Spatial Data Sciences

Master 1 MIASHS (2022-2023)

Colin Fourment & Ugo Zennaro
Anthropologie, Sciences Sociales et Politiques (ASSP)
Université de Lyon, Université Lumière Lyon 2

Enseignant : Julien Ah-Pine

Table des matières

1 Introduction

Ce projet a pour but d'étudier un jeu de donnée de l'entreprise VERSO avec les méthodes vues en cours de Spatial Data Sciences. Le jeu de données étudié est un découpage de la France en tuiles, des polygones rectangulaires, issues d'un découpage d'un planisphère en $2^{15} * 2^{15}$. Pour chacune de ces tuiles on a les coordonnées longitudinales, les coordonnées X, Y qui représentent la position de la tuiles et abscisse et en ordonnée du planisphère, des quantités de routes par type, le nombre de magasins, de bâtiments et le taux de d'espace recouvert des bâtiments. Avec ces informations quantitatives chaque tuile présente une classification d'urbanisme selon trois modalités (rural, urbain et urbain dense) et un encodage des toutes le variables quantitatives (en dehors des variables de coordonnées) par un taux obtenu avec un auto-encodeur en 2 dimensions avec une fonction d'activation softmax. On a gardé la probabilité sur laquelle les tuiles plus urbaines ont une plus forte valeur. Dans un contexte où l'entreprise veut optimiser des trajets et des tournées il est important d'avoir de bonnes estimations de la vitesse et nous comptons utiliser nos estimations de l'urbanisme pour le faire. Ici nous implémenterons à ces résultats des méthodes qui prennent en compte l'environnement spatial, cet aspect nous permettras de prendre en compte le voisinage de chaque tuile et de s'émanciper de la granularité choisie, on espère ici pouvoir rendre compte d'une évaluation plus fine de l'urbanisme et de l'intuition qu'on peut avoir que la vitesse est ralentie quand l'urbanisme environnant est dense et d'autant plus quand celui-ci se trouve dans un agglomérat d'urbanisme.

2 Méthodes et motivations

2.1 Carte choroplèthe

Cette première partie visera à visualiser le taux d'urbanisme sur l'ensemble du territoire et nous permet d'avoir un a priori des structures spatiales présentes et sur les méthode à envisager couplée avec les intuitions et les connaissances qu'on peut avoir sur notre contexte et sur les données. Aussi on pourra valider ou invalider le bien fondé de l'utilisation de ce taux obtenu par encodage.

2.2 Lissage des taux d'urbanisme et dépendance spatiale

Secondement, nous allons lisser les valeurs de l'encodage. Pour cela on affecte à chaque tuile la valeur moyenne des huit plus proches voisins avec une pondération normale par rapport à la distance à celle-ci.

Ensuite nous vérifierons l'hypothèse de dépendance spatiale avec le graphe et l'indice de Moran. Cet indicateur compare le positionnement de la valeur d'urbanisme par rapport à la moyenne avec la valeur lissée par rapport à sa moyenne. Ainsi, si les valeurs extrêmes le restent après lissage, alors on pourra conclure à une dépendance spatiale. Instinctivement, par rapport à la façon de découper notre espace qui ne repose pas sur des structures existantes avec des frontières interprétables, on peut se dire que les tuiles proches appartiennent à une même structure spatiale (nous aborderons cette problématique plus en profondeur dans la section 2.3) et qu'on peut imaginer qu'il y ai cette corrélation spatiale. Aussi, après une analyse experte de tuiles, on comprend bien que la distinction entre un lissage qui affaiblie la valeur de l'encodage d'urbanisme de la tuile et un lissage qui augmente sa valeur -quand ces écarts sont considérables- est importante. Par exemple, une tuile dans laquelle passe un fleuve et où il y a un parc mais qui est en plein centre ville pourrait avoir un encodage de l'urbanisme qui nous indique mal la vitesse observée et dont on voudrait que le voisinage corrige à la hausse l'urbanisme, ce que le lissage nous aiderait à faire. En revanche une tuile dans laquelle on observerait une petite ville, un urbanisme qui pourrait impacter la vitesse observée, mais avec des tuiles environnantes dans lesquelles on observerait des champs et donc un environnement rural, on devrait pouvoir ne pas perdre l'information initiale que nous coûterait le lissage. Nous étudierons donc ensuite ces deux cas et proposerons un nouveau taux d'urbanisme où nous remplacerons, pour chaque tuile, la valeur de l'encodage par la valeur lissée si cette dernière est plus grande.

