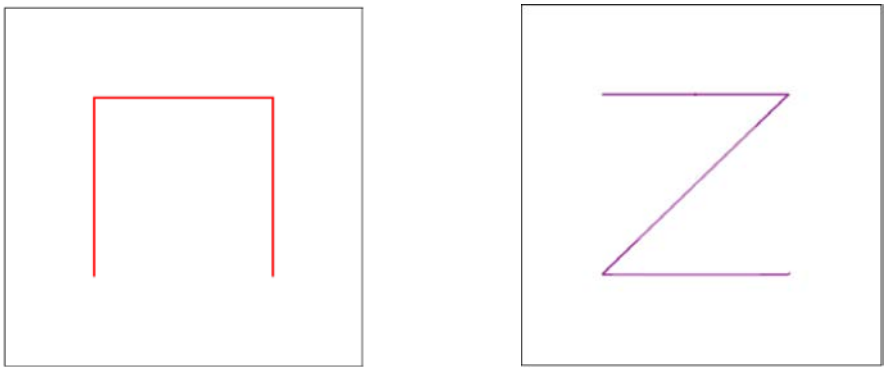
D'une manière générale, les courbes remplissant l'espace sont appelées *courbes de Peano*, et peuvent être définies comme des courbes planes paramétrées par une fonction continue sur l'intervalle $[0, 1]$, surjective dans le carré $[0, 1] \times [0, 1]$. Ces courbes sont rangées dans la catégorie des fractales.

Le motif de départ d'une courbe remplissant l'espace dépend notamment du nombre de subdivisions que subit le carré unité à l'étape 1. Ainsi, pour une subdivision en 4, les courbes sont nommées *courbes de Peano binaires*, pour une subdivision en 9, il s'agit de *courbes de Peano ternaires*. Mais il existe également des motifs adaptés à une subdivision en triangles.

Depuis 1890, de nombreux motifs furent présentés pour construire les courbes remplissant l'espace. Tout d'abord, David Hilbert [43] en 1891, présenta un nouveau motif dont nous pouvons voir les trois premières itérations à la figure 3.3(a). Vint par la suite, Moore en 1900 [74], Lebesgue en 1905 [61], ou Sierpinski en 1912 [89] dont le motif et les premières itérations sont présentés à la figure 3.3(b), ces motifs, et d'autres, peuvent être retrouvés dans l'ouvrage de Hans Sagan [87].

Les courbes remplissant l'espace sont utilisées pour linéariser l'espace, et sont couramment employées dans le domaine du traitement de l'image, ou de l'analyse multi-résolution [102, 80, 79, 31].

Les motifs de la courbe de Hilbert (figure 3.4(a)), ou de la courbe de Lebesgue (figure 3.4(b)) sont simples contrairement au motif de la courbe de Sierpinski (figure 3.3(b)), ces premiers sont donc plus faciles à implanter. De plus, la courbe de Lebesgue présente l'avantage de ne faire subir aucune transformation au motif de base lors des différentes



(a) Hilbert (b) Lebesgue

FIGURE 3.4 – Motifs pour la construction d'une courbe remplissant l'espace

étapes de construction. Cependant, elle présente des *sauts* dans sa linéarisation (figure 3.3(c)) c'est-à-dire, qu'il existe des lieux proches sur la courbe qui ne le sont pas dans l'espace initial. La courbe de Lebesgue ne respectant pas la propriété de localité, nous utiliserons le motif de Hilbert (figure 3.4(a)) pour construire les courbes remplissant l'espace.

3.2.2 Courbe de Hilbert

La courbe de Hilbert est basée sur un motif simple, présenté sur la figure 3.4(a), mais elle peut aussi être construite avec trois autres motifs construits par rotation ($\frac{\pi}{2}$, π et $\frac{3\pi}{2}$) de ce dernier. Lors de la construction de la courbe de Hilbert, ce motif subit des rotations et des symétries qui diminuent les sauts de linéarité et préservent ainsi la propriété de localité. La courbe de Hilbert construite ainsi est notamment utilisée pour le partitionnement de domaines d'études [47, 46, 88]. Cette courbe est construite récursivement, comme le montre la figure 3.3(a). C'est cette construction récursive, couplée aux diverses rotations et symétries qui donne à la courbe ses propriétés de localité.

3.2.3 Chemin de Hilbert adaptatif

La courbe de Hilbert permet d'utiliser des méthodes linéaires afin d'analyser un phénomène spatial. Toutefois, si ce phénomène spatial présente des parties vides étendues sur la zone d'étude et d'autres parties plus denses, la courbe créée peut résulter d'une itération très grande et ainsi, présenter une longue portion sans information sur le phénomène étudié. C'est pourquoi, nous présentons maintenant, une construction adaptée

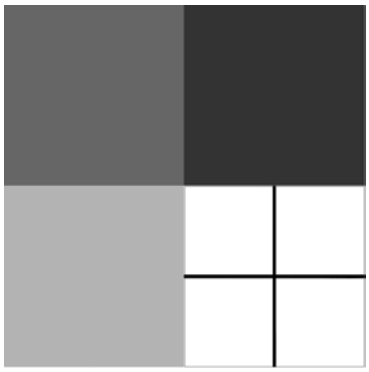


FIGURE 3.5 – Exemples de cases créées par division d'une cellule carrée

de la courbe, connue en tant que *Chemin de Hilbert Adaptatif*, en nous inspirant des travaux de Quinqueton et Berthod [84] pour la mise en place de sa construction. Par la suite, nous emploierons indifféremment les termes *courbe de Hilbert adaptative*, *chemin de Hilbert adaptatif* et *CHA*.

Avant de présenter le processus de fabrication d'un tel chemin, nous posons d'abord la définition des termes d'étude. Tout d'abord, nous appelons *cellule* le carré initial sur lequel le phénomène spatial doit être étudié. Les réalisations du phénomène spatial sur une cellule sont appelées *objets* (par exemple : point, mesure, etc). Les divisions carrées uniques de la cellule sont des *cases*. La cellule est une case si elle n'a pas été divisée. La *division* d'une case produit exactement quatre cases, elle s'opère en prenant la médiatrice de chaque côté du carré formant la case. Dans l'exemple de la figure 3.5, les carrés en gris sont des cases, mais le grand carré blanc n'est pas une case car il a été divisé au moins une fois.

Présentation

Le chemin de Hilbert adaptatif est une variante de la courbe de Hilbert qui permet une réduction de dimension des données tout en contrôlant le nombre de cases créées dans le plan.

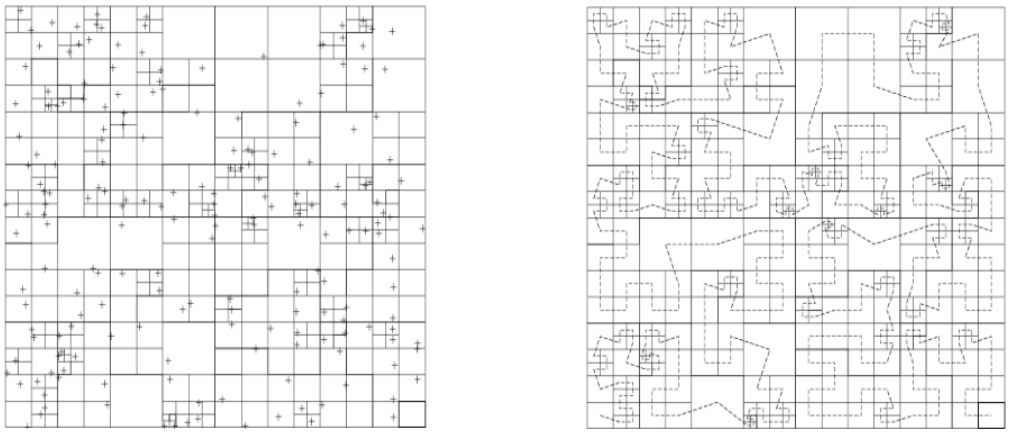
Pour présenter la méthode de création d'un chemin de Hilbert adaptatif, nous supposons que le domaine d'étude est une cellule carré contenant plus d'un objet, notée $\mathcal{C}_0$ dans un plan de $\mathbb{R}^2$. Nous définissons le chemin de Hilbert adaptatif obtenu comme un ensemble ordonné de cases, noté $CHA\{\mathcal{C}_0\} = \{\kappa_i\}_{i\in\mathbb{N}}$. L'ensemble est ordonné suivant i, et permet de faire correspondre les cases avec leur place sur le chemin.

Le processus se déroule comme suit : tout d'abord, la cellule est définie comme l'unique case du chemin de Hilbert adaptatif initial (ligne 1, algorithme 1). Ensuite,

Algorithm 1: Création du chemin de Hilbert adaptatif

Data: $\mathcal{C}_0$: Cellule
Result: $\{\kappa_i\}_{i\in\mathbb{N}}$: Ensemble ordonné de cases
begin
1 $CHA\{\mathcal{C}_0\} = \{\kappa_0\} = \mathcal{C}_0$
2 **while** $\{\ \exists\ \kappa_k \in CHA\{\mathcal{C}_0\}$ *qui contient plus d'un objet* $\}$ **do**
3 $\kappa_k = (\kappa_{k_1}, \kappa_{k_2}, \kappa_{k_3}, \kappa_{k_4})$
4 $CHA\{\mathcal{C}_0\} = \{\kappa_0, ..., \kappa_{k-1}, \kappa_{k_1}, \kappa_{k_2}, \kappa_{k_3}, \kappa_{k_4}, \kappa_{k+1}, ..., \kappa_n\}$
5 **return** $CHA\{\mathcal{C}_0\} = \{\kappa_i\}_{i\in\mathbb{N}}$

tant qu'il existe une case contenant plus d'un objet spatial dans le chemin de Hilbert adaptatif (ligne 2, algorithme 1), celle-ci est divisée en quatre cases (3, algorithme 1) et les cases obtenues sont numérotées pour s'insérer logiquement dans le chemin (ligne 4, algorithme 1). À la fin, le chemin de Hilbert adaptatif créé à partir de $\mathcal{C}_0$ et du phénomène spatial est l'ensemble ordonné des cases construites $\{\kappa_i\}_{i\in\mathbb{N}}$ (ligne 5, algorithme 1). Finalement, l'algorithme a permis de construire le chemin de Hilbert adaptatif de la cellule initiale en prenant en compte les zones vides d'objets.



(a) Division en cases de la cellule

(b) Parcours des cases issues de la cellule

FIGURE 3.6 – Exemple de construction de chemin de Hilbert adaptatif dans une cellule, d'après un nuage de point.

Sur la figure 3.6(a), un exemple de découpe d'une cellule suivant un nuage de point est présenté. Nous pouvons remarquer la disparité de taille dans les cases créées,

confirmant l'adaptation de la division à la densité du phénomène observé. Le parcours de ce découpage est présenté sur la figure 3.6(b). Le point de départ du chemin est situé dans la case en bas à gauche et le point d'arrivée dans la case en bas à droite.

3.3 Méthodes de Markov

Pour analyser et générer l'information spatiale, les modèles de Markov (cachés ou non) ont également été largement utilisés. Ils présentent de réels avantages, comme leur facilité de mise en place, et la possibilité d'utiliser les mêmes modèles pour la phase d'apprentissage et celle de génération.

3.3.1 Rappels

Soit S_i, $i = 0, ..., t$, une suite de variables aléatoires. Nous notons la suite de réalisation $S_0 = s_0, S_1 = s_1, ..., S_t = s_t$ par $S_0^t = s_0^t$. Une chaîne de Markov d'ordre 1 est un processus stochastique à temps discret et à espace d'états discrets caractérisé par la relation de dépendance :

$$P(S_t = s_t | S_0^{t-1} = s_0^{t-1}) = P(S_t = s_t | S_{t-1} = s_{t-1})$$

Ce qui peut se traduire par : tout le passé du processus est résumé dans l'état précédent. La loi jointe des variables aléatoires $S_0, S_1, ..., S_t$ peut alors se factoriser de la manière suivante.

$$\begin{aligned} P(S_0^t = s_0^t) &= P(S_t = s_t | S_0^{t-1} = s_0^{t-1}) P(S_0^{t-1} = s_0^{t-1}) \\ &= (\prod_{\nu=0}^{t-1} P(S_{t-\nu} = s_{t-\nu} | S_{t-\nu-1} = s_{t-\nu-1})) P(S_0 = s_0) \end{aligned}$$

On appelle probabilité de transition la probabilité suivante :

$$P(S_t = j | S_{t-1} = i)$$

Dans le cas où, pour tout t,

$$P(S_t = j | S_{t-1} = i) = p_{ij}$$

on parle de chaîne de Markov homogène dans le temps. Ainsi une chaîne de Markov à J états d'ordre 1, homogène dans le temps, est définie par les paramètres suivants :

— Probabilités initiales $\pi_j = P(S_0 = j)$, $j = 0, ..., J-1$ Les probabilités initiales constituent une loi de probabilité.

— Probabilités de transition p_{ij}

On peut construire de la même façon une chaîne de Markov d'ordre r à partir de la relation de dépendance suivante :

$$P(S_t = s_t | S_0^{t-1} = s_0^{t-1}) = P(S_t = s_t | S_{t-r}^{t-1} = s_{t-r}^{t-1})$$

L'état du processus à l'instant t ne dépend donc que des r états précédents.

Le caractère discret de l'espace d'état peut être un obstacle pour traiter certains problèmes. Pour cela, il existe les processus de Markov à temps discret et à espace d'états continus. Dans ce cas, les probabilités de transition dans la propriété markovienne sont remplacées par des densités de probabilité.

3.3.2 Utilisation des modèles de Markov

De nombreuses variantes de Modèles de Markov apparaissent dans la littérature. Mais dans tous les cas, leur utilisation intervient dès lors que nous avons la volonté de révéler un lien entre l'estimation d'états et la segmentation de séries chronologiques (d'après Kehagias [49]).

Les modèles de Markov sont utilisés dans des domaines variés tels que la génétique, la robotique, l'analyse de texte ou la reconnaissance de parole [45, 85]. Par exemple, la première application importante des modèles de Markov cachés fut en reconnaissance de la parole, dont le but était de diviser un signal de parole en segments, chaque segment correspondant à un *phonème*. Les premiers articles [5, 6] traitent des séries chronologiques discrètes ; des extensions aux séries chronologiques continues ont également été produites par la suite[48, 50, 8]. L'utilisation des modèles de Markov cachés dans le cas de données spatiales et temporelles s'avère plus courante car ils permettent de mieux capter les états mixtes en temps, et en espace. Les modèles de Markov cachés sont massivement utilisés notamment en reconnaissance de formes [7], en intelligence artificielle [13],en science du vivant [1, 69] ou encore en traitement automatique du langage naturel [85]. L'utilisation des modèles de Markov s'est également beaucoup répandue dans le domaine de la segmentation d'images, et donc, sur une forme de données spatiales. Nous pouvons citer le travail de Pieczynski [83] qui présente une compilation de méthodes répandues dans ce domaine.

Les modèles de Markov présentent l'avantage de pouvoir se coupler facilement avec d'autres méthodes. Ainsi, nous pouvons citer l'exemple de Kluszczynski [52]. Dans cet article, les auteurs présentent l'utilisation des champs de Markov polygonaux multi couleurs pour la segmentation des images. Cet outil est construit à partir des méthodes

de Arak-Clifford-Surgailis, qui sont bien adaptées pour la segmentation. Les auteurs présentent la segmentation de l'image comme un problème d'estimation statistique par modification des champs de Markov polygonaux sous-jacent. Des techniques de Monte Carlo, des méthodes de Gibbs et le modèle d'Arak sont utilisés pour estimer les paramètres du modèle et trouver la partition optimale de l'image.

Les modèles de Markov peuvent également être couplés avec l'utilisation des courbes de type Peano, comme nous l'avons évoqué dans la partie précédente. Benmiloud [9] a travaillé sur le couplage de ces deux méthodes pour la segmentation statistique non supervisée d'images. Dans cet article, les auteurs cherchent à estimer des paramètres dans les chaînes de Markov cachées, afin d'étudier la segmentation statistique non supervisée d'images. Ils proposent deux algorithmes d'estimation basés sur les algorithmes Iterative Conditional Estimation (ICE) et Stochastic Expectation Maximisation (SEM). Ces algorithmes sont utilisés dans le cas de segmentations d'images grâce à l'utilisation des courbes de Peano afin de transformer les processus bi-dimensionnels en processus uni-dimensionnels. D'après Benboudjema [7], les modèles de Markov cachés sont très utilisés dans l'analyse et la segmentation d'images car ils sont faciles à mettre en place et ont un potentiel de capture des éléments sur les images de qualité. Mais les méthodes couplées sont plus rapides que celles utilisant des modélisations par champs de Markov cachés et la perte de l'efficacité est acceptable, selon Benmiloud [9]. Les nouveaux algorithmes qu'ils proposent, permettent de concevoir de nombreux algorithmes de segmentation non supervisée spatio-temporelle d'images. Les modèles de Markov sont aussi utilisés pour l'analyse des données agronomiques. Nous pouvons citer deux solutions pour l'analyse de données d'utilisation des terres agricoles par les agronomes. Il s'agit des systèmes CarrotAge [58] et son évolution, ARPEnTAge [67]. Tous deux sont des boîtes à outils d'exploration de données pour l'exploitation des données temporelles. Un volet spatial est aussi présent dans ARPEnTAge.

CarrotAge est un système d'extraction de connaissances basé sur des modèles de Markov cachés de grand ordre, il est utilisé pour la fouille et l'analyse des données d'utilisation de sol. Le Ber et *al* [58], dans cette étude, adoptent l'approche présentée par Fayyad [30] sur l'extraction de connaissances (extraire les objets réguliers suivant des critères prédéfinis sans connaître complètement les connaissances à trouver). Pour extraire les motifs, les auteurs utilisent un modèle de Markov cachés de second ordre. Ils partent de l'hypothèse que l'utilisation des terres au temps t dépend de celle aux temps *t-1* et *t-2*. Les états du modèle de Markov cachés permettent de capturer un comportement stationnaire, ils représentent une classe (*i.e.* une culture ou un motif cultural).

Dans le cas d'une utilisation agronomique des modèles de Markov, nous pouvons citer les travaux de Aurbacher et Dabbert [3]. Ici, les modèles de Markov sont utilisés pour générer des utilisations de sols. Ils utilisent des chaînes de Markov de premier ordre. Dans ce cas, la matrice de transition est la probabilité de passer d'un état de culture à un autre, calculée à partir des données. Ils utilisent cette même matrice pour générer des nouveaux états de cultures. Cette utilisation à la fois pour la modélisation et pour la génération est une force pour les modèles de Markov. C'est pour cela que nous avons mis en place, nous aussi, une méthode d'apprentissage basée sur les chaînes de Markov. Une approche similaire est décrite dans [18].

Nous utiliserons également les modèles de Markov. Cependant, dans notre cas, nous n'utiliserons pas de modèles cachés. De plus, dans notre étude, les données ne présentent pas de dimension temporelle, seulement spatiale.

Deuxième partie

Pré-traitements et fouille de données spatiales

Chapitre 4

Données et pré-traitements

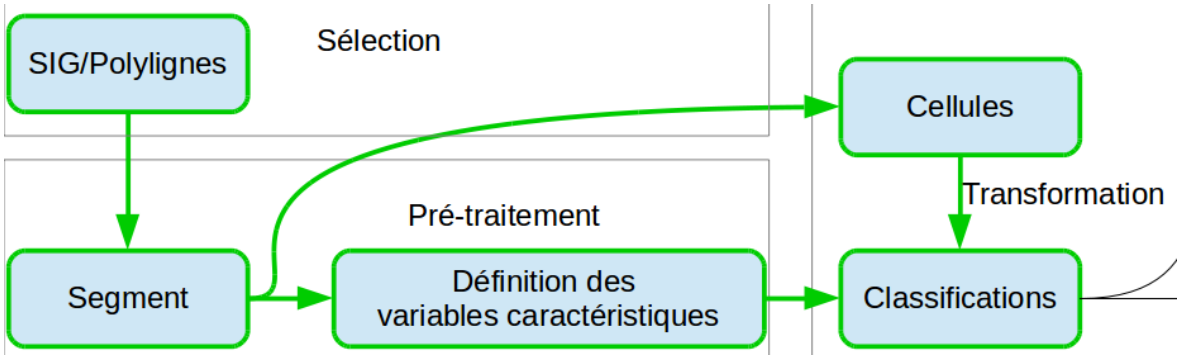


FIGURE 4.1 – Étapes décrites au chapitre 4

Introduction

Nous travaillons sur deux jeux de données, issus de deux territoires français distincts. Le premier, (données A), est fourni par l'équipe Écologie de la Production Intégrée (EPI) de l'unité PSH (Plantes et Systèmes de culture Horticoles) à l'INRA PACA, antenne d'Avignon. Ces données sont à notre disposition dans le cadre des études menées sur le rôle des haies dans la dynamique des populations de ravageurs. Le second (données B) nous est fourni par l'UMR COSTEL de Rennes, il est choisi au regard de sa différence marquée avec le premier tant sur le plan géographique que sur le plan agricole. Ce chapitre présente les premières étapes du travail sur les données et suit le cheminement présenté à la figure 4.1. Dans ce chapitre, nous détaillons les deux jeux de données (section 4.1), afin d'appréhender complètement leurs différences. Nous présentons ensuite les pré-traitements qui leur ont été appliqués à deux niveaux :

extraction des segments et découpage en cellules (section 4.2). Les classifications effectuées ensuite donneront une première vision globale de ces données (section 4.3). Ces pré-traitements ont pour but de donner une première caractérisation des données et de les transformer afin de mettre en place les méthodes de fouille de données décrites dans les chapitres 5 et 6.

4.1 Présentation des données

4.1.1 Description des zones d'études

Les zones d'où proviennent les jeux de données sont très différentes. Les données A fournies par l'unité PSH représentent un paysage dans le Sud-Est de la France, au nord du département des Bouches du Rhône (Basse vallée de la Durance : coordonnées en WGS84 : de 43°46'27" N à 43°51'23" N et de 4°51'12" E à 4°57'34" E). Il s'étend sur près de 70 km^2 et comprend majoritairement des vergers (pommiers et poiriers) pour 70% de la Surface Agricole Utile (SAU), ainsi que du maraîchage. Cette région est caractérisée par une forte densité de haies brise-vent, essentiellement composées de cyprès ou peupliers. Ces haies protègent les cultures du mistral, un vent fort, d'orientation Nord-Sud. Les autres haies présentes dans le paysage sont composées de platanes en bord de route, ou sont composites. Les données contiennent également les informations relatives aux réseaux routiers et hydriques fournies par l'IGN.

En Bretagne, les données B ont été collectées sur la Zone Atelier Armorique et la zone étudiée couvre environ 120 km^2 dans la région de Pleine Fougère, au nord-est de Rennes (de 48°25'32"N à 48°34'06"N et de 1°31'39"O à 1°39'07"O). Nous avons à notre disposition le fichier contenant les coordonnées et caractéristiques des linéaires de haies et, comme pour l'autre région, ceux concernant les réseaux routiers et hydriques produits également par l'IGN. Le sud de la zone est constitué de bocage historique contrairement au nord remembré durant les dernières années. Les haies sont réparties en quatre catégories, à savoir : végétation éparse, haie arbustive, haie arborée discontinue et haie arborée continue.

4.1.2 Production des données

En basse vallée de la Durance (données A), les haies ont été numérisées manuellement à partir d'une orthophoto aérienne (source IGN) dans un système d'informations géographiques. Les données représentent donc l'implantation réelle des haies, avec prise en compte des discontinuités. Le fichier vectoriel est constitué de polylignes. Pour le

territoire breton, les limites administratives des parcelles ont été dessinées à partir du cadastre numérique puis chacune d'elle a été notée soit comme vide, soit comme appartenant à une des quatre catégories de haies définies précédemment grâce à une photo prise lors d'un vol ULM (2006). Le fichier vectoriel ainsi formé comporte également des polylignes.

4.2 Préparation des données

4.2.1 Découpe en segments

Nous avons fait le choix de segmenter les polylignes en segments afin de pouvoir travailler sur la longueur et la direction de chaque haie. En effet, il n'est pas naturel de définir l'angle d'une polyligne. Nous avons segmenté de la même façon les polylignes des haies et des réseaux routiers et hydriques afin de garantir l'homogénéité de la forme des données. Chaque segment est attaché à une catégorie (H = haie, R = route ou C = canal).

Le découpage en segments a eu des effets différents selon les régions : il y a seulement 56 segments de plus que de polylignes en Basse vallée de la Durance alors qu'à Pleine-Fougères, le nombre de segments des haies est environ le double du nombre de polylignes (Tableau 4.1).

Données	A		B	
Type de Données	Polylignes	Segments	Polylignes	Segments
Haies	11501	11557	7561	14819
Canaux	183	2014	328	9152
Routes	796	3737	2516	43416

TABLE 4.1 – Effet du découpage des polylignes en segments sur le nombre d'éléments par catégorie

En ce qui concerne les routes et les canaux, moins nombreux que les haies, l'effet du découpage en segments est encore plus marqué, le *ratio* nombre de segments sur nombre de polylignes allant de 5 (routes de A) à 28 (canaux de B). Notre jeu de données est donc maintenant constitué d'un ensemble de segments de trois types.

4.2.2 Découpe en cellules

Nous définissons le rectangle circonscrit à l'ensemble des segments comme étant le domaine d'étude initial. Celui-ci est le plus petit rectangle contenant les données et

dont les côtés sont parallèles aux axes Nord-Sud et Est-Ouest. Une fois ce domaine défini, nous avons choisi de le découper en cellules carrées pour diverses raisons :

— Mettre en évidence l'hétérogénéité dans chacun des jeux de données étudiés.
— Réduire la taille des données et donc, le temps de calcul.
— Faciliter la mise en œuvre des méthodes de partitionnement de l'espace présentées dans la section 3.2.

Nous avons découpé le domaine d'étude initial en cellules carrées en créant un chevauchement de toutes les cellules voisines égal à 10% de leur taille. Il est nécessaire, pour la suite, de bien choisir la taille des cellules. Pour cela nous avons, sur les données A, procédé au découpage pour des tailles de cellules variant entre $200m$ et $1300m$ de côté et, nous avons déterminé pour chaque valeur la distribution du nombre de segments de chaque type par cellule. Nous avons sélectionné le découpage en cellules de taille $1100m \times 1100m$ car il présentait un nombre de segments de haies suffisant par cellule, ainsi qu'un nombre de cellules suffisamment grand pour montrer la diversité des situations dans les données traitées. La variable *Nombre de segments de haies par cellule* possède une moyenne et une variance suffisante dans cette configuration. Nous obtenons ainsi un bon compromis entre le nombre de segments de type H par cellule et le nombre de cellules totales.

Ce découpage génère 121 cellules pour les données A et 140 pour les données B (Figure 4.2). Pour la suite, chaque cellule sera citée avec la codification suivante

$$L - X\_Y$$

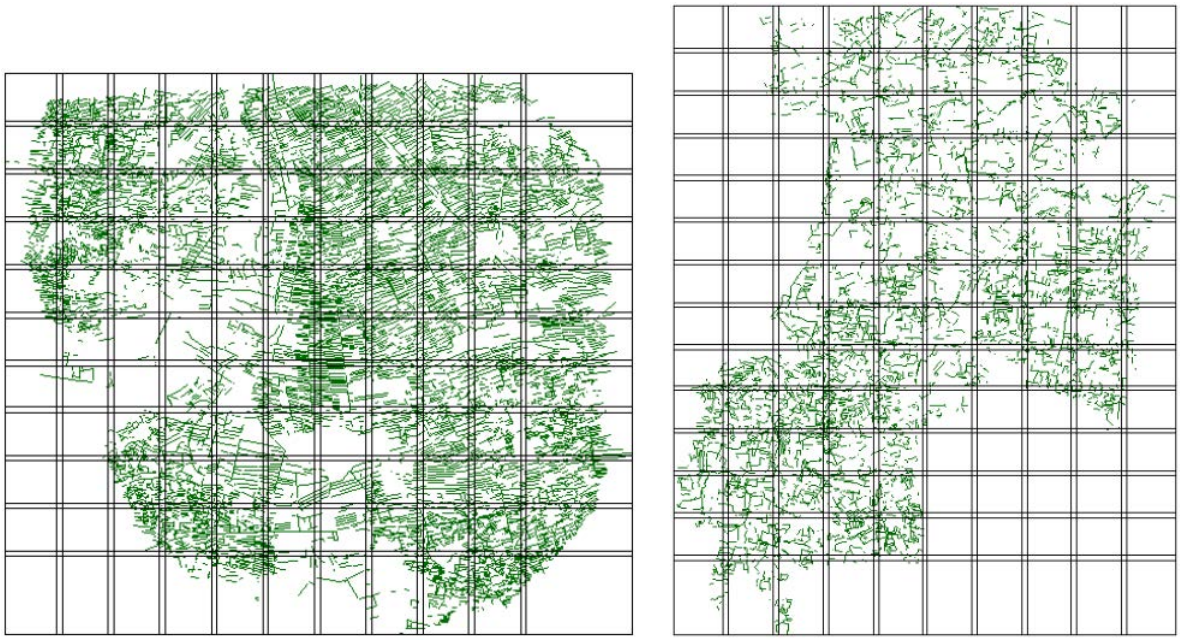
où

— L est la lettre rattachée à la zone d'origine des données
— X et Y sont l'abscisse et l'ordonnée de la cellule par rapport à l'ensemble des cellules.

Le repère attaché est constitué de la ligne horizontale en haut, et de la ligne verticale à gauche. La première cellule du repère, en haut à gauche, sera donc de coordonnées 0_0.

Exemples

La cellule à l'extrémité supérieure gauche sur la figure 4.2(a) sera citée sous la référence $A - 0\_0$. La cellule à l'extrémité inférieure droite sur la figure 4.2(b) sera citée sous la référence $B - 13\_9$.



(a) Données A (b) Données B

FIGURE 4.2 – Géographie des cellules pour les deux jeux de données

4.2.3 Classification des cellules

Nous avons, en découpant le domaine initial en cellules, considéré l'hétérogénéité possible de zones sur l'ensemble des données. À présent, nous allons classer les cellules obtenues afin d'obtenir des zones géographiques homogènes dans leur ensemble. Les cellules permettent de regrouper les zones semblables à l'échelle du segment et la classification permet de regrouper des zones semblables à l'échelle du domaine d'étude.

Définition : Zone Cellulaire de X_Y

La zone cellulaire de X_Y est composée de neuf cellules : la cellule X_Y et les huit cellules directement voisines. La représentation d'une zone cellulaire peut être vue à la figure 4.3.

La classification doit permettre de visualiser l'impact des caractéristiques de la cellule sur les caractéristiques des voisinages des haies. Afin de procéder à une classification des cellules, nous allons mettre en œuvre une classification hiérarchique selon quatre variables :

— *Nombre de segments de haies dans la cellule* X_Y,
— *Nombre de segments de route dans la cellule* X_Y,
— *Nombre de segments de haies dans la zone cellulaire de* X_Y,
— *Nombre de segments de route dans la zone cellulaire de* X_Y,

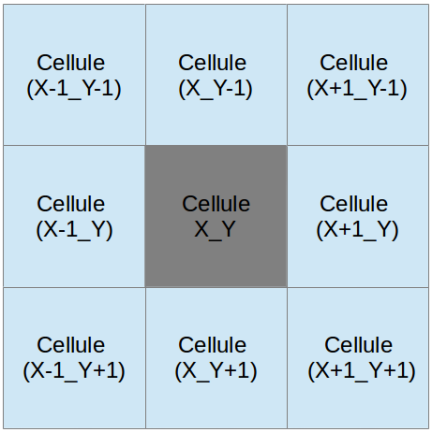


FIGURE 4.3 – Représentation d'une zone cellulaire, la cellule cible au centre (en gris foncé), les voisines autour (en bleu clair)

où X et Y sont les coordonnées des cellules.

Comme nous pouvons le voir sur la figure 4.2, une partie des cellules situées dans le domaine d'étude ne contient pas ou peu d'informations sur les segments. Ceci est dû au fait que les haies n'ont pas été numérisées sur la totalité du domaine d'étude. Pour éviter les biais liés à ce défaut de numérisation, nous n'allons pas effectuer la classification hiérarchique sur l'ensemble des cellules, mais seulement sur 64 cellules pour les données A et 35 cellules pour les données B. Les cellules retenues respectent deux critères : tout d'abord, l'ensemble des linéaires qu'elles renferment a été numérisé ; ensuite, la zone cellulaire attachée à chaque cellule retenue ne contient que des cellules présentant une proportion suffisante de linéaires numérisés.

4.3 Analyse des données

4.3.1 Dans leur globalité

L'objectif, à cette échelle, est de construire des classes de segments à partir des variables considérés, ici la longueur et l'orientation des segments. Cette classification intervient pour deux raisons. D'une part, nous avons la possibilité de différencier les segments car nous pensons qu'il existe des groupes ayant des conditions d'implantation spécifique. D'autre part, le regroupement en classes permet de discrétiser les variables continues étant donné que nous envisageons, pour la suite, d'appliquer les méthodes de Markov dans le cas de variables discrètes. *Longueur* est une variable simple à obtenir une fois les données produites ; *Orientation* pour un segment est définie par rapport à l'axe de référence Nord-Sud.

Nous nous intéressons principalement à la variable *Orientation*. La distribution de la variable *Longueur* n'a pas permis d'obtenir des classes ayant une réalité sur le terrain. Sur la figure 4.4(a), nous présentons l'histogramme circulaire des angles pour les données A. Les angles étant calculés par rapport à l'axe Nord-Sud, nous avons effectué sur la figure 4.4(a) une rotation de $\frac{\pi}{2}$ afin d'avoir une représentation des angles équivalente à la disposition des haies sur le terrain. Nous constatons que la distribution des angles n'est pas uniforme et qu'il se dégage 3 classes d'angles différentes. Tout d'abord, une première classe proche de l'horizontale, ayant pour direction l'axe (ouest-nord-ouest – est-sud-est). Celle-ci peut être considérée comme la classe des haies brise-vent (notée HV). La seconde classe (notée HP) est perpendiculaire à la première, et s'oriente selon l'axe (nord-nord-est – sud-sud-ouest). La dernière classe est composée du reste des valeurs. Nous avons donc défini les trois classes d'angle suivantes $[0, \frac{\pi}{9}] \cup [\frac{8\pi}{9}, \pi]; [\frac{\pi}{3}, \frac{5\pi}{9}];]\frac{\pi}{9}, \frac{\pi}{3}[\cup]\frac{5\pi}{9}, \frac{8\pi}{9}[$(modulo π).

