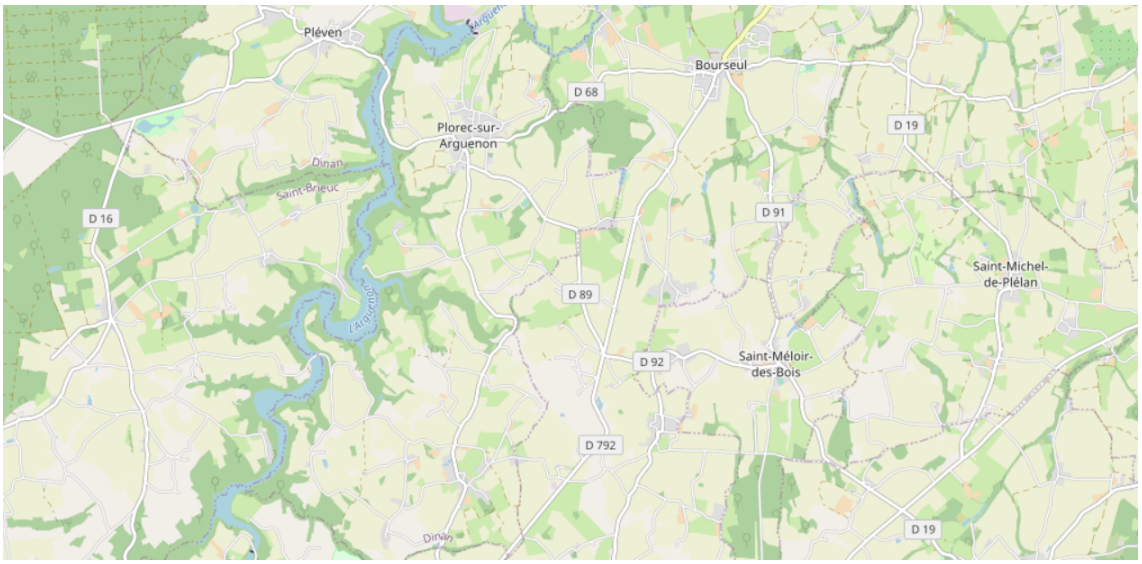
Projet Long de Recherche :

Fouille de Données Spatiale

Étude et caractérisation des linéaires
dans les paysages agricoles



Cyprien MANGEOT - Arslane MEDJAHDI

M2 IMSD 2022-2023

Table des matières

1 Contexte

Un paysage agricole est un espace de terre cultivée constitué d'éléments tels que des terres arables, des routes, des haies et des canaux d'irrigation qui ont des fonctions différentes. La composition et la configuration de ces éléments créent une structure unique pour chaque paysage.

La gestion du paysage agricole est un enjeu majeur de notre société en termes de préservation de la biodiversité, de productivité agronomique, de protection des espèces animales mais aussi de qualité des sols. La structure et l'agencement des ses différents éléments a une influence sur les processus qui se déroulent dans les paysages agricoles, tels que la dynamique des populations animales ou végétales, ou encore la propagation de maladies sur les végétaux.

Les haies sont des éléments majeurs dans la caractérisation des paysages agricoles et jouent plusieurs rôles pour les espèces qui évoluent sur ces paysages. Elles offrent des habitats pour les espèces végétales et animales spécifiques, forment des couloirs de circulation pour les individus et peuvent protéger les sols et maintenir la qualité de l'eau.

Des études ont porté sur les déplacements des insectes sur les paysages agricoles. Ces études ont attiré une attention particulière sur la compréhension du lien entre la distribution spatiale des éléments semi-naturels (tels que des haies) et l'abondance (et la répartition) de certaines espèces.

2 Introduction

C'est dans ce contexte que s'inscrit notre projet de fouille de données spatiale. L'objectif de ce projet long est d'étudier et de caractériser la distribution des linéaires d'un paysage agricole.

Plus précisément, notre travail se concentre sur l'analyse de données spatiales de haies et d'autres linéaires (routes et canaux) dans un paysage agricole. Nous cherchons à extraire des caractéristiques spatiales pour comprendre la construction de ces paysages.

Pour atteindre notre objectif, nous commencerons par pré-traiter nos données. La statistique descriptive nous aidera à prendre en main nos données. Ensuite, nous ajouterons de nouvelles données à notre étude. Cela nous permettra de mieux caractériser nos linéaires.

Pour finir, nous utiliserons des algorithmes de machine learning pour faire du clustering hiérarchique. Le but sera de regrouper les linéaires ayant les mêmes caractéristiques mais aussi de comprendre l'utilité des haies en fonction des paramètres qui les décrivent.

Grâce à ces étapes, nous pourrons extraire des caractéristiques sur la répartition spatiale de nos linéaires et donc mieux comprendre la construction d'un paysage agricole.

3 Pré-traitement des données

3.1 Présentation des données

Dans le contexte donné précédemment, M. DA SILVA nous a fourni deux datasets contenant des informations relatives à des linéaires (route, canaux, haies) dans deux paysages agricoles différents. Les deux datasets sont structurés exactement de la même façon. Nous allons donc présenter le contenu d'un des deux datasets.

Ces deux datasets ont été produits par M. DA SILVA lors de sa thèse. Il avait des données sur les linéaires sous forme de polylignes et a décidé de découper chaque polyligne en segments pour pouvoir les étudier. Plusieurs segments peuvent donc former une seule et même linéaire sur le terrain.

Nous allons décrire les variables de ces deux dataset.

	Landscape	row	col	type	id	bool1	X1	Y1	X2	Y2	bool2	XG	YG	Lng	Ang	typeH
0	B	0	1	H	11609	1	306802.749855	2.402633e+06	306883.406015	2.402616e+06	1	306843.077935	2.402625e+06	82.430064	1.778632	HV
1	B	0	1	H	11610	0	306883.406015	2.402616e+06	306960.748447	2.402601e+06	1	306922.077231	2.402608e+06	78.879489	1.768532	HV
2	B	0	1	H	11674	0	306928.189087	2.402702e+06	306960.748447	2.402601e+06	0	306944.468767	2.402652e+06	106.830799	2.831890	HP
3	B	0	1	H	11678	1	306796.095039	2.402712e+06	306928.189087	2.402702e+06	1	306862.142063	2.402707e+06	132.400726	1.638872	HV
4	B	0	1	H	11731	1	306893.498271	2.402800e+06	306928.189087	2.402702e+06	1	306910.843679	2.402751e+06	103.254287	2.798953	HP

Comme nous pouvons le voir sur le screen ci-dessus, les dataset décrivent des linéaires d'un paysage agricole.