2.3 Clustering avec DBSCAN

Dans cette partie, nous allons essayer d'utiliser la relative finesse du découpage de la carte pour définir les contours des structures d'urbanisme présentes sur le territoire en faisant du clustering des tuiles dont la valeur de l'encodage est supérieur à $0,5$, ce seuil a été choisi après avoir observé les histogrammes 7. Pour cela nous allons utiliser l'algorithme DBSCAN qui propose des regroupements d'individus qui sont suffisamment proches, ce regroupement s'étend tant qu'il existe suffisamment d'individus assez proches lui. Avec cette méthode on espère pouvoir ajouter de l'information robuste sur une tuile sans avoir de connaissance du territoire étudié. Aussi on aimerait que la population de ces clusters soit particulière et on imaginerait qu'elle soit constituée d'une population plus urbaines. C'est donc en observant la différence entre la population d'un cluster et celle de la population générale qu'on va choisir les paramètres utilisés par la méthode DBSCAN.

ϵ	Nombre minimum de voisin
1	4
6	45
4	45

TABLE 1 – Liste des paramètres utilisés

La valeur ϵ qui définit le rayon du cercle dans lequel doivent se trouver des points pour être dans le même cluster que le centre. Cette distance sera définit grâce aux coordonnées X, Y qui dénombres simplement les tuiles sur deux axes, cela nous permettra d'utiliser une valeur d'ϵ qui soit compréhensible et dont une unité correspond à la distance d'une tuile à sa voisine. Le premier couple

de paramètres va chercher 4 tuiles parmi les tuiles voisines (voisinage au sens ROOK, c'est à dire celles qui partagent une frontière) et on aura donc un cluster pour chaque tuile entourée de tuile suffisamment urbaine. Ensuite, on a un couple de paramètre qui fera du clustering sur une échelle plus large puisque l'algorithme cherchera 45 voisins une distance de 6 tuiles de la tuile centrale. Enfin, le denier couple de paramètres cherchera aussi 45 voisins mais avec $\epsilon = 4$.

2.4 Régression

Dans la dernière partie, nous allons faire des régressions selon des régimes spatiaux sur la valeur encodée. Cette approche est contre-intuitive pour plusieurs raisons. Tout d'abord car cette valeur est obtenue avec un auto-encodeur qui utilise les mêmes indicateurs qui seront ici utilisés pour faire nos régressions et l'entreprise n'a en pratique pas besoin de deviner cette valeur. Ensuite parce que l'entreprise souhaiterait d'abord utiliser la classification en trois modalité pour estimer les vitesses, ces ces variables là qui devraient ici être prédite. Les méthodes utilisées là n'ont donc pas vocation à l'être par l'entreprise, néanmoins ce travail sera utile pour comprendre, de manière détournée, comment les variables -du jeu de données initial et celle créées dans ce travail- interagissent entre elles et si elles sont susceptible de proposer une information utile quand il s'agira de faire de la classification. Ainsi, pour chaque modalité d'urbanisme, nous étudierons une régression linéaire, nous verrons sa propension à exprimer la variance de la variable cible. Nous verrons également si des variables sont plus ou moins importantes selon les modalité en regardant la t-statistic et la probabilité d'indépendance, si la t-statistic est très forte et qu'on a une probabilité faible (inférieur à 0.05 typiquement) alors on pourra affirmer que la variable est importante dans la régression.

Aussi nous étudierons des régressions des niveaux d'urbanismes moyens et des taux de chaque modalité d'urbanisme dans chacun des clusters pour voir si la taille d'un cluster peut permettre de définir les tuiles qui le compose.