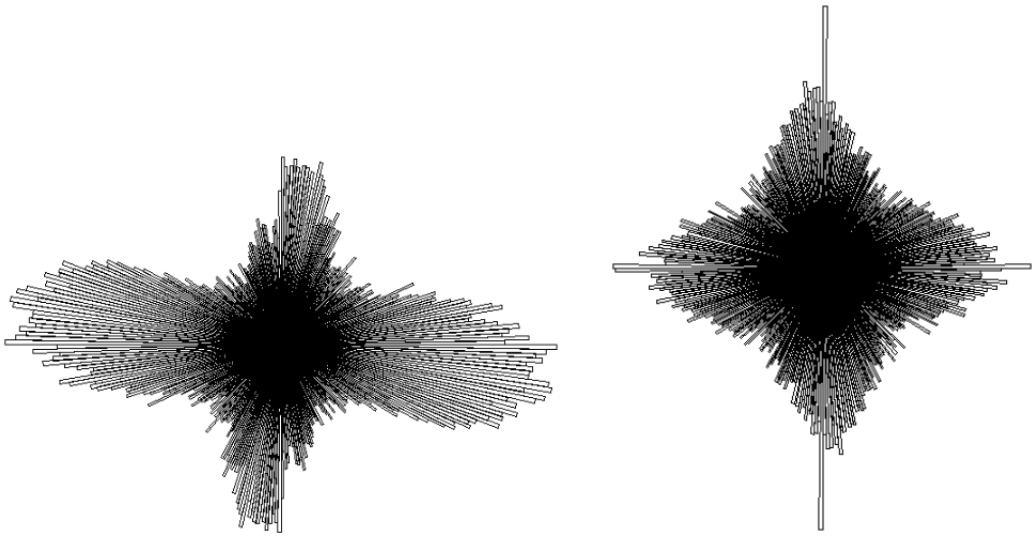
Nous avons effectué les mêmes études pour les données B. Bien qu'il n'existe pas un vent fort dominant dans cette région de Bretagne, l'orientation des haies présente elle aussi des classes (figure 4.4(b)). Tout d'abord, il existe une première classe d'angles autour de l'axe nord-sud avec un nombre élevé de haies. La seconde classe se situe autour de l'axe est-ouest. La troisième classe d'angle contient le reste. En accord avec les données A, nous notons les haies de la première classe (HP) et celles de la seconde (HV). Nous avons donc, pour les données B, les trois classes d'angles suivantes $[0, \frac{\pi}{9}] \cup [\frac{8\pi}{9}, \pi]$; $[\frac{7\pi}{18}, \frac{11\pi}{18}]$; $]\frac{\pi}{9}, \frac{7\pi}{18}[\cup]\frac{11\pi}{18}, \frac{8\pi}{9}[$ (modulo π).

4.3.2 Par classe

La procédure de classification de cellules permet de définir cinq classes dans la basse vallée de la Durance et six en Bretagne.

Nous pouvons remarquer, sur la figure 4.3.2, que pour chaque jeu de données, les classes présentent des zones distinctes et séparées les unes des autres. Il existe un gradient par rapport à la localisation des classes pour les données A. Les classes 1, 5, 4 et 3 sont disposées selon l'axe Nord-Est/Sud-Ouest, dans cet ordre. Tandis que pour les données B, les classes 1 et 2 sont exclusivement situées dans la zone Sud, qui ne comporte d'ailleurs pas d'autres classes. La partie Nord est divisée selon les autres classes, mais toujours avec des zones préférentielles pour chaque classe, la classe 4 en haut, la classe 6 en bas, la classe 5 à gauche. La classe 3 sert, plus ou moins, de zone de transition entre les autres classes.

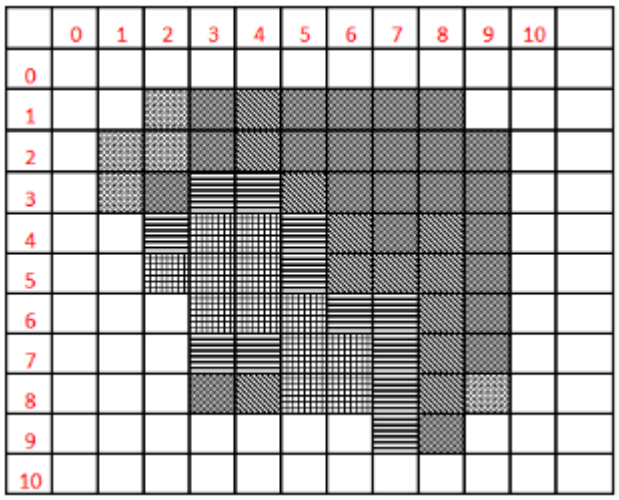
Ces classes comprennent un nombre moyen de segments de haies différent, de 58 à



(a) Données A (b) Données B

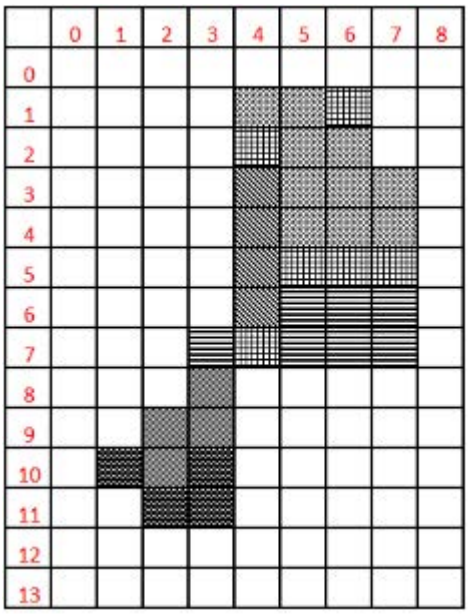
FIGURE 4.4 – Histogrammes circulaires des angles pour les deux jeux de données

237 par cellule dans le paysage de la Durance et de 167 à 406 par cellule dans le paysage de Bretagne. Les classes comprennent également un nombre moyen de segments de route différent, de 8 à 57 par cellule dans le paysage de la Durance et de 417 à 601 par cellule dans le paysage de Bretagne, mais diffèrent peu en nombre de segments de canaux par cellule. Le rapport de segments de haies HV à des segments de haies HP varie de 1,6 à 5 pour les cellules dans la vallée de la Durance et est plus stable, dans le paysage Bretagne, variant de 0,9 à 1,3 (tableau 4.2).



Sample	Landscape	
12	23	Cat. 1
2	5	Cat. 2
5	12	Cat. 3
5	12	Cat. 4
5	12	Cat. 5
29 Cells	64 Cells	

(a) Données A



Sample	Landscape	
2	4	Cat. 1
2	4	Cat. 2
3	6	Cat. 3
5	10	Cat. 4
2	4	Cat. 5
3	7	Cat. 6
17 Cells	35 Cells	

(b) Données B

FIGURE 4.5 – Visualisation des classes obtenues par classification hiérarchique pour les deux jeux de données

		Haies				Routes		Canaux	
		Totalité		HP	HV	Totalité		Totalité	
		Cellule	Zone	Cellule	Cellule	Cellule	Zone	Cellule	Zone
Basse vallée de la Durance (A)									
Classe	1	165 ±33	1191 ±71	25 ±17	84 ±20	42 ±18	330 ±57	21 ±15	161 ±50
	2	237 ±79	1528 ±115	66 ±36	106 ±16	57 ±20	405 ±36	24 ±6	163 ±29
	3	58 ±29	543 ±88	10 ±8	29 ±19	8 ±7	108 ±35	24 ±10	153 ±22
	4	95 ±26	808 ±56	18 ±7	55 ±22	22 ±12	172 ±50	12 ±8	141 ±30
	5	142 ±32	988 ±44	18 ±14	93 ±28	31 ±14	257 ±64	17 ±11	124 ±40
Bretagne (B)									
Classe	1	406 ±64	2495 ±151	130 ±55	121 ±33	601 ±54	4116 ±98	109 ±51	662 ±89
	2	309 ±50	2285 ±146	90 ±30	86 ±10	469 ±138	3517 ±178	87 ±65	612 ±198
	3	177 ±30	1516 ±129	46 ±27	60 ±16	444 ±43	3666 ±90	92 ±61	641 ±352
	4	167 ±37	1287 ±113	47 ±14	46 ±17	555 ±151	3951 ±119	38 ±53	347 ±178
	5	174 ±67	1233 ±176	57 ±24	62 ±24	417 ±45	3225 ±41	77 ±107	430 ±227
	6	241 ±79	1669 ±132	91 ±39	82 ±37	427 ±73	3359 ±97	106 ±66	926 ±145

TABLE 4.2 – Nombre de segments (moyenne±écart type) de chaque type (haies, routes, canaux) dans une cellule et sa zone cellulaire, et nombre de segments de haies HP et HV dans une cellule pour chaque classe de cellules dans les deux paysages étudiés (basse vallée de la Durance (A) et Bretagne (B))

Chapitre 5

Caractérisation des structures spatiales de segments par l'étude de leurs voisinages

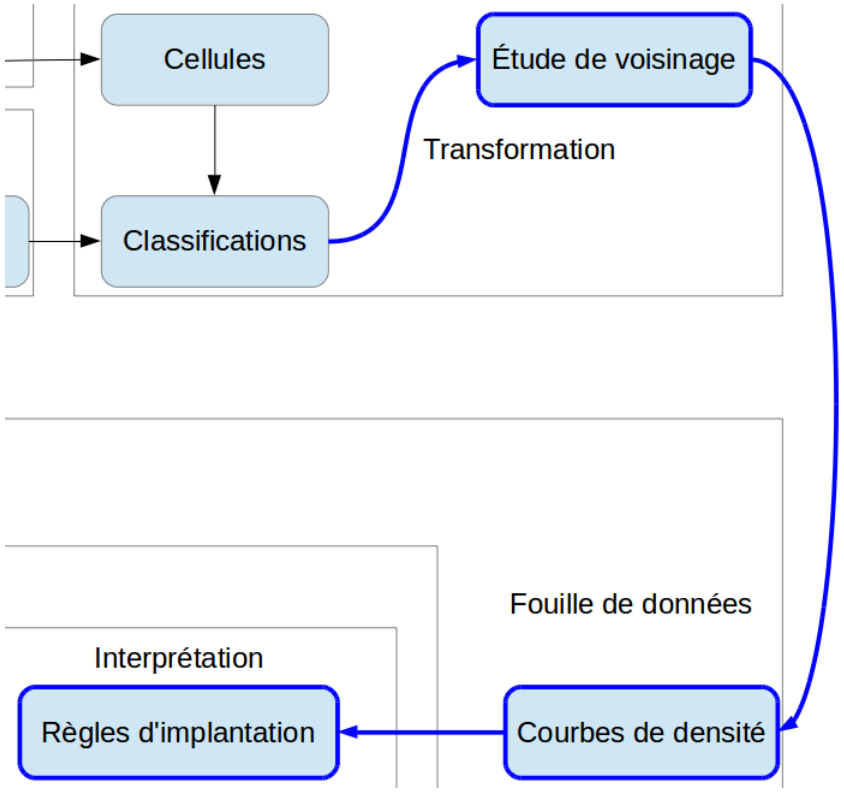


FIGURE 5.1 – Étapes décrites au chapitre 5

Introduction

Dans la suite, nous appelons données A (respectivement B) l'ensemble des segments du paysage A (respectivement B). Ces données regroupent des segments de différents types (haies, routes et canaux), et nous allons étudier les liaisons qui existent entre eux. Ce chapitre présente la première méthode développée pour la caractérisation des structures spatiales des segments et suit le cheminement présenté à la figure 5.1. Dans cette méthode, nous utilisons une approche fondée sur le voisinage des segments qui permet de calculer un indice de caractérisation du paysage s'appuyant sur les densités de segments de différents types dans ce voisinage. Nous présenterons tout d'abord l'approche développée pour cela dans la section 5.1. Puis dans la section 5.2 nous présenterons les résultats obtenus sur les données A et B afin d'en déduire des règles d'implantation des segments de haies dans les paysages agricoles.

5.1 Études de voisinages et densité de segments dans l'espace

5.1.1 Outils développés

Distance entre segments

Rappel : Distance entre deux segments (cf. sous-section 3.1.2)

Soit S et S' deux segments de $\mathbb{R}^2$, soit d la distance euclidienne entre deux points du plan, nous déterminons la distance entre deux segments S et S', notée $\texttt{DiSt}(S, S')$, par

$$\texttt{DiSt}(S, S') = min\{d(x, y), \forall x \in S, \forall y \in S'\}$$

Afin de calculer la distance entre deux segments, nous devons tout d'abord déterminer l'équation dans le plan de la droite support, notée F_1 (respectivement F_2), du segment S_1 (respectivement S_2)(ligne 1, procédure 8). Ensuite, nous déterminons le projeté de chacun des points extrémités A_1 et A_2 du segments S_1 sur la droite F_2 (ligne 2, procédure 8) et, de la même façon, nous déterminons le projeté de chacun des points extrémités A_3 et A_4 du segments S_2 sur la droite F_1 (ligne 3, procédure 8). Cette étape est illustrée sur la figure 5.2. Après, pour chaque projeté, nous vérifions s'il est inclus dans l'un des deux segments (ligne 4, procédure 8). Si c'est le cas, nous calculons la distance entre le point et son projeté (ligne 5, procédure 8) et sinon, nous fixons la

Procedure DiSt - Calcul de la distance entre deux segments

Data: Soit $S_1 = [(x_{A_1}; y_{A_1}), (x_{A_2}; y_{A_2})]$ et $S_2 = [(x_{A_3}; y_{A_3}), (x_{A_4}; y_{A_4})]$ deux segments de $\mathbb{R}^2$

Result: $\texttt{DiSt}(S_1; S_2)$

begin

for $k = 1, 2$ **do**

1 Déterminer l'équation de la droite F_k support du segment S_k

for $k = 1, 2$ **do**

2 Déterminer les coordonnées du projeté A'_k du point A_k sur la droite F_2

for $k = 3, 4$ **do**

3 Déterminer les coordonnées du projeté A'_k du point A_k sur la droite F_1

for $E \in \{A_1, A_2, A_3, A_4\}$ **do**

E' son projeté orthogonal sur F_1 ou F_2 **if** $E' \in S_1 \cup S_2$ **then**

5 $EE' = \sqrt{(x_E - x_{E'})^2 + (y_E - y_{E'})^2}$

else

6 $EE' = +\infty$

7 Calculer les longueurs $A_k A_j \mid k \in \{1; 2\}$ et $j \in \{3; 4\}$

8 $\texttt{DiSt}(S_1, S_2) \longleftarrow \min(A_1A'_1, A_2A'_2, A_3A'_3, A_4A'_4, A_1A_3, A_1A_4, A_2A_3, A_2A_4)$

distance entre le point et son projeté égale à $+\infty$ (ligne 6, procédure 8). Dans le cas exposé sur la figure 5.2, les projetés A'^1, A'^2 et A'^3 sont inclus dans les segments, la distance $A^iA'^i$ sera calculée pour $i = 1, 2$ `et` 3, comme A'^4 n'est sur aucun segment, alors $A^4A'^4 = +\infty$ Enfin, nous calculons pour chaque extrémité du segment S_1, la distance à chaque extrémité du segment S_2 (ligne 7, procédure 8). Finalement, la distance entre les segments S_1 et S_2 est définie comme le minimum entre toutes les distances calculées précédemment (ligne 8, procédure 8).

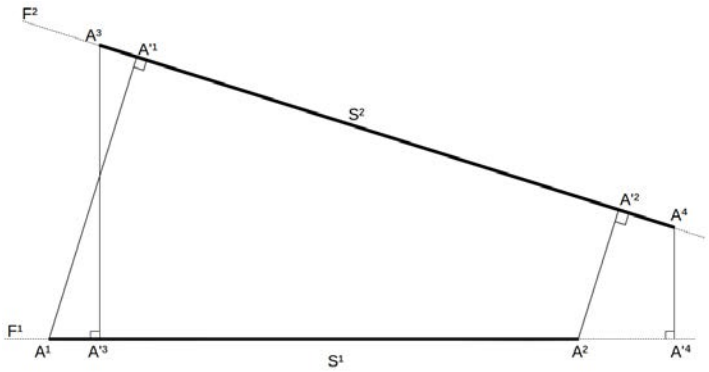


FIGURE 5.2 – Schéma explicatif pour le calcul de la distance entre deux segments

Voisinage d'un segment

Nous appelons voisinage de taille b d'un segment S de longueur L, la zone définie par l'ensemble des points se situant à une distance inférieure à b de S. Son aire, notée $B(L, b)$, est égale à :

$$B(L, b) = \pi \times b^2 + 2 \times L \times b$$

Le voisinage autour d'un segment de longueur L prend la forme décrite dans le dessin 5.3.

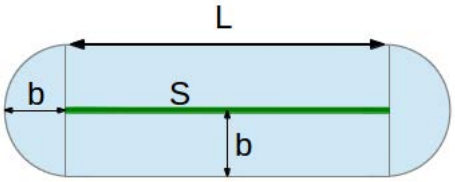


FIGURE 5.3 – Visualisation du voisinage à une distance b d'un segment S de longueur L, pour la distance `DiSt`

Angle entre segments

Le domaine de valeurs des angles $[0; \pi]$ est discrétisé en trois classes (*Perpendiculaire*, *Parallèle*, *Autre*), respectivement notées ($C(\theta_\perp)$, $C(\theta_\parallel)$, $C(\theta_A)$) et présentées dans le tableau 5.1. Ces dernières sont définies en accord avec les caractéristiques des données (cf. sous-section 4.3.1).

Intervalles de valeurs des angles	$[0; \frac{2\pi}{18}[$	$[\frac{2\pi}{18}; \frac{7\pi}{18}[$	$[\frac{7\pi}{18}; \frac{11\pi}{18}[$	$[\frac{11\pi}{18}; \frac{16\pi}{18}[$	$[\frac{16\pi}{18}; \pi[$
Nature de la liaison entre segments	Parallèle	Autre	Perpendiculaire	Autre	Parallèle

TABLE 5.1 – Récapitulatif des classes d'angle entre deux segments

Densité relative de différents types de segments dans le voisinage d'un segment

La densité relative de segments S_I de type I dans le voisinage d'un segment référence S est calculée comme le rapport entre le nombre de segments de type I situés à une distance inférieure à b du segment S et le nombre de segments de type I attendus dans ce voisinage, compte tenu de la densité des segments de type I sur la zone cellulaire considérée.

Le nombre de segments de type I voisins d'un segment S à une distance inférieure à b, noté $N_o^I(S, b)$ est défini par,

$$N_o^I(S, b) = |\{S_I | \texttt{DiSt}(S, S_I) \leqslant b\}|$$

où S_I est un segment de type I.

Le nombre de segments attendus dans un voisinage est défini pour un segment S de longueur L et une distance b, il est donné de la façon suivante. Soit A l'aire de la zone cellulaire dont la cellule centrale contient S, T_A^I le nombre de segments de type I dans cette zone, $\frac{T_A^I}{A}$ est la densité moyenne de segments de type I dans la zone cellulaire. Le nombre attendu de segments de type I dans ce voisinage est alors

$$N_t^I(S, b) = \frac{B(L, b) \times T_A^I}{A}$$

La densité relative de segments de type I dans le voisinage d'un segment S, notée

$D_r^I(S,b)$, est alors donnée par

$$D_r^I(S,b) = \frac{N_o^I(S,b)}{N_t^I(S,b)}$$

Pour un segment S, ce rapport entre le nombre observé et le nombre attendu de voisins de type I, à une distance inférieure à b, prend ses valeurs dans $\mathbb{R}^+$. Si ce *ratio* est supérieur à 1 alors nous pouvons dire que les voisins de type I sont sur-représentés, s'il est inférieur à 1 alors ceux-ci sont sous-représentés et dans le cas où il est égal à 1, alors la densité de segments de type I au voisinage de S est identique à la densité de segments de type I présents sur la zone.

Nous avons également défini une densité relative orientée. Celle-ci est définie pour des segments de type I qui sont dans le voisinage du segment S et qui forment un angle particulier avec celui-ci. Ainsi, en reprenant les classes d'angles entre segments vues dans la sous-section 5.1.1, nous pouvons adapter les valeurs précédemment introduites comme suit :

$$N_o^I(S,b) \text{ devient } N_o^I(S,b,C(\theta))$$

et

$$D_r^I(S,b) \text{ devient } D_r^I(S,b,\theta) = \frac{N_o^I(S,b,C(\theta))}{N_t^I(S,b)}$$

avec $N_o^I(S,b,C(\theta)) = |\{S_I \mid \texttt{DiSt}(S,S_I) \leqslant b \text{ et } \widehat{SS_I} \in C(\theta)\}|$ et $C(\theta)$ représente une classe d'angle (*perpendiculaire*, *parallèle*, *autre* définie dans le tableau 5.1) ; entre le segment S et son voisin. Par contre, nous ne pouvons pas calculer de nombre théorique pour une direction donnée, de ce fait, la valeur $N_t^I(S,b)$ ne dépend pas de θ.

Une relation existe entre $D_r^I(S,b)$ et $D_r^I(S,b,\theta)$, en effet :

$$D_r^I(S,b) = D_r^I(S,b,C(\theta_\perp)) + D_r^I(S,b,C(\theta_\parallel)) + D_r^I(S,b,C(\theta_A))$$

Il n'est plus possible à présent de comparer les densités relatives orientées à 1 mais en divisant celles-ci par la densité relative orientée de la même classe d'angles à une distance b suffisamment grande, nous pouvons alors comparer ce rapport à 1 afin de vérifier une sur ou sous représentation de voisins de type I dans le voisinage orienté de S.

Densité relative, à l'échelle de la cellule, de différents types de segments dans le voisinage d'un segment

Une fois que nous avons déterminé pour chaque segment S d'une cellule $\mathcal{C}$ la densité relative de différents types de segments dans son voisinage orienté $D_r^I(S, b, C(\theta))$, nous pouvons généraliser cela à une densité relative par cellule, que nous noterons $D_r^I(\mathcal{C}, b, C(\theta))$. Son calcul se fait grâce à la formule suivante :

$$D_r^I(\mathcal{C}, b, C(\theta)) = \sum_{S \in \mathcal{C}} \frac{D_r^I(S, b, C(\theta))}{NS\mathcal{C}}$$

où $NS\mathcal{C}$ est le nombre de segments S dans la cellule $\mathcal{C}$.

Afin de pouvoir comparer les résultats sur l'ensemble des cellules et ceci, peu importe le paysage, nous définissons pour chaque cellule la densité relative normalisée, notée $D_{rn}^I(\mathcal{C}, b, C(\theta))$ pour la distance b et la classe d'angle $C(\theta)$. Celle-ci est déterminée comme le rapport entre la densité relative $D_r^I(\mathcal{C}, b, C(\theta))$ à une distance b, et la densité relative $D_r^I(\mathcal{C}, 500, C(\theta))$ à la distance $500m$.

$$D_{rn}^I(\mathcal{C}, b, C(\theta)) = \frac{D_r^I(\mathcal{C}, b, C(\theta))}{D_r^I(\mathcal{C}, 500, C(\theta))}$$

5.1.2 Tests

Dans ce qui précède, nous avons défini des densités relatives pour caractériser les structures de voisinage. Nous avons défini deux tests de permutations pour tester des hypothèses quant aux valeurs prises par ces densités.

Test des facteurs qui impactent les densités relatives

Le premier test permet de vérifier l'existence d'effet sur les densités relatives normalisées des voisins pour les voisinages de tailles croissantes, des caractéristiques suivantes :

— la classe de la cellule
— le type de segments de haies
— le type de segments voisins
— l'orientation relative de l'un part rapport à l'autre

Pour ce faire, nous analysons dans un premier temps les effets de ces facteurs et de leurs interactions simples sur les densités relatives dans un modèle linéaire. Les données n'étant pas indépendantes, nous avons besoin de construire les distributions des statistiques de Fisher sous l'hypothèse $H0$ d'une absence d'effet de ces différents

facteurs. Ces distributions sont obtenues en appliquant à nouveau ces modèles sur 100 permutations des densités relatives associées à chaque combinaison "type de segments de haies × type de segments voisins × orientation relative" au sein de chaque cellule. Si la valeur observée se situe en deçà du quantile 0,95 de la distribution, l'hypothèse $H0$ est conservée, si elle se situe au-delà de cette valeur, elle est rejetée.

Test de significativité des densités relatives observées

Le second test permet de vérifier si la densité relative normalisée de différents types de voisins au voisinage des segments de haies, dans un voisinage croissant et avec une orientation relative donnée est plus forte que celle attendue sous l'effet du hasard.

Pour ce faire nous avons généré aléatoirement, dans une zone cellulaire, l'implantation de segments de haies en respectant le nombre de segments par cellule et leur orientation. Les segments de routes et canaux n'ont pas été modifiés. Ensuite, nous avons émis l'hypothèse que la densité relative normalisée de la cellule réelle était une réalisation de la distribution obtenue d'après les simulations aléatoires, et cela pour chaque distance, chaque classe d'angle, chaque type de voisins et chaque type de haies.

5.2 Densité relative au voisinage des segments de haies

5.2.1 Le plus proche voisin d'un segment de haie

Comme nous le voyons sur la figure 5.4, le voisin le plus proche d'un segment de haies, défini grâce à la distance `DiSt` (cf. sous-section 5.1.1), est principalement un autre segment de haies, et ceci, dans les deux paysages. Cela représente, en moyenne, respectivement 69%(±14%) et 72%(±7%) des voisins des segments des haies HP et HV, dans la basse vallée de la Durance et, respectivement 71%(±13%) et 67%(±12%) des voisins des segments des haies HP et HV, en Bretagne. Les segments de routes sont globalement des voisins plus fréquents en Bretagne que dans la basse vallée de la Durance, tandis que les segments de canaux sont les voisins les moins fréquents. Ce résultat est cohérent avec la plus grande proportion de segments de routes par rapport aux segments de haies en Bretagne. En outre, les segments de routes en Bretagne sont plus souvent les voisins les plus proches quand ils sont perpendiculaires aux segments de haies que lorsqu'ils sont parallèles à eux. La tendance est moins claire dans la basse vallée de la Durance.

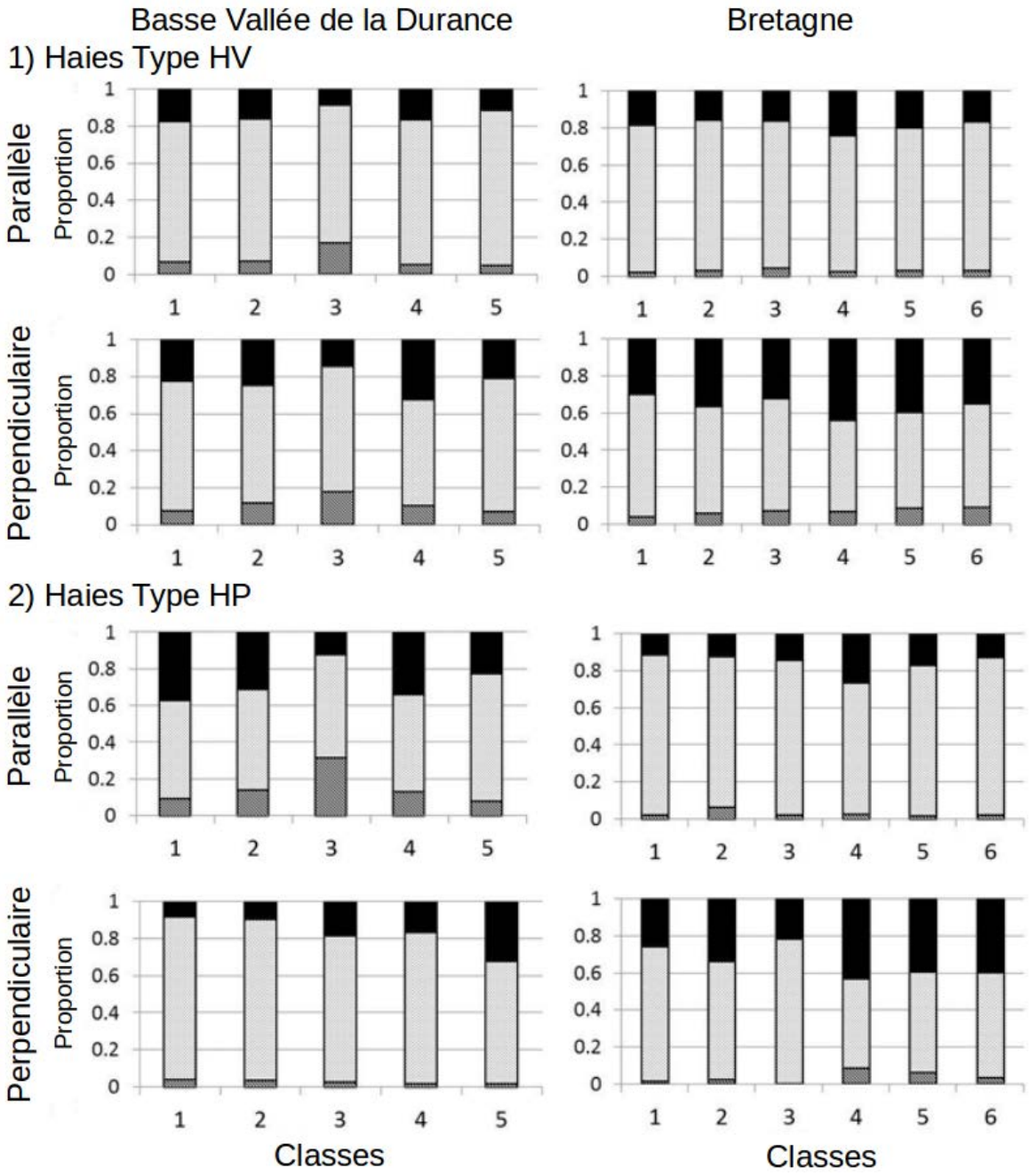


FIGURE 5.4 – Proportion de chaque type de voisin le plus proche pour chaque classe de cellule. Le résultat est présenté pour chaque paysage (à gauche : basse vallée de la Durance, à droite : Bretagne), pour chaque type de haie (1) : haies HV, 2) haies HP) et pour les orientations parallèles et perpendiculaires. Le type de voisin est route (noir), haie (gris clair) ou canal (gris foncé)

5.2.2 Densité relative de segments de haies, routes et canaux dans les voisinages de segments de haies

Distributions des voisins sur deux cellules typiques

Nous allons présenter les résultats sur deux cellules typiques, dont la géométrie est présentée par la figure 5.5.

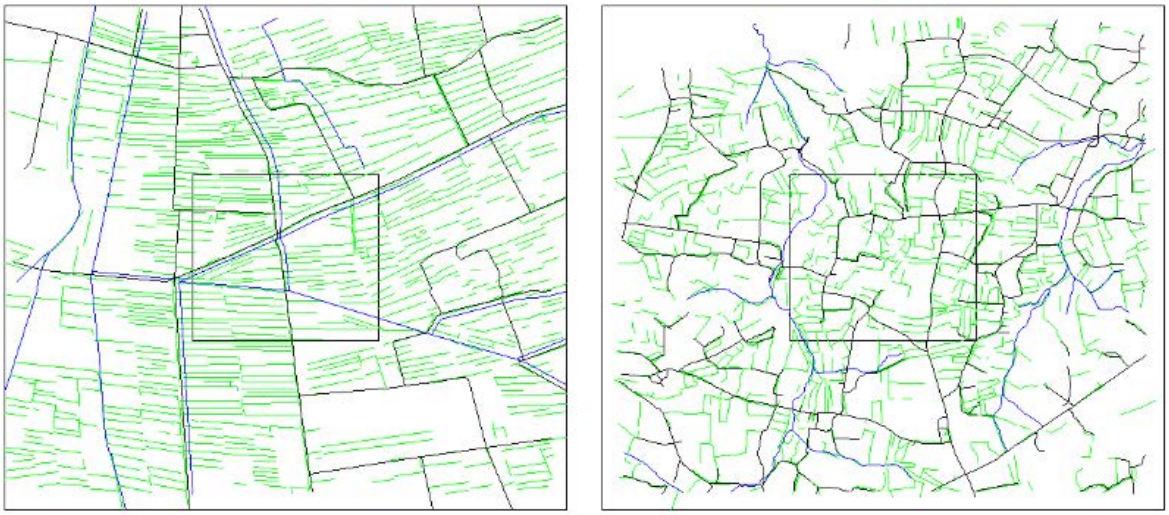


FIGURE 5.5 – Deux cellules typiques et leur zone cellulaire dans la basse vallée de la Durance ($A-5\_6$ à gauche) et en Bretagne ($B-9\_2$ à droite). Le carré central représente la cellule de dimension $1100m \times 1100m$. Les lignes vertes représentent les haies, les noires représentent les routes et les bleues représentent les canaux.

Sur la base de l'examen de l'ensemble des segments de haies dans ces deux cellules typiques, les densités relatives $D_r^H(\mathcal{C}, b), D_r^R(\mathcal{C}, b), D_r^C(\mathcal{C}, b)$ de chaque type de segment (haies, routes et canaux) ont été représentées à la figure 5.6. Si l'on considère l'ensemble des haies, nous constatons que les densités relatives des trois types de segments diminuent dans les 150 premiers mètres vers le seuil correspondant à la proportion de segments de ce type qui sont soit parallèles, soit perpendiculaires aux haies. La plupart des types de segments sont ainsi sur-représentés à une courte distance des segments de haies, indiquant une association à courte distance (inférieure à $150m$) avec eux (et en dessous de $50m$ pour les segments de haies perpendiculaires). La seule exception concerne la densité relative de segments de haie parallèles aux segments de haies $D_r^H(\mathcal{C}, b, C(\theta_{\|}))$ dans la cellule $A-5\_6$. Ce nombre est presque constant, avec un pic situé à environ $100m$, conforme à la répartition apparemment régulière de haies dans la figure 5.5. La sur-représentation à courtes distances concerne principalement les haies dans la cellule $B-9\_2$. Il existe une forte association entre les segments de haies pour

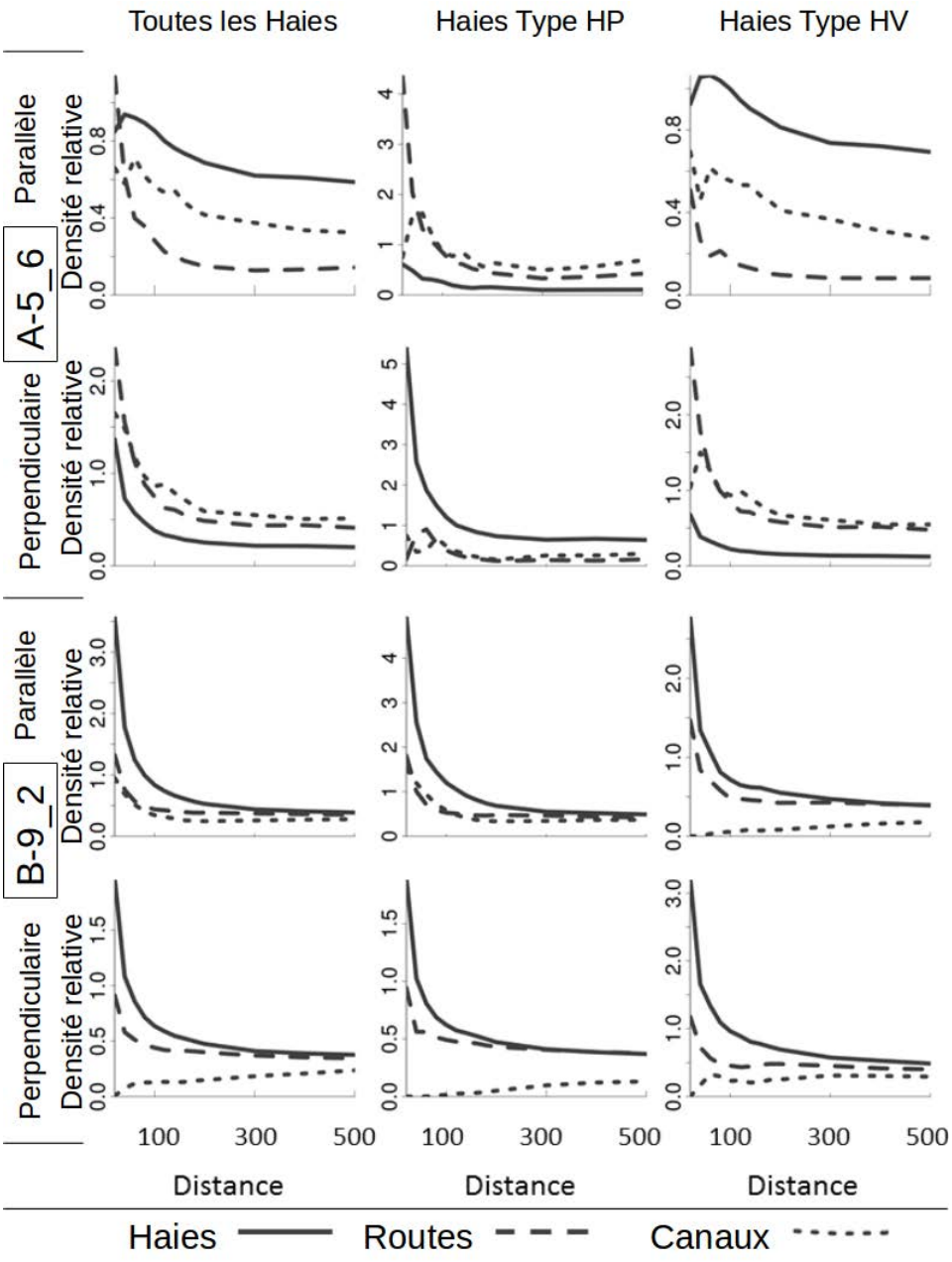


FIGURE 5.6 – Densité relative $D_r^I(\mathcal{C}, b, C(\theta))$ de chaque type d'éléments (haies, routes, canaux) dans un voisinage croissant (de $20m$ à $500m$) autour des segments de haies pour deux cellules typiques ($A-5\_6$ pour la basse vallée de la Durance et $B-9\_2$ pour la Bretagne). $D_r^I(\mathcal{C}, b, C(\theta))$ est donné pour les deux orientations relatives (parallèle et perpendiculaire) et dans le voisinage de tous les segments de haies (à gauche), des haies HP (au centre) ou des haies HV (à droite).

la cellule $B-9\_2$ alors qu'il s'agit surtout d'une association entre les segments de routes et les segments de haies dans la cellule $A-5\_6$. Les valeurs de seuil à longues distances diffèrent dans la cellule $A-5\_6$ entre les directions parallèle et perpendicu-

laire, avec les haies comme éléments le plus fréquent dans la direction parallèle et le moins fréquent dans la direction perpendiculaire. Cette tendance n'a pas été observée dans $B-9\_2$.