- La variable **Landscape** indique le lieux où se trouve la linéaire (B pour Bretagne).
- Les variables **row** et **col** indiquent la zone sur laquelle se trouve la haie.
- La variable **type** donne le type de la linéaire : H pour haie, R pour route, C pour canaux.
- La variable **id** donne à chaque linéaire un identifiant.
- Les variables **X1** et **Y1** donnent les coordonnées de "départ" de la haie en format Lambert.
- Les variables **X2** et **Y2** donnent les coordonnées d'"arrivée" de la haie en format Lambert.
- Les variables **XG** et **YG** donnent les coordonnées de l'isobarycentre de la haie en format Lambert.
- La variable **bool1** est un booléen qui vaut 1 si la haie est entièrement contenu dans la cellule (donnée par row et col), 0 sinon.
- La variable **bool2** est un booléen qui vaut 1 si l'isobarycentre de la haie se trouve dans la cellule (donnée par row et col), 0 sinon.
- La variable **Lng** donne la longueur de la haie.
- La variable **Ang** donne l'angle de la haie par rapport à l'axe Nord-Sud
- La variable **typeH** est une variable que M. DA SILVA a créée pendant sa thèse, nous ne nous en sommes par servi, il est donc inutile de détailler ici sa signification.

Dans le listing précédent, il est à noter que la variable Landscape est inutile. Elle est invariable pour chaque dataset. Sur le dataset des linéaires de Bretagne, elle vaudra toujours B. Sur le dataset de Provence, elle vaudra toujours A (pour Avignon).

Par ailleurs, nous n'avons pas utilisé les variables bool1 et bool2 car nous avons appris juste avant la fin du projet leur signification. Nous n'avons pas non plus utilisé la variable typeH car elle a été construite par M. DA SILVA lors de sa thèse.

Enfin, nous vous avons parlé des variables row et col. Ces variables découpent la zone qu'occupent nos linéaires en un quadrillage de petites zones. Nous avons affiché différentes cellules sur QGIS avec un code couleur.

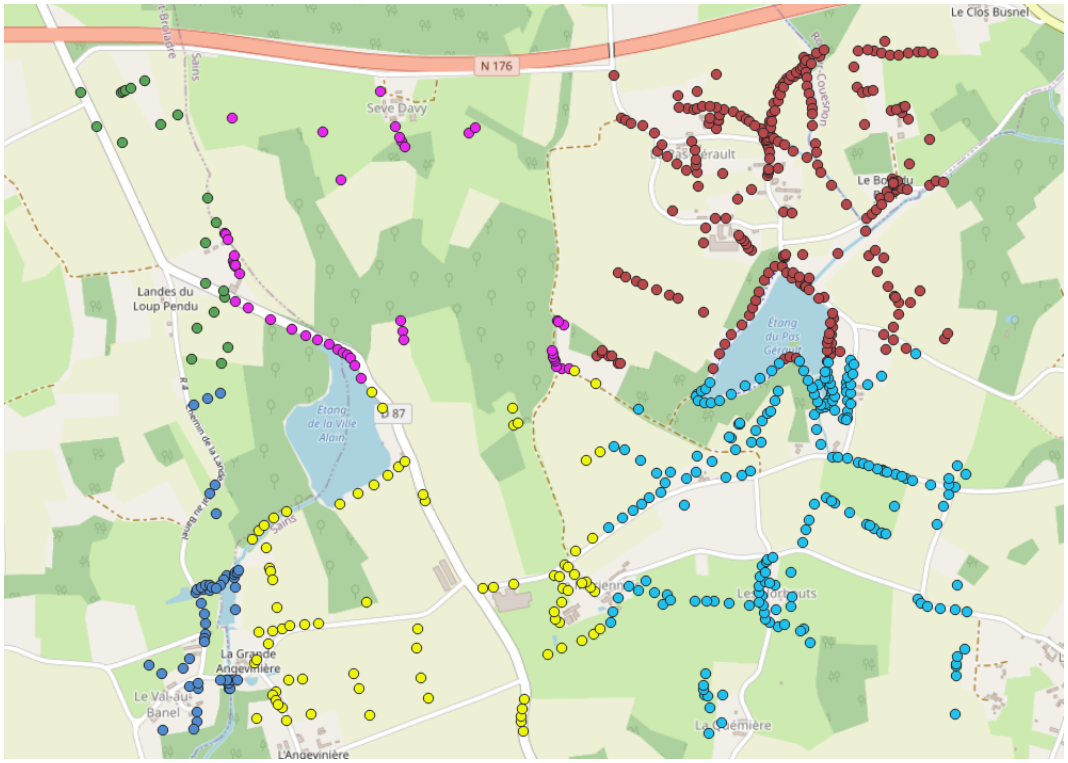


FIGURE 1 – Visualisation d'une partie des cellules (données par row et col) sur QGIS

Sur la figure ci-dessous, nous avons représenté 6 cellules (il y en a bien d'autres) en prenant $row \in \{0, 1\}$ et $col \in \{1, 2, 3\}$. Chaque point est l'isobarycentre d'une linéaire. Deux points d'une même couleur se trouvent dans la même cellule.

Nous avons remarqué que plusieurs haies avaient le même identifiant. Après avoir étudié la question, nous nous sommes rendu compte que lorsqu'une haie était à cheval sur deux cellules, elle était doublée sur le dataset. Toutes les variables étaient identiques, sauf row et col. Nous avons donc décidé de supprimer les doublons.

3.1.1 Dataset de Bretagne

Pour le dataset de Bretagne, nous avions 67.222 linéaires (dont 18.955 haies, 6.497 canaux et 41.770 routes) avant la suppression des linéaires redondantes. Après suppression, nous obtenons 53.772 linéaires (dont 14.819 haies, 5.197 canaux et 33.756 routes).

Voici un aperçu des données du dataset de Bretagne via QGIS

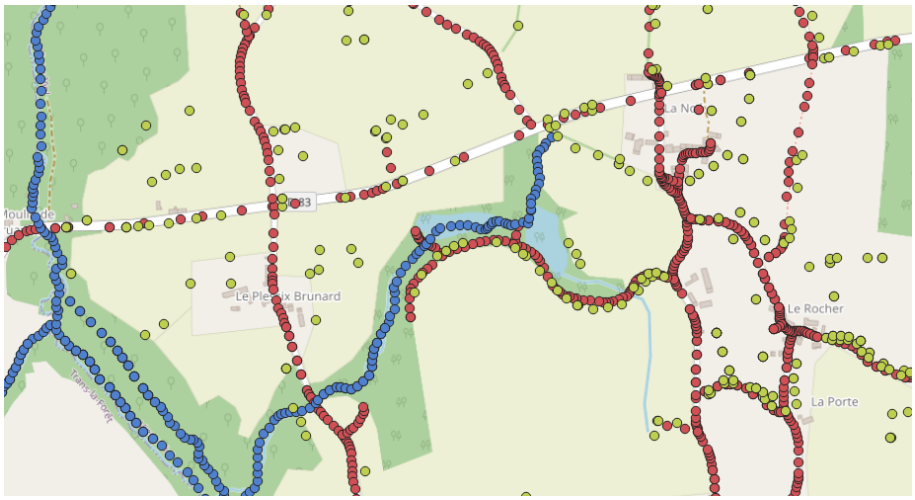


FIGURE 2 – Visualisation des routes, des canaux et des haies sur QGIS

Sur la figure ci-dessus, les points correspondent aux isobarycentres des linéaires. Les routes sont modélisées en rouge, les canaux en bleu et les haies en vert.