3 Résultats

3.1 Carte choroplèthe

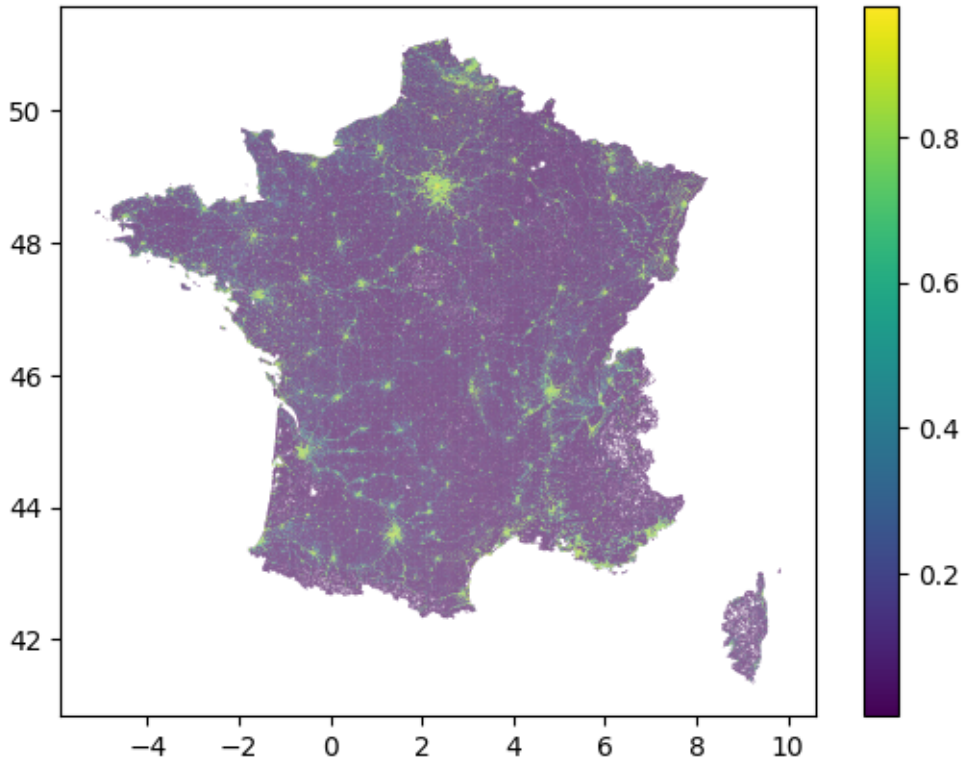


FIGURE 1 – Carte choroplèthe de l'encodage

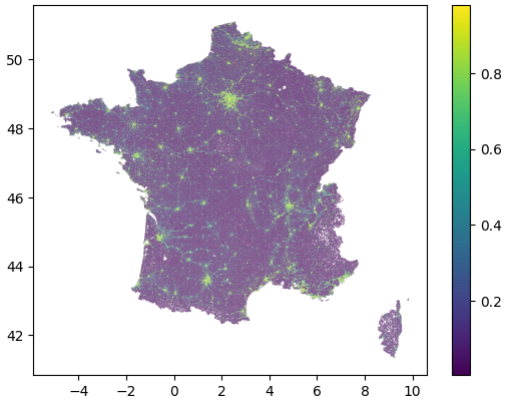
La carte choroplèthe de l'encodage semble montrer de fortes concentrations de tuiles avec un fort taux d'urbanisme sur les grandes villes de France, visuellement, on peut confirmer que cet encodage correspond bien à un taux d'urbanisme.

Confirmons maintenant qu'il y a la dépendance spatiale qu'on peut voir.

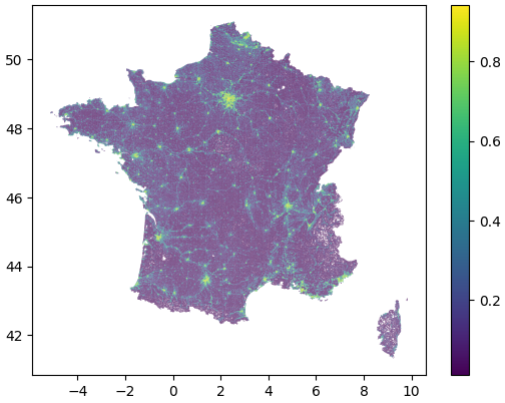
3.2 Lissage des taux d'urbanisme

En comparant la carte choroplèthe de l'encodage on observe les mêmes zones d'urbanisme dense. Même si on observe que le lissage a pu atténuer certaines disparités puisque la carte lissée semble avoir, sur les zones denses, un jaune (qui correspond à un taux proche de 1) qui devient vert bleuté (plus proche d'environ 0.6), et sur les zones rurales un violet qui semble s'éclaircir. Le graphe de Moran correspondant propose une corrélation positive entre les deux estimateurs et l'indice de Moran valant 0.52, la p-value que celui-ci soit égal à 0 et que nos tuiles soient spatialement indépendante est de 0.01 et on peut donc rejeter cette hypothèse. On peut maintenant étudier la forme de ce nuage de points qui n'est pas parfaitement résumé par la régression linéaire affichée.