L'examen plus spécifique des segments de haies HV et HP montre les différences entre ces types de segments, principalement dans la cellule $A-5\_6$. Dans celle-ci, les deux types de haies sont situés près des routes, mais les segments de haies HV sont principalement perpendiculaires à ces routes ($P < 0,05$ pour les distances inférieures à $150m$ en orientation perpendiculaire, mais pas en orientation parallèle), alors que les segments de haies HP sont parallèles à des routes ($P < 0,05$ pour les distances inférieures à $150m$ de l'orientation parallèle). En outre, les segments de haies HP sont préférentiellement situés perpendiculairement à des segments de haie proches ($P < 0,05$ pour les distances inférieures à $100m$), tandis que les segments de haies HV ne le sont pas. Les différences entre les segments de haies HP et HV sont moins apparentes dans $B-9\_2$. Une particularité de $B-9\_2$ est que les canaux ont une orientation Nord-Sud et sont donc sous-représentés en tant que voisins parallèles des segments de haies HV ainsi qu'en tant que voisins perpendiculaires des segments de haies HP.

Synthèse en considérant l'ensemble des cellules

Nous utilisons la densité relative normalisée, comme définie dans la sous-section 5.1.1, afin de pouvoir comparer les valeurs sur les différentes cellules. Comme dans les cellules $A-5\_6$ et $B-9\_2$, le même motif général est trouvé lorsque toutes les cellules sont considérées ensemble : les segments de haies et leurs voisins sont fortement associés jusqu'à une distance d'environ $150m$, sauf pour les segments de haies parallèles aux segments de haies HV dans la basse vallée de la Durance qui montrent un modèle plus régulier (Figure 5.7).

Effet de l'orientation relative des voisins Dans les deux paysages, la densité relative de voisins d'un segment de haies donné, pour chaque distance, dépend à la fois du type de voisin, c'est-à-dire si c'est une haie, une route ou un canal, et de son orientation par rapport à ce segment de haies (effet significatif de *V.Types * R.Orient* dans le tableau 5.2). L'effet du type de voisin en Bretagne résulte de la faible association des haies et des canaux, tandis que l'association entre les haies et, soit d'autres haies soit des routes, est plus forte. Dans la basse vallée de la Durance, cette interaction significative résulte d'une association plus faible des haies entre elles qu'avec les canaux et les routes. L'effet de l'orientation relative du voisinage est particulièrement clair en Bretagne où la densité relative des voisins parallèles est supérieure à celle des voisins

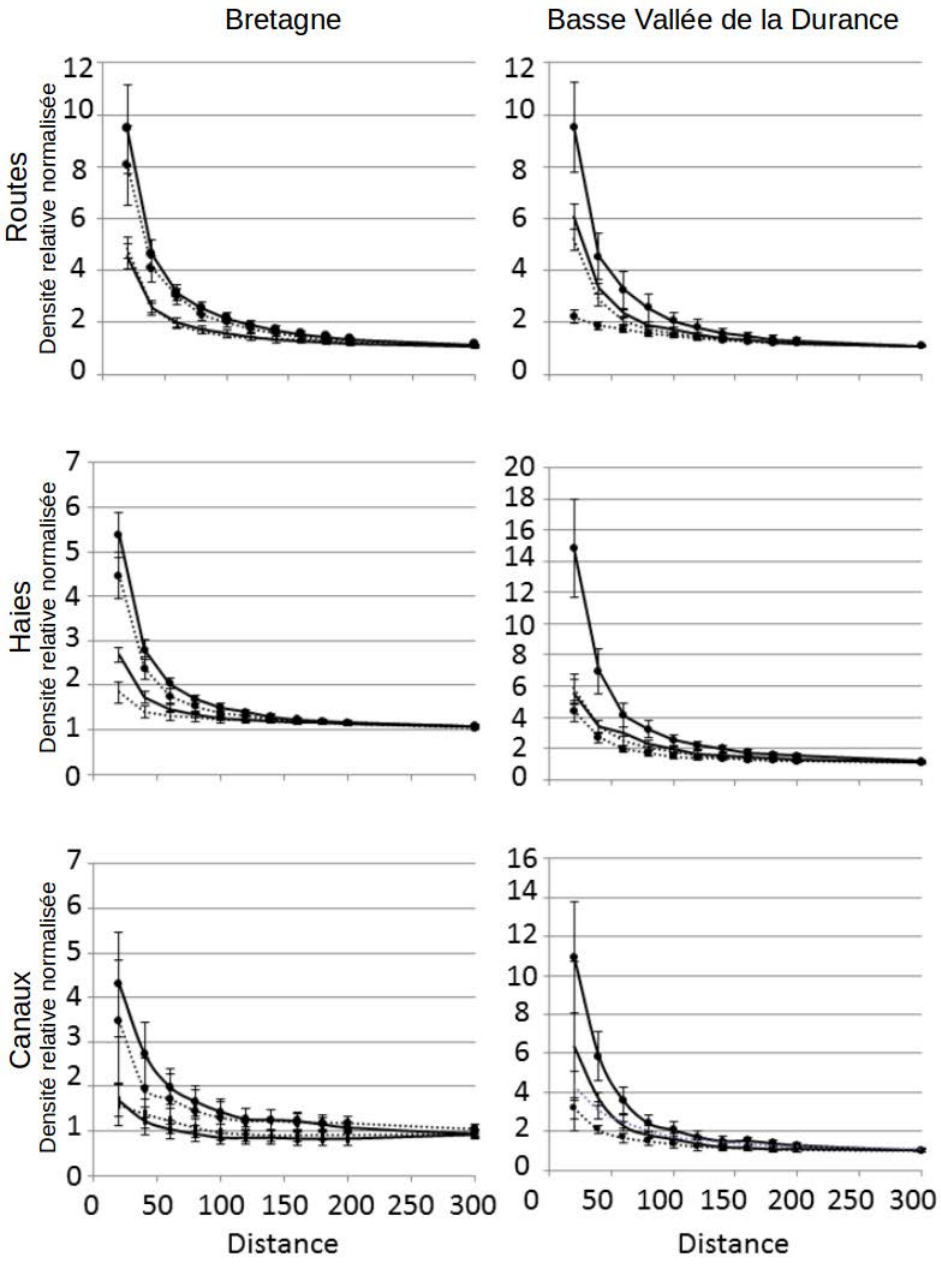


FIGURE 5.7 – Densité relative normalisée de chaque type de segments (en haut : routes, au milieu : haies, en bas : canaux) à des distances croissantes des segments de haies dans chaque paysage (Bretagne à gauche et basse vallée de la Durance à droite). Les traits pleins représentent les segments de haies HP, les traits en pointillés représentent les segments de haies HV, les ronds noirs sont présents sur les courbes concernant l'orientation parallèle du voisinage et les courbes sans ronds noirs concernent l'orientation perpendiculaire du voisinage

perpendiculaires quel que soit le type de voisins (Figure 5.7). Dans la basse vallée de la Durance, l'effet *R.Orient* est significatif comme un terme d'interaction à la fois avec

le type de haies (HP ou HV) et le type de voisin (H,R,C) comme détaillé dans le paragraphe suivant.

Taille du voisinage (m)	20	40	60	80	100	120	140	160	180	200
Basse vallée de la Durance										
Class	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
H.Type	**	**	**	**	**	*	*	ns	ns	ns
V.Type	**	**	**	**	**	**	**	**	**	**
R.Orient	**	**	**	**	**	**	**	**	**	**
Class × H.Type	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Class × V.Type	*	ns	ns	*	**	**	**	**	**	**
Class × R.Orient	*	**	**	ns	ns	ns	ns	ns	ns	ns
H.Type × V.Type	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
H.Type × R.Orient	**	**	**	**	*	*	*	ns	*	*
V.Type × R.Orient	**	**	ns	ns	ns	ns	ns	ns	ns	ns
Bretagne										
Class	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
H.Type	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
V.Type	**	**	**	**	**	**	**	**	**	**
R.Orient	**	**	**	**	**	**	**	**	**	**
Class × H.Type	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Class × V.Type	**	**	**	**	**	**	**	**	**	**
Class × R.Orient	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
H.Type × V.Type	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
H.Type × R.Orient	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
V.Type × R.Orient	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns

TABLE 5.2 – Test de permutation pour étudier l'effet des classes de cellules (Class), du type de segments de haies (H.type), du type de segments voisins (V.Type), de l'orientation relative de l'un part rapport à l'autre (R.Orient) et toutes les interactions doubles sur les densités relatives normalisées des voisins pour les voisinages de tailles croissantes. L'absence de significativité est notée ns, sa présence est notée, selon sa grandeur * ou **.

Comportement différent des segments de haie de type HP et HV Une différence majeure entre les deux paysages est l'effet du type de haies (i.e. orientation Est / Ouest ou Nord / Sud qui correspond à des haies brise-vent ou non en basse vallée de la Durance). Cet effet est significatif à la fois comme *effet principal* et en *interaction*

avec d'autres caractéristiques dans la basse vallée de la Durance, mais pas en Bretagne (tableau 5.2).

Dans la basse vallée de la Durance, le type de voisins qui montre la plus forte association à courte distance selon leur orientation relative dépend du type de haie. Mais cette association peut perdurer jusqu'à $300m$. Ainsi, les segments de haies HP sont significativement parallèles aux routes ou aux canaux au moins jusqu'à $300m$ (colonnes Routes et Canaux, ligne 4, tableau 5.3) et ont également des haies perpendiculaires dans leur proche voisinage, jusqu'à $100m$ (colonne Haies, ligne 3, tableau 5.3).

Au contraire, les segments de haies HV ont un niveau intermédiaire d'association à courtes distances avec des routes et des haies perpendiculaires. Les segments de haies HV ont, dans leur proche voisinage perpendiculaire (entre $20m$ et $60m$), les trois types de segments (ligne 1, tableau 5.3). Mais cette relation avec les segments de haies ne perdurent pas, ce qui suggère que deux segments de haies ne sont proches que s'ils sont perpendiculairement orientés l'un à l'autre (colonne Haies, ligne 1, tableau 5.3). Les segments de haies HV présentent une distribution régulière en ce qui concerne les haies parallèles voisines. Il est à noter la répétition de la liaison, les segments de haies HV ont préférentiellement, dans leur voisinage parallèle, des haies aux distances $100m$ et $200m$, ce qui suggère la répétition d'un motif d'implantation (colonne Haies, ligne 2, tableau 5.3). Cet effet du type de haies disparaît à une distance d'environ $300m$.

		Haies						Routes						Canaux					
Taille du voisinage (m)		20	60	100	140	200	300	20	60	100	140	200	300	20	60	100	140	200	300
Type	Orientation																		
1 HV	Perpendiculaire	*	*	ns	ns	ns	ns	*	*	*	*	*	*	ns	*	*	*	*	*
2 HV	Parallèle	ns	ns	*	ns	*	ns	*	ns	ns	ns	ns	ns	ns	ns	*	*	*	*
3 HP	Perpendiculaire	*	*	*	ns	ns	ns	ns	*	*	ns	ns	ns	ns	ns	ns	ns	ns	ns
4 HP	Parallèle	ns	*	*	ns	ns	ns	*	*	*	*	*	*	ns	*	*	*	*	*
5 All	Perpendiculaire	*	*	*	ns	ns	ns	*	*	*	ns	*	*	*	*	*	*	*	*
6 All	Parallèle	ns	ns	*	ns	ns	ns	*	*	*	*	ns	ns	ns	*	*	*	*	*

TABLE 5.3 – Valeur significative (c'est-à-dire plus forte qu'attendue sous l'effet du hasard) de la densité normalisée de différents types de voisins (H, R, C) au voisinage des segments de haies pour la basse vallée de la Durance, dans un voisinage croissant de $20m$ à $300m$ et avec une orientation relative donnée (Perpendiculaire ou Parallèle). La significativité est représentée par *, son absence par ns. Le type All correspond à la totalité des haies, sans distinction.

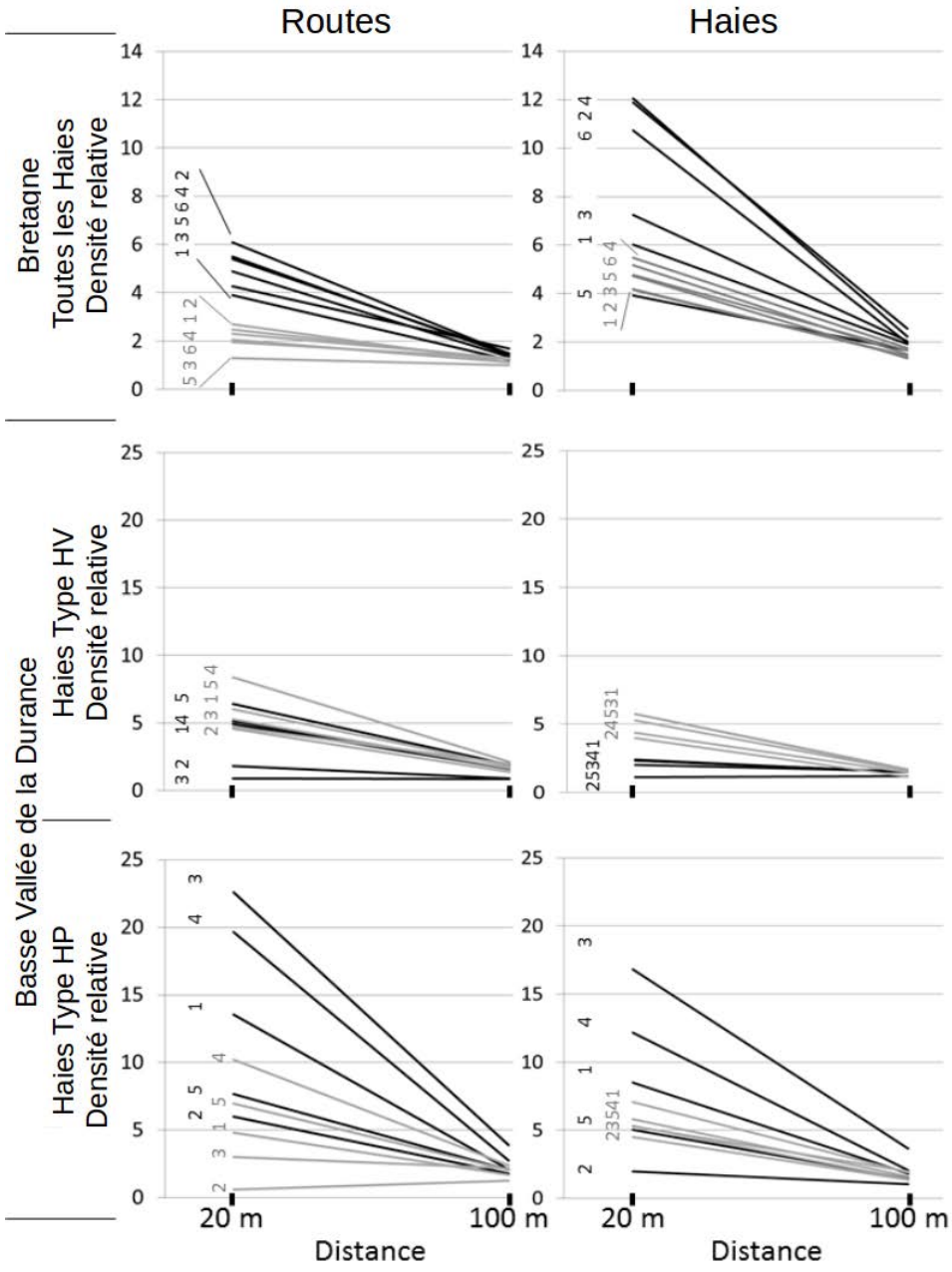


FIGURE 5.8 – Densité relative normalisée de voisins du type routes ou haies, pour deux tailles de voisinages (20 m et 100 m de distance). Les résultats sont présentés pour tous les segments de haies du paysage de Bretagne (en haut), et pour les segments de haies HV (au milieu) ou de haies HP (en bas) pour le paysage de la basse vallée de la Durance. Les lignes grises correspondent aux tendances du changement pour les voisins avec l'orientation relative perpendiculaire et les lignes noires pour les voisins avec une orientation relative parallèle. Chaque ligne correspond à une classe avec son numéro correspondant indiqué sur la gauche.

Hétérogénéité au sein des paysages : l'effet "classe" Il existe encore une certaine hétérogénéité dans le paysage concernant les résultats de la figure 5.8. L'effet du type de voisin, en particulier, dépend de la classe de la cellule (interaction de *V.Types * Class*). En Bretagne, par exemple, les haies sont très fortement associées à d'autres segments de haies, mais moins avec les routes à courte distance dans certaines classes uniquement (par exemple dans les classes 2, 4 et 6 pour les voisins parallèles).

Dans le paysage de la basse vallée de la Durance, les associations de courtes distances sont par ailleurs très fortes avec les voisins perpendiculaires mais pas avec les voisins parallèles dans certaines classes (par exemple, les classes 3 et 4 pour les haies HP) alors qu'il n'y a aucun effet de l'orientation relative des voisins dans les autres classes (Figure 5.8), nous constatons effectivement un effet significatif de l'interaction *Class * R.Orient* à courte distance, jusqu'à $80m$ environ, dans le tableau 5.2. Les classes 3 et 4, qui se différencient des autres classes concernant le comportement des segments de haies HP, sont constituées du plus petit nombre de segments de haies et de routes (tableau 4.2).

Chapitre 6

Apprentissage sur les structures spatiales pour les générer

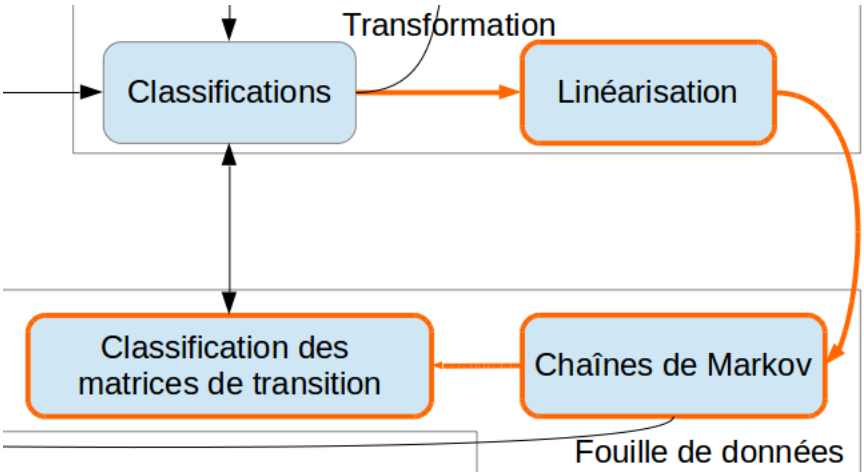


FIGURE 6.1 – Étapes décrites au chapitre 6

Introduction

Ce chapitre présente la deuxième méthode développée pour la caractérisation des structures spatiales des segments et suit le cheminement présenté à la figure 6.1. Nous avons choisi d'utiliser une méthode commune d'apprentissage et de génération, il s'agit des chaînes de Markov. Nous présenterons son principe à la section 6.2 et son application aux données à la section 6.4. Avant, à la section 6.1, nous présentons la méthode de linéarisation choisie pour permettre l'emploi des chaînes de Markov, et sa mise en place sur les données à la section 6.3.

6.1 Linéarisation de l'information spatiale avec le chemin de Hilbert adaptatif

La distribution des segments dans les cellules peut être captée par un chemin de Hilbert adaptatif. Selon le type de la cellule (densité, répartition et orientation des segments), nous nous attendons à des chemins différents. Pour caractériser le chemin de Hilbert adaptatif effectué sur les cellules des données A et B, nous introduisons quelques définitions.

6.1.1 Définition de case et profondeur de découpe dans le chemin de Hilbert adaptatif

Définition : Une case C d'un chemin de Hilbert adaptatif est définie comme une division unique convexe de la cellule initiale C^i obtenue à l'issue du processus.

Pour une case C, nous définissons la profondeur de découpe.

Définition : La profondeur de découpe PdD_C associée à une case C d'un chemin de Hilbert adaptatif CHA de cellule initiale C^i est définie grâce à l'aire A_{C^i} de la cellule initiale C^i et de l'aire A_C de la case C par :

$$PdD_C = \frac{\ln(A_{C^i}) - \ln(A_C)}{2 \times \ln(2)}$$

La cellule initiale possède une profondeur de découpe égale à 0. A chaque fois qu'une case C_a est divisée, les quatre cases obtenues C_b ont une profondeur de découpe égale à la profondeur de découpe de la case mère augmentée de un.

$$PdD_{C_a} + 1 = PdD_{C_b}$$

D'une façon générale, nous pouvons montrer que la profondeur de découpe entre deux cases est reliée par la formule suivante :

$$PdD_{C_{t+n}} = PdD_{C_t} + n$$

où C_{t+n} est de profondeur de découpe $t + n$ et C_t est de profondeur de découpe t, avec $0 \leqslant n$.

Démonstration :

$$\begin{aligned} PdD_{C_t} &= \frac{\ln(A_{C^i}) - \ln(A_{C_t})}{2 \times \ln(2)} \\ &= \frac{\ln(A_{C^i}) - \ln(2\exp^{2n} \times A_{C_b})}{2 \times \ln(2)} \\ &= \frac{\ln(A_{C^i}) - \ln(2\exp^{2n}) - \ln(\times A_{C_b})}{2 \times \ln(2)} \\ &= \frac{\ln(A_{C^i}) - \ln(\times A_{C_b})}{2 \times \ln(2)} - \frac{\ln(2\exp^{2n})}{2 \times \ln(2)} \\ &= PdD_{C_{t+n}} - n \end{aligned}$$

ce qui donne

$$PdD_{C_t} + n = PdD_{C_{t+n}}$$

6.1.2 Définition de temps de parcours et temps d'attente dans le CHA

Nous faisons l'hypothèse que le parcours sur le CHA s'effectue à vitesse constante (arbitrairement fixé à 1). Pour chaque case, nous écrivons cette vitesse comme :

$$\frac{\text{longueur du chemin dans la case}}{\text{temps de parcours dans la case}}$$

Nous pouvons ainsi définir le temps de parcours d'une case, en fonction de la taille du chemin à l'intérieur de celle-ci.

Définition : *Temps de parcours*

Le temps de parcours TdP_C associé à une case C de profondeur de découpe PdD_C est défini par :

$$TdP_C = \frac{1}{2^{PdD_C}}$$

Nous pouvons suite à cela définir le temps de parcours sur tout ou partie du chemin de Hilbert adaptatif.

Extension : Temps de parcours pour un ensemble de cases

Soit $S = \{C_i | i \in [\![1, n]\!]\}$ une suite de cases d'un chemin de Hilbert adaptatif et $\{PdD_i\}_{i \in [\![1,n]\!]}$ l'ensemble des profondeurs de découpe respectives, nous pouvons écrire le temps de parcours pour S comme :

$$TdP_S = \sum_{i=1}^{n} \frac{1}{2^{PdD_{C_i}}}$$

Le temps de parcours d'une case, ainsi défini, garantit la cohérence des temps de parcours lors du calcul pour une suite de cases. En effet, deux cases de profondeur de découpe PdD auront un temps de parcours égal au temps de parcours d'une case de profondeur de découpe $(PdD - 1)$. Ceci permet de respecter la condition de vitesse constante.

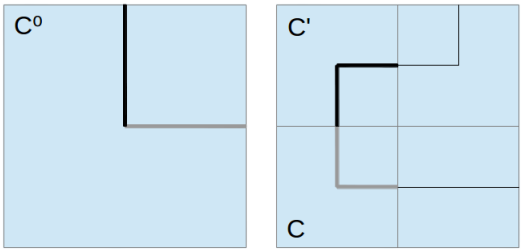


FIGURE 6.2 – Forme additive du temps de parcours

Exemple

Soit C et C' deux cases de profondeur de découpe 3, nous pouvons vérifier que le temps de parcours de ces deux cases est bien égal au temps de parcours de la case C^0 dont elles sont issues, qui est de profondeur de découpe 2

$$TdP_{CC'} = \sum_{1}^{2} \frac{1}{2^3} = \frac{2}{2^3} = \frac{1}{2^2} = TdP_{C^0}$$

Comme nous pouvons le voir sur la figure 6.2, le segment gris dans la case C^0 est de même longueur que le segment gris dans la case C et le segment épais noir dans la case C^0 est de même longueur que le segment épais noir dans la case C'.

Définition : Temps d'attente

Le temps d'attente TdA entre deux cases différentes C et C', ou deux objets appartenant à deux cases différentes $m \in C$ et $m' \in C'$, est défini par

$$TdA_{CC'} = TdA_{mm'} = \frac{TdP_C}{2} + TdP_{CC'} + \frac{TdP'_C}{2}$$

Le temps d'attente entre deux éléments est donc défini comme le temps de parcours entre les deux cases contenant les deux éléments auxquelles nous ajoutons la moitié du temps de parcours de chaque case contenant les éléments.

6.2 Apprentissage par chaînes de Markov

6.2.1 Markov sur le chemin de Hilbert adaptatif

Le parcours d'une cellule par un chemin de Hilbert adaptatif permet d'ordonner les isobarycentres des segments le long du chemin. A chacun de ces isobarycentres peuvent être attachées les valeurs des trois variables, *Longueur du segment*, *Angle du segment* et *Temps d'attente*. Dans ce cas, le *Temps d'attente* représente le temps, sur le chemin, entre les isobarycentres successifs. Ces trois variables, pour une cellule, peuvent être représentées par une chaîne de valeur comme illustrée sur le tableau 6.1.

	Segment 1	Segment 2	Segment 3	...	Segment n
Longueur du segment	152,2	78,5	415,1	...	98,2
Angle du segment	1,23	2,32	1,4	...	0,42
Temps d'attente	0,014	0,0025	0,0625	...	0,125

TABLE 6.1 – Exemple de séquences de valeurs des variables *Longueur du segment*,*Angle du segment* et *Temps d'attente* le long d'un chemin de Hilbert adaptatif sur une cellule

À partir de chaque séquence de valeurs des variables *Longueur du segment*,*Angle du segment* et *Temps d'attente*, nous calculons la matrice de transition associée à chacune d'elles. Nous obtenons ainsi une matrice de transition pour chaque cellule et pour la variable.

6.2.2 Distance entre matrices de transition

Nous avons choisi de déterminer une nouvelle classification basée sur les matrices de transition des variables *Longueur du segment*, *Angle du segment* ou *Temps d'attente*. Pour ce faire, nous utilisons une distance entre matrice capable de prendre en compte le fait que les vecteurs d'états permettant la création de ces matrices sont possiblement très différents l'un de l'autre. La distance ainsi définie est capable de capter la proximité des états et le cas échéant leurs différences.

Notons $M_{\mathcal{C}_1}$ la matrice de transition associée à une des variables *Longueur du segment*, *Angle du segment* et *Temps d'attente* dans la cellule $\mathcal{C}_1$, et $\{E_i^1\}_{i\in\mathbb{N}}$ l'ensemble des états de cette variable dans la cellule. Pour la même variable, notons $M_{\mathcal{C}_2}$ la matrice de transition associée, pour la cellule $\mathcal{C}_2$, et $\{E_i^2\}_{i\in\mathbb{N}}$ l'ensemble des états de cette variable dans la cellule.

Tout d'abord, nous déterminons le nombre d'états communs à l'ensemble des états, $\{E_i^1\}_{i\in\mathbb{N}}$ et $\{E_i^2\}_{i\in\mathbb{N}}$, notés n_c. Notons respectivement $A = \{a_{i,j}\}_{i,j\in[\![1,n_c]\!]^2}$ et $B = \{b_{i,j}\}_{i,j\in[\![1,n_c]\!]^2}$,

les sous matrices de $M_{\mathcal{C}_1}$ et $M_{\mathcal{C}_2}$, composées des états communs. Nous calculons ensuite la différence entre ces deux matrices, notée $Diff$, par la formule :

$$\texttt{Diff} = \sqrt{\sum_{j=1}^{n_c}\sum_{i=1}^{n_c}(a_{ij} - b_{ij})^2}$$

Nous définissons finalement la distance entre les deux matrices, notée `DistM` par :

$$\texttt{DistM} = \texttt{Diff} \times \frac{n_{E^1} + n_{E^1}}{2 \times n_c}$$

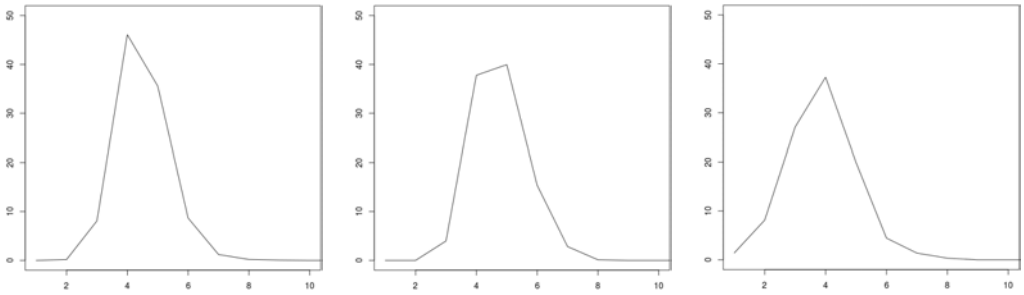
où n_{E^1} (respectivement n_{E^2}) est le nombre d'états associés à la matrice $M_{\mathcal{C}_1}$ (respectivement $M_{\mathcal{C}_2}$).

6.3 Création du chemin de Hilbert adaptatif pour les données A et B

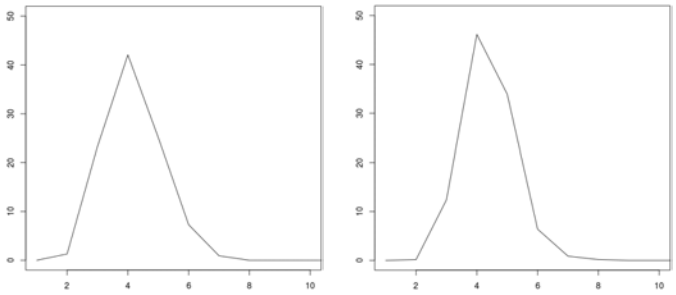
Le chemin de Hilbert adaptatif a été mis en œuvre sur l'ensemble des cellules des données A et B en considérant l'isobarycentre des segments de haies auquel nous avons rattaché les valeurs de longueurs et d'angles du segment associé.

6.3.1 Caractérisation des chemins de Hilbert adaptatifs par les profondeurs de découpe

La variable *Profondeur de Découpe* prend ses valeurs dans le segment $[\![1, 9]\!]$ pour l'ensemble des données A (figure 6.3), et dans le segment $[\![2, 11]\!]$ pour les données B (figure 6.4). La profondeur de découpe est donc un peu plus élevée dans les cellules des données B que dans les cellules des données A. Cette différence peut s'expliquer par le nombre de points plus important dans les cellules construites à partir des données B que dans celles construites à partir des données A. Le nombre moyen de segments de haies dans une cellule, est compris entre 58 et 237 pour les données A et entre 167 et 406 pour les données B, selon les classes (cf. tableau 4.2). Nous déterminons pour chaque classe la distribution de la variable *Profondeur de Découpe* pour une cellule moyenne, représentant la classe. Cette distribution est calculée en déterminant la distribution sur chaque cellule de la classe et en la moyennant sur l'ensemble des cellules. Cette cellule moyenne sera utilisée pour déterminer la valeur de la variable *Profondeur de Découpe* associé à la classe.



(a) Classe 1 (b) Classe 2 (c) Classe 3



(d) Classe 4 (e) Classe 5

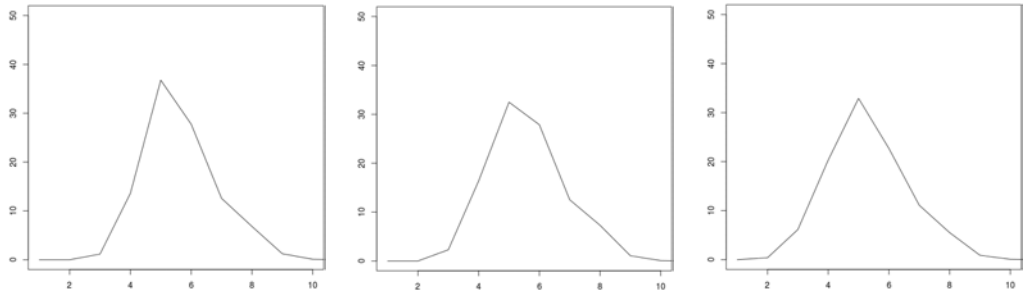
FIGURE 6.3 – Proportion de la distribution de la variable *Profondeur de Découpe* pour une cellule moyenne, selon chaque classe, pour les données A. La profondeur de découpe est en abscisse. L'ordonnée est exprimée en pourcentage, par rapport à l'ensemble des valeurs prises par la variable *Profondeur de Découpe* sur l'ensemble de la cellule moyenne.