Nous avons également représenté les linéaires via la librairie folium de Python. La visualisation est plus agréable que sur QGIS puisqu'on affiche la linéaire en entier et non seulement son isobarycentre. On applique le même code couleur que précédemment.

FIGURE 3 – Visualisation des routes, des canaux et des haies via Python

Les coordonnées des linéaires de ce dataset sont au format Lambert II étendu.

3.1.2 Dataset de Provence

Pour le dataset de Provence, nous avions 22.207 linéaires (dont 15.151 haies, 2.552 canaux et 4.504 routes) avant la suppression des linéaires redondantes. Après suppression, nous obtenons 17.048 linéaires (dont 11.612 haies, 1.944 canaux et 3.492 routes).

Nous avons cherché à afficher les linéaires sur une carte folium comme précédemment. Les coordonnées sont au format Lambert 93. Pour ce qui est des routes et des canaux, tout se passait très bien, comme on peut le voir sur le figure ci-dessous.

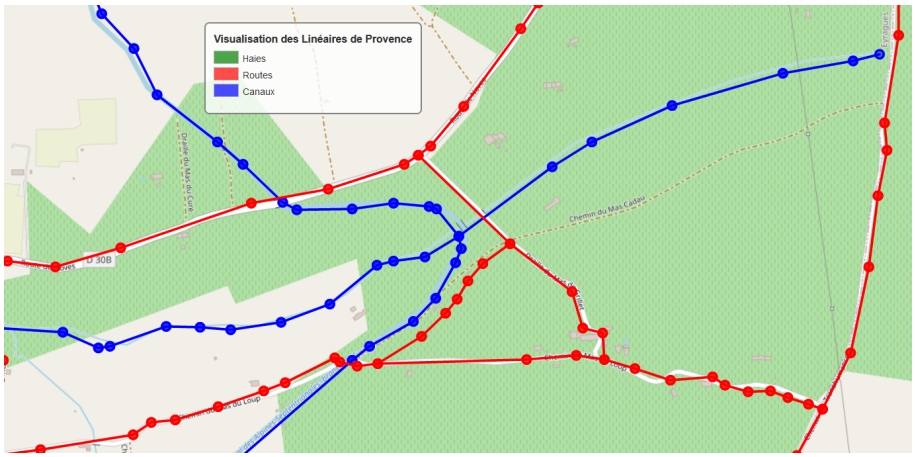


FIGURE 4 – Visualisation des routes et des canaux pour le dataset de Provence

En revanche, pour ce qui est des haies, c'est une autre histoire. Si leur position semble parfois très correcte (figure de gauche), il existe des endroits où leur agencement devient complètement absurde (figure de droite).

FIGURE 5 – Visualisation des haies pour le dataset de Provence

Nous pensons que les données sont faussées. En bref, nous considérons que ce dataset est inutilisable. Dans la suite de ce projet, nous traiterons exclusivement les données liées au dataset de Bretagne.

3.2 Étude préliminaire des données

Dans le but d'avoir une vue d'ensemble de nos données, nous décidons de faire un peu de statistique descriptive. Voici un tour d'horizon des résultats que nous avons obtenus :

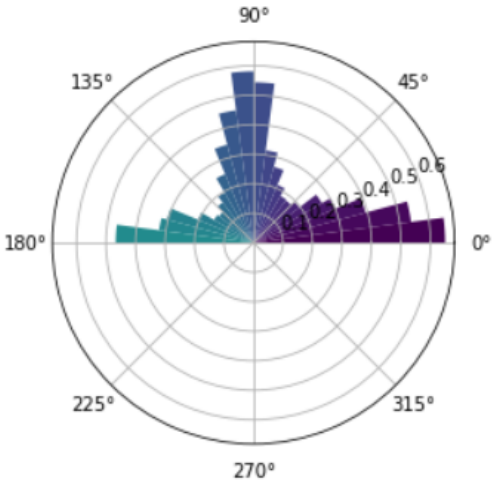


FIGURE 6 – Histogramme circulaire de distribution des angles pour les haies du dataset de Bretagne

Nous constatons que la grande majorité des haies sont orientées sur les axes Nord-Sud et Est-Ouest.

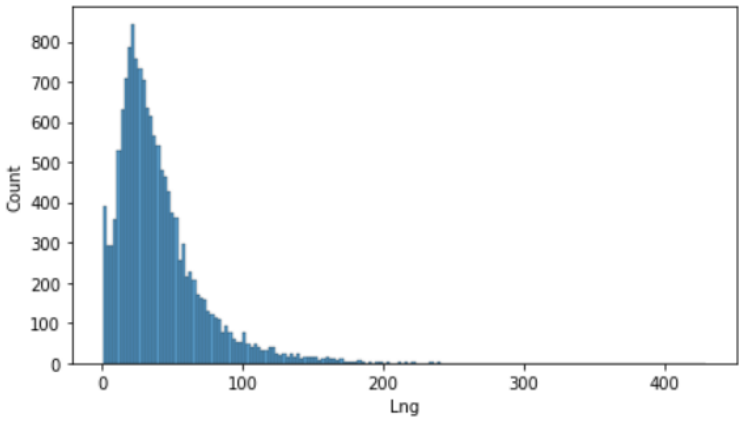


FIGURE 7 – Histogramme de répartition de la longueur des haies du dataset de Bretagne

La longueur est codée en décimètres. Nous observons qu'il y a un nombre non négligeable de haies de moins de 1 mètres. La plupart des longueurs étant comprises entre 3 et 5 mètres.

3.3 Agrégation de linéaires

Lors de l'étude préliminaire de nos données, et particulièrement lors de la visualisation de nos linéaires via Python, nous remarquons un nombre conséquent de haies dont la longueur semble anormalement faible. Nous questionnons donc la pertinence de ces haies : sont-elles des artéfacts de construction ? Sont-elles issues de mauvaises mesures ? Sont-elles des résidus d'autres haies ?

Lors de cette même visualisation, nous remarquons qu'un nombre important de haies donnent l'impression d'avoir la même disposition que leur voisine : angle équivalent, longueur similaire et agencement semblable. Nous soulevons donc une interrogation sur ces haies : ne seraient-elles pas dans la réalité des faits une seule et unique haie découpée en plusieurs parties de manière informatique ?