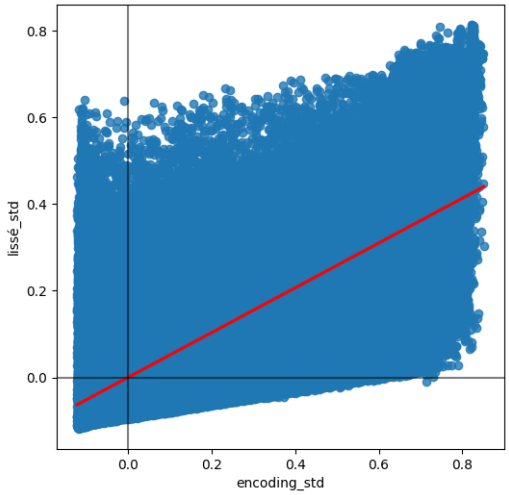
D'abord, en observant le cas où la valeur lissée est supérieure à la valeur initiale, on peut déjà dire que la régression proposée dans le graphe de Moran donne une courbe dont le coefficient directeur semble assez proche de 1, et donc que le lissage n'augmente en générale que très peu ces cas là. Aussi on observe que de très rares cas où la valeur lissée est vraiment plus forte que l'encodage initiale, et on peut penser que ces cas sont les cas qui correspondent à une mauvaise interprétation de l'urbanisme



(a) Carte choroplète de l'encodage

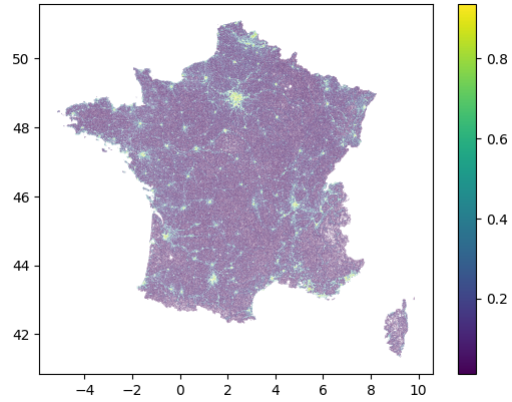


(b) Carte choroplèthe de l'encodage lissé

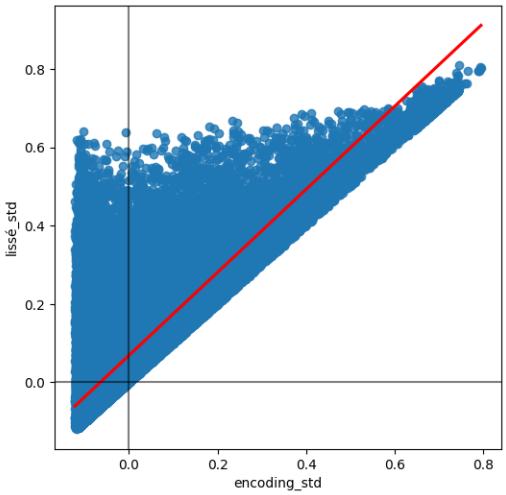


(c) Graphe de Moran

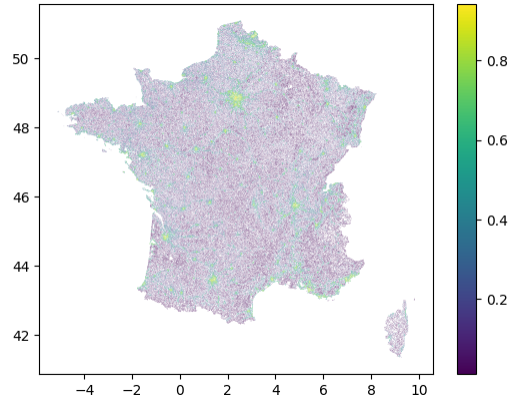
FIGURE 2 – Corrélation spatiale



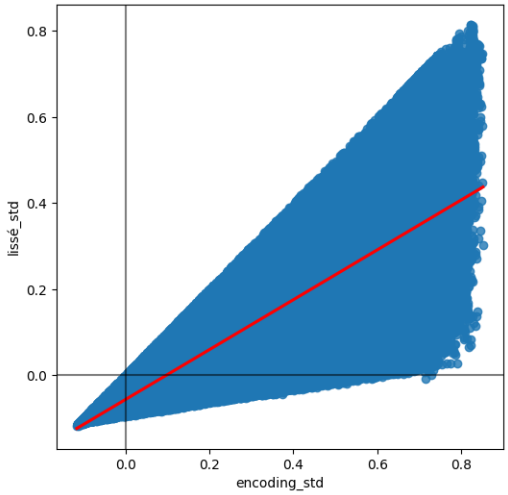
(a) Carte choroplèthe où la valeur lissée est supérieure à la valeur encodée



(b) Graphe de Moran où la valeur lissée est supérieure à la valeur encodée



(c) Carte choroplèthe où la valeur lissée est inférieure à la valeur encodée



(d) Graphe de Moran où la valeur lissée est inférieure à la valeur encodée