Nous précisons que le mode de la variable *Profondeur de Découpe* dont nous parlons ensuite, est aussi égal à sa médiane dans toutes les cellules et toutes les classes. Pour rappel, le mode correspond à la valeur de la variable *Profondeur de Découpe* la plus présente.

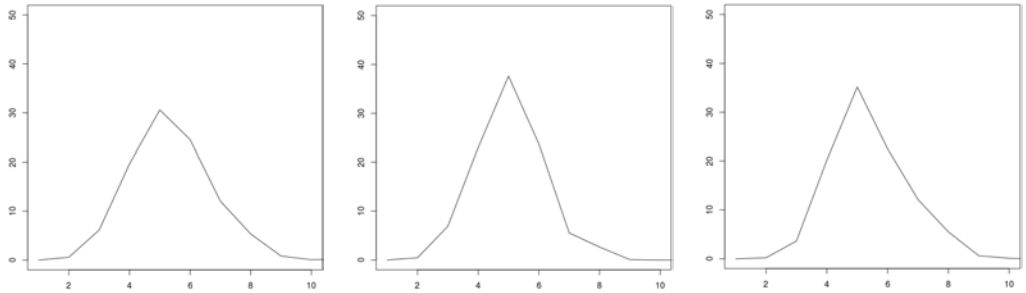
Dans la figure 6.3, nous remarquons que le mode de la variable *Profondeur de Découpe* ne prend que 2 valeurs différentes (4 et 5) dans le segment $[\![1, 9]\!]$ pour l'ensemble des classes des données A. L'examen de la figure 6.4 montre que le mode de la variable *Profondeur de Découpe* ne prend qu'une valeur (5) dans le segment $[\![2, 11]\!]$ pour l'ensemble des cellules des données B.

Plus particulièrement, sur les 35 cellules des données B, il n'existe que 5 cellules pour lesquelles la variable *Profondeur de Découpe* a un mode à 6, pour les autres, le mode est à 5. Il semble donc qu'il n'y ait que peu de variabilité entre les cellules. Au contraire, sur les 64 cellules des données A, le mode de la variable *Profondeur de Découpe* est plus étalé. En effet, 10 (respectivement 2 et 1) cellules ont un mode à 5 (respectivement 3 et 2) pour la variable *Profondeur de Découpe*.

Pour les données A, seule la classe 4 contient des cellules qui ont toute le même



(a) Classe 1 (b) Classe 2 (c) Classe 3



(d) Classe 4 (e) Classe 5 (f) Classe 6

FIGURE 6.4 – Proportion de la distribution de la variable *Profondeur de Découpe* pour une cellule moyenne, selon chaque classe, pour les données B. La profondeur de découpe est en abscisse. L'ordonnée est exprimé en pourcentage, par rapport à l'ensemble des valeurs prises par la variable *Profondeur de Découpe* sur l'ensemble de la cellule moyenne.

mode pour la variable *Profondeur de Découpe*. Si nous associons la valeur du mode le plus fréquent, parmi les cellules d'une classe, à cette classe, nous pouvons dire que les classes 1, 3, 4 et 5 ont une profondeur de découpe égale à 4 tandis que la classe 2 possède une profondeur de découpe égale à 5. Pour les données B, toutes les cellules possèdent le même mode pour la variable *Profondeur de Découpe*. Nous pouvons donc dire que toutes les classes ont une profondeur de découpe égale à 5. Les profondeurs de découpe, associées à chaque classe, ne semblent pas élevées. Des profondeurs de découpe de valeurs 4 ou 5 correspondent en effet respectivement à des cases de largeurs égales à $\frac{1}{16^{ème}}$ et $\frac{1}{32^{ème}}$ de la largeur de la cellule initiale (environ $70m$ et $35m$ de coté).

Mise à part la classe 2 pour les données A (figure 6.3(b)), les différentes classes de chaque jeu de données possèdent une valeur identique de profondeur de découpe. Nous pouvons supposer que cela correspond à une implantation des segments propre à chaque jeu de données. Ainsi, le chemin de Hilbert adaptatif offre une autre caractérisation de la différence entre le paysage de Bretagne et celui de la basse vallée de la Durance.

6.3.2 Caractérisation des chemins de Hilbert adaptatifs par les temps d'attente

Nous déterminons, pour chaque cellule, la médiane de la variable *temps d'attente*. Ensuite, pour chaque classe, nous calculons la moyenne et l'écart type des médianes de la variable *temps d'attente* des cellules de la classe. Nous récapitulons ces résultats dans les tableaux 6.2 et 6.3. Nous remarquons tout d'abord que les valeurs sont très différentes entre les deux jeux de données, allant du simple au double entre les données A et B.

Plus particulièrement, la variable *temps d'attente*, pour les données A, prend 129 valeurs différentes comprises entre $6,25\times10^{-3}$ et $2,15\times10^{-2}$. La moyenne de la variable *temps d'attente*, pour les classes 1, 2, 4, et 5 est égale à $\frac{1}{2^4}$ (Ligne 1 dans le tableau 6.2). Ceci correspond au temps de parcours d'une case ayant une valeur de *Profondeur de Découpe* égale à 4. Pour la classe 3, le mode de la variable *temps d'attente* est égale à $\frac{1}{2^3}$ (Ligne 1 dans le tableau 6.2). Ce qui correspond au temps de parcours d'une case ayant une valeur de *Profondeur de Découpe* égale à 3. Ce qui veut dire que, pour le paysage A, les centres de haies sont éloignés, majoritairement, d'une case de profondeur de découpe 3 ou 4. Il s'agit là d'un possible indice de caractérisation du paysage A, permettant la comparaison entre paysages.

Données A	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Valeur du mode de *temps d'attente* ($\times10^{-2}$)	6,25	6,25	12,5	6,25	6,25
Moyenne de *temps d'attente* ($\times10^{-2}$)	8,08	6,88	14,6	11,2	8,98
Écart type de *temps d'attente* ($\times10^{-2}$)	1,51	1,87	4,96	2,20	1,78

TABLE 6.2 – Récapitulatif du nombre de valeur, du mode, de la moyenne et de l'écart type de *temps d'attente*, selon les classes dans les données A

La classe 2 des données A possède la plus faible moyenne de la variable *temps d'attente*, les cellules de cette classe présentent donc une densité moyenne de points sur le chemin de Hilbert adaptatif plus élevée que les cellules des autres classes. Ceci peut être expliqué par la différence de type de cultures dans les cellules de la classe 2 et les autres. En effet, le maraîchage est le type de cultures prépondérant dans les cellules de la classe 2 alors que les autres surfaces sont utilisées majoritairement pour cultiver des vergers.

Nous remarquons également que les classes 1 et 5 (tableau 6.2) ont des valeurs similaires pour les indicateurs de la variable *temps d'attente*, ce qui suggère un motif équivalent dans les cellules de ces deux classes. Nous pouvons imaginer que ces cellules possèdent une implantation des haies similaires.

Pour l'ensemble des indicateurs présents dans le tableau 6.2, les autres classes sont rangées dans un ordre identique. Si nous prenons l'ordre croissant, les classes s'organisent comme ceci : 1, 5, 4, 3. Cet ordre peut se comparer à celui observé sur l'implantation géographique des cellules de chaque classe (cf figure 4.5(a) section 4.2). Nous avons un gradient entre la classe 1, principalement située dans le Nord-Est du domaine d'études et la classe 3, située dans le Sud-Ouest du domaine d'étude, en passant par la classe 5 puis la classe 4. Ceci confirme la caractérisation possible des différences entre classes de cellules par l'utilisation du chemin de Hilbert adaptatif.

La variable *temps d'attente*, pour les données B, prend 293 valeurs différentes comprises entre $0,49 \times 10^{-3}$ et $1,02$. Le mode de la variable *temps d'attente* a la même valeur pour toutes les classes : $3,13 \times 10^{-2} \approx \frac{1}{2^5}$ (Ligne 1 dans le tableau 6.3). Ce qui correspond au temps de parcours d'une case ayant une valeur de *Profondeur de Découpe* égale à 5. Donc, pour le paysage B, les centres de haies sont éloignés, majoritairement, d'une case de profondeur de découpe 5.

Données B	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Valeur du mode de *temps d'attente* ($\times 10^{-2}$)	3,13	3,13	3,13	3,13	3,13	3,13
Moyenne de *temps d'attente* ($\times 10^{-2}$)	4,31	4,83	5,73	5,42	5,86	5,25
Écart type de *temps d'attente* ($\times 10^{-2}$)	0,37	0,70	0,4	1,25	0,78	0,74

TABLE 6.3 – Récapitulatif du nombre de valeur, du mode, de la moyenne et de l'écart type de *temps d'attente*, selon les classes dans les données B

D'après le tableau 6.3, les classes 1 et 2 possèdent une moyenne sur la variable *Temps d'attente* inférieure à 5×10^{-2}. En visualisant la position géographique des cellules de ces classes (cf figure 4.5(b) section 4.2), nous remarquons qu'elles sont voisines et situées dans la zone Sud-Ouest du domaine d'étude, correspondant à une zone de bocage traditionnel. Les cellules se situant dans la zone remembrée (Centre et Nord) possèdent une moyenne sur la variable *Temps d'attente* supérieure à 5×10^{-2}.

Nous pouvons donc dire que le paysage A et le paysage B ne se comportent pas de la même manière face au découpage par le chemin de Hilbert adaptatif. Même si la différence semble minime, ces éléments montrent que les deux paysages n'appartiennent pas au même type général, considérant l'implantation des haies.

6.4 Utilisation des chaînes de Markov sur les informations linéarisées pour les données A et B

Sur les chemins obtenus, peuvent être appliquées les chaînes de Markov. Les trois variables *Temps d'attente*, *Angle du segment* et *Longueur du segment* sont affectées aux observations séquentielles construites.

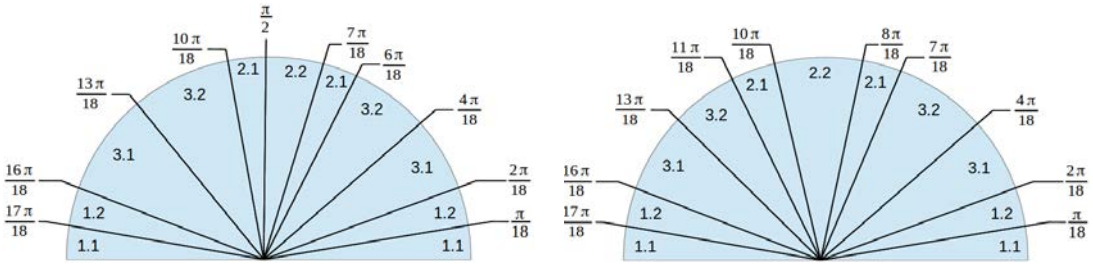
6.4.1 Calcul des matrices de transition

Les variables *Temps d'attente*, *Angle du segment* et *Longueur du segment* ne sont pas corrélées, tant d'un point de vue global qu'au sein de chaque cellule (cf. chapitre 4). Nous avons donc décidé de créer une matrice de transition pour chacune de ces variables et pour chaque cellule. Pour chaque cellule, les réalisations de la variable *Temps d'attente* sont en nombre conséquent pour des valeurs discrètes, ce qui permet de construire une matrice de transition contenant un niveau d'information acceptable. Nous n'avons pas choisi de modifier cela, en faisant des classes de valeurs pour la variable *Temps d'attente* afin de ne pas risquer d'écraser la variabilité intra-cellule.

En revanche, les variables *Angle du segment* et *Longueur du segment* sont à valeurs continues dans leur intervalle d'existence, et ainsi le nombre de réalisations pour chaque valeur prise est faible, voire égal à un. De ce fait, les matrices de transition issues des variables continues peuvent être évidentes, et ne présenter que des 0 et des 1. C'est pourquoi, nous avons décidé de discrétiser ces données. Pour ce faire, nous avons repris les classes d'angles introduites dans la section 4.3 pour traiter la variable *Angle du segment*, et nous avons utilisé les quantiles pour discrétiser les valeurs de la variable *Longueur de segment*. Nous pouvons voir sur le tableau 6.4 le résumé de ces informations et la création des six classes pour la variable *Longueur du segment*.

Les figures 6.5(a) et 6.5(b) permettent de visualiser les classes d'angles suivant le jeu de données. Les classes d'angles portent un numéro de la forme $x.y$ correspondant au type de haies associées et à l'intervalle considéré. Ainsi, les classes de la forme $1.y$, respectivement $2.y$ et $3.y$, concernent respectivement les haies HP, les haies HV et les autres haies. Nous choisissons de séparer chaque intervalle décrivant les classes d'angles

définies dans la section 4.3 en deux intervalles égaux. Nous notons les nouvelles classes d'angle de la forme $x.1$ quand il s'agit des angles les plus proches de π ou $-\pi$, et de la forme $x.2$ sinon.



(a) Données A (b) Données B

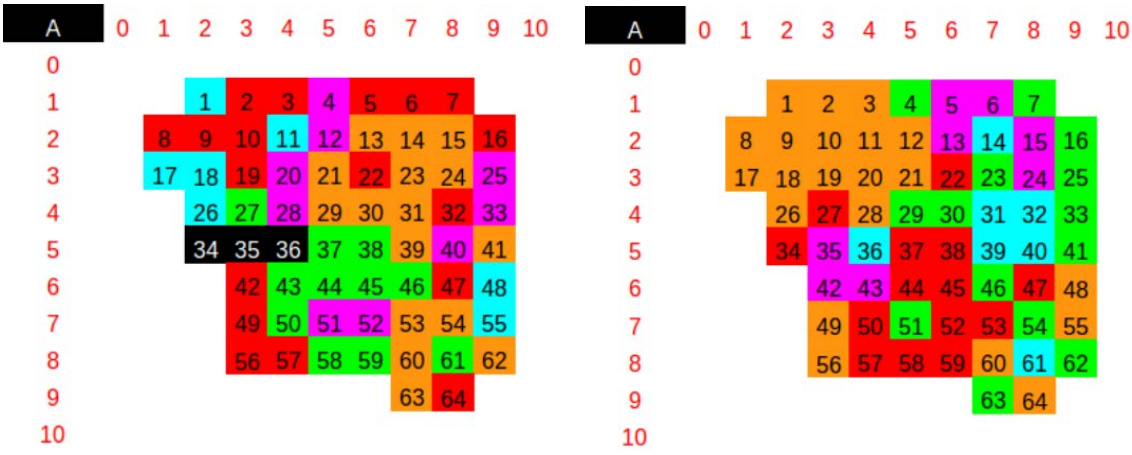
FIGURE 6.5 – Représentation des classes d'angle pour les données.

Données A Intervalle des quantiles	Classe des longueurs	Données B Intervalle des quantiles
$[0,27; 36,1[$	1	$[0,36; 5,08[$
$[36,1; 57,6[$	2	$[5,08; 7,57[$
$[57,6; 82,6[$	3	$[7,57; 10,7[$
$[82,6; 1,12 \times 10^2[$	4	$[10,7; 15,2[$
$[1,12 \times 10^2; 1,61 \times 10^2[$	5	$[15,2; 23,7[$
$[1,61 \times 10^2; 1,23 \times 10^3[$	6	$[23,7; 3,75 \times 10^2[$

TABLE 6.4 – Tableau récapitulatif des valeurs de classes pour la variable *Longueur du segment*

6.4.2 Classification des matrices de transition

Cette classification s'appuie sur la distance présentée à la sous-section 6.2.2. Les résultats concernant la variable *Temps d'attente* ne sont pas affichés car ils ne présentent pas de motif particulier.



(a) Variable *Longueur du segment* (b) Variable *Angle du segment*

FIGURE 6.6 – Représentation graphique du résultat de la classification des matrices de transition suivant les deux variables, pour les données A. Chaque couleur correspond à une classe.

Variable *Longueur du segment*

— Pour le paysage A :

D'après la figure 6.6(a), la classification selon la variable *Longueur du segment* permet de distinguer cinq types de cellules pour les données A. Les cellules numérotées 34, 35 et 36 apparaissent comme des cellules atypiques dans la classification. Tout d'abord, les cellules de couleur verte se concentrent au Sud-Ouest de la zone d'étude, tandis que les cellules de couleur orange se concentrent à l'Est de la zone d'étude, et plus particulièrement au Nord. Les cellules de couleur rouge semblent ceinturer la zone d'étude, surtout au Nord et à l'Ouest. Les cellules de couleur bleue ne sont présentes qu'aux extrêmes Nord-Ouest et Sud-Est de la zone. Finalement, les cellules de couleur magenta ne présentent pas de positions particulières.

Afin d'étudier cette classification plus avant, nous reprenons la classification des cellules définie dans le chapitre 4, et présentée, pour les données A, sur la figure 4.5(a). Nous constatons que les deux classifications possèdent une séparation entre le Nord-Est et le Sud-Ouest de la zone. Mais si pour la classification présentée dans le chapitre 4, le gradient est faible, pour celle présentée au dessus, il est plus élevé, sans classes

servant d'intermédiaire. Les deux classifications distinguent également une classe aux extrémités Nord-Ouest et Sud-Est de la zone, mais celle-ci est moins nette pour la seconde classification. Enfin, la seconde classification crée une classe regroupant des cellules aux frontières de la zone d'étude (de couleur rouge), ce que ne fait pas la première classification.

— Pour le paysage B :

La nouvelle classification sur les matrices de transition de la variable *Longueur du segment* crée six classes de cellules (Figure 6.7(a)). Tout d'abord, les cellules de couleur verte se situent exclusivement au Sud de la zone d'étude, dans la partie du bocage historique (cf. sous-section 4.1.1), qui est majoritairement couverte par une seule classe. Seules les cellules numérotées 34 et 35, en bordure de zone, ne sont pas dans la même classe, mais elles font parties des cellules dont toutes les haies n'ont pas été numérisées. Le reste des classes ne semble pas être positionné de façon spécifique.

Afin d'étudier cette classification plus avant, nous reprenons la classification présentée, pour les données B, sur la figure 4.5(b). La seule similitude entre les deux classifications vient du traitement de la partie bocagère historique du Sud, partie très caractéristique. Dans les deux cas, les cellules de cette partie appartiennent à des classes qui ne sont pas présentes sur le reste de la zone. La partie remembrée, quant à elle, ne semble pas être traitée de la même façon par les deux classifications.

Variable *Angle du segment*

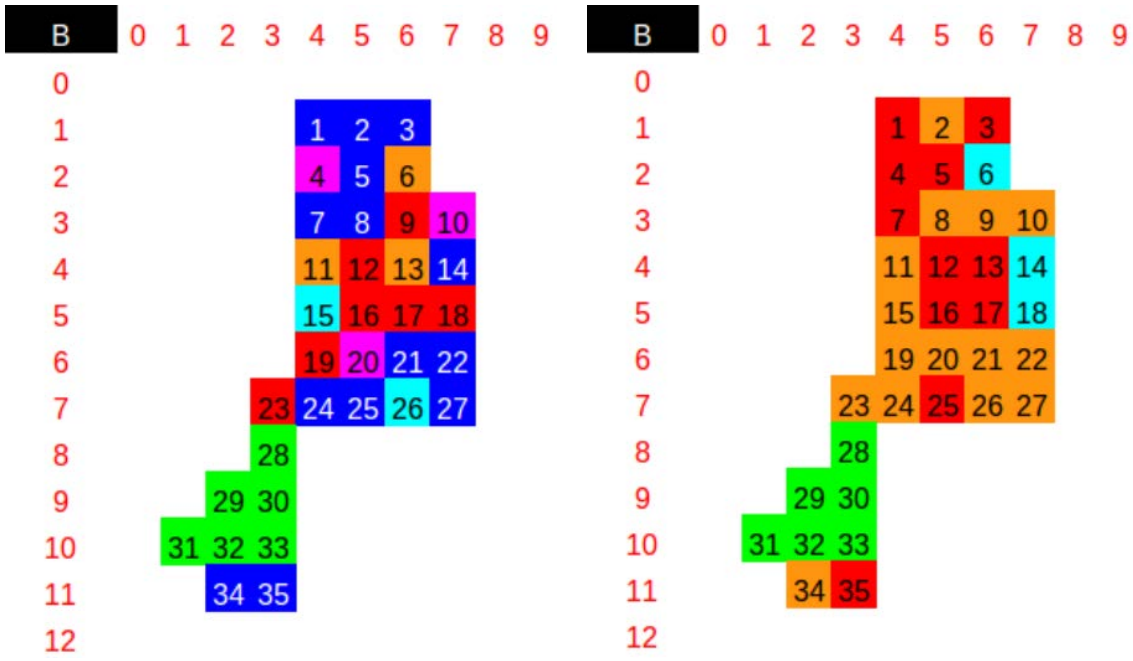
— Pour le paysage A :

Sur la figure 6.6(b), la classification selon la variable *Angle du segment* permet également de distinguer des zones d'implantation de cellules ayant les mêmes caractéristiques. Il y a là, cinq classes de cellules. Tout d'abord, les cellules de couleur orange sont situées essentiellement au Nord-Ouest de la zone d'étude, et les cellules de couleur rouge au Sud (Centre) de la zone. Quant aux cellules de couleur bleue et celles de couleur verte, elles se mélangent au Nord-Est de la zone.

Si nous comparons ces résultats avec la classification de la figure 4.5(a), nous constatons tout d'abord que la zone Nord-Ouest n'est pas réservée à une petite classe comme pour la première classification. De plus, le gradient Nord-Est Sud-Ouest a quasiment disparu. Il n'existe pas de similitudes fortes entre ces deux classifications, même si la disposition générale des classes ne semble pas trop éloignée entre les deux.

— Pour le paysage B :

D'après la figure 6.7(b), la nouvelle classification sur les matrices de transition de la variable *Angle du segment* crée quatre classes de cellules. Tout comme précédemment,



(a) Variable *Longueur du segment* (b) Variable *Angle du segment*

FIGURE 6.7 – Représentation graphique du résultat de la classification des matrices de transition suivant les deux variables, pour les données B. Chaque couleur correspond à une classe.

la zone Sud abritant le bocage historique est composée d'une seule classe, qui n'est présente qu'ici. Le même constat apparaît pour les cellules numérotées 34 et 35. Mais contrairement à la classification précédente, une démarcation se fait entre les autres classes. Tout d'abord, une classe se situe préférentiellement au centre de la zone d'étude (cellules de couleur orange), c'est-à-dire, au Sud de la partie remembrée. Une classe se situe plutôt au Centre et au Nord de la partie remembrée (cellules de couleur rouge). Une troisième classe très réduite se situe sur la frontière Est de la zone (cellules de couleur bleue).

Si nous comparons ces résultats avec la classification de la figure 4.5(b), nous constatons comme précédemment, la similitude des deux classifications dans la zone Sud. La nouvelle classification présente, comme l'ancienne mais dans une moindre mesure, une distinction entre le Nord et le Sud de la partie remembrée. Le Sud et l'Ouest de la zone remembrée sont classées dans une même classe dans la nouvelle classification mais dans deux classes distinctes dans l'ancienne. Le Nord de la partie remembrée, homogène avec l'ancienne classification, ne l'est plus avec la nouvelle, et abrite trois classes.

En conclusion, suite à ces nouvelles classifications suivant les matrices de transitions des variables *Longueur du segment* et *Angle du segment*, nous constatons qu'elles

peuvent créer des classes homogènes de cellules, situées dans des zones géographiques restreintes. Selon la variable choisie, les classes ne sont pas les mêmes, mais la relation géographique subsiste. Même si les nouvelles classifications se rapprochent de l'ancienne pour les données A mais pas pour les données B, il est important de noter que dans tous les cas, la partie Sud des données B est traitée de manière identique. Il semblerait donc, que l'utilisation des chemins de Hilbert adaptatifs couplés avec les chaînes de Markov, permet de créer un indicateur utile à la caractérisation de certains motifs spatiaux.

Troisième partie

Simulation et évaluation

Chapitre 7

Génération de structures de segments dans l'espace

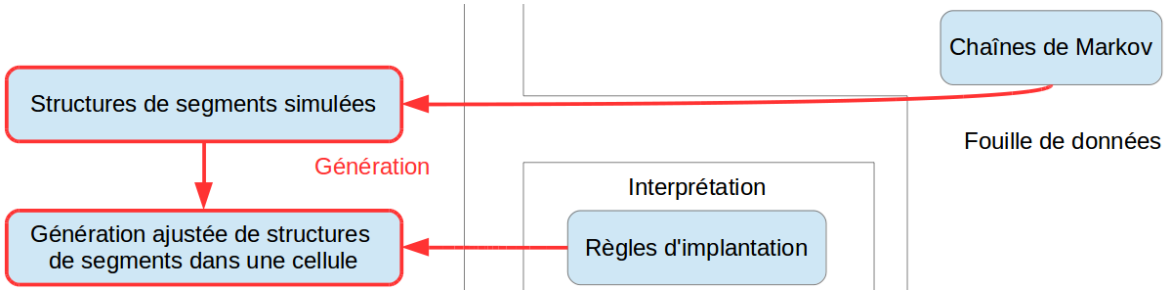


FIGURE 7.1 – Étapes décrites au chapitre 7

Introduction

Deux méthodes ont été présentées pour caractériser des structures de voisinages. Ces caractéristiques, apprises sur des paysages réels, servent à la création de paysages virtuels. Dans ce chapitre, nous présentons une méthode couplant des chemins de Hilbert classiques et des chaînes de Markov afin de générer des segments suivant les règles apprises sur les cellules issues de données réelles. Nous suivons le cheminement présenté à la figure 6.1. Nous présentons à la section 7.1 la première stratégie que nous avons développée pour la génération de segments de haies dans un paysage. Nous présentons ensuite à la section 7.2 des résultats de comparaison entre des cellules simulées et les cellules réelles initiales. La comparaison est faite tout d'abord sur des indicateurs statistiques (sous-section 7.2.1), puis d'après les chemins de Hilbert adaptatifs (sous-section 7.2.2). Nous présenterons finalement à la section 7.3 une nouvelle méthode de

génération, ajustée à partir des connaissances du domaine, issues des caractérisations effectuées au chapitre 5.

7.1 Stratégie

L'idée générale est de reproduire dans une cellule vide, un chemin suivant les caractéristiques d'une classe de cellules puis de générer des segments de haies en utilisant les modèles appris sur cette même classe.

Processus

Données

Soit un paysage réel défini comme l'ensemble des segments présent sur un territoire. Soit $\mathbb{P} = (\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n)$ l'ensemble des cellules créées à partir de ce paysage réel (méthode chapitre 4). Soit $\mathbb{H} = (\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_l)$ l'ensemble des l classes de cellules créées à partir de $\mathbb{P}$ (méthode chapitre 4). Soit $PdD_{\mathcal{H}_j}$ la profondeur de découpe associée à la classe $\mathcal{H}_j$ (méthode chapitre 5)

Nous définissons alors, pour chaque classe $\mathcal{H}_j$, un triplet, base du processus de génération, noté

$$\{\mathcal{H}_j, PdD_{\mathcal{H}_j}, \{\mathcal{C}_{1_j}, \ldots, \mathcal{C}_{i_j}\}\} \in \mathcal{P}(\mathbb{P}) \times \mathbb{N} \times \mathcal{P}(\mathbb{P})$$

où $i \in [\![1; n]\!]$ et $j \in [\![1; l]\!]$

Étape 1 : Création d'une cellule vide

La première étape consiste à mettre en place le support de la génération. Pour cela, nous déterminons la *probabilité de présence* de chaque classe dans le paysage (ligne 1, algorithme 2), notée *pp*,

$$\forall j \in [\![1; l]\!] \; ; \; pp(\mathcal{H}_j) = \frac{cardinal(\mathcal{H}_j)}{cardinal(\mathbb{P})}$$

Nous choisissons aléatoirement une classe $\mathcal{H}$ dans l'ensemble $\{\mathcal{H}_j\}_{j \in [\![1;l]\!]}$ suivant les *probabilités de présence* associées $pp(\mathcal{H}_i)$ (ligne 2, algorithme 2). Une fois la classe $\mathcal{H}$ choisie, nous utilisons la valeur de la *Profondeur de Découpe* $PdD_{\mathcal{H}}$ associée à la classe $\mathcal{H}$ afin de créer un chemin de Hilbert classique, noté CHC, sur une cellule vide $\mathcal{C}_0$, de taille identique aux cellules de $\mathbb{P}$(ligne 3, algorithme 2).

Algorithm 2: Création de $\mathcal{C}_0$

Data: $\{\mathcal{H}_j, PdD_{\mathcal{H}_j}, \{\mathcal{C}_1, \ldots, \mathcal{C}_{p_j}\}\} \in \mathcal{P}(\mathbb{P}) \times \mathbb{N} \times \mathcal{P}(\mathbb{P})$
Result: $\mathcal{H}$; $\mathcal{C}_0$
begin
  for $\{\ j = 1\ ;\ j \leqslant l\ ;\ j{+}{+}\ \}$ **do**
1 $pp(\mathcal{H}_i) = \frac{card(\mathcal{H}_i)}{card(\mathbb{P})}$
2 $\mathcal{H} \longleftarrow$ `Random`$(\{\mathcal{H}_j\}_j, \{pp(\mathcal{H}_j)\}_j)$
   $PdD \longleftarrow (\mathcal{H}; PdD_{\mathcal{H}})$
3 $\mathcal{C}_0 \longleftarrow CHC(PdD_{\mathcal{H}})$

Étape 2 : Simulation du positionnement du milieu des segments

Dans la classe $\mathcal{H}$, nous choisissons, suivant une loi uniforme, une cellule $\mathcal{C}$. Chaque cellule est associée à un triplet de matrice (cf. méthode dans la sous-section 6.4.1), noté :

$$\mathcal{T}_M^{\mathcal{C}} = (M_{TdP}^{\mathcal{C}}, M_{\theta}^{\mathcal{C}}, M_L^{\mathcal{C}})$$

où

- $M_{TdP}^{\mathcal{C}}$ est la matrice de transition de la variable *Temps de Parcours* sur le CHA de $\mathcal{C}$. Nous notons $V_{TdP}^{\mathcal{C}} = \{TdP_1^{\mathcal{C}}, TdP_2^{\mathcal{C}}, \ldots, TdP_k^{\mathcal{C}}\}$, le vecteur des états de la variable, avec k le nombre d'états de la variable *Temps de Parcours* sur $\mathcal{C}$.
- $M_{\theta}^{\mathcal{C}}$ est la matrice de transition de la variable *Angle* sur le CHA de la cellule $\mathcal{C}$. Les états de la variable sont les classes d'angles définies à la sous-section 4.3.1. Nous notons le vecteur d'états par $V_{\theta}^{\mathcal{C}} = \{\theta_1^{\mathcal{C}}, \theta_2^{\mathcal{C}}, \ldots, \theta_s^{\mathcal{C}}\}$ avec s le nombre d'états de la variable *Angle*.
- $M_L^{\mathcal{C}}$ est la matrice de transition de la variable *Longueur* sur le CHA de la cellule $\mathcal{C}$. Les états de la variable sont les classes de longueur définies à la sous-section 6.4.1. Nous notons le vecteur d'états par $V_L^{\mathcal{C}} = \{L_1^{\mathcal{C}}, L_2^{\mathcal{C}}, \ldots, L_a^{\mathcal{C}}\}$ avec a le nombre d'états de la variable *Longueur*.

A la première étape nous choisissons, suivant une loi uniforme, une valeur TdP_1 dans le vecteur d'états des temps de parcours $V_{TdP}^{\mathcal{C}}$(ligne 1, algorithme 3).

Pour placer le point généré S_1 milieu du segment, nous parcourons le CHC sur $\mathcal{C}_0$

d'une longueur équivalente à TdP_1 depuis le début de CHC (ligne 2, algorithme 3). Une fois la distance parcourue, nous obtenons l'emplacement d'une case de CHC, il existe alors deux méthodes pour placer le point :

— le point est placé exactement le long du chemin qui relie le centre des cases ;
— le point est placé aléatoirement suivant une loi uniforme dans la case.

Pour les étapes suivantes, à partir du temps de parcours attaché TdP_c précédemment tiré, nous choisissons TdP_{c+1} d'après les probabilités de transition attachées à TdP_c (ligne 4, algorithme 3) et ceci, tant que la somme des temps de parcours est plus petite que le temps de parcours total du CHC, noté TdP_{CHC} (ligne 3, algorithme 3). Puis, nous parcourons le CHC d'une distance équivalente à TdP_{c+1} depuis le point courant S_c (ligne 6, algorithme 3). Ceci nous indique une case du CHC et nous plaçons le nouveau point S_{c+1} suivant la même méthode qu'à la première étape.

Nous obtenons ainsi un vecteur de point $V_S^{\mathcal{C}_0} = (S_1^{\mathcal{C}_0}, S_2^{\mathcal{C}_0}, \ldots, S_t^{\mathcal{C}_0})$ représentant l'isobarycentre des t segments générés dans la cellule $\mathcal{C}_0$.

Remarque : Dans le cas particulier où le temps de parcours tiré à l'étape i ne possède que des probabilités de transitions nulles, le temps de parcours à l'étape $i+1$ est tiré sur l'ensemble des états selon une loi uniforme (comme pour le choix initial).

Étape 3 : Simulations indépendantes pour les variables *Longueur du segment* et *Angle du Segment*

Les états des matrices de transitions $M_\theta^{\mathcal{C}}$ et $M_L^{\mathcal{C}}$ étant des classes de valeurs, cette étape permet de simuler une valeur de classe, pour les variables *Longueur du segment* et *Angle du Segment*.

Nous notons indifféremment $C(.)^{\mathcal{C}}$ pour les valeurs de classe qu'il s'agisse des variables *Longueur du segment* ou *Angle du Segment* :

— $1^{ère}$ étape : Pour $S_1^{\mathcal{C}_0}$, nous choisissons suivant une loi uniforme (lignes 1 et 2, algorithme 4) une valeur $C(.)^{\mathcal{C}_0}$ parmi le vecteur d'état associé (cf. méthode dans la sous-section 6.4.1)
— étapes suivantes : nous supposons que la valeur $C(.)_i^{\mathcal{C}_0}$ a été attribuée, par simulation, au point $S_i^{\mathcal{C}_0}$. Nous choisissons, suivant les probabilités de transitions (dans la matrice de transition associée) attachées à l'état $C(.)_i^{\mathcal{C}_0}$ (exemple figure 7.2), une valeur $C(.)_{i+1}^{\mathcal{C}_0}$ (lignes 3 et 4 , algorithme 4). Elle sera attribuée au point $S_{i+1}^{\mathcal{C}_0}$. Cette opération s'effectue jusqu'au point $S_{t-1}^{\mathcal{C}_0}$.