C'est donc pour répondre à ces deux problèmatiques que nous proposons d'implémenter un algorithme pour fusionner certaines haies. Il y a cependant une manière de les fusionner afin de ne pas engendrer de nouvelles haies qui n'auraient plus aucun sens physique. En effet, l'intêret premier de ce projet est d'étudier la disposition spatiale de nos haies, nous cherchons donc à garder une réprésentation physique qui fait sens pour pouvoir appliquer par la suite nos méthodes de fouilles de données et avoir des résultats exploitables.

Enfin, notre algorithme ne répond qu'à deux règles simples : fusionner deux haies seulement si ces dernières possèdent une extrémité en commun et seulement si la différence de leur angle est négligeable (nous considèrons que cette différence permet de garder le sens physique de nos haies si elle est inférieure à 10 degrés).

Nous obtenons alors, après fusion, les résultats suivants :

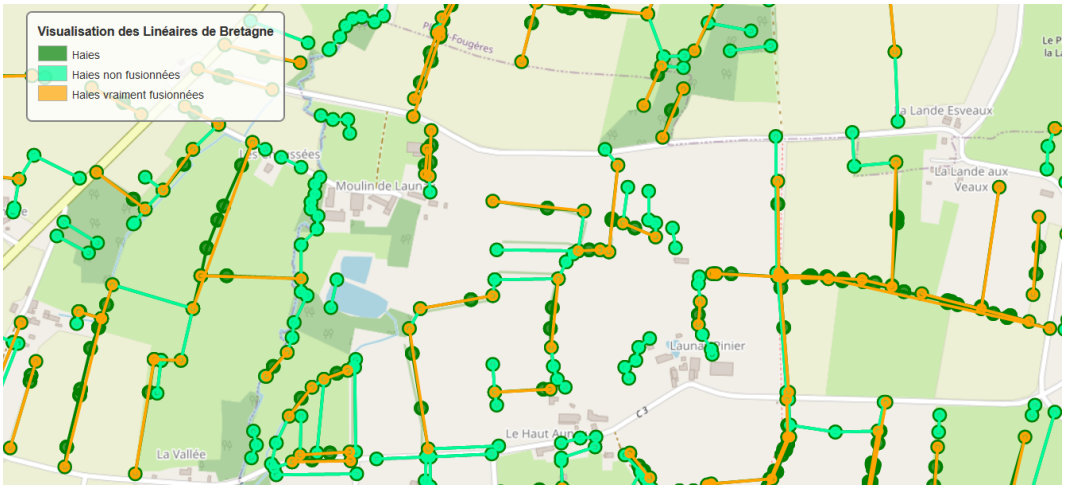


FIGURE 8 – Visualisation des haies avant/après fusion pour le dataset de Bretagne

On constate que notre programme ne fait pas de fusion abusive. Il se contente de fusionner les haies ayant quasiment le même angle, quitte à ne pas fusionner des haies qui pourrait l'être. C'est un choix de notre part pour ne pas fusionner des haies qui ne devrait pas l'être.

Nous pouvons regarder la répartition de la longueur des haies avant et après cette fusion.

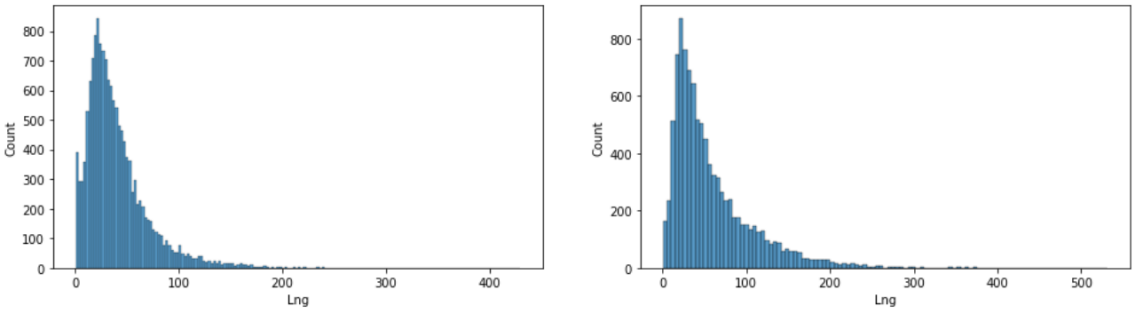


FIGURE 9 – Histogramme de répartition de la longueur des haies avant/après la fusion pour le dataset de Bretagne

Nous observons que les haies de petites tailles sont fusionnées. D'où la forte diminution des haies de taille inférieure à 1 mètre.

Dans une même logique, nous soulevons la possibilité de fusionner nos routes et nos canaux afin de diminuer le nombre de ces dernières et donc d'accélérer nos algorithmes (*cf. Partie "Ajout de variables"*). À noter que pour la fusion des routes, il faut baisser l'angle qui décide s'il y a fusion ou non, sinon la fusion n'a plus de sens.

Cependant après avoir fusionné les routes et les canaux, nous remarquons une perte d'informations concernant l'encadrement de nos haies. En effet, le nombre de routes et de canaux ayant diminuer, leur densité, de manière informatique seulement, a subi une diminution similaire ce qui entraîne une plus grande difficulté à déterminer l'environnement des haies. Nous faisons donc finalement le choix de ne pas fusionner les routes et les canaux pour garantir une meilleure précision du voisinage de nos haies.

Nous allons tout de même vous montrer le résultat de la fusion de nos routes et de nos canaux pour le dataset de Bretagne.

FIGURE 10 – Visualisation des canaux avant/après fusion pour le dataset de Bretagne

FIGURE 11 – Visualisation des routes avant/après fusion pour le dataset de Bretagne

3.4 Ajout de variables

Pour l'instant nos haies sont caractérisées par leur angle, leur longueur, leur emplacement dans le territoire (grâce aux variables row et col). Mais qu'en est-il de ce qui les entoure ? Sont-elles proches de canaux, de routes, d'autres haies ou totalement isolées du reste du paysage ? Nous nous intéresserons aussi à l'orientation des ces dernières par rapport à leur environnement : sont-elles perpendiculaires par rapport aux canaux ou aux routes dont elles sont proches ? Nous proposons donc d'introduire de nouvelles variables pour décrire l'environnement de nos haies mais aussi leur orientation par rapport à leur environnement proche.

Nous considérons qu'une haie est proche d'un canal si la distance qui les séparent n'excède pas 3 mètres. On choisit 3 mètres car la largeur d'un canal est d'environ 3 mètres. Si nous considérons que nos "canaux informatiques" sont le milieu des "canaux réels", alors nous pouvons raisonnablement dire qu'une haie est proche d'un canal si elle se trouve à moins de 3 mètres de celui-ci.

Par un raisonnement similaire, la notion de proximité entre une haie et une route est définie par une distance maximale de 4 mètres. Nous sommes conscients que les distances que nous avons fixé sont assez faibles. Nous avons préféré la précision, quitte à manquer des haies potentiellement proches de nos routes/canaux.

Sur le schéma suivant, nous pouvons observer les haies séléctionnées comme étant proches d'un canal.

FIGURE 12 – Visualisation des haies proches d'un canal pour le dataset de Bretagne

Maintenant que nous avons défini les notions de "haies proches d'un canal" et de "haies proches d'une route", nous pourrions essayer de classifier ces haies pour dire si elles sont perpendiculaires ou parallèles à ce qui les jouxtent.

Nous considèrerons qu'une haie suit un canal/une route si la valeur absolue des différences des angles est inférieure à 45 degrés. Par ailleurs, nous considèrerons qu'une haie est perpendiculaire à une route/un canal si la valeur absolue de la différence des angles est comprise entre 45 et 105 degrés.

Parmi les haies proches d'un canal que nous avons montré précédemment, voici les haies considérées comme étant perpendiculaires au canal, et les haies considérées comme parallèles au canal.

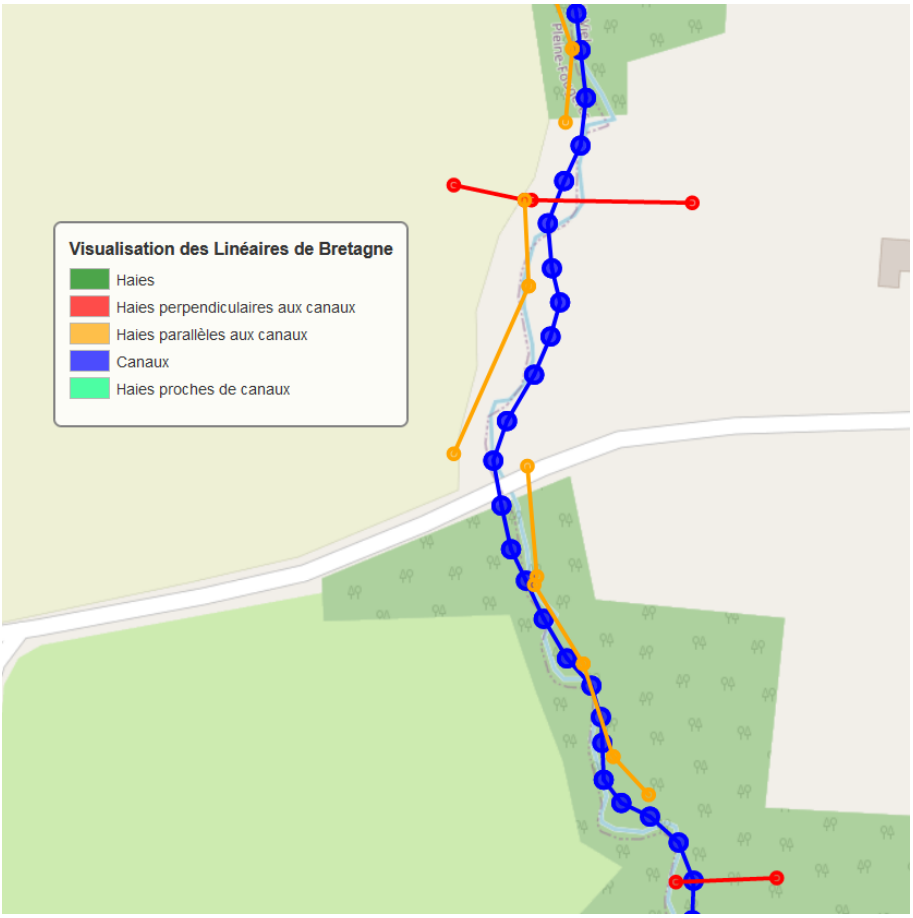


FIGURE 13 – Visualisation des haies proches d'un canal pour le dataset de Bretagne

La logique est exactement la même pour les haies proches de routes. Grâce à notre programme, nous avons pu ajouter des variables descriptives à nos haies. Cela nous aidera à mieux comprendre la répartition spatiale de nos haies.

4 Recherche de nouvelles données

4.1 Données cadastrales

Dans l'optique de continuer à décrire nos haies, nous nous sommes rendu sur la plateforme ouverte des données publiques du gouvernement français. Et nous avons téléchargé les données cadastrales des régions qui nous intéressaient.

Un cadastre est l'ensemble des documents qui recensent et évaluent les propriétés foncières de chaque commune. En d'autres termes, un cadastre décrit toutes les parcelles qu'il contient. Il serait donc intéressant dans le cadre de notre étude de mettre en perspective les données cadastrales et les positions de nos haies.

Dans un premier temps, il était intéressant de visualiser les parcelles de chaque cadastre via le site suivant : `https://cadastre.data.gouv.fr/map?style=ortho#13.57/48.49615/-1.59761`.

Dans le cadre du dataset de Bretagne, nous avons récupéré les données cadastrales des communes dans lesquelles se trouvaient nos haies. Puis nous avons affiché sur une carte (via la librairie folium de Python) les parcelles des tous les cadastres sélectionnés, ainsi que les haies.

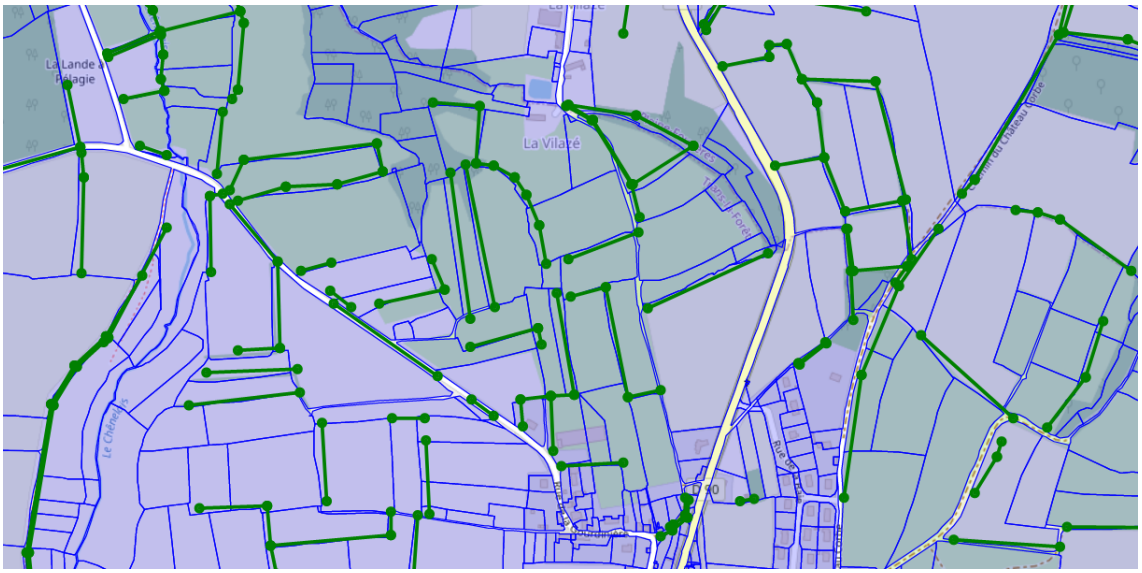


FIGURE 14 – Visualisation des données cadastrales, et des haies, pour le dataset de Bretagne