Algorithm 3: Simulation de la position du milieu des segments dans $\mathcal{C}_0$

Data: $\{(M^{\mathcal{C}\prime}_{TdP}; V^{\mathcal{C}\prime}_{TdP}) \mid \mathcal{C}\prime \in \mathcal{H}\}$; $\mathcal{C}_0$; $\{\mathcal{C}_j\}_j = \{\mathcal{C} \mid \mathcal{C} \in \mathcal{H}\}$; *Méthode*
Result: $\mathcal{C}$; $V_S^{\mathcal{C}_0} = (S_1^{\mathcal{C}_0}, S_2^{\mathcal{C}_0}, \ldots, S_t^{\mathcal{C}_0})$
begin
    initialisation
    $k = 1$
1   $TdP_1 \longleftarrow$ `Random`$(V^{\mathcal{C}}_{TdP}, \mathbb{1})$
2   *Depuis* $S_0 = (0,0)$ *parcourir* TdP_1 *donne l'emplacement d'une case* $Case_{CHC}$
    `Placer`$(S_1, Case_{CHC}, méthode)$
3   **while** $\{TdP_c < TdP_{CHC}\}$ **do**
4       $TdP_k \longleftarrow$ `Random`$(V^{\mathcal{C}}_{TdP}, Prob_{TdP_{k-1}})$
        $TdP_c = TdP_c + TdP_k$
5   **if** $TdP_c \leqslant TdP_{CHC}$ **then**
6   *Depuis* $S_0 = (0,0)$ *parcourir* TdP_c
            donne l'emplacement d'une case $Case_{CHC}$
            `Placer`$(S_k, Case_{CHC}, méthode)$

Étape 4 : Attribution de valeurs pour les variables *Longueur du segment* et *Angle du segment*

Pour tous les points de $V_S^{\mathcal{C}}$, nous remplaçons les valeurs de classe des couples $(C(L), C(\theta))$ associés à S par une valeur de *longueur du segment* et d'*angle du segment*. Pour les deux variables, nous remplaçons la valeur de classe par une valeur choisie suivant une loi uniforme dans les valeurs de la classe.

Dans le cas de la variable *Longueur du segment*, chaque classe est constituée des valeurs réelles sur l'ensemble du paysage (i.e. $C(L_i) = \{l_1, l_2, \ldots, l_{m_i}\}$). De ce fait, nous attribuons des longueurs réelles aux points $S^{\mathcal{C}}$.

Dans le cas de la variable *Angle du segment*, chaque classe est délimitée par deux

Function Placer

Data: S est le point du parcours $\in$ $Case_{CHC}$ de temps d'attente TdP_S; *méthode*
Result: $(x_H; y_H)$: Isobarycentre du segment de haie
begin

REM $Case_{CHC}$ est un carré défini par 4 points
$(x_{min}; y_{min})$; $(x_{max}; y_{min})$; $(x_{max}; y_{max})$; $(x_{min}; y_{max})$
S est le point courant du parcours de coordonnées (x_S, y_S)
$G = (x_G, y_G)$ est le barycentre de $Case_{CHC}$

switch *méthode* **do**
case *Sur le CHC*
$x_H \longleftarrow x_G$
$y_H \longleftarrow y_G$
case *Aléatoirement*
$x_H \longleftarrow$ `Random(`$[x_{min}; x_{max}], \mathbb{1}$`)`
$y_H \longleftarrow$ `Random(`$[y_{min}; y_{max}], \mathbb{1}$`)`

Algorithm 4: Simulation des classes de L et θ pour tous les points de $V_S^{\mathcal{C}_0}$

Data: $V_S^{\mathcal{C}_0} = (S_1^{\mathcal{C}_0}, S_2^{\mathcal{C}_0}, \ldots, S_t^{\mathcal{C}_0})$; $(M_\theta^{\mathcal{C}}; V_\theta^{\mathcal{C}})$; $(M_L^{\mathcal{C}}; V_L^{\mathcal{C}})$
Result: $\{(S_k, C(L_k), C(\theta_k))\}_{k \in [\![1;t]\!]} \in V_S^{\mathcal{C}_0} \times V_L^{\mathcal{C}_0} \times V_\theta^{\mathcal{C}_0}$
begin
for *k = 1* **do**
1 $C(\theta)_1 \longleftarrow$ `Random(`$V_\theta^{\mathcal{C}}, \mathbb{1}$`)`
2 $C(L)_1 \longleftarrow$ `Random(`$V_L^{\mathcal{C}}, \mathbb{1}$`)`
for *{ k = 2 ; k ⩽ t ; k++ }* **do**
3 $C(\theta)_k \longleftarrow$ `Random(`$V_\theta^{\mathcal{C}}, Prob_{\theta_{k-1}}$`)`
4 $C(L)_k \longleftarrow$ `Random(`$V_L^{\mathcal{C}}, Prob_{L_{k-1}}$`)`

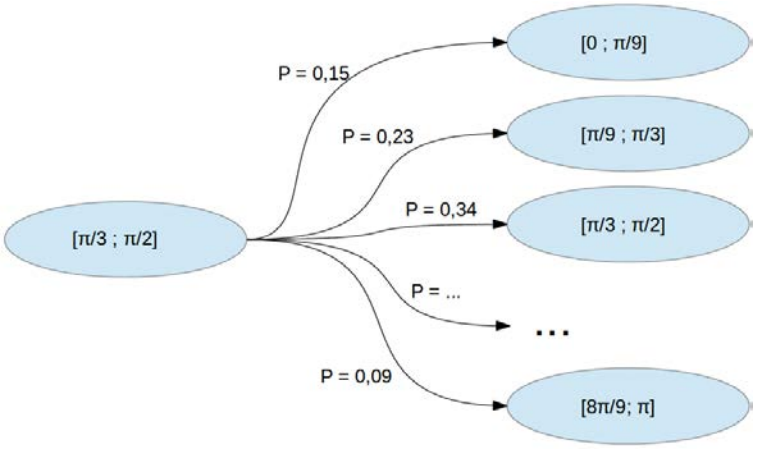


FIGURE 7.2 – Exemple de probabilité de transition d'une variable *Angle* pour la *classe* $[\frac{\pi}{3}; \frac{\pi}{2}]$.

valeurs c et d et nous attribuons aux points $S^{\mathcal{C}}$ une valeur choisie selon une loi uniforme sur le segment $[c, d]$

Algorithm 5: Simulation des valeurs L et θ à partir des valeurs de classes

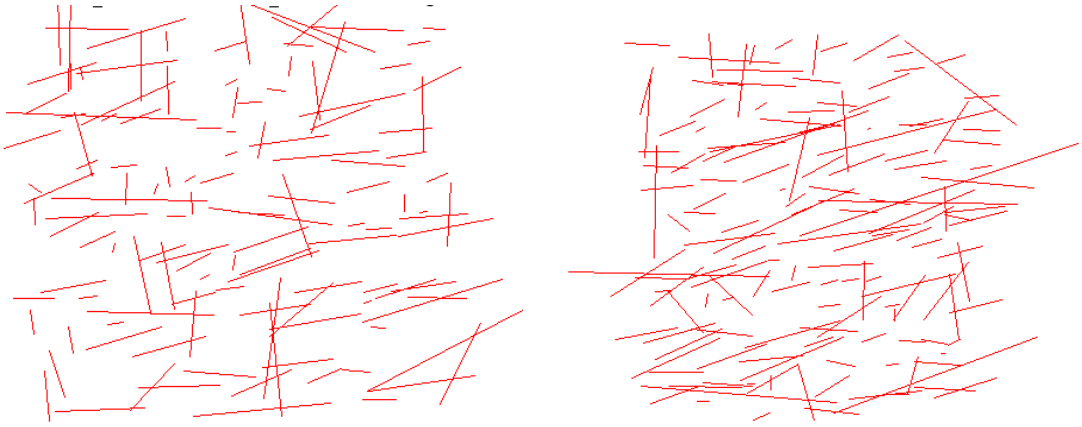
Data: $\{(S_k, C(L_k), C(\theta_k))\}_{k\in[\![1;t]\!]} \in V_S^{\mathcal{C}_0} \times V_L^{\mathcal{C}_0} \times V_\theta^{\mathcal{C}_0}$; $C(\theta)$; $C(L)$
Result: $\{(S_k, L_k, \theta_k)\}_{k\in[\![1;t]\!]} \in V_S^{\mathcal{C}_0} \times \mathbb{R}^+ \times [0; \pi]$
begin
  for $\{\ k = 1\ ;\ k \leqslant t\ ;\ k{+}{+}\ \}$ **do**
1 $\theta_k \longleftarrow$ `Random`$(C(\theta)_k, \mathbb{1})$
2 $L_k \longleftarrow$ `Random`$(C(L)_k, \mathbb{1})$

Étape 5 : Simulation de la cellule suivante

Dans le cas de la génération d'un paysage complet, nous utilisons, pour l'instant, la même méthode qu'à l'étape 1, pour choisir la classe associée à la cellule suivante.

7.2 Cellules simulées par génération de structures de segments dans l'espace - Comparaison au réel

Nous avons simulé des cellules grâce au processus précédemment présenté. Ces cellules simulées présentes une disposition que ne semble pas être semblable aux cellules réelles. Nous pouvons d'ailleurs voir des exemples de cellules simulées à la figure 7.3. Nous allons cependant les comparer avec les cellules réelles sur des critères non-visuels.



(a) (b)

FIGURE 7.3 – Exemples de cellules simulées

Tout d'abord, nous ferons la comparaison (dans le paragraphe 7.2.1) selon les trois indicateurs de statistiques descriptives suivant

- le nombre de segments de haies, noté Nb^g_{HH} dans le cas de données générées et Nb^g_{HH} dans le cas de données réelles
- la proportion de segments typiques i.e. :
  1. la proportion de segments de type HP (Segments de haies d'orientation Nord-Sud, défini au paragraphe 4.3.1) par rapport au nombre de segments de haies de la cellule, notée PrP^g_{HP} dans le cas de données générées et PrP^r_{HP} dans le cas de données réelles
  2. la proportion de segments de type HV (Segments de haies "brise-vent", défini au paragraphe 4.3.1) par rapport au nombre de segments de haies de la cellule, notée PrP^g_{HV} dans le cas de données générées et PrP^r_{HV} dans le cas de données réelles
- la distribution des longueurs des segments de haies par cellules

Puis, nous comparerons (dans le paragraphe 7.2.2) les résultats de la création des chemins de Hilbert adaptatif sur les cellules simulées et sur les cellules réelles.

7.2.1 Indicateurs de statistiques descriptives

Nous avons généré des segments de haies sur 100 cellules d'après la méthode décrite à la section 7.1 pour chaque cellule réelle, puis nous avons calculé les indicateurs pour l'ensemble des cellules, simulées et réelles. Pour la restitution sur les figures, nous avons choisi de représenter les valeurs des indicateurs des cellules simulées par des boîtes à moustaches dont la couleur dépend de la classe de la cellule (classe 1 en vert, classe 2 en bleu, classe 3 en cyan, classe 4 en rose et classe 5 en jaune). La valeur des indicateurs des cellules réelles est représentée par un point rouge.

Indicateur 1 : Nombre de segments de haies par cellules

Pour le paysage A : Pour les cellules des classes 1, 2 et 5, Nb^g_{HH} est proche de Nb^r_{HH} (figure 7.4). Pour les classes 3 et 4 (figure 7.4), Nb^g_{HH} est systématiquement trop élevé, parfois très nettement, comme dans le cas de la cellule $A-7\_6$. Si nous reprenons le schéma 4.2 de la section 4.2, nous pouvons voir que les cellules de la classe 3 et 4 sont celles où les haies sont les moins nombreuses, car plus proches des zones d'habitation. Le faible nombre de haies dans ces cellules s'explique par la perte des terrains agricoles au profit de l'urbanisation. Cela implique donc une perturbation dans la structure des haies.

Pour le paysage B : Nb^g_{HH} est proche de Nb^r_{HH} pour les classes 3, 4, 5 et 6 (figure 7.5) mais il est systématiquement trop faible pour les classes 1 et 2 (figure 7.5). Les cellules faisant partie de ces deux classes sont situées dans la partie Sud-Ouest de la zone d'étude. Pour cette partie constituée du bocage historique, la génération n'est pas parvenue à capter la densité élevée des segments de haies, au contraire des autres classes.

En prenant en compte les deux paysages, il semblerait que la méthode de simulation parvienne à générer un nombre de segments de haies proche du nombre de segments de haies dans la cellule réelle si celui-ci est compris entre 100 et 250. Une mise à l'échelle du CHC semble être nécessaire dans les autres cas.

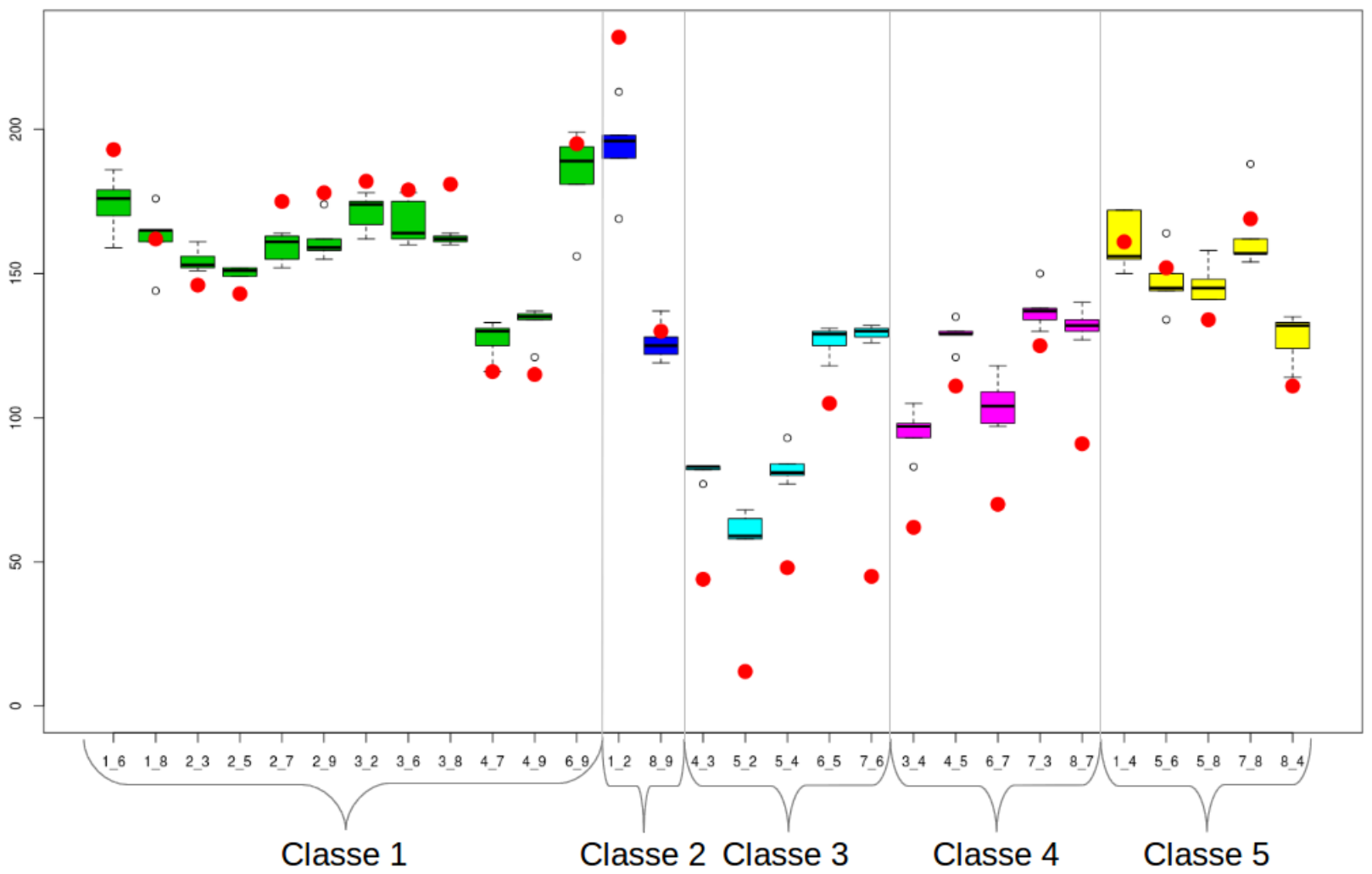


FIGURE 7.4 – Nombre de segments de haies par cellule pour les données A. En abscisse : le code des cellules du paysage A et la classe dans laquelle elles sont. En ordonnée : Le nombre de segments de haies.

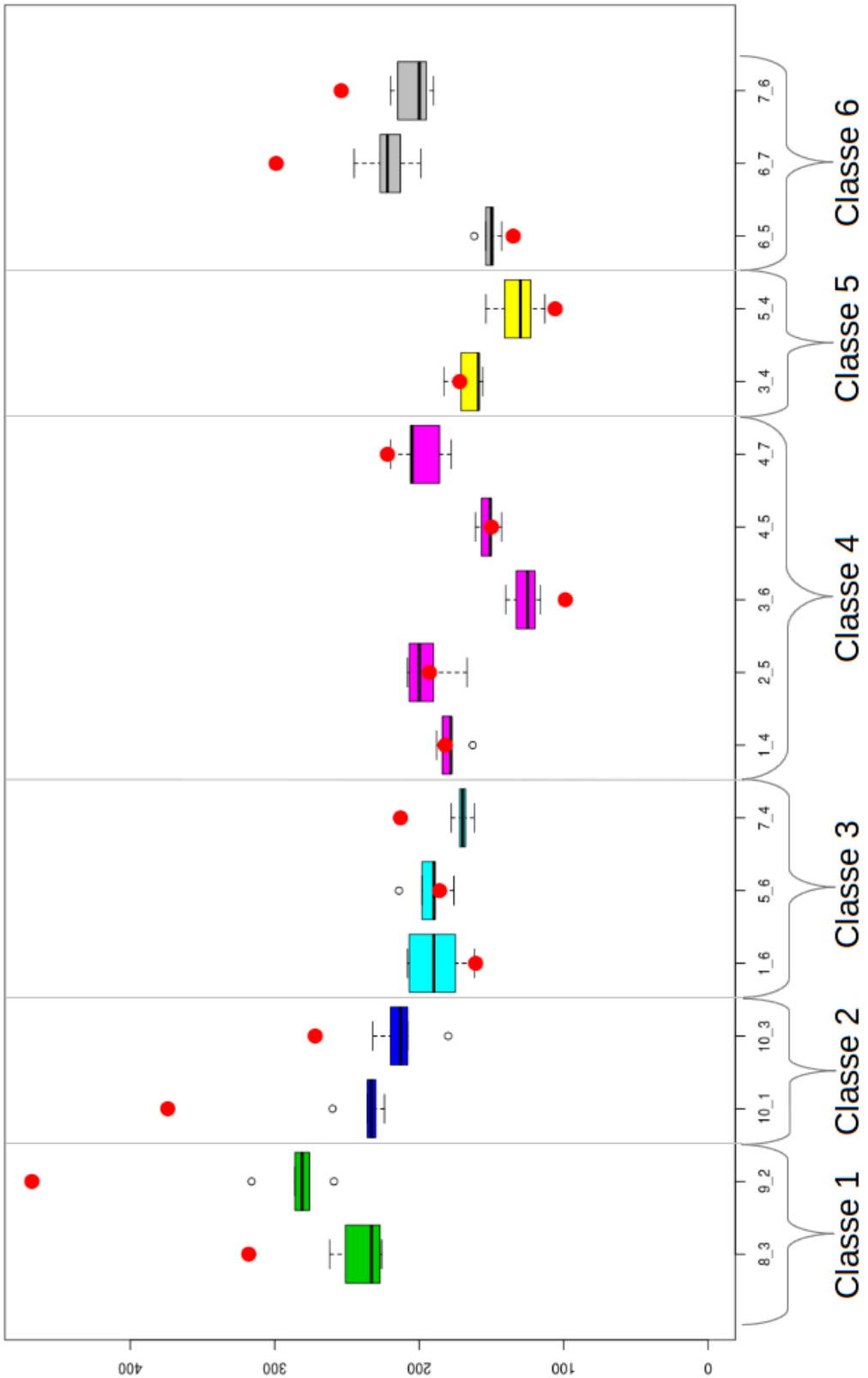


FIGURE 7.5 – Nombre de segments de haies par cellule pour les données B. En abscisse : le code des cellules du paysage B et la classe dans laquelle elles sont. En ordonnée : Le nombre de segments de haies.

Indicateur 2 : Proportion des segments de type HV et HP par cellules

Le même calcul est effectué par type de haies. Les résultats sont présentés aux figures 7.6, 7.7, 7.8 et 7.9.

De type HV dans le paysage A : Dans la figure 7.6, nous constatons qu'il n'y a pas de valeur unique PrP^r_{HV} par classe. Les valeurs PrP^r_{HV} s'étalent de 20% à 80% pour l'ensemble. PrP^g_{HV} est de l'ordre de grandeur de PrP^r_{HV} pour l'ensemble des cellules, même si nous ne pouvons pas dire que cela soit bien restitué pour une classe en particulier. Les résultats obtenus ne sont pas très satisfaisants si nous souhaitons pouvoir générer des cellules comportant une proportion de segments de haies de type HV. Nous pouvons cependant noter que les générations, pour chaque cellule, ne font pas apparaître une trop grande variabilité de cette proportion, mis à part pour la cellule $A-5\_2$.

De type HV dans le paysage B : Dans la figure 7.7, nous constatons que d'une manière générale, PrP^g_{HV} est proche de PrP^r_{HV} pour toutes les cellules. Pour l'ensemble des classes, PrP^r_{HV} est située dans la boite construite à partir de PrP^g_{HV} sur les cellules simulées correspondantes. La proportion réelle de segments de haies de type HV varie entre 15% et 45% sur l'échantillon de cellules représentées.

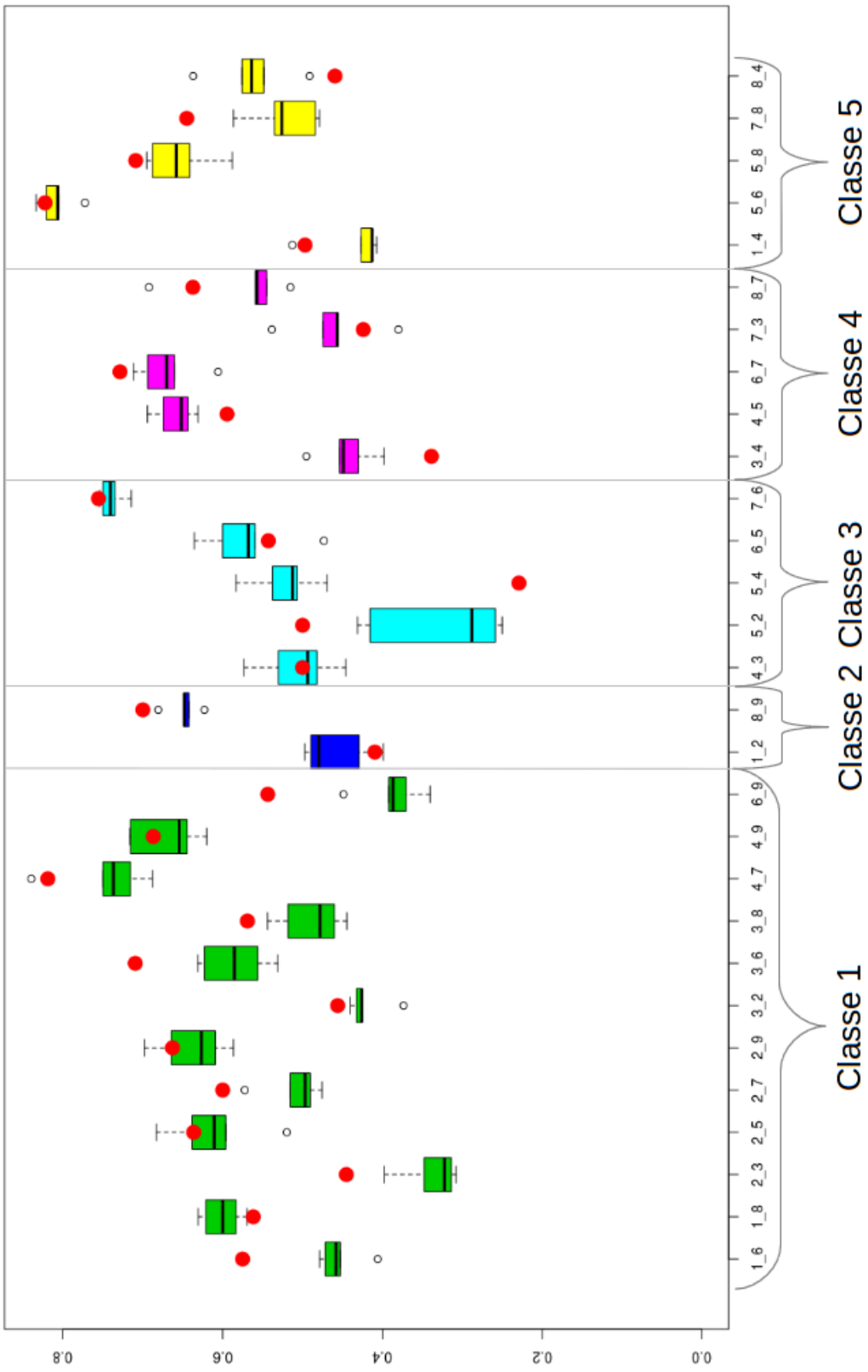


FIGURE 7.6 – Proportion des segments de type HV pour un échantillon de cellules issues du paysage A. En abscisse : le code des cellules du paysage A et la classe dans laquelle elles sont.

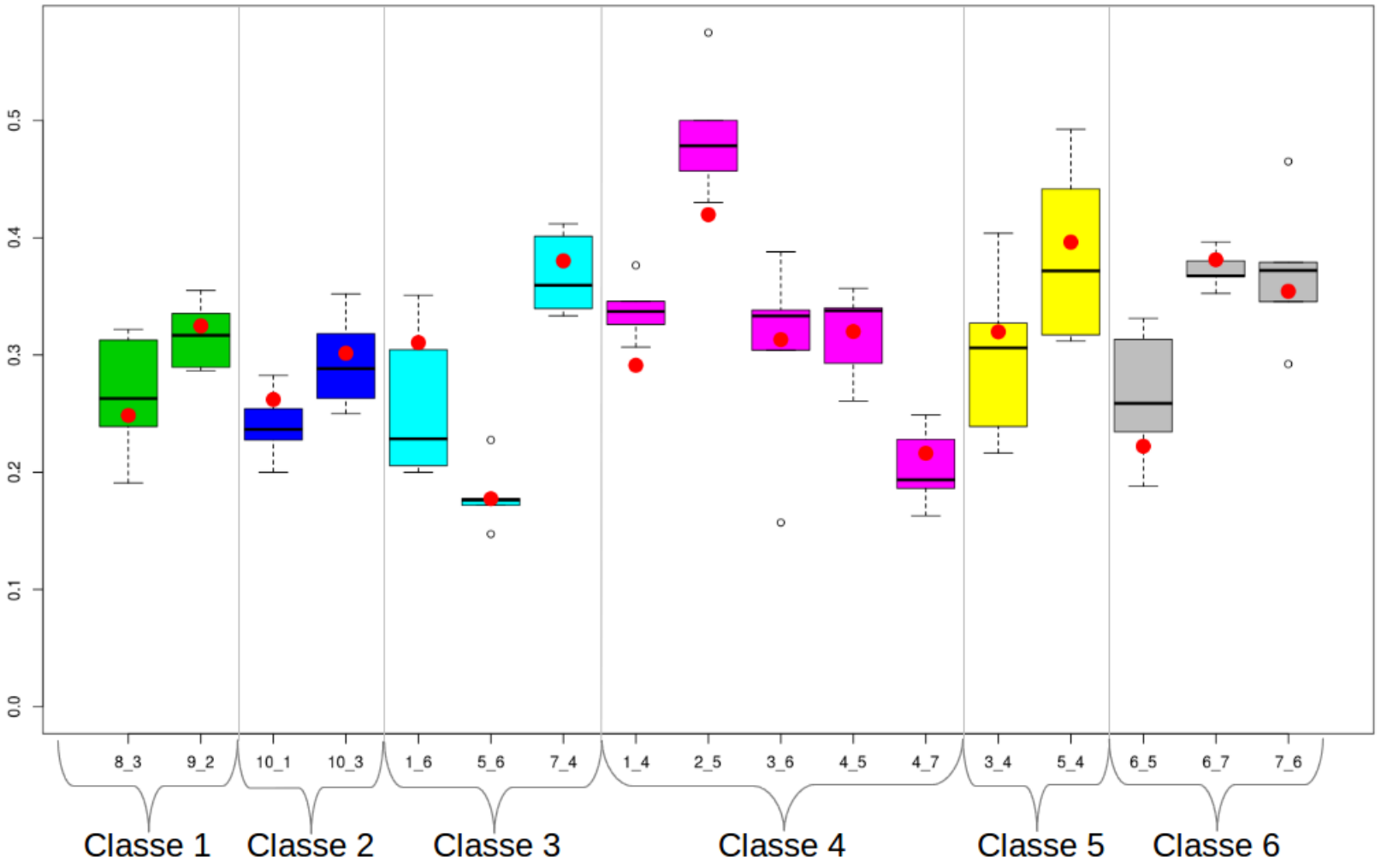


FIGURE 7.7 – Proportion des segments de type HV pour un échantillon de cellules issues du paysage B. En abscisse : le code des cellules du paysage B et la classe dans laquelle elles sont.

De type HP dans le paysage A : Dans la figure 7.8, PrP^r_{HP} se situe entre 0% et 40%. Pour l'ensemble des cellules de notre échantillon, PrP^g_{HP} est très proche de PrP^g_{HP}. Pour les classes 2 et 4, toutes les cellules réelles ont une valeur PrP^r_{HP} très proche de la médiane de PrP^g_{HP} pour les cellules simulées correspondantes. Pour les classes 1, 3 et 5, la médiane de PrP^g_{HP} est moins proche de PrP^r_{HP} mais les proportions entre réelles et simulées restent proches l'une de l'autre, excepté pour la cellule $A - 5_2$ au profil atypique.

De type HP dans le paysage B : PrP^g_{HP} est très proche de PrP^r_{HP} pour toutes les cellules des classes 1, 2, et 5 (figure 7.9). Pour les classes 3, 4 et 6, il existe au moins une cellule pour laquelle PrP^r_{HP} se situe en dehors de la boîte construite à partir des PrP^g_{HP} des cellules correspondantes.

Pour les deux paysages, la proportion de segments de haies de type HP est bien représentée. Cela permet de garantir la proportionnalité de la composition des cellules générées par rapport aux cellules réelles. Cette proportion est atteinte indépendamment du nombre de segments de haies ou de la classe de la cellule de départ.

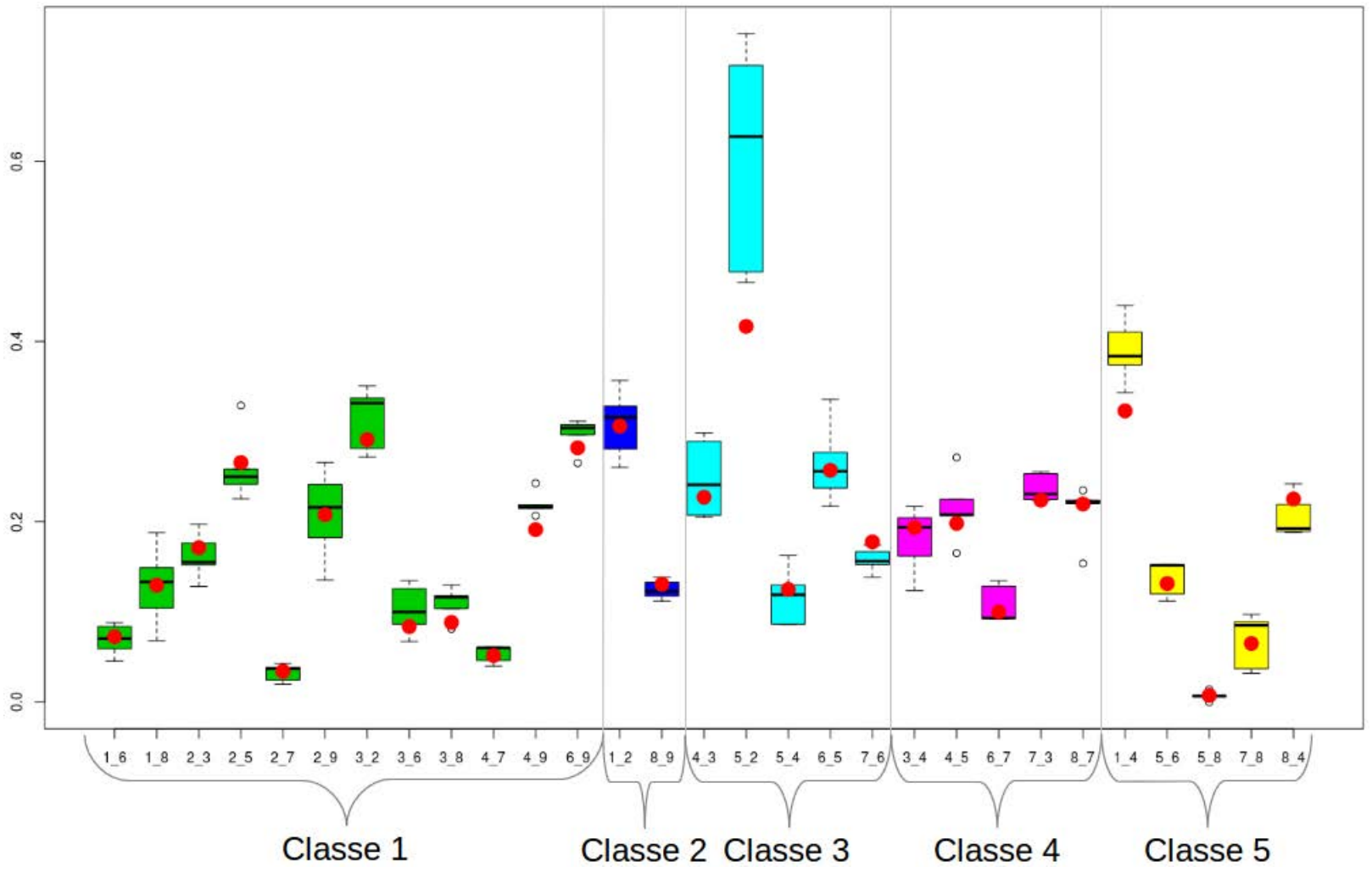


FIGURE 7.8 – Proportion des segments de type HP pour un échantillon de cellules issues du paysage A. En abscisse : le code des cellules du paysage A et la classe dans laquelle elles sont.

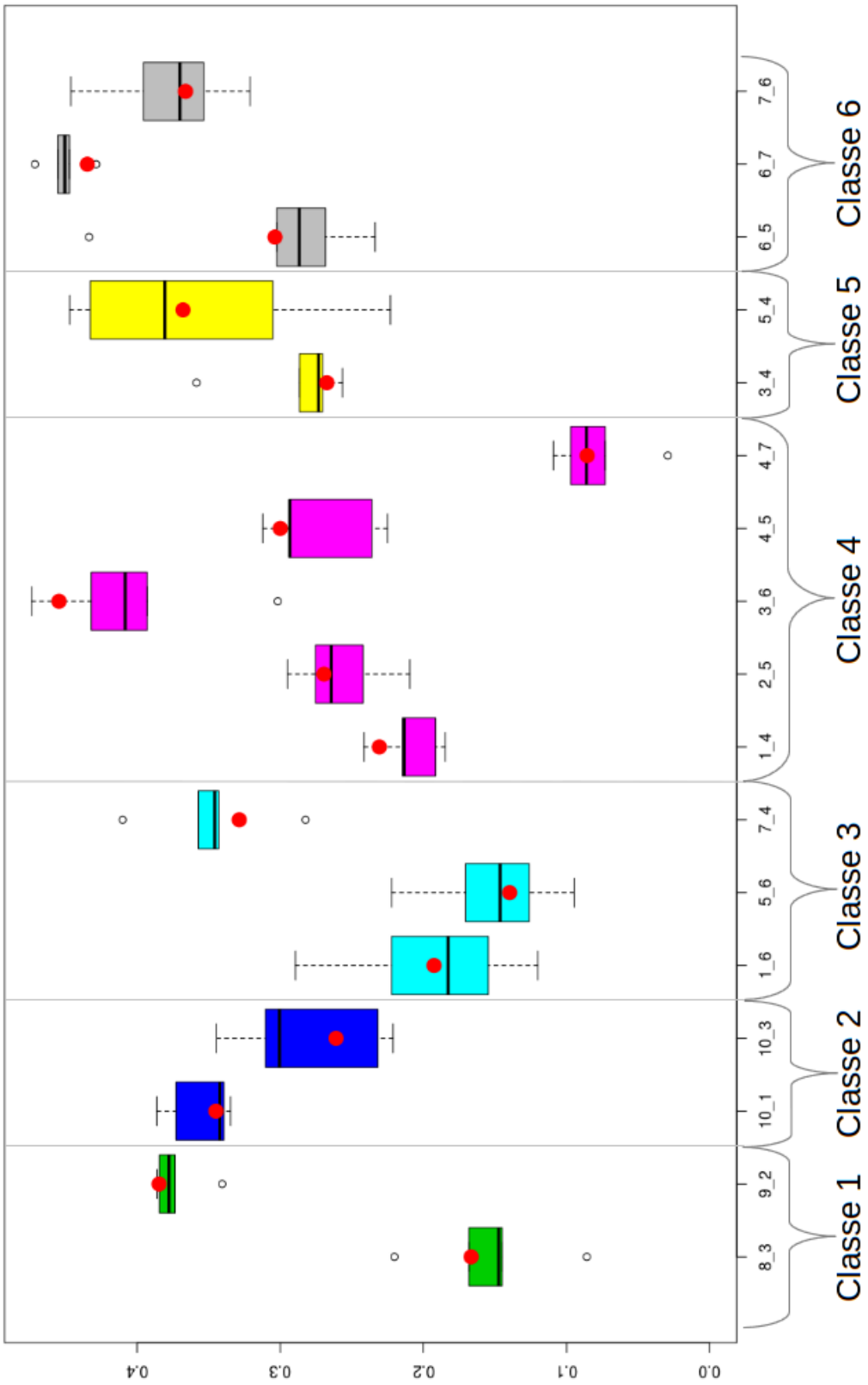


FIGURE 7.9 – Proportion des segments de type HP pour un échantillon de cellules issues du paysage B. En abscisse : le code des cellules du paysage B et la classe dans laquelle elles sont.

Indicateur 3 : Longueur des segments de haies par cellules

Afin de comparer les longueurs obtenues dans les cellules générées à celles mesurées dans la cellule réelle, nous avons choisi de raisonner sur les classes de longueurs définies dans la sous-section 6.4.1 et de calculer la proportion de chaque classe de longueurs pour chaque cellule réelle (cercle rouge sur la figure 7.10). Pour les 100 cellules simulées, nous avons déterminé la proportion moyenne de chaque classe de longueur (cercle noir sur la figure 7.10). Nous pouvons, grâce à cela, tracer la figure 7.10. Pour chaque cellule, exceptée $A-5\_2$, la proportion moyenne de chaque classe sur les cellules générées est proche de celle de la cellule réelle, la différence maximale étant de l'ordre de 4%, pour la cellule $A-4\_5$. Ce résultat très favorable quant à la génération des proportions des classes de longueur reste vrai pour les cellules du paysage B. Les résultats ne sont pas présentés ici afin de ne pas surcharger le chapitre.

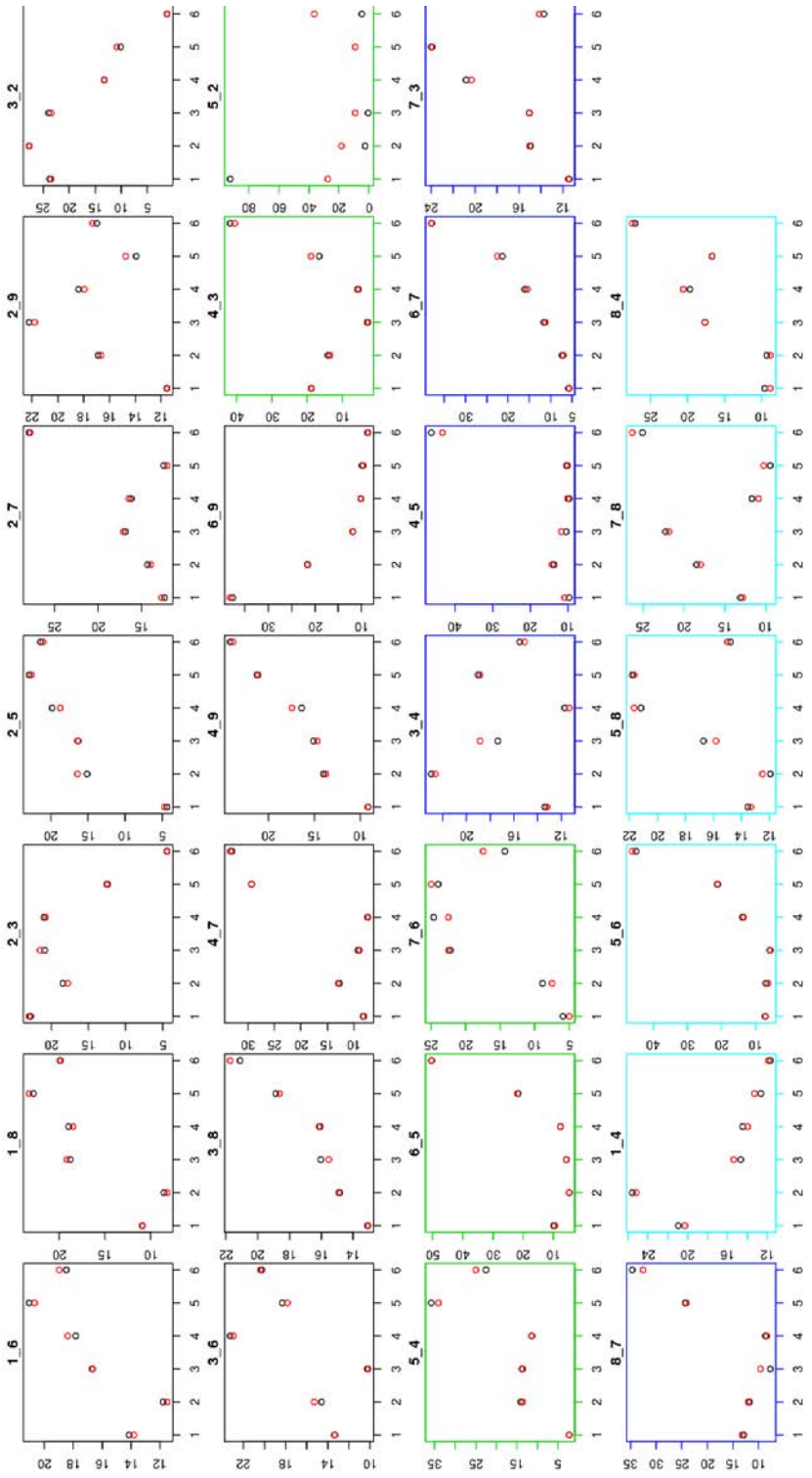


FIGURE 7.10 – Pour chaque cellule, proportion de chaque classe de longueur, en moyenne pour les 100 cellules générées (cercle noir), et pour la cellule réelle (cercle rouge). Les cellules sont issues du paysage A.

7.2.2 Chemins de Hilbert adaptatifs

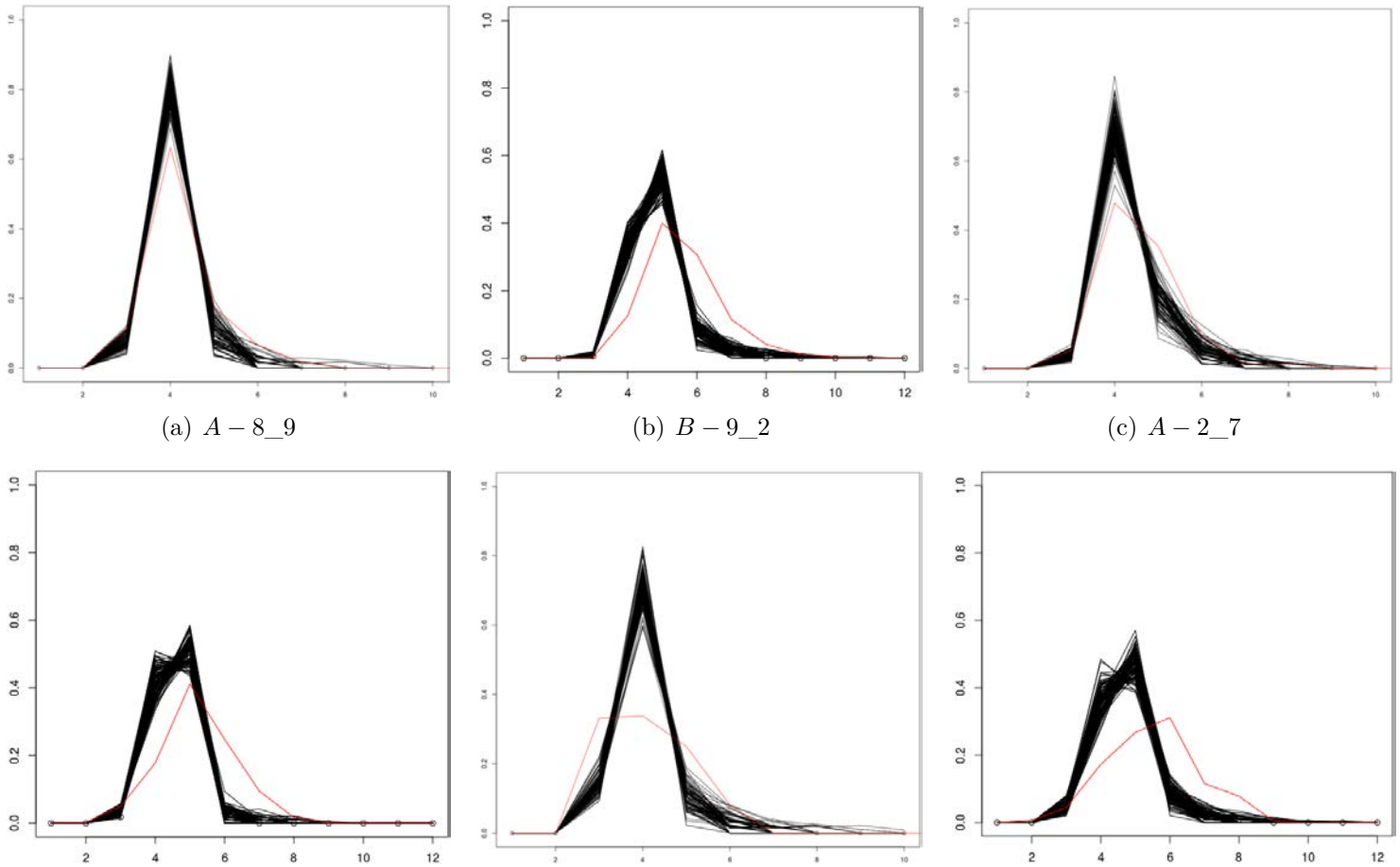
Nous déterminons le chemin de Hilbert adaptatif sur les cellules simulées afin de pouvoir les comparer avec les cellules réelles. La comparaison s'effectuera au moyen de la variable *Profondeur de Découpe*.

Nous avons effectué une évaluation pour toutes les classes de cellules. Aucune d'entre elles n'est bien reproduite en totalité, mais pour chacune, il existe des cellules bien reproduites, exceptée pour la classe 2 du paysage B. Pour le paysage A, nous pouvons voir sur la figure 7.11(a) que les cellules simulées ont une distribution de la variable *Profondeur de Découpe* proche de celle de la cellule réelle, un exemple avec le paysage B est présenté sur la figure 7.11(b).

Un exemple de cas moyen est présenté à la figure 7.11(c) pour le paysage A et à la figure 7.11(d) pour le paysage B. Nous pouvons voir que le mode de la variable est respecté, mais que sa variabilité ne l'est pas, ainsi les cellules simulées possèdent une distribution bimodale pour la cellule $B - 4\_7$ tandis que pour la cellule $A - 2\_7$, la prépondérance du mode de la variable *Profondeur de Découpe* dans les cellules simulées réduit la variabilité sur les valeurs adjacentes.

Enfin, avec les figures 7.11(e) et 7.11(f), nous voyons un exemple de cas défavorable. Dans le cas du paysage A, le mode de la variable est beaucoup trop élevé par rapport à sa valeur dans la cellule réelle et dans le cas du paysage B, le mode de la variable dans le cas des cellules simulées n'est pas le même que celui de la cellule réelle.

Nous pouvons dire que la méthode, permet de bien reproduire certaines cellules mais qu'elle ne rend pas justice aux cellules avec des modes peu marqués ou avec des densités plus importantes. Ceci est lié à la construction, qui s'appuie sur un chemin de Hilbert classique, dont le pas est le mode de la classe choisie. Il pourrait s'avérer intéressant de ne pas simuler les cellules avec un chemin de Hilbert classique ayant un pas identique à toute la classe, mais plutôt, d'attacher un valeur de pas, à chaque cellule réelle.



(a) $A-8\_9$ (b) $B-9\_2$ (c) $A-2\_7$

(d) $B-4\_7$ (e) $A-6\_7$ (f) $B-2\_5$

FIGURE 7.11 – Distribution de la variable *Profondeur de Découpe* pour 100 cellules simulées (en noir) et pour la cellule réelle correspondante (en rouge). En ordonnées : la proportion de chaque valeur de profondeur de découpe sur chaque cellule.

7.3 Génération améliorée - Ajustement à partir des connaissances du domaine

Les représentations graphiques des cellules simulées, à la figure 7.3, semblent indiquer que le positionnement des segments de haies avec le précédent processus de génération n'est pas satisfaisant bien que les résultats de la comparaison effectuée à la section 7.2 montrent que le processus de génération permet la conservation d'indicateurs intéressants dans les cellules. Nous mettons alors en place une méthode de génération améliorée issue de la méthode précédente mais ajustée à partir des connaissances du domaine que nous avons obtenues au chapitre 5.

Pour ce faire, nous adaptons le processus pour fusionner les étapes décrites à la section 7.1. Dans le processus initial, il existe trois étapes pour la génération de segments, une étape plaçant la totalité des points dans la cellules simulées (Algorithme 3) puis une étape attribuant une valeur de classe de longueurs et d'angles aux points (Algorithme 4) et finalement, une étape attribuant une valeur de longueurs et d'angles aux points (Algorithme 5). Dans le processus modifié, dès qu'un temps de parcours est tiré (ligne 6, algorithme 6), nous déterminons son positionnement dans le plan (lignes 9 et 10, algorithme 6), et y associons immédiatement une valeur de classe de longueur et d'angle (lignes 12 et 13, algorithme 6) puis une valeur de longueur et d'angle (lignes 15 et 16, algorithme 6). Ceci permet de décrire un segment dans la cellule, avant de l'implanter, nous vérifions si sa position ne transgresse pas les règles établies dans le chapitre 5. Si ce n'est pas le cas, nous positionnons le segment dans la cellule (ligne 20, algorithme 6). Sinon, nous effectuons un retour sur les fonctions de générations dans l'ordre inverse, d'abord les valeurs d'angle et de longueur (ligne 14, algorithme 6), puis si après n tentatives, le segment ne convient toujours pas, nous générons de nouveau un couple de classes d'angle et de longueur (ligne 11, algorithme 6) durant m tentatives et de même, si cela n'aboutit pas, nous regénérons un positionnement de l'isobarycentre de la haie (ligne 8, algorithme 6) durant t tentatives et du temps de parcours durant r tentatives (ligne 5, algorithme 6). Si l'ensemble des tentatives ont abouti à des échecs de placement, nous positionnons dans la cellule le premier segment généré (ligne 19, algorithme 6), précédemment sauvegardé (ligne 18, algorithme 6). Un schéma simplifié de cet algorithme est représenté sur la figure 7.12. Les seuils notés n, m, t, r sont fixés de manières empiriques afin de parcourir au mieux les valeurs de longueurs et d'angles mais ils devraient faire l'objet d'une étude particulière que nous n'avons pu mener.

Remarque : TdP_{CHC} représente le temps de parcours de la totalité du chemin de Hilbert sur la cellule.

Cette génération ajustée a permis d'obtenir des résultats qui sont visuellement proches des structures de segments dans les paysages agricoles des données A, comme nous pouvons le constater sur la figure 7.13. Ce résultat n'est pas vérifié pour les données B, où les cellules simulées ne ressemblent pas aux cellules réelles (Figure 7.14). Cet écart peut s'expliquer par le caractère moins structuré du paysage B.

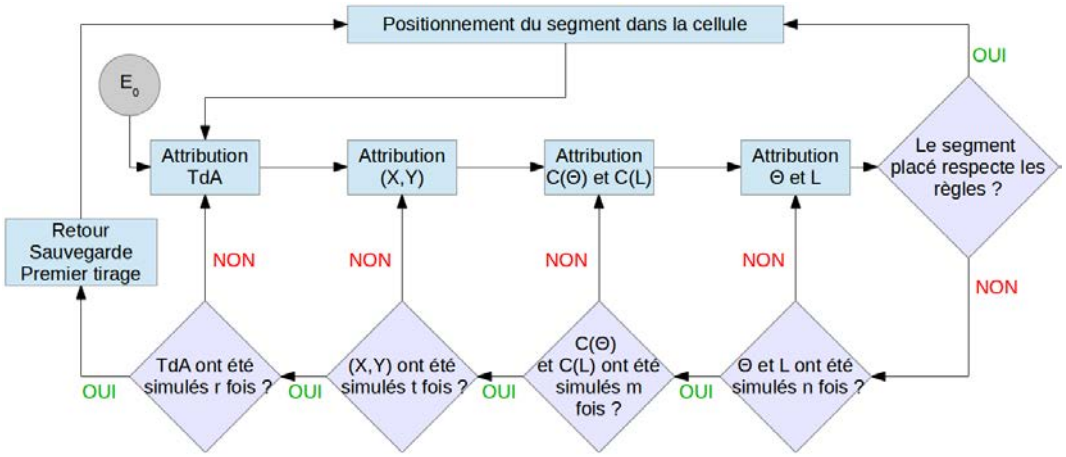
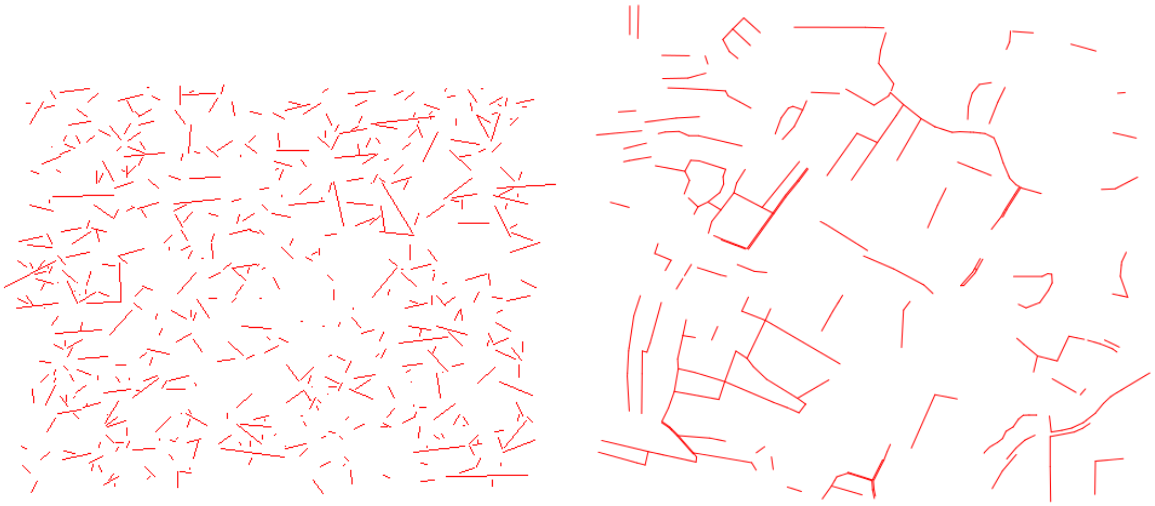


FIGURE 7.12 – Processus ajusté à partir des connaissances du domaine avec retour sur trace.

(a) *simulation* (b) *réel*

FIGURE 7.13 – Résultat d'une cellule simulée (à gauche) pour la cellule réelle issue des données A (à droite) par la génération ajustée avec connaissance du domaine.

Ces résultats sont encourageants, il reste maintenant à comparer ces nouvelles simulations avec les cellules réelles afin de savoir si le gain opéré sur le positionnement



(a) *simulation* (b) *réel*

FIGURE 7.14 – Résultat d'une cellule simulée (à gauche) pour la cellule réelle issue des données B (à droite) par la génération ajustée avec connaissance du domaine.

n'a pas eu un impact négatif trop important sur les indicateurs de densité, de longueurs et d'angle des segments de haies.

Algorithm 6: Simulation de la position des segments dans $\mathcal{C}_0$ avec le processus de génération ajustée

Data: $\{\{(M^{\mathcal{C}}_{TdP}; V^{\mathcal{C}}_{TdP})\ ;\ (M^{\mathcal{C}}_{\theta}; V^{\mathcal{C}}_{\theta})\ ;\ (M^{\mathcal{C}}_{L}; V^{\mathcal{C}}_{L})\} \mid \mathcal{C} \in \mathcal{H}\}$; $\mathcal{C}_0$; Règles d'implantation

Result: $\{(S_k, L_k, \theta_k)\}_{k\in[\![1;t]\!]} \in \mathbb{R}^2 \times \mathbb{R}^+ \times [0;\pi]$ segments dans $\mathcal{C}_0$

begin

initialisation $k = 1$

1 $TdP_1 \longleftarrow$ `Random`$(V^{\mathcal{C}}_{TdP}, \mathbb{1})$; $C(\theta)_1 \longleftarrow$ `Random`$(V^{\mathcal{C}}_{\theta}, \mathbb{1})$;
$C(L)_1 \longleftarrow$ `Random`$(V^{\mathcal{C}}_{L}, \mathbb{1})$; $\theta_1 \longleftarrow$ `Random`$(C(\theta)_1, \mathbb{1})$;
$L_1 \longleftarrow$ `Random`$(C(L)_1, \mathbb{1})$

2 *Depuis* $S_0 = (0,0)$ *parcourir* TdP_1 *donne l'emplacement d'une case* $Case_{CHC}$

3 `Placer`$(S_1, Case_{CHC}, Aléatoirement)$

$TdP_c = TdP_1$

4 **while** $(TdP_c < TdP_{CHC})$ **do**

$Regle = FALSE$; $l = 0$

5 **while** $(!Regle \ \&\&\ l \leqslant r)$ **do**

6 $TdP_k \longleftarrow$ `Random`$(V^{\mathcal{C}}_{TdP}, Prob_{TdP_{k-1}})$

$j = 0$; $l = l + 1$

7 *Depuis* $S_0 = (0,0)$ *parcourir* $(TdP_c + TdP_k)$ *donne l'emplacement d'une case* $Case_{CHC}$

8 **while** $(!Regle \ \&\&\ j \leqslant t)$ **do**

9 $x^j_H \longleftarrow$ `Random`$([x^{Case}_{min}; x^{Case}_{max}], \mathbb{1})$

10 $y^j_H \longleftarrow$ `Random`$([y^{Case}_{min}; y^{Case}_{max}], \mathbb{1})$

$i = 0$; $j = j + 1$

11 **while** $(!Regle \ \&\&\ i \leqslant m)$ **do**

12 $C(\theta)^i_k \longleftarrow$ `Random`$(V^{\mathcal{C}}_{\theta}, Prob_{\theta_{k-1}})$

13 $C(L)^i_k \longleftarrow$ `Random`$(V^{\mathcal{C}}_{L}, Prob_{L_{k-1}})$

$h = 0$; $i = i + 1$

14 **while** $(!Regle \ \&\&\ h \leqslant n)$ **do**

15 $\theta^h_k \longleftarrow$ `Random`$(C(\theta)_k, \mathbb{1})$

16 $L^h_k \longleftarrow$ `Random`$(C(L)_k, \mathbb{1})$

17 **if** $(\{(x^j_H; y^j_H); \theta^h_k; L^h_k\}$ *respecte les règles d'implantation*$)$ **then**
$Regle = TRUE$

18 **if** $(l = 0 \ \&\&\ j = 0 \ \&\&\ i = 0 \ \&\& h = 0\ )$ **then**
$Sauvegarde = \{(x^0_H; y^0_H); \theta^0_k; L^0_k\}$

$h = h + 1$

19 **if** $(r < l \ \&\&\ t < j \ \&\&\ m < i \ \&\&\ n < h \ \&\&\ !Regle\ )$ **then**
$\{(x^k_H; y^k_H); \theta_k; L_k\} = Sauvegarde$

20 $Implanter\{(x^k_H; y^k_H); \theta_k; L_k\}$

21 $TdP_c = TdP_c + TdP_k$

$k = k + 1$;

Quatrième partie

Conclusion

Chapitre 8

Conclusion et perspectives

Dans cette thèse, la problématique initiale était d'origine agronomique. Sans revenir sur l'ensemble des raisons exposées dans le chapitre 1, nous avons montré qu'il était important de pouvoir modéliser et générer les haies dans les paysages agricoles. Nous avons traité cette problématique en optant pour une approche d'extraction de connaissances à partir de données (ECBD), comme proposé par Fayyad et al. [30]. En effet, la masse de données disponible au départ, et le caractère imprévisible des résultats attendus se prêtaient bien à cette approche. Nous avons donc mis en œuvre une démarche d'ECBD appliquée à des données spatiales agricoles constituées de segments dispersés dans un plan.

D'un point de vue agronomique, l'originalité dans ce travail vient de la façon d'envisager les haies comme éléments structurant le paysage et non pas comme éléments structurés par lui. Nous avons donc été amenés à les étudier seules, sans étudier de processus agro-écologiques connexes. Pour cela, nous avons développé une approche nouvelle dans l'exploration de données spatiales linéaires, en nous focalisant uniquement sur les aspects géométriques de celles-ci.

D'un point de vue méthodologique, nous avons développé et mis en œuvre deux méthodes de fouilles de données distinctes, qui ont permis d'analyser les données à plusieurs niveaux et produit des résultats complémentaires.

La première méthode permet d'analyser la structure de certains segments par rapport aux segments de même type et aux structures d'autres types de segments. Nous avons créé un indice permettant de caractériser une structure spatiale constituée de segments. Cet indice est déterminé en s'appuyant sur la densité de voisins de certains types, contenus dans le voisinage de segments. Ensuite, nous avons mis en place une étude statistique de cet indice sur l'ensemble des structures réelles de deux paysages agricoles. Nous avons ainsi obtenu des règles d'implantation pour ces structures dans

le plan.

La seconde méthode permet de caractériser la structure des segments sur la zone qu'ils recouvrent mais sans s'intéresser aux autres structures de segments. Nous avons considéré les segments de haies comme des points disposés de manière non homogène dans le plan. Nous avons eu recours aux courbes de type Hilbert-Peano afin de linéariser ces données spatiales. Plus précisément nous avons utilisé les chemins adaptatifs de Hilbert, qui tiennent compte de la densité locale des données et limitent la perte d'information et l'information creuse. Les données ainsi linéarisées permettent d'utiliser de manière aisée les chaînes de Markov, pour la caractérisation des structures de segments dans le plan.

L'avantage des chaînes de Markov, en plus d'être faciles d'usage, est d'être utilisable à la fois pour l'apprentissage et pour la génération de données. Outre l'usage exploratoire qui peut en être fait, la génération de données est un outil de validation. On peut en examinant les paysage générés, vérifier que les chaînes de Markov ont capté les bonnes caractéristiques des données.

Finalement les méthodes développées ont fait l'objet d'une implantation dans un logiciel qui permet aux non spécialistes de pouvoir les utiliser de façon indépendante.

Les méthodes développées ont été testées sur deux paysages aux caractéristiques bien différentes. Toutefois, il est important de pouvoir disposer d'autres données réelles afin de pouvoir généraliser les méthodes et leur implantation logicielle. De plus, la première méthode, faisant intervenir un nouvel indice caractérisant le paysage, semble être dépendante de l'existence d'une orientation préférentielle sur le paysage. Il serait intéressant d'étudier cet indice sur des paysages exempts d'orientation ou en se fondant sur d'autres caractéristiques des haies. La seconde méthode, fondée sur la linéarisation des données, a montré une limite lorsque ces dernières ont une densité élevée et des implantations spatiales trop proches. En effet, cela implique un pas trop petit sur le chemin de Hilbert, et un temps de calcul élevé qui nuit alors à l'efficacité de la méthode.

Au delà du travail effectué, plusieurs développements sont envisageables. Tout d'abord il faut effectuer les calculs d'indices sur la totalité des cellules du paysage et non plus sur un échantillon comme dans cette thèse. Dans le même esprit, nous préconisons de déterminer la linéarisation des données et de calculer les matrices de modèle de Markov sur la totalité des paysages. Ceci pourrait permettre de caractériser d'une façon nouvelle les structures linéaires dans le plan et donc les paysages agricoles. Nous pourrions ainsi affiner les méthodes de génération présentées au chapitre 7. Elles seraient basées sur cette nouvelle classification des structures et sur la fusion des modèles de Markov appris sur une même classe et non plus sur un seul élément de la classe.

De même, une nouvelle façon d'envisager la génération de structures sur un paysage agricole serait d'utiliser des règles sur les relations de voisinage entre les cellules dans le paysage et non plus par un choix effectué aléatoirement parmi celles-ci.

A court terme, le premier travail à réaliser sera d'étudier les résultats obtenus par la génération ajustée avec les connaissances du domaine afin de savoir si les cellules simulées sont cohérentes avec les cellules réelles. Ces travaux, une fois réalisés, doivent permettre d'inclure toutes les méthodes développées dans le logiciel Genexp-LandSiTes [2]. Celui-ci sera alors capable, en plus de simuler des parcellaires agricoles, de générer des structures de haies ou d'autres linéaires sur lesquelles viendraient s'appuyer les parcellaires. Ainsi, nous pourrons disposer de paysages virtuels aux caractéristiques contrôlées permettant de tester des scénarios paysagers pour la protection des cultures.

Au-delà de cette étude à visée agronomique, les méthodes développées se veulent plus générale. Elles peuvent s'appliquer à tout problème pouvant se ramener à l'étude d'un ensemble de segments d'étiquettes variées dans l'espace. Par exemple les données de géolocalisation pour étudier, grâce à l'indice de densité relative, les liens entre le déplacement des personnes et des zones spécifiques (par exemple, des zones d'intérêt). Le déplacement des personnes peut être considéré comme des polylignes, fragmentable en segment. Les centres d'intérêt peuvent être vu comme un ensemble de segment connecté, tandis que les zones spécifiques peuvent être un segment, représentant, par exemple, un ensemble de panneaux publicitaires dans une ville. Le second type de données pourrait être l'étude des coups de pinceaux dans les peintures, comme le font Li *et al.* [62]. Dans cet article, les auteurs présentent une méthode pour fouiller les coups de pinceaux suivant une approche statistiques afin de comparer le style de Vincent van Gogh à ses contemporains. Les auteurs transforment les coups de pinceaux en informations séquentielles et étudient les liens qui peuvent exister entre les différentes séquences pour chaque tableau. Nos deux approches pourraient être utilisées dans ce cas pour fouiller ces données, cependant, le volet générationnel de la seconde ne serait pas exploité.

2. `http://engees.unistra.fr/~fleber/Landsites/`

Cinquième partie

Bibliographie

Bibliographie

[1] Vadim Alexandrov and Mark Gerstein. Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures. *BMC Bioinformatics*, 5(1) :2, 2004.

[2] Graham Wallis Arnold. The influence of ditch and hedgerow structure, length of hedgerows, and area of woodland and garden on bird numbers on farmland. *Journal of applied ecology*, 20(3) :731–750, 1983.

[3] Joachim Aurbacher and Stephan Dabbert. Generating crop sequences in land-use models using maximum entropy and Markov chains. *Agricultural Systems*, 104(6) :470–479, 2011.

[4] Jean-Stéphane Bailly, Pascal Monestiez, and Philippe Lagacherie. Modelling spatial variability along drainage networks with geostatistics. *Mathematical Geology*, 38(5) :515–539, 2006.

[5] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6) :1554–1563, 1966.

[6] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic. *The Annals of Mathematical Statistics*, 41(1) :164–171, 1970.

[7] Dalila Benboudjema and Wojciech Pieczynski. Unsupervised Statistical Segmentation of Nonstationary Images Using Triplet Markov Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8) :1367–1378, 2007.

[8] Yoshua Bengio. Markovian models for sequential data. *Neural computing surveys*, pages 1–30, 1999.

[9] Btissam Benmiloud and Wojciech Pieczynski. Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images. *Traitement du signal*, 12(5) :433–454, 1995.

[10] Andrew F Bennett, James Q Radford, and Angie Haslem. Properties of land mosaics : implications for nature conservation in agricultural environments. *Biological Conservation*, 133(2) :250–264, 2006.

[11] A Bhattacharya, M Roux, Henri Maître, Ian H Jermyn, Xavier Descombes, Josiane Zerubia, and Others. Computing statistics from man-made structures on the earth's surface for indexing satellite images. *International journal of simulation modelling*, 6(2) :73–83, 2007.

[12] Jean-François Blanc. Deux paysages en terrasses de l'Ardèche. *Revue de géographie de Lyon*, 56(4) :391–409, 1981.

[13] Philippe Bouché and Marc Le Goc. Analyse stochastique de séquences d'événements discrets pour la découverte de signatures. In *EGC*, pages 103–114, 2005.

[14] Françoise Burel. Landscape structure effects on carabid beetles spatial patterns in western France. *Landscape Ecology*, 2(4) :215–226, 1989.

[15] Françoise Burel. Hedgerows and Their Role in Agricultural Landscapes. *Critical Reviews in Plant Sciences*, 15(2) :169–190, 1996.

[16] Françoise Burel and Jacques Baudry. *Écologie du paysage. Concepts, méthodes et applications*. Tec & doc edition, 1999.

[17] Georg Cantor. Ein Beitrag zur Mannigfaltigkeitslehre. *Journal für die reine und angewandte Mathematik*, 84 :242–258, 1877.

[18] M.S. Castellazzi, J.N. Perry, N. Colbach, H. Monod, K. Adamczyk, V. Viaud, and K.F. Conrad. New measures and tests of temporal and spatial pattern of crops in agricultural landscapes. *Agriculture, Ecosystems & Environment*, 118(1-4) :339–349, January 2007.

[19] Hal Caswell. Community structure : a neutral model analysis. *Ecological monographs*, pages 327–354, 1976.

[20] Rebecca Chaplin-Kramer, Megan E O'Rourke, Eleanor J Blitzer, and Claire Kremen. A meta-analysis of crop pest and natural enemy response to landscape complexity. *Ecology Letters*, 14(9) :922–932, 2011.

[21] David Coeurjolly, Annick Montanvert, Jean-Marc Chassery, and Others. *Géométrie discrète et images numériques*. 2007.

[22] Noel A C Cressie. *Statistics for Spatial Data : Revised Edition*. John Wiley & Sons, 1993.

[23] Samuel A Cushman, Andrew Shirk, and Erin L Landguth. Separating the effects of habitat area, fragmentation and matrix resistance on genetic differentiation in complex landscapes. *Landscape ecology*, 27(3) :369–380, 2012.

[24] Zoe G Davies and Andrew S Pullin. Are hedgerows effective corridors between fragments of woodland habitat ? An evidence-based approach. *Landscape ecology*, 22(3) :333–351, 2007.

[25] Michael John De Smith, Michael F Goodchild, and Paul A Longley. *Geospatial analysis : a Comprehensive Guide to Principles, Techniques and Software Tools*, volume 1. Troubador, 2007.

[26] D Degenhardt. Description of Tree Distribution Patterns and Their Development Through Marked Gibbs Processes. *Biometrical Journal*, 41(4) :457–470, 1999.

[27] Xavier Descombes and E Zhizhina. The Gibbs fields approach and related dynamics in image processing. *Condensed Matter Physics*, 11(2) :293–312, 2008.

[28] Elias Egho, Nicolas Jay, Chedy Raïssi, Dino Ienco, Pascal Poncelet, Maguelonne Teisseire, and Amedeo Napoli. A contribution to the discovery of multidimensional patterns in healthcare trajectories. *Journal of Intelligent Information Systems*, 42(2) :283–305, March 2014.

[29] Amro Elfeki and Michel Dekking. A Markov chain model for subsurface characterization : Theory and applications. *Mathematical Geology*, 33(5) :569–589, 2001.

[30] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3) :37, 1996.

[31] Joseph E Flaherty, Raymond M Loy, Mark S Shephard, Boleslaw K Szymanski, James D Teresco, and Louis H Ziantz. Adaptive local refinement with octree load balancing for the parallel solution of three dimensional conservation laws. *Journal of Parallel and Distributed Computing*, 47 :139–152, 1997.

[32] Richard T T Forman and Jacques Baudry. Hedgerows and hedgerow networks in landscape ecology. *Environmental Management*, 8(6) :495–510, 1984.

[33] Richard T T Forman and Michel Godron. Patches and structural components for a landscape ecology. *BioScience*, pages 733–740, 1981.

[34] Carlo Gaetan and Xavier Guyon. *Modélisation et statistique spatiales*, volume 63. Springer, 2008.

[35] Robert Gardner, Bruce Milne, Monica Turnei, and Robert O'Neill. Neutral models for the analysis of broad-scale landscape pattern. *Landscape Ecology*, 1(1) :19–28, 1987.

[36] Cédric Gaucherel, D Fleury, Daniel Auclair, and P Dreyfus. Neutral models for patchy landscapes. *Ecological Modelling*, 197(1–2) :159–170, 2006.

[37] Flavia Geiger, Felix L Wäckers, and Felix J J A Bianchi. Hibernation of predatory arthropods in semi-natural habitats. *Biocontrol*, 54(4) :529–535, 2009.

[38] Guillaume Grégoire and Hugues Chaté. Onset of collective and cohesive motion. *Physical review letters*, (2) :1–4, 2004.

[39] Jeroen C J Groot, André Jellema, and Walter A H Rossing. Designing a hedgerow network in a multifunctional agricultural landscape : Balancing trade-offs among ecological quality, landscape character and implementation costs. *European Journal of Agronomy*, 32(1) :112–119, 2010.

[40] Yann Guédon, D. Barthélémy, Y. Caraglio, and E. Costes. Pattern Analysis in Branching and Axillary Flowering Sequences. *Journal of Theorical Biology*, 212(4) :481–520, 2001.

[41] Thomas Guyet. Visualisation de données relationnelles. *Conférence Internationale de Géomatique et d'Analyse Spatiale (SAGEO)*, 2013.

[42] Gérard Guyot. Climatologie de l'environnement, Edition Dunod. *Paris. France*, 1999.

[43] David Hilbert. Ueber die stetige abbildung einer linie auf ein flächenstück. *Mathematische Annalen*, 38(3) :459–460, 1891.

[44] Shelley A Hinsley and Paul E Bellamy. The influence of hedge structure, management and landscape context on the value of hedgerows to birds : a review. *Journal of environmental management*, 60(1) :33–49, 2000.

[45] Ruihong Huang and Christina Kennedy. Uncovering Hidden Spatial Patterns by Hidden Markov Model. In Thomas Cova, Harvey Miller, Kate Beard, Andrew Frank, and Michael Goodchild, editors, *Geographic Information Science*, volume 5266 of *Lecture notes in Computer Science*, pages 70–89. Springer Berlin / Heidelberg, 2008.

[46] Jan Hungershöfer and Jens-Michael Wierum. On the quality of partitions based on space-filling curves. In Peter M.A. Sloot, Alfons G. Hoekstra, C.J. Kenneth Tan, and Jack J. Dongarra, editors, *Computational Science—ICCS 2002*, volume 2331 of *Lecture Notes in Computer Science*, pages 36–45. Springer Berlin / Heidelberg, 2002.

[47] Hosagrahar V Jagadish. Analysis of the Hilbert curve for representing two-dimensional space. *Information Processing Letters*, 62(1) :17–22, 1997.

[48] Bing-Hwang Juang, Stephene Levinson, and M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains. *Information Theory, IEEE Transactions on*, 32(2) :307–309, 1986.

[49] Athanasios Kehagias. A hidden Markov model segmentation procedure for hydrological and environmental time series. *Stochastic Environmental Reseau*, 18 :117–130, 2004.

[50] P. Kenny, M. Lennig, and P. Mermelstein. A linear predictive HMM for vector-valued observations with applications to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(2) :220–225, 1990.

[51] Kiên Kiêu, Katarzyna Adamczyk-Chauvat, Hervé Monod, and Radu S. Stoica. A completely random -tessellation model and Gibbsian extensions. *Spatial Statistics*, 6 :118–138, November 2013.

[52] R Kluszczynski, M N Van Lieshout, and Tomasz Schreiber. Image segmentation by polygonal Markov fields. *Annals of the Institute of Statistical Mathematics*, 59(3) :465–486, 2007.

[53] Caroline Lacoste, Xavier Descombes, and Josiane Zerubia. Point processes for unsupervised line network extraction in remote sensing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10) :1568–1579, 2005.

[54] Florent Lafarge, Georgy Gimel'farb, and Xavier Descombes. Geometric Feature Extraction by a Multimarked Point Process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9) :1597–1609, 2010.

[55] Christian Lantuéjoul. *Geostatistical simulation : models and algorithms*. Number 1139. Springer, 2002.

[56] Federica Larcher and Jacques Baudry. Landscape grammar : a method to analyse and design hedgerows and networks. *Agroforestry Systems*, 87(1) :181–192, June 2013.

[57] Claire Lavigne, Etienne K. Klein, Jean-Franois Mari, Florence Le Ber, Katarzyna Adamczyk, Hervé Monod, and Frédérique Angevin. How do genetically modified (GM) crops contribute to background levels of GM pollen in an agricultural landscape ? *Journal of Applied Ecology*, 45(4) :1104–1113, August 2008.

[58] Florence Le Ber, Marc Benoît, Céline Schott, Jean-François Mari, and Catherine Mignolet. Studying crop sequences with CarrotAge, a HMM-based data mining software. *Ecological Modelling*, 191(1) :170–185, 2006.

[59] Florence Le Ber, Claire Lavigne, Katarzyna Adamczyk, Frédérique Angevin, Nathalie Colbach, Jean-François Mari, and Hervé Monod. Neutral modelling of agricultural landscapes by tessellation methods - Application for gene flow simulation. *Ecological Modelling*, 220(24) :3536–3545, 2009.

[60] Florence Le Ber, Claire Lavigne, Jean-François Mari, Katarzyna Adamczyk, and Frédérique Angevin. GenExP, un logiciel pour simuler des paysages agricoles, en vue de l'étude de la diffusion de transgènes. pages 1–12, 2006.

[61] H Lebesgue. Sur les fonctions représentables analytiquement. *Journal de mathematiques pures et appliquees*, 1 :139–216, 1905.

[62] Jia Li, Lei Yao, Ella Hendriks, and James Z Wang. Rhythmic brushstrokes distinguish van gogh from his contemporaries : findings via automated brushstroke extraction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6) :1159–1176, 2012.

[63] Fabien Liagre. *Les haies rurales : rôles, création, entretien.* France Agricole Editions, 2006.

[64] BC Lovell. Hidden markov models for spatio-temporal pattern recognition and image segmentation. *international Conference on Advances in Pattern Recognition*, 2003.

[65] Roland Maier and Volker Schmidt. Stationary iterated tessellations. *Advances in Applied Probability*, 35(2) :337–353, 2003.

[66] Nikhil Mantrawadi, Mais Nijim, and Young Lee. Object identification and classification in a high resolution satellite data using data mining techniques for knowledge extraction. *2013 IEEE International Systems Conference (SysCon)*, pages 750–755, April 2013.

[67] Jean François Mari, El Ghali Lazrak, and Marc Benoît. Time space stochastic modelling of agricultural landscapes for environmental issues. *Environmental Modelling & Software*, 46 :219–227, August 2013.

[68] Jean-François Mari and Florence Le Ber. Temporal and spatial data mining with second-order hidden markov models. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 10(5) :406–414, 2006.

[69] Jean-François Mari, Florence Le Ber, El-Ghali Lazrak, Marc Benoît, Catherine Eng, Annabelle Thibessard, Pierre Leblond, and Others. Using Markov Models to Mine Temporal and Spatial Data. *New Fundamental Technologies in Data Mining*, pages 561–584, 2011.

[70] Dominique Mazzi and Silvia Dorn. Movement of insect pests in agricultural landscapes. *Annals of Applied Biology*, 160(2) :97–113, 2012.

[71] Burghard C Meyer, Torsten Wolf, and Ralf Grabaum. A multifunctional assessment method for compromise optimisation of linear landscape elements. *Ecological Indicators*, 22(0) :53–63, 2012.

[72] André Meynier. Les paysages agraires. *Les paysages agraires*, page 199, 1958.

[73] Pascal Monestiez, Jean-Stéphane Bailly, Philippe Lagacherie, and Marc Voltz. Geostatistical modelling of spatial processes on directed trees : Application to fluvisol extent. *Geoderma*, 128(3-4) :179–191, 2005.

[74] E Moore. On Certain Crinkly Curves. 1900.

[75] Werner Nagel, Joseph Mecke, Joachim Ohser, and Viola Weiss. A tessellation model for crack patterns on surfaces. 2008.

[76] Werner Nagel and Viola Weiss. Crack STIT Tessellations : Characterization of Stationary Random Tessellations Stable with Respect to Iteration. *Advances in Applied Probability*, 37(4) :859–883, 2005.

[77] Annie Ouin, Manuel Martin, and Françoise Burel. Agricultural landscape connectivity for the meadow brown butterfly Maniola jurtina. *Agriculture, Ecosystems & Environment*, 124(3) :193–199, 2008.

[78] Alessandro Paletto and Chincarini, Marta. Heterogeneity of linear forest formations : differing potential for biodiversity conservation. A case study in Italy. *Agroforestry Systems*, 86(1) :83–93, March 2012.

[79] Manish Parashar, James C Browne, Carter Edwards, and Kenneth Klimkowski. A common data management infrastructure for adaptive algorithms for PDE solutions. In *Supercomputing, ACM/IEEE 1997 Conference*, page 56. IEEE, 1997.

[80] Abani Patra and J.Tinsley Oden. Problem decomposition for adaptive hp finite element methods. *Computing Systems in Engineering*, 6(2) :97–109, April 1995.

[81] Giuseppe Peano. Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen*, 36(1) :157–160, 1890.

[82] Michel Perigord. Essai de traitement géographique des paysages : l'exemple du Limousin. *Norois*, 162(1) :235–256, 1994.

[83] Wojciech Pieczynski. Modèles de Markov en traitements d'images Markov models in image processing. *Traitement du Signal*, 20(3) :255–278, 2003.

[84] Joël Quinqueton and Marc Berthod. A locally adaptive Peano scanning algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (4) :403–412, 1981.

[85] Gerhard Rigoll, Andreas Kosmala, J Rattland, and C Neukirchen. A comparison between continuous and discrete density hidden Markov models for cursive handwriting recognition. In *Pattern Recognition, 1996, Proceedings of the 13th International Conference*, volume 2, pages 205–209, 1996.

[86] KH Riitters, RV O'neill, CT Hunsaker, and A. A factor analysis of landscape pattern and structure metrics. *Landscape Ecology*, 10(1) :23–39, 1995.

[87] Hans Sagan. *Space-filling curves*. Springer-Verlag New York, 1994.

[88] Stefan Schamberger and Jens-Michael Wierum. Partitioning finite element meshes using space-filling curves. *Future Generation Computer Systems*, 21(5) :759–766, 2005.

[89] Wacław Sierpiński. *Sur une nouvelle courbe continue qui remplit toute une aire plane*. 1912.

[90] Radu Stoica. *Processus ponctuels pour l'extraction de réseaux linê\ "iques dans les images satellitaires et aériennes*. PhD thesis, Université de Nice-Sophia Antipolis, Nice, France, 2001.

[91] Dietrich Stoyan, Wilfrid S Kendall, Joseph Mecke, and L Ruschendorf. *Stochastic geometry and its applications*, volume 2. Wiley Chichester, 1995.

[92] Dietrich Stoyan and Antti Penttinen. Recent applications of point process methods in forestry statistics. *Statistical Science*, 15(1) :61–78, 2000.

[93] Dietrich Stoyan and Helga Stoyan. Non-Homogeneous Gibbs Process Models for Forestry — A Case Stud. *Biometrical Journal*, 40(5) :521–531, 1998.

[94] Hakon Tjelmeland and Julian Besag. Markov Random Fields with Higher-order Interactions. *Scandinavian Journal of Statistics*, 25(3) :415–433, 1998.

[95] Monica G Turner and Robert H Gardner. Quantitative met hods in landscape ecology. In *Ecological Studies Analysis and Synthesis : Quantitative Methods in Landscape Ecology : The Analysis and Interpretation of Landscape Heterogenetiy*, chapter 1. Springer-Verlag, New York, 1991.

[96] Rebecca Tyson, Howard Thistlewood, and Gary J R Judd. Modelling dispersal of sterile male codling moths,Cydia pomonella, across orchard boundaries. *Ecological modelling*, 205(1) :1–12, 2007.

[97] Emma H. van der Zanden, Peter H. Verburg, and Caspar a. Mücher. Modelling the spatial distribution of linear landscape elements in Europe. *Ecological Indicators*, 27 :125–136, April 2013.

[98] Marie-Colette N M van Lieshout. *Markov point processes and their applications*. Imperial College Press, 2000.

[99] Clémence Vannier. Analyse spatiale de structures paysagères en contexte agricole bocager. *Cybergeo : European Journal of Geography*, 2012.

[100] Clémence Vannier and Laurence Hubert-Moy. Wooded hedgerows characterization in rural landscape using very high spatial resolution satellite images. In *2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 347–350, 2010.

[101] Andrea Veres, Sandrine Petit, Cyrille Conord, and Claire Lavigne. Does landscape composition affect pest abundance and their control by natural enemies ? A review. *Agriculture, Ecosystems & Environment*, 166 :110–117, 2013.

[102] Michael S Warren and John K Salmon. A parallel hashed oct-tree n-body algorithm. In *Proceedings of the 1993 ACM/IEEE conference on Supercomputing*, pages 12–21, 1993.

[103] Thorsten Wiegand, Kirk A Moloney, Javier Naves, and Felix Knauer. Finding the missing link between landscape structure and population dynamics : a spatially explicit perspective. *The American Naturalist*, 154(6) :605–627, 1999.

[104] Chuanrong Zhang and Weidong Li. Markov Chain Modeling of Multinomial Land-Cover Classes. *GIScience & Remote Sensing*, 42(1) :1–18, March 2005.

Annexe A

Manuel utilisateur de l'Outil

L'outil présenté ici rassemble les méthodes décrites dans la thèse (chapitre 4, 5, 6 et 7). Il a été développé avec l'aide de deux étudiants de l'école TELECOM à Nancy, Léo Demangeon et Maxime Pallez. La totalité du développement a été fait en langage C.

A.1 Interface et utilisation

L'interface graphique (Figure A.1) a été pensée de façon à être intuitive, elle n'est constituée que d'une seule fenêtre, décomposée en plusieurs parties, sous forme de *expander*. Les différents paramètres à renseigner par l'utilisateur sont ceux nécessaires à la bonne exécution des fonctions.

Dans la suite du chapitre, le nom des paramètres sera écrit en italique et sera suivi par le type (int, double, ...) entre parenthèses.

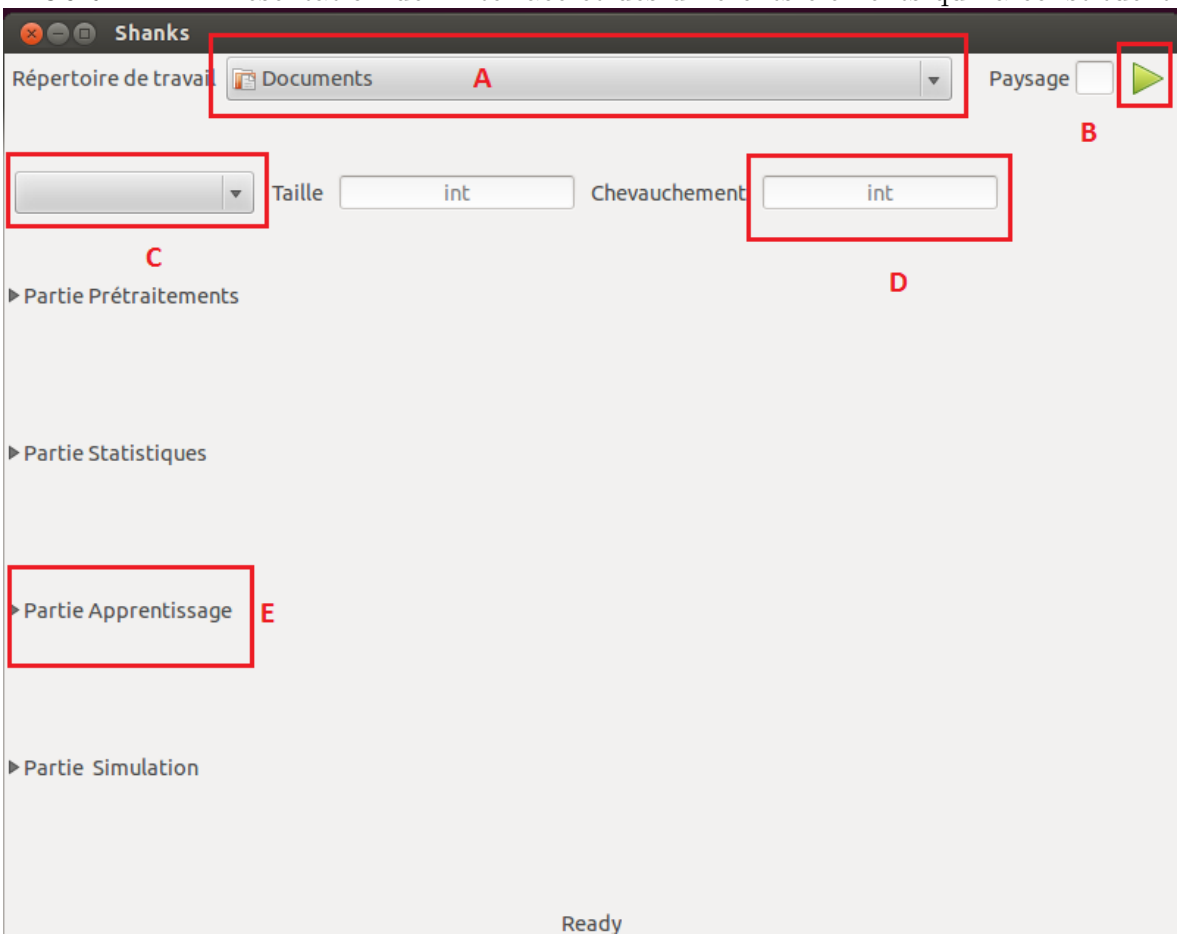
Exemple :

$$niveau\ de\ découpe(int)$$

Il a été choisi, lors du développement du logiciel, de laisser la possibilité à l'utilisateur de choisir son répertoire de travail, il devra le spécifier dans le sélecteur présent en haut de l'interface (Encart A : Figure A.1). Il n'est pas nécessaire pour l'utilisateur de créer l'arborescence des dossiers, ceci se fait automatiquement.

Une fois le dossier choisi, l'utilisateur devra remplir l'encart **Paysage** à côté de la flèche verte (Encart B : Figure A.1). Celui-ci sert à indiquer la lettre avec laquelle le paysage sera nommé par la suite. L'utilisateur devra également spécifier différents paramètres, correspondant aux paramètres de découpe (cf. paragraphe A.1.1), créant

FIGURE A.1 – Présentation de l'interface et des différents éléments qui la constituent



ainsi un dossier propre à chaque quadruplet de paramètres choisis :

$$mode\ de\ découpe, taille\ de\ cellule(int), chevauchement(int), paysage(char)$$

Ces paramètres ne sont pas directement utilisés par la fonction de découpe (cf. paragraphe A.1.1), ils permettent simplement de créer un dossier avec un nom paramétré, facilitant ainsi la distinction des dossiers de travail de l'utilisateur. Cependant, ils doivent être en permanence renseignés.

Une fois ces cinq paramètres indiqués, en cliquant sur le bouton **run** (Encart B : A.1), l'utilisateur créera un dossier du type

$$/repertoire/paysage - taille\_de\_cellule - chevauchement - mode\_de\_decoupe$$

où répertoire correspond au répertoire spécifié par l'utilisateur dans le sélecteur de dossier. Ce répertoire contiendra tous les fichiers créés lorsque l'utilisateur appellera

les fonctions du logiciel [3].

L'interface graphique a été décomposée en quatre parties qui sont

Prétraitements, Statistiques, Apprentissage, Simulation

Pour chacune de ces parties, nous allons présenter toutes les fonctions appelées, voir leur utilisation ainsi que les paramètres nécessaires à leur bon fonctionnement et enfin, étudier les fichiers de sortie.

L'interface graphique a été réalisée en gardant la nature séquentielle des fonctions. Pour la plupart d'entre elles, l'utilisateur aura besoin des fichiers créés lors de l'exécution des fonctions dans les onglets précédents. Il doit donc vérifier que les fichiers nécessaires à la bonne exécution des fonctions soient présents dans le dossier de travail [4].

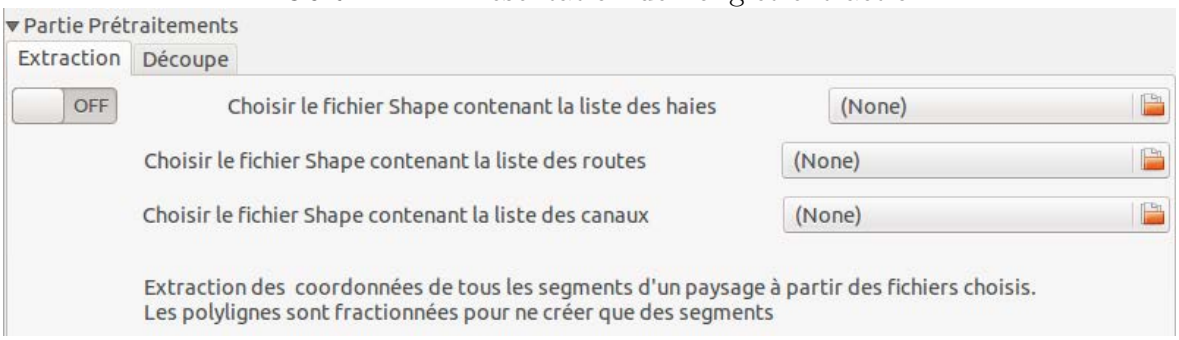
Par la suite, les différents éléments de l'interface seront appelés comme suit :

— Sélecteur de dossier : Encart A Figure A.1
— Bouton Run : Encart B Figure A.1
— Liste de choix : exemple Encart C Figure A.1
— Partie : exemple Encart E Figure A.1
— Onglets : Les différents onglets que l'on peut choisir dans les parties

A.1.1 Partie Prétraitement

Onglet : Extraction

FIGURE A.2 – Présentation de l'onglet extraction



L'**Extraction** (Figure A.2) se base sur des résultats récupérés sous forme de fichier Shape (.shp). L'utilisateur a la possibilité de choisir trois fichiers, correspondant aux

3. sauf la partie Extraction
4. Étape optionnelle si l'utilisateur n'interrompt pas son étude

fichiers Shape des Haies, Routes et Canaux pour un paysage donné. Attention, l'ensemble des fichiers déterminant la base de données en Shape doit être présent (fichier en *.prj, .qix, .qml, .shx*). Pour lancer cette fonction, l'utilisateur doit activer le bouton (**ON/OFF**). Un dossier *"files/Extraction"* sera alors créé dans le répertoire choisi précédemment par l'utilisateur. En fonction des boutons activés, les fichiers *"Haies-**Paysage**.csv", "Canaux-**Paysage**.csv", "Routes-**Paysage**.csv"* seront créés. Ici, **Paysage** correspond à la lettre renseignée précédemment par l'utilisateur dans l'encart à côté de la flèche **run**. Ces fichiers contiennent une liste de coordonnées qui correspondent aux coordonnées des extrémités des segments construites à partir du fichier Shape.

Soient X_0^i, Y_0^i, X_1^i et Y_1^i les coordonnées du i^{eme} segment, alors il est possible que l'utilisateur trouve des valeurs de i telles que $(X_1^i, Y_1^i) = (X_0^{i+1}, Y_0^{i+1})$. Dans ce cas, cela correspond à une polyligne dont les parties ont été considérées comme des segments.

Cette fonction crée un fichier du type

/répertoire/Extraction/{Haies, Canaux, Routes}-Paysage.csv

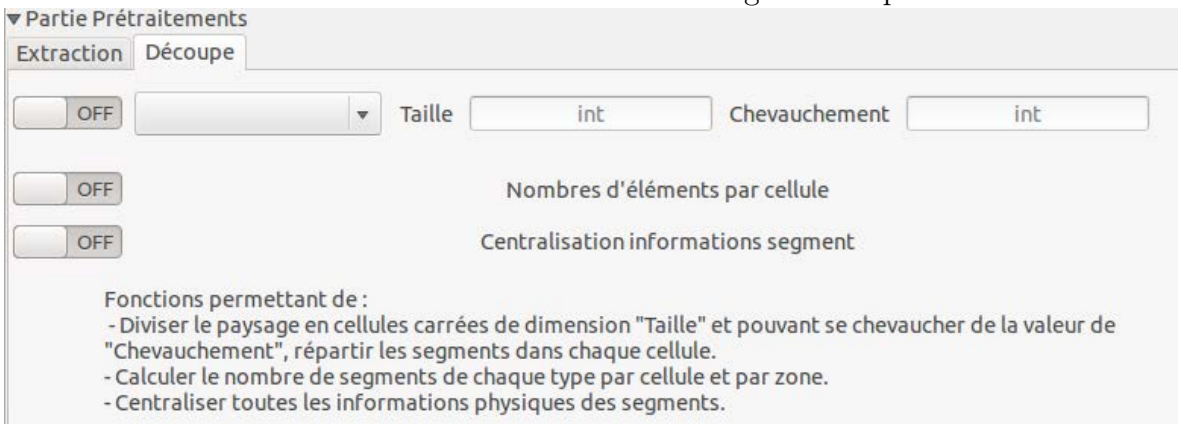
Dans ce fichier, les données sont de la forme :

CoordSeg

où **CoordSeg** est la liste des coordonnées de chaque segment avec (X_0, Y_0) pour une extrémité et (X_1, Y_1) pour l'autre. Le fichier C correspondant à cette fonction est *extraction.c*

Onglet : Découpe

FIGURE A.3 – Présentation de l'onglet découpe



L'onglet **découpe** (Figure A.3) comporte trois boutons permettant de lancer trois fonctions différentes. On retrouve trois paramètres :

$$methode\ de\ decoupe, taille\ de\ cellule(int), chevauchement(int)$$

Ces paramètres diffèrent de ceux présents en haut de la fenêtre car ils servent à la fonction, tandis que ceux présents en haut servent à créer ou choisir le répertoire de travail.

Découpe La première fonction permet de découper un paysage grâce à des cellules carrées recouvrant le rectangle circonscrit au paysage selon deux méthodes :

- A l'origine : les cellules sont créées en partant du coin en haut à gauche
- Centrée : les cellules sont centrées à l'intérieur du rectangle circonscrit au paysage

Les cellules créées ne recouvrant pas exactement le rectangle circonscrit, la différence entre les deux méthodes tient à l'emplacement des excédents par rapport au rectangle, à droite et en bas pour la première ; tout autour de l'ensemble des cellules pour la seconde. La découpe se matérialise en la répartition des segments dans les cellules, selon leur type (Haie, Route, Canal).

Le premier bouton permet d'effectuer une découpe des données précédemment extraites selon la taille de cellule renseignée. Chaque segment, représenté par ses coordonnées, est réparti dans les cellules. La fonction crée des fichiers, pour chaque cellule, du type

*/répertoire/Cells/Cell-**Paysage-Ligne_Colonne**.csv*

Dans le cas où un segment est contenu dans plusieurs cellules, il est écrit dans toutes les cellules où il est présent.

Dans ce fichier, les données sont de la forme :

Paysage Ligne Colonne Type Id Bool CoordSeg

où

- Paysage représente la lettre qui désigne notre paysage
- Ligne représente la position de la cellule (contenant le segment) par rapport au point origine (en haut à gauche) suivant l'axe des abscisses (commence à 0)
- Colonne représente la position de la cellule (contenant le segment) par rapport au point origine (en haut à gauche) suivant l'axe des ordonnées (commence à 0)

— Type représente la nature du segment (H pour Haie, R pour Route, C pour Canal)
— Id représente l'identifiant numérique du segment assigné par la fonction **découpe**
— Bool représente un booléen indiquant si le segment est entièrement contenu dans la cellule ou non
— CoordSeg est la liste des 4 coordonnées du segment : X_0, Y_0, X_1, Y_1

Un fichier du type

*répertoire/ResumDecoup-**Paysage**.csv*

est créé et contient l'information sur la position spatiale des cellules créées.

Dans ce fichier, les données sont de la forme :

Paysage Ligne Colonne x0 y0 x1 y1

où

— $x0$ $y0$ représente le point extrême en haut à gauche de la cellule
— $x1$ $y1$ représente le point extrême en bas à droite de la cellule

Enfin, le dernier fichier créé se nomme

*/répertoire/PaysageInit-**Paysage**.csv*

et contient les coordonnées du rectangle circonscrit au paysage.

Ces fichiers sont indispensables au fonctionnement des traitements suivants. Le fichier C correspondant à cette fonction est *decoupe.c*

Nombre d'éléments par cellule Cette fonction utilise les fichiers créés avec la fonction découpe. Elle va permettre de compter le nombre de segments de chaque type dans une cellule ainsi que dans sa zone (la cellule et les 8 cellules voisines directes). Le fichier produit par cette fonction stockera également, pour chaque cellule, le nombre de haies de type HV et le nombre de haies de type HP.

Cette fonction crée un fichier du type :

*/répertoire/NbElCell-**Paysage**.csv*

Dans ce fichier, les données sont de la forme :

Paysage Numéro Ligne Colonne NbH_C NbR_C NbC_C ...
... NbH_Z NbR_Z NbC_Z NbH_Para_C NbH_Perp_C

où

— Numéro représente le numéro unique de la cellule
— NbH_C représente le nombre de haies dans la cellule
— NbR_C représente le nombre de routes dans la cellule
— NbC_C représente le nombre de canaux dans la cellule
— NbH_Z représente le nombre de haies dans la zone
— NbR_Z représente le nombre de routes dans la zone
— NbC_Z représente le nombre de canaux dans la zone
— NbH_HP_C représente le nombre de haies de type HP dans la cellule
— NbH_HV_C représente le nombre de haies de type HV dans la cellule

Le fichier C correspondant à cette fonction est *nb_elements_by_cell.c*

Résumé des informations En activant, le dernier bouton (**ON/OFF**), l'utilisateur obtiendra un fichier contenant le résumé des informations connues jusqu'à présent sur le paysage. Ainsi, l'accès à une liste complète des segments contenus dans le paysage sera rendu possible. Cette fonction a besoin, en entrée, des fichiers issus de la fonction découpe.

Elle crée un fichier du type :

*/répertoire/SummarySegment-**Paysage**.csv*

Dans ce fichier, les données sont de la forme :

Paysage Ligne Colonne Type Id Bool CoordSeg ...
... Bool_1 Coord_Bar Lng Ang Orientation

où

— Bool est un booléen valant 1 si le segment est entièrement contenu dans la cellule
— CoordSeg représente un vecteur contenant les coordonnées des extrémités du segment
— Bool_1 est un booléen valant 1 si l'isobarycentre du segment est contenu dans la cellule
— Coord_Bar représente un vecteur contenant les coordonnées de l'isobarycentre du segment
— Lng représente la longueur du segment
— Ang représente l'angle du segment

A.1.2 Partie Statistiques

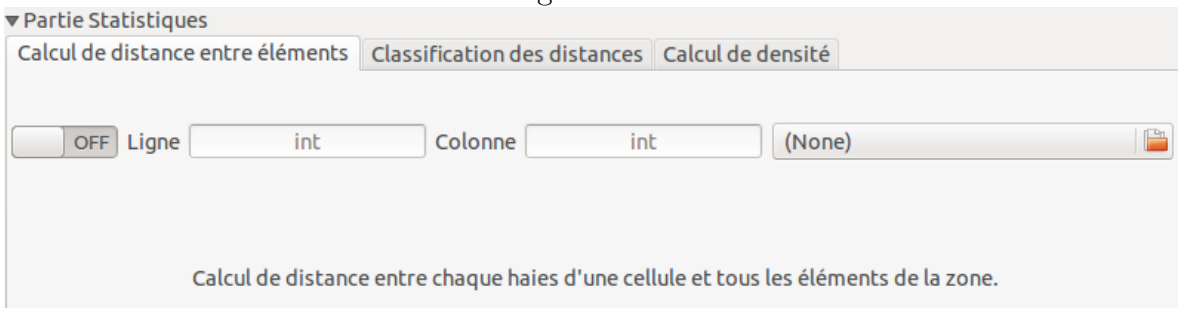
L'utilisateur a la possibilité d'effectuer les calculs sur une cellule représentée par un couple $(Ligne; Colonne)$ ou sur un ensemble de cellules. Dans la seconde option, l'utilisateur doit, à partir de l'outil réservé à cet effet, choisir un fichier de type *.csv* dans lequel les données doivent être de la forme :

Paysage Ligne Colonne ...

En réalité, seules les colonnes 2 et 3 seront utilisées pour cette fonction, mais le programme a été développé de façon à ce que ce type de fichier, permettant d'accélérer les traitements sur plusieurs cellules, soit réutilisé.

Onglet : Calcul de distance entre éléments

FIGURE A.4 – Présentation de l'onglet de calcul de distance entre les éléments



La fonction **Calcul de distance entre éléments** (Figure A.4) permet de calculer pour chaque haie de la cellule ciblée, toute une série de mesure par rapport aux segments des cellules de sa zone (cellule centrale et ses 8 cellules voisines directes). La fonction permet de calculer la distance entre les isobarycentres des segments, la distance DiSt entre les segments ainsi que la différence d'angle en radian entre les segments. Cette fonction peut présenter un temps de calcul de l'ordre de quelques secondes. Elle produit un fichier du type

*/repertoire/Distance/Dist-**Paysage-Ligne_Colonne**.csv*

Dans ce fichier, les données sont de la forme :

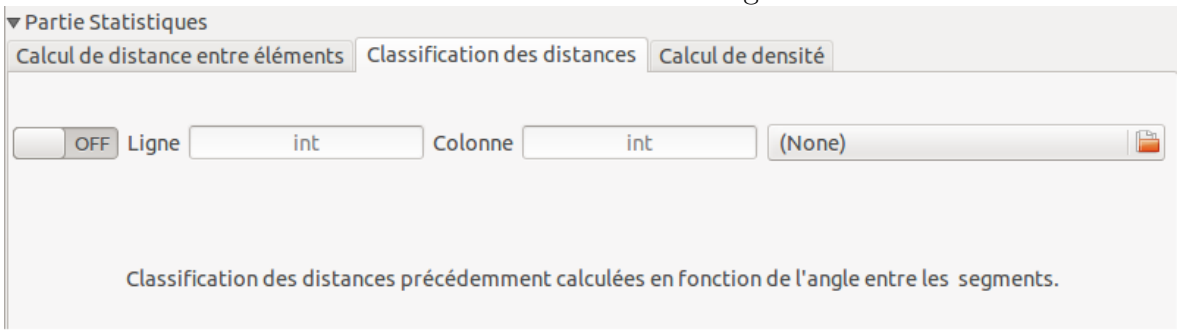
Ligne_1 Colonne_1 Type_1 Id_1 Ligne_2 Colonne_2 Type_2 Id_2 ...
... DistBary(S_1, S_2) **DiSt**(S_1, S_2) **DiffAngle**(S_1, S_2)

où

— Les paramètres se terminant par _1 correspondent à la haie de la cellule étudiée.
— Les paramètres se terminant par _2 correspondent aux segments se trouvant dans la zone cellulaire.
— $DistBary(S_1, S_2)$ représente la distance entre les isobarycentres des segments S_1 et S_2.
— $DiSt(S_1, S_2)$ représente la distance entre les segments S_1 et S_2.
— $DiffAngle(S_1, S_2)$ représente la différence d'angle entre les segments S_1 et S_2 en prenant l'axe vertical comme référentiel, et toujours dans le sens inverse trigonométrique.

Onglet : Classification des distances

FIGURE A.5 – Présentation de l'onglet distance



La fonction **Classification des distances** (Figure A.5) permet, à partir du fichier de distance créé par la fonction précédente, de créer de nouveaux fichiers et de trier les couples de segments en fonction de l'angle entre les segments. Trois fichiers seront créés, correspondant aux angles **Parallèles, Perpendiculaire, Autres**. Pour la cellule choisie (ou des cellules concernées par le fichier sommaire choisi), chaque ligne va être consultée et écrite dans un fichier différent en fonction de la valeur de l'angle entre les segments.

Cette fonction produit trois fichiers du type

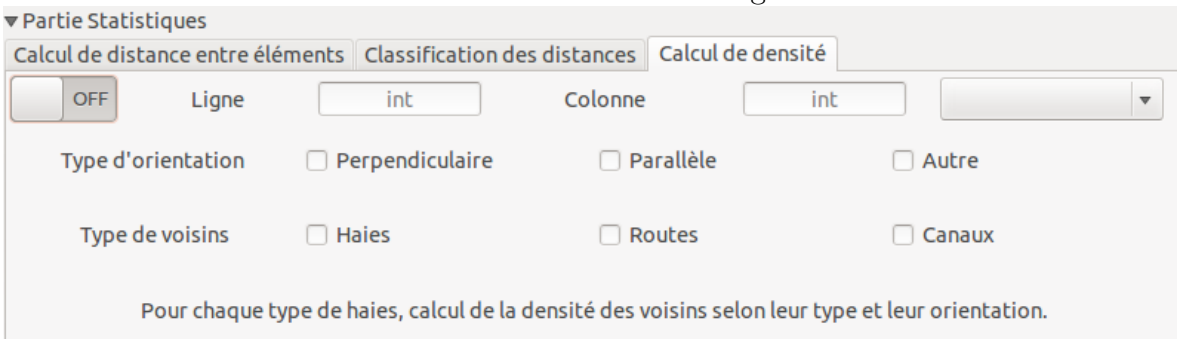
*/répertoire/Distance/Dist-**Paysage-Ligne__Colonne**-{Para, Perp, Autr}.csv*

Ces fichiers contiennent les renseignements sous la même forme que le fichier parent.

Onglet : Calcul de densité

La fonction **Calcul de densité** (Figure A.6) permet de calculer, pour une cellule $\mathcal{C}$, la densité relative $D_r^I(\mathcal{C}, b, C(\theta))$. Le voisinage se détermine autour d'un type de haie

FIGURE A.6 – Présentation de l'onglet densité



(type HP, type HV, type Autres, tout type) en fonction de la distance b, et de l'angle entre segments $C(\theta)$. La zone d'étude est la zone cellulaire.

L'utilisateur doit avant tout renseigner les encarts *Ligne(int) Colonne(int)* correspondant à la cellule ciblée par les calculs. Il devra ensuite choisir le type de haies qu'il souhaite considérer pour son étude (type HP, type HV, type Autres, tout type), le type de voisins (Haies, Routes, Canaux) et l'angle par rapport aux haies qu'il souhaite considérer (Perpendiculaire, Parallèle, Autre). L'utilisateur peut choisir toutes les orientations ou tous les types de voisins, mais il faut savoir que cette fonction est très longue à l'exécution (de l'ordre de 10-15 min pour une cellule de 200 haies dans une zone contenant environ 6000 segments).

Ici les classes de distances sont définies en dur dans le code. Pour les changer, il faudra donc modifier dans le fichier *dens_all.c* la fonction

compute_relative_density_for_all_classes

A.1.3 Partie Apprentissage

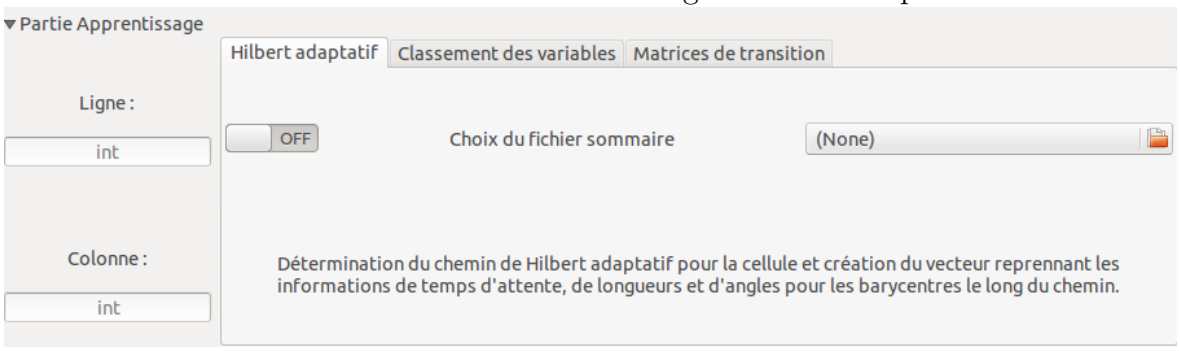
Pour cette partie, dans chacun des onglets, l'utilisateur pourra choisir entre spécifier une cellule par un couple $(Ligne; Colonne)$ ou bien sélectionner un fichier sommaire, permettant d'appliquer les fonctions de la partie apprentissage à toutes les cellules renseignées dans ce fichier. Le choix du fichier est prioritaire sur le choix du couple $(Ligne; Colonne)$ c'est à dire que si l'utilisateur renseigne d'un coté une cellule par une ligne et une colonne et de l'autre un fichier sommaire, les calculs ne seront effectués que sur la liste de cellules contenues dans le fichier sommaire.

Celui-ci doit être construit comme suit :

Paysage Ligne Colonne ...

Onglet : Hilbert adaptatif

FIGURE A.7 – Présentation de l'onglet Hilbert adaptatif



Pour la partie apprentissage, l'utilisateur doit tout d'abord utiliser la fonction Hilbert adaptatif (Figure A.7), permettant de classer les segments des cellules réelles suivant un chemin de Hilbert adaptatif. Cette fonction va permettre à partir des isobarycentres d'une cellule donnée de construire le chemin de Hilbert adaptatif. Il est adaptatif car la découpe de la cellule s'adapte aux nombres de segments. Si pour un niveau de découpe, plusieurs isobarycentres sont dans une case, cette case sera divisée mais si une case ne contient qu'au plus un barycentre, celle-ci ne sera plus divisée.

Cette fonction crée de nombreux fichiers temporaires qui ne seront pas tous détaillés ici, mais le fichier résultat principal de cette fonction est du type

*/répertoire/H-A/CMC-**Paysage-Lignes_Colonne**.csv*

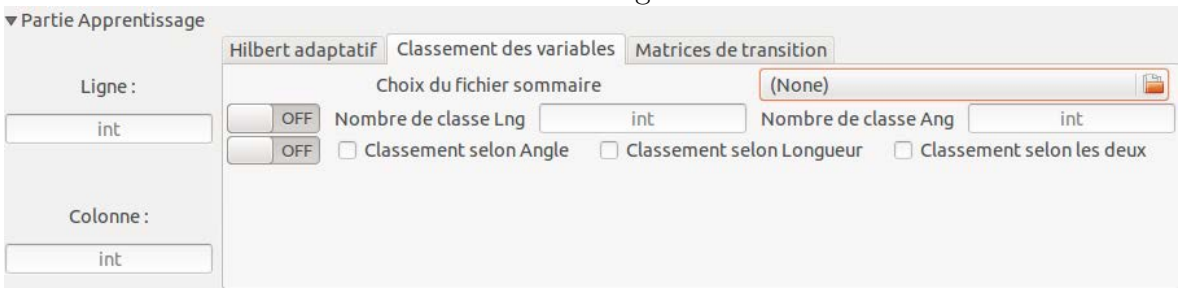
Dans ce fichier, les données sont de la forme

NuméroBary NiveauDecoupe Lng Ang TempsAttente

où

- NuméroBary représente l'Id du segment dont est issu l'isobarycentre.
- NiveauDecoupe représente la taille de la case créée par le chemin.
- TempsAttente représente le temps d'attente pour arriver à cet isobarycentre.
- Lng représente la longueur du segment dont est issu l'isobarycentre.
- Ang représente l'angle entre la verticale et le segment dont est issu l'isobarycentre.

FIGURE A.8 – Présentation de l'onglet classement des variables



Onglet : Classement des variables

Cet onglet (Figure A.8) comporte deux boutons (**ON/OFF**) permettant de lancer deux fonctions complémentaires.

Création du fichier résumé de classe La première fonction permet de créer un fichier résumé des classes utilisé pour classer les variables angle et longueur. L'utilisateur peut spécifier deux nombres de classes différents pour les longueurs et les angles mais il n'a pas à spécifier un fichier sommaire ou un couple $(Ligne; Colonne)$ puisque les classes concernent un paysage tout entier.

Les fonctions suivantes utilisent ce fichier résumé pour classer les longueurs et les angles afin de créer des matrices de transitions sur les variables classées.

Les classes sont créées à partir du fichier *SummarySegment-**Paysage**.csv* (cf. section A.1.1.0.0). En effet, une liste contenant les valeurs d'angle et une autre contenant les valeurs de longueur sont créées, et le nombre de classes voulues par l'utilisateur détermine le nombre de partitions de ces vecteurs de valeurs.

Cette fonction produit un fichier du type

$$textit/repertoire/H - A/ResumClass - \textbf{Paysage}.csv$$

Le contenu du fichier est :

— Première ligne : Liste des classes de longueurs

— Deuxième ligne : Liste des classes d'angles

Classification des valeurs La deuxième fonction permet de classer les variables "*Longeur, Angle*" dans les fichiers créés par la fonction Hilbert adaptatif selon les classes que l'utilisateur vient de spécifier. Cette fonction va donc remplacer les valeurs réelles des longueurs et des angles contenues dans les fichiers des cellules choisies par les numéros des classes correspondants.

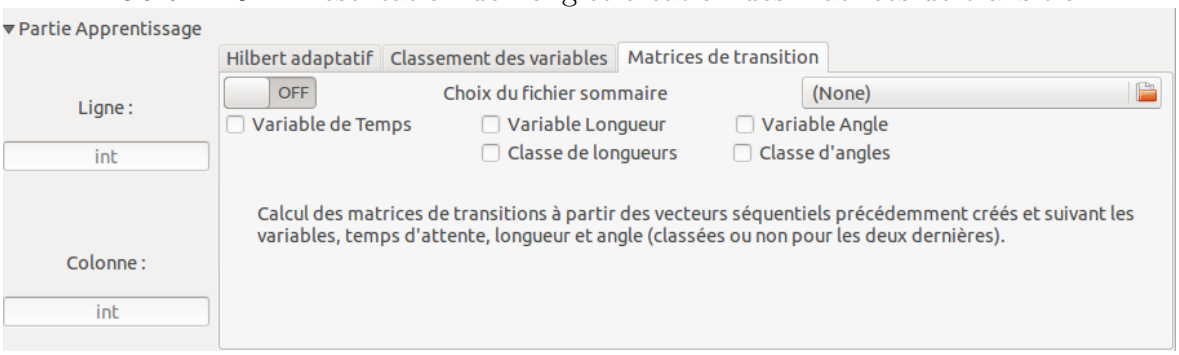
Les fichiers produits sont du type

*/répertoire/CMC-Sort-**Variable_Classée-Paysage-Ligne_Colonne**.csv*

Comme dit précédemment, ces fichiers contiennent les mêmes informations que le fichier produit par la fonction Hilbert adaptatif, les valeurs des longueurs (ou angles ou les deux) sont remplacées par les valeurs des classes correspondantes.

Onglet : Matrices de transition

FIGURE A.9 – Présentation de l'onglet création des matrices de transition



La fonction **Création de Matrice de transition** (Figure A.9) va permettre de créer les matrices de transition, à partir des fichiers précédemment créés, outils indispensables pour effectuer les calculs dans la partie Simulation (Chapitre A.1.4) . La matrice de transition des états correspond aux probabilités de passer d'un état à un autre. L'utilisateur doit spécifier pour quelle variable il souhaite créer la matrice de transition, il peut créer les trois matrices à la fois. Par défaut, les matrices créées pour les longueurs et les angles ne prennent pas en compte les classes de valeurs. Si l'utilisateur souhaite utiliser les classes de valeur qu'il a créées précédemment, il lui faudra cocher la case réservée à cet effet.

Cette fonction produit un fichier du type

*/répertoire/Markov/Matrix-**Variable-Paysage-Ligne_Colonne-Param**.csv*

où

— Variable appartient à {Tps, Ang, Lng}

— Param indique si l'on a utilisé un fichier classé ou non

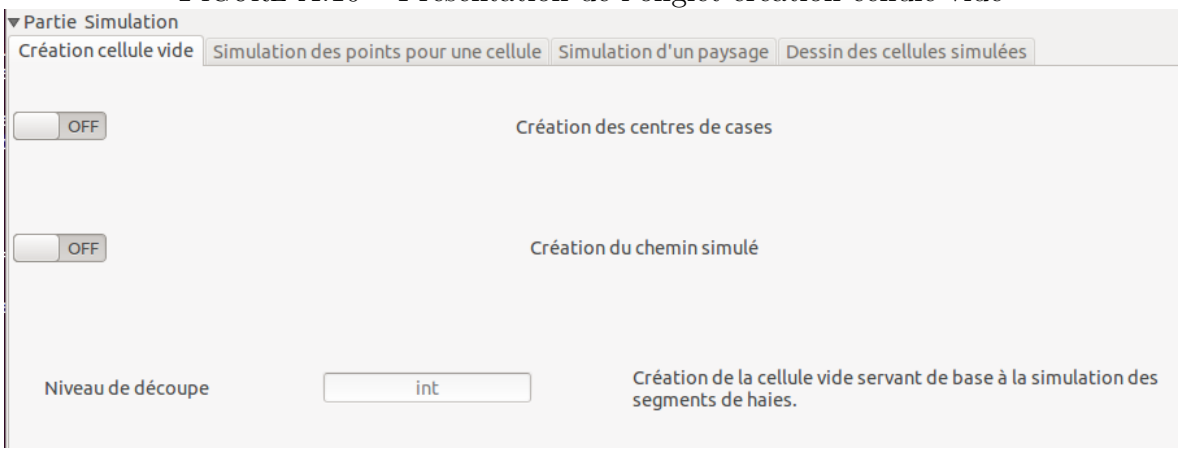
Dans ce fichier, les données doivent être consultées sous forme de tableau. Le premier chiffre, en haut à gauche, sera toujours 0. Ensuite, le reste de la première ligne et de

la première colonne reprendra la liste des états de la variable dans la cellule. Enfin, chaque chiffre du tableau représente la probabilité de passer de l'état écrit en début de ligne à l'état écrit en début de colonne.

A.1.4 Partie Simulation

Onglet : Création du chemin

FIGURE A.10 – Présentation de l'onglet création cellule vide



Cet onglet (Figure A.10) comporte deux boutons (**ON/OFF**) permettant de lancer deux fonctions complémentaires.

Création de la liste des centres de cases La première fonction permet de créer une liste de centres de cases à partir d'une taille de cellule et d'un niveau de découpe. Le niveau de découpe correspondant au nombre de fois que la fonction doit découper la cellule afin d'obtenir les cases souhaitées. La taille d'une cellule correspond au paramètre précédemment rentré par l'utilisateur tout en haut de l'interface graphique (entre les encart C et D dans la figure A.1). Le niveau de découpe est quant à lui à renseigner dans l'encart correspondant (*niveauDeDecoupe*(int)).

Pour gagner du temps, l'utilisateur a la possibilité de rentrer une plage de valeur pour le niveau de découpe, permettant d'effectuer les calculs sur cette plage. Ainsi, s'il souhaite effectuer les calculs pour les niveaux de découpe de 1 à 7, il lui suffira d'entrer $1-7$. L'exécution de cette fonction est presque instantanée, il faudra cependant être prudent si l'on souhaite mettre des niveaux de découpes supérieurs à 8. Par exemple, le *Niveau de découpe* 10 implique la création de $4^{10} = 1048576$ points.

Cette fonction crée un fichier du type

/répertoire/list_centre_case-taille_max-niveauDeDecoupe.csv

Dans ce fichier, les données sont de la forme :

X Y

où

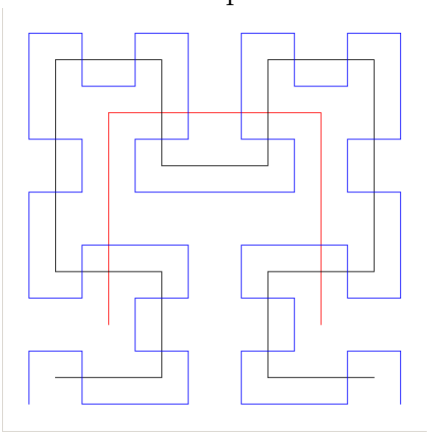
— X représente l'abscisse du centre de la case.

— Y représente l'ordonnée du centre de la case.

Le fichier C correspondant à cette fonction est *creer_centre_case.c*.

Création de la liste des centres de cases classés La deuxième fonction permet de procéder au classement des centres de cases selon le chemin de Hilbert. Elle prend en paramètre le fichier précédemment créé ainsi que le niveau de découpe et la taille de la cellule. Comme pour la fonction précédente, le niveau découpe peut être spécifié comme appartenant à une plage de valeur.

FIGURE A.11 – Les trois premières itérations pour la construction d'un chemin de Hilbert

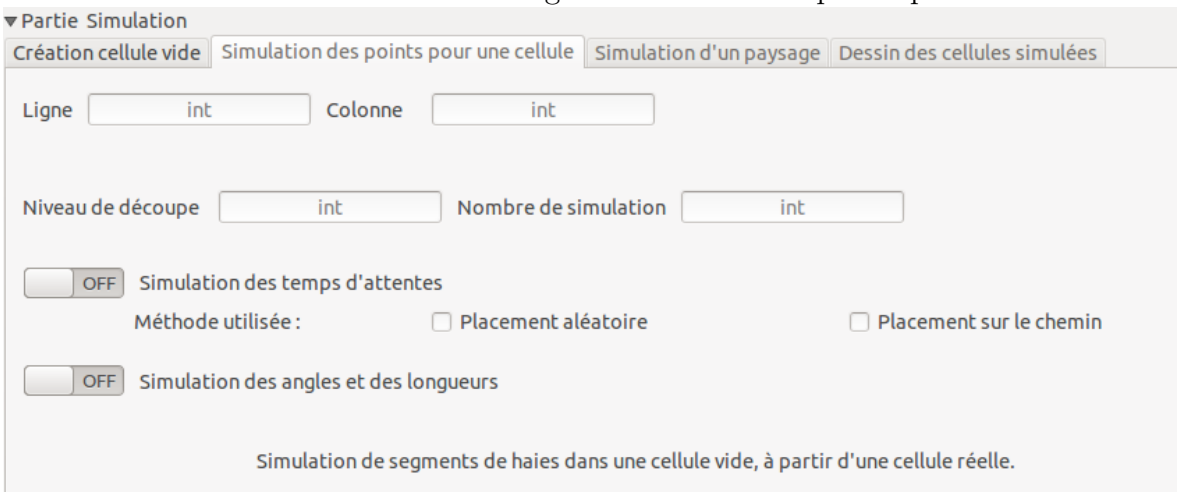


A partir de la liste des centres de cases, le chemin de Hilbert va être construit (suivant le motif initial présenté, en rouge, dans la figure A.11) correspondant au niveau de découpe renseigné par l'utilisateur. Une liste de centres de cases classés selon le chemin sera obtenue, permettant ensuite d'effectuer des simulations.

Les fichiers C correspondant à cette fonction sont *hilbert_chemin_simule.c* et *fenetre_init.c*.

Onglet : Simulation pour une cellule

FIGURE A.12 – Présentation de l'onglet Simulation des points pour une cellule



Dans cet onglet (Figure A.12), la simulation est effectuée à partir d'un couple $(Ligne; Colonne)$, c'est-à-dire une cellule réelle, spécifiée par l'utilisateur. Ce dernier doit choisir entre deux méthodes de simulation :

- *Placement aléatoire* où les points sont placés dans les cases de la cellule selon une loi uniforme en abscisse et en ordonnée.
- *Placement sur le chemin* où les points sont placés exactement sur le chemin de Hilbert qui parcourt les cases.

L'utilisateur doit choisir aussi un niveau de découpe, qui correspondra au choix du fichier de liste de centres de cases classés, ce niveau de découpe est à mettre en relation avec le mode du niveau de découpe calculé sur la cellule réelle. De plus, l'utilisateur a la possibilité de choisir le nombre de simulation de placement des points à effectuer.

A partir de la liste de centres de cases classés et de la matrice de transition des temps d'attente, les point seront placés au fur et à mesure dans la cellule. Après chaque point placé, un temps d'attente est déterminé grâce à la matrice de transition en fonction du temps d'attente précédent, et celui-ci va représenter l'avancement, dans le parcours des cases de la cellule, jusqu'au prochain point.

Cette fonction produit un fichier du type

*/répertoire/Simulation/Sim-**Date-Heure**/**NumSimul**-SimCell-**NiveauDecoupe-Paysage-Ligne__Colonne-Parametre**.csv*

où

— **Date-Heure** correspond à la date et à l'heure auxquelles l'utilisateur lance la simulation.
— NumSimul représente le numéro de la simulation.
— Parametre correspond à la méthode utilisée pour placer les isobarycentres.

Le deuxième bouton permet à l'utilisateur de simuler des valeurs d'angles et de longueurs pour chaque point de la cellule, c'est à dire d'associer à chaque point une longueur réelle appartenant à la classe et une valeur d'angle tirée aléatoirement entre les bornes de la classe.

La fonction crée un premier fichier du type

*/répertoire/Simulation/Sim-**Date-Heure**/**NumSimul**-SimCell-**NiveauDecoupe-Paysage-Ligne_Colonne-Parametre**-final.csv*

où les données sont de la forme :

X Y Ang Lng

où

— **X** et **Y** sont les coordonnées de l'isobarycentre du segment.
— Lng est la longueur associée à ce segment.
— Ang est l'angle associé à ce segment.

et un second du type

*/répertoire/Simulation/Sim-**Date-Heure**/**NumSimul**-SimCell-**NiveauDecoupe-Paysage-Ligne_Colonne-Parametre**-final-seg.csv*

où les données sont de la forme :

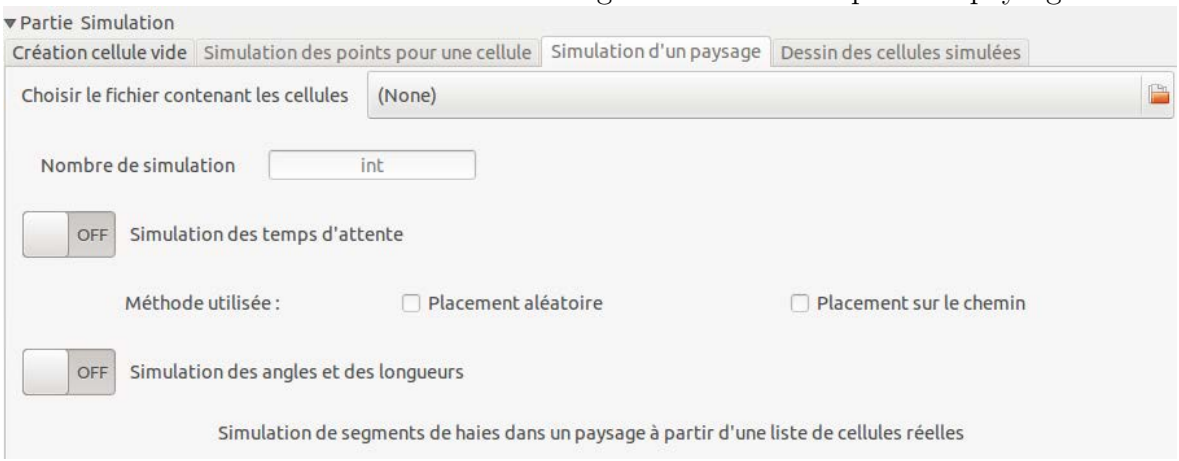
CoordSeg

qui représente les coordonnées des haies simulées.

Onglet : Simulation pour une classe

Cet onglet (Figure A.13) permet d'effectuer les mêmes calculs que l'onglet précédent mais à partir d'un fichier sommaire, choisi par l'utilisateur, qui contiendra une liste de cellules d'un même paysage. La fonction tirera alors aléatoirement une classe dans ce fichier sommaire, et associera le niveau de découpe de la cellule vide au niveau de découpe spécifié dans le fichier.

FIGURE A.13 – Présentation de l'onglet de simulation pour un paysage



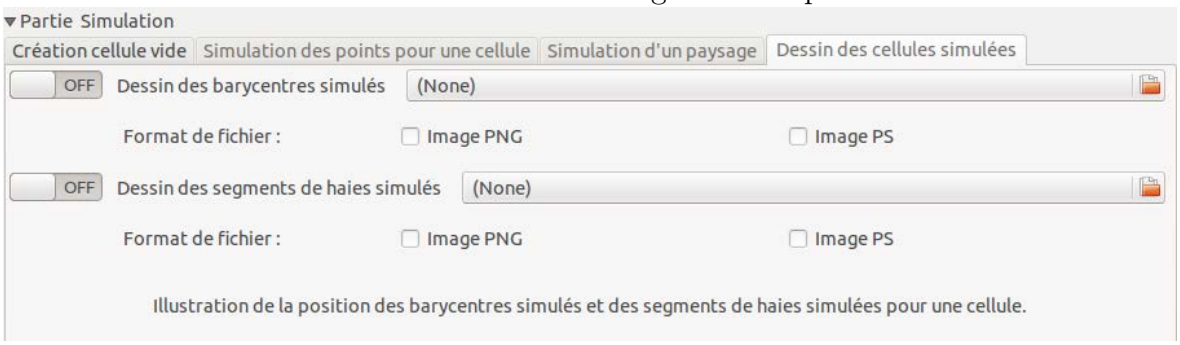
Le fichier sommaire doit alors être construit de cette façon :

Paysage Ligne Colonne Classe NiveauDecoupe

où **NiveauDecoupe** correspond au niveau de découpe que l'utilisateur veut associer aux cellules contenues dans une même classe.

Onglet : Dessin pour une cellule

FIGURE A.14 – Présentation de l'onglet dessin pour une cellule



Cet onglet (Figure A.14) comporte deux boutons (**ON/OFF**) permettant de lancer deux fonctions.

A.1.5 Dessin des barycentres simulés

La première fonction (Premier bouton (**ON/OFF**) dans la figure A.14) donne la possibilité à l'utilisateur de dessiner la liste des isobarycentres obtenus après la simulation. Il doit choisir un fichier créé par l'une des deux fonctions précédentes, et contenant la liste des points simulés, mais sans les valeurs d'angles et de longueurs.

Cette fonction produit un fichier du type :

*/répertoire/Simulation/Sim-**Date-Heure**/**NumSimul**-SimCell-**NiveauDecoupe-Paysage-Ligne_Colonne-Parametre**-final-seg.**Format***

où

— Format est PNG ou PS en fonction du choix de l'utilisateur

A.1.6 Dessins des segments de haies simulés

La seconde fonction (Second bouton (**ON/OFF**) dans la figure A.14) permet à l'utilisateur de dessiner les haies simulées dans une cellule. Il doit choisir le fichier se terminant par *-seg* créé à partir de l'onglet "Simulation pour une cellule" de la section A.1.4, contenant la liste des coordonnées des extrémités des segments.

Abstract

This thesis is part of a partnership between INRA and INRIA in the field of knowledge extraction from spatial databases. The study focuses on the characterization and simulation of agricultural landscapes. More specifically, we focus on linears that structure the agricultural landscape, such as roads, irrigation ditches and hedgerows. Our goal is to model the spatial distribution of hedgerows because of their role in many ecological and environmental processes. We more specifically study how to characterize the spatial structure of hedgerows in two contrasting agricultural landscapes, one located in south-eastern France (mainly composed of orchards) and the second in Brittany (western France, *bocage*-type). We determine if the spatial distribution of hedgerows is structured by the position of the more perennial linear landscape features, such as roads and ditches, or not. In such a case, we also detect the circumstances under which this spatial distribution is structured and the scale of these structures. The implementation of the process of Knowledge Discovery in Databases (KDD) is comprised of different preprocessing steps and data mining algorithms which combine mathematical and computational methods. The first part of the thesis focuses on the creation of a statistical spatial index, based on a geometric neighborhood concept and allowing the characterization of structures of hedgerows. Spatial index allows to describe the structures of hedgerows in the landscape. The results show that hedgerows depend on more permanent linear elements at short distances, and that their neighborhood is uniform beyond 150 meters. In addition different neighborhood structures have been identified depending on the orientation of hedgerows in the South-East of France but not in Brittany. The second part of the thesis explores the potential of coupling linearization methods with Markov methods. The linearization methods are based on the use of alternative Hilbert curves : Hilbert adaptive paths. The linearized spatial data thus constructed were then treated with Markov methods. These methods have the advantage of being able to serve both for the machine learning and for the generation of new data, for example in the context of the simulation of a landscape. The results show that the combination of these methods for learning and automatic generation of hedgerows captures some characteristics of the different study landscapes. The first simulations are encouraging despite the need for post-processing. Finally, this work has enabled the creation of a spatial data mining method based on different tools that support all stages of a classic KDD, from the selection of data to the visualization of results. Furthermore, this method was constructed in such a way that it can also be used for data generation, a component necessary for the simulation of landscapes.

Keywords: Data mining, Geomatic, Spatial analysis, Spatial statistics, Geometric modeling 2D, Machine learning, Spatial information, Hilbert adaptative curve, Hedgerows

Résumé

Cette thèse s'inscrit dans un partenariat entre l'INRA et l'INRIA et dans le champs de l'extraction de connaissances à partir de bases de données spatiales. La problématique porte sur la caractérisation et la simulation de paysages agricoles. Plus précisément, nous nous concentrons sur des lignes qui structurent le paysage agricole, telles que les routes, les fossés d'irrigation et les haies. Notre objectif est de modéliser les haies en raison de leur rôle dans de nombreux processus écologiques et environnementaux. Nous étudions les moyens de caractériser les structures de haies sur deux paysages agricoles contrastés, l'un situé dans le sud-est de la France (majoritairement composé de vergers) et le second en Bretagne (Ouest de la France, de type bocage). Nous déterminons également si, et dans quelles circonstances, la répartition spatiale des haies est structurée par la position des éléments linéaires plus pérennes du paysage tels que les routes et les fossés et l'échelle de ces structures. La démarche d'extraction de connaissances à partir de base de données (ECBD) mise en place comporte différentes étapes de prétraitement et de fouille de données, alliant des méthodes mathématiques et informatiques. La première partie du travail de thèse se concentre sur la création d'un indice spatial statistique, fondé sur une notion géométrique de voisinage et permettant la caractérisation des structures de haies. Celui-ci a permis de décrire les structures de haies dans le paysage et les résultats montrent qu'elles dépendent des éléments plus pérennes à courte distance et que le voisinage des haies est uniforme au-delà de 150 mètres. En outre différentes structures de voisinage ont été mises en évidence selon les principales orientations de haies dans le sud-est de la France, mais pas en Bretagne. La seconde partie du travail de thèse a exploré l'intérêt du couplage de méthodes de linéarisation avec des méthodes de Markov. Les méthodes de linéarisation ont été introduites avec l'utilisation d'une variante des courbes de Hilbert : les chemins de Hilbert adaptatifs. Les données spatiales linéaires ainsi construites ont ensuite été traitées avec les méthodes de Markov. Ces dernières ont l'avantage de pouvoir servir à la fois pour l'apprentissage sur les données réelles et pour la génération de données, dans le cadre, par exemple, de la simulation d'un paysage. Les résultats montrent que ces méthodes couplées permettant un apprentissage et une génération automatique qui capte des caractéristiques des différents paysages. Les premières simulations sont encourageantes malgré le besoin d'un post-traitement. Finalement, ce travail de thèse a permis la création d'une méthode d'exploration de données spatiales basée sur différents outils et prenant en charge toutes les étapes de l'ECBD classique, depuis la sélection des données jusqu'à la visualisation des résultats. De plus, la construction de cette méthode est telle qu'elle peut servir à son tour à la génération de données, volet nécessaire pour la simulation de paysage.

Mots-clés: Fouille de données, Géomatique, Analyse spatiale, Statistiques spatiales, Modélisation géométrique 2D, Apprentissage, Information spatiale, Chemin de Hilbert adaptatif, Haies