Nous observons que la majorité des haies se situent le long des limites parcellaires. Ceci soulève la question de savoir si ces haies ont été intentionnellement plantées sur ces limites parcellaires ou si ce sont les parcelles qui ont été dessinées à partir de la position des haies déjà présentes sur le terrain.

La création des parcelles cadastrales ont, pour la plupart, été effectué par des géomètres à l'aide d'instruments comme le théodolite. Cet instrument permettait de mesurer les angles et de lever des plans.

Il est probable que ces géomètres aient également utilisé des éléments naturels tels que les haies, les canaux, les arbres ou encore les montagnes et collines pour délimiter les parcelles.

Pour résumer, il n'y a pas de réponse formelle à cette question. Il est possible que des haies préexistantes aient été utilisées pour tracer les parcelles cadastrales, tout comme il est possible que des haies aient été plantées sur les limites parcellaires pour mieux délimiter les parcelles.

4.2 Données d'occupation des sols

QGIS nous permet assez facilement d'afficher l'occupation des sols sur le territoire considéré en Bretagne à partir du fichier shapefile correspondant. Nous récupérons alors la couche suivante, où, chaque couleur coïncide avec un type de sol différent :

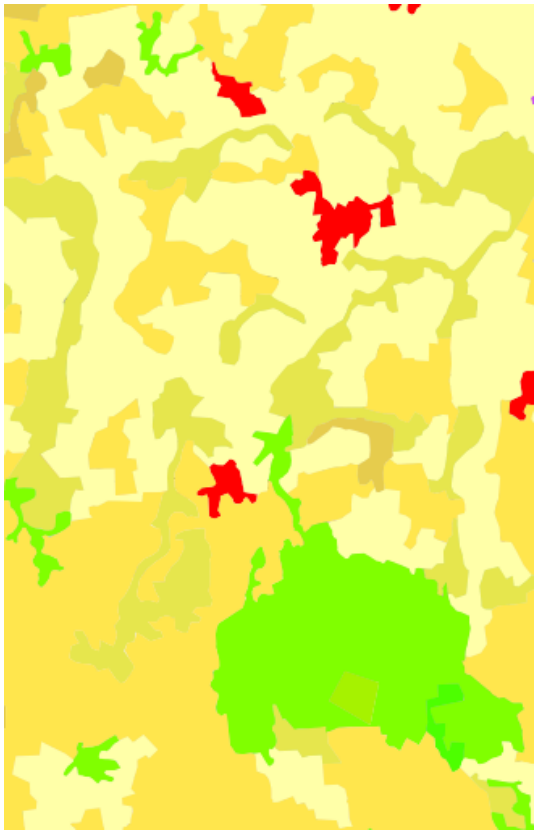


FIGURE 15 – Visualisation de l'occupation des sols sur le territoire de Bretagne

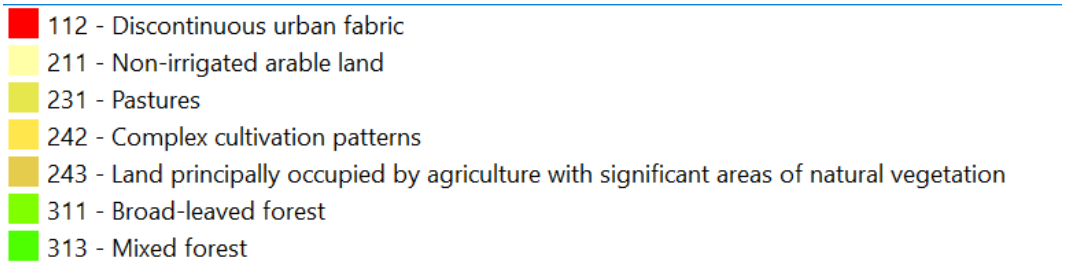


FIGURE 16 – Légende correspondante pour l'occupation des sols du territoire de Bretagne

Nous constatons alors que le territoire considéré ne contient que 7 types de sols distincts et dont 5 seulement nous intéressent pour caractériser nos haies. En effet, nous ne retrouvons aucune haie en forêt. Par conséquent les sols de type 311 et 313 ne nous servent à rien.

QGIS nous permet alors de croiser les multipolygones qui definissent chacun de nos sols avec la couche qui définit nos haies : ce qui nous permet de récupérer, pour chaque occupation des sols, les haies qui se trouvent dessus. Ce processus est assez long mais nous préférions suivre cette méthode et être sûrs d'avoir des résultats, plutôt que d'essayer d'implémanter ceci sur python et de ne rien avoir à la fin.

Nous nous retrouvons alors avec 5 nouvelles variables pour caractériser nos haies.

- Les haies se trouvant sur un sol de type *Tissu urbain discontinu* (112).
- Les haies se trouvant sur un sol de type *Terres arables hors périmètres d'irrigation* (211).
- Les haies se trouvant sur un sol de type *Prairies et autres surfaces toujours en herbe à usage agricole* (231).
- Les haies se trouvant sur un sol de type *Systèmes culturaux et parcellaires complexes* (242).
- Les haies se trouvant sur un sol de type *Surfaces essentiellement agricoles, interrompues par des espaces naturels importants* (243).

Ces nouvelles données nos ont permis de rajouter des variables descriptives à nos haies. Nous commençons désormais à avoir une bonne base de travail. Nous pourrons nous servir de ces nouvelles caractéristiques pour mieux comprendre la construction du paysage agricole étudié. Notamment via des algorithme de clustering hierarchique.

Voici un aperçu des variables que nous avons créées :

proche canal	perp route	perp canal	paral route	paral canal	ville	pastures	complex_cult	natural_vegetation	non_irrigated
0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	1	0	0

FIGURE 17 – Aperçu des nouvelles variables descriptives

5 Fouille de Données

5.1 Clustering Hiérarchique

C'est donc en ayant terminé le pré-traitement de nos données que nous pouvons nous intéresser aux méthodes de fouille, et plus particulièrement au clustering hiérarchique implémenté sur python. Nous avons donc à disposition pas moins de 16 variables pour décrire nos haies, nous nous proposons alors de réaliser plusieurs clustering sur ces variables. Notamment pour différencier les haies en fonction de leur proximité, à la fois aux routes et aux canaux.

Nous remarquons assez vite que les haies qui sont proches des canaux ne semblent pas avoir d'angle particulier, contrairement aux haies qui suivent les limites parcellaires. En effet, les canaux sont des éléments naturels qui ont donc imposé la position des limites parcellaires qui les jouxtent. Enfin les haies proches des canaux semblent se trouver quasiment exclusivement sur des sols de paturage ou de cultivation complexe, attestant donc de leur utilité pour l'agriculture. Elles sont aussi plus courtes en moyennes que les autres haies (2,9 m contre 3,9 m), dû au fait qu'elles doivent suivre les courbes des canaux, ce qui laisse peu de place pour de grandes haies.

Concernant les haies qui sont proches des routes, elles ont une longueur moyenne de 3,5 m, ce qui n'est donc pas sensiblement différent de la longueur moyenne de toutes les haies. Nous constatons aussi que la plupart de ces haies se trouvent soit dans l'axe Nord-Sud soit dans l'axe Est-Ouest. En fait, 50% des haies qui sont orientées horizontalement ou verticalement sont situées à proximité d'une route. Nous nous interrogeons sur les raisons de cette observation, serait-ce dû au fait que les routes sont orientées horizontalement ou verticalement ? Cela ne semble pas être le cas (en Bretagne) car seulement une route sur cinq suit l'un de ces deux axes.

Finalement, nous constatons que les haies proches des routes ne se trouvent pas sur un type de sol en particulier. D'une part, ce ne sont pas des haies que l'on retrouve en ville, car il n'y a quasiment aucune haie en ville. D'autre part, elles ne se trouvent pas non plus dans les pâturages, ou dans les terres agricoles car les routes ne se trouvent pas dans ces zones.

Nous terminons donc par nous intéresser aux haies qui ne sont ni proches des canaux ni proches des routes. Nous en distinguons principalement 3 types.
Le premier identifie les haies qui entourent les pâturages et servent donc comme haies de bocage. Elles ont une longueur moyenne de plus de 5 mètres. Le second type que nous distinguons sont les haies agricoles, principalement les haies qui entourent les cultures. Elles sont en moyennes d'une longueur de 4,7 mètres, soit légèrement plus courtes que les haies de bocage. Elles sont néanmoins beaucoup plus nombreuses que ces dernières et représentent près de 80 % des haies qui ne sont ni proches des routes ni proches des canaux. Le dernier type de haie se trouve sur les terres non irriguées. Ces haies sont les plus longues, en moyennes près de 6 mètres.

6 Conclusion

Notre projet de fouille de données spatiale nous a permis d'analyser la distribution des linéaires, en se concentrant sur les haies, les routes et les canaux d'irrigation. Les différentes étapes de ce projet étaient toutes enrichissantes.

Après avoir pré-traité les données, nous avons pu faire de la fouille de données sur les linéaires de type haie. Nous sommes ensuite allés à la recherche de nouvelles données. Nous avons pu ajouter des données cadastrales à notre étude, ainsi que des données relatives à l'occupation des sols. Ces dernières nous ont permis de pouvoir mieux caractériser nos haies. Mais également de pouvoir mieux comprendre l'environnement d'une haie.

Certaines variables n'ont malheureusement pas pu être exploitées. Nous n'avons pas pu mettre à profit les variables bool1 et bool2 car nous ne connaissions pas leur signification. Pour ce qui est des variables row et col, nous aurions aimé les exploiter d'avantage. Nous aurions pu considérer, pour une haie donnée, le nombre de haies contenues dans sa cellule. Peut-être aussi considérer les répartitions des angles et des longueurs par cellules, et mettre cela en parallèle avec l'occupation des sols.

Une fois la caractérisation de nos haies effectuée, nous avons eu recours à la méthode de clustering hiérarchique. Cette dernière nous a alors permis de mettre en lumière différents types de haies : les haies qui bordent les canaux et les routes servent souvent à délimiter ceux-ci. Les autres haies (qui ne sont ni proches des routes ni proches des canaux) servent de haies de bocages pour délimiter à la fois les pâturages et les cultures agricoles. Ces haies sont en moyenne beaucoup plus grandes que les autres.

En somme, notre projet nous a permis d'identifier certaines caractéristiques d'un paysage agricole breton. Ces caractéristiques pourraient encore être améliorées. Comprendre l'agencement d'un paysage agricole est primordial pour préserver la biodiversité et la productivité agricole. Ces résultats pourraient être utiles pour les gestionnaires du paysage, les agriculteurs ou encore les scientifiques. Cela permettrait de mieux comprendre les interactions entre les différents éléments du paysage et d'élaborer des stratégies de gestion plus durables et efficaces.

7 Manuel utilisateur du Projet

L'ensemble des fichiers créées lors de la réalisation de ce projet long (fichiers de code python ou fichiers QGIS) sont disponibles à l'adresse suivante :
`https://github.com/Arslane18/Projet-Long`

8 Gestion de Projet

8.1 Méthode de travail

- M. DA SILVA nous a inculqué des méthodes de gestion de projet que nous avons appliqué. Ces dernières consistaient à résumer par écrit chaque rendez-vous. Nous devions tenir un journal de bord qui répertoriait chaque rendez-vous, la date associée, les discussions majeures du rendez-vous, et les avancées du projet suite à ce rendez-vous.
- Nous avons effectué 7 rendez-vous avec M. DA SILVA. Chaque rendez-vous avait pour but de parler de nos avancées mais aussi nos difficultés. De ces rendez-vous naissaient des pistes de réflexions pour continuer l'avancement du projet et surmonter les difficultés rencontrées.
- Le 8-ème et dernier rendez-vous était l'occasion de rendre notre compte-rendu à M. DA SILVA et de faire une soutenance blanche. La façon dont M. DA SILVA nous a encadré était très appréciable. C'était stimulant et bienveillant.

8.2 Chronologie

Compte rendu du 11/01/23 : RdV N°1

- La première réunion a eu lieu le 11/01/2023 et servait d'initialisation au projet long de recherche.
- M. DA SILVA nous a présenté les deux datasets sur lesquels nous avons travaillé. Ces deux datasets comprennent des coordonnées où se trouvent des linéaires. Pour chaque linéaire, nous avons le type de linéaire (routes, canaux, haies). Chaque dataset correspond à une zone géographique (A = Avignon, B = Bretagne).
- Nous avons ensuite établi les premiers objectifs à atteindre. On devra commencer par décrire le dataset de façon rigoureuse. Nous allons ensuite nous munir du logiciel QGIS pour charger les données et tenter de les afficher sur une carte.
- Nous avons ensuite fixé des rendez-vous hebdomadaires avec M. DA SILVA pour faire des check-up réguliers sur l'avancement du projet. Prochain rendez-vous le 24/01/23 à 10h.

Suite à ce rendez-vous :

- Nous avons affiché les barycentres de chaque linéaire sur une carte via QGIS.

Compte rendu du 24/01/23 : RdV N°2

- Lors de ce rendez-vous, nous avons préciser le but du projet, mais également les étapes majeures du projet. Nous avons évoqué un maximum de pistes pour commencer à exploiter les données, et pouvoir mieux les comprendre.
- Nous avons également présenté nos avancées à M. DA SILVA.
- Comme les datasets sont très très volumineux, on a évoqué l'idée de nettoyer les données pour avoir des datasets plus efficaces encore.
- M. DA SILVA nous a suggéré de chercher de nouvelles données sur `https://www.data.gouv.fr/fr/` et d'effectuer une visualisation des données.

Suite à ce rendez-vous :

- Nous avons recalculé la longueur des haies. Nos résultats coïncident bien avec la variable Lng que nous avions. On a fait des histogrammes sur la longueur des haies.
- On a remarqué que certaines haies avaient le même id. Après avoir fait des recherches, on a découvert que c'était des doublons. On a donc supprimé toutes les haies qu'on avait en double.
- Nous avons développé un programme pour sélectionner les haies proches des canaux et des routes.
- On voudrait trouver un moyen de tracer les haies pour pouvoir mieux les visualiser.
- Nous avons commencé à développer un programme pour fusionner certaines haies, beaucoup d'interrogations se posent.
- Nous avons également récupérer de nouvelles données sur le site du gouvernement concernant l'occupation des sols. Nous avons visualisé ces données via QGIS.

Compte rendu du 30/01/23 : RdV N°3

- Nous avons présenté à M. DA SILVA nos deux algorithmes de détection de haies proches des canaux et des routes. Ceux-ci semblent bien fonctionner.
- En revanche, l'algorithme de fusion de haies avait des défauts. M. DA SILVA nous a proposé des pistes pour résoudres ces problèmes. Nous avons également convenu de ne pas fusionner toutes les haies, mais uniquement celles ayant le même type et des angles similaires.

Suite à ce rendez-vous :

- Nous avons effectué des modifications sur l'algorithme de fusion de haies, qui semble mieux fonctionner.
- Nous avons tenté de modéliser les haies sur une carte, mais sans succès.
- Nous avons rassemblé les deux algorithmes de détection de haies proches des canaux et des routes en un seul algorithme.

Compte rendu du 06/02/2023 : RdV N°4

- Nous avons présenté notre nouvel algorithme pour sélectionner les haies proches des canaux et des routes.
- Nous avons discuté de l'algorithme de fusion et avons encore proposé des pistes d'amélioration.
- M. DA SILVA nous a aidés à trouver un moyen d'afficher les haies sur une carte pour mieux réfléchir à la construction de l'algorithme de fusion.

Suite à ce rendez-vous :

- Nous avons réussi à représenter les haies sur une carte folium via Python. Cette visualisation nous a permis de mieux comprendre ce que faisait notre programme de fusion des haies. Nous avons donc pu encore améliorer l'algorithme de fusion.
- Nous avons utilisé cette visualisation pour observer les haies avant et après fusion.

Compte rendu du 03/03/2023 : RdV N°5

- Nous avons re-clarifié l'objectif du projet long.
- Le but serait maintenant de trouver un maximum de caractéristiques sur les haies (longueur, angle, proche d'un canaux (si oui, orientation par rapport à ce dernier), proche d'une route (si oui, orientation par rapport à cette dernière), occupation du sol sur lequel se trouve la haie, données météorologiques, type de cadastre, densité des haies par région, type de haies (particulier, bocage, bois de chauffage, accueille d'oiseaux, ...) etc
- Une fois que nous aurons un maximum de caractéristiques sur nos haies, on pourra appliquer des algorithmes de clustering pour caractériser les haies.

Suite à ce rendez-vous :

- Nous avons modifié l'algorithmes pour trouver les haies proches des canaux et des routes. Celui-ci rajoutent maintenant des colonnes "proche route" et "proche haie" dans le dataset haies. Mais également les variables "perpendiculaire canal", "parallèle canal", idem pour les routes.
- Nous avons décidé de fusionner également les routes et les canaux. Puis de visualiser ces fusions sur une carte folium. Pour fusionner les routes, il faut réduire l'angle "acceptable" pour que la fusion ait du sens.

Compte rendu du 13/03/2023 : RdV N°6

- Nous avons fait du clustering sur les haies en utilisant la longueur, puis l'angle.
- Nous avons constaté que les haies sont pour la grande majorité sur les axes Nord-Sud et Est-Ouest. Sauf les très petites haies où l'angle est complètement aléatoire. Cela nous permet de déduire que les petites haies sont uniquement là pour faire la liaison entre 2 grandes haies.

Suite à ce rendez-vous :

- Nous avons utilisé la fonction *Select by Location* de QGIS pour séléctionner les haies positionnées sur un même type de sol. Nous avons fait cela pour chaque type de sol à la main en attendant de trouver comment automatiser le processus.
- Nous avons réussit à afficher les parcelles cadastrales sur une carte folium. On constate que les haies suivent les limites des parcelles. M. DA SILVA nous a proposé de répondre à la question suivante : Est-ce que les haies ont été positionnées sur les limites parcellaires ? Ou est ce que les parcelles ont étés construites sur la position des haies ?
- Nous avons commencé à afficher les données d'occupation des sols sur une carte folium, mais il nous manque certains éléments.

Compte rendu du 22/03/2023 : RdV N°7

- Ce rendez-vous servait de conclusion au projet long. M. DA SILVA nous a donné de précieux conseils pour construire notre rapport.
- Nous reverrons M. DA SILVA le 28/03/2023 pour lui rendre le rapport et faire une soutenance blanche.

9 Bibliographie

Documentation Scikit-Learn. *Supervised learning.*
https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

Sabeur ARIDHI, *Fouille de Données et Extraction de Connaissance.* Université de Lorraine, 2022

Thèse de Monsieur Sébastien DA SILVA, *Fouille de données spatiales et modélisation de linéaires de paysages agricoles.* Université de Lorraine, 2014

Plateforme ouverte des données publiques françaises.
https://www.data.gouv.fr/fr/.

GeoFree, La Boîte à Outils Géographique,
https://geofree.fr/gf/projguess.asp

Coordonnées GPS,
https://www.coordonnees-gps.fr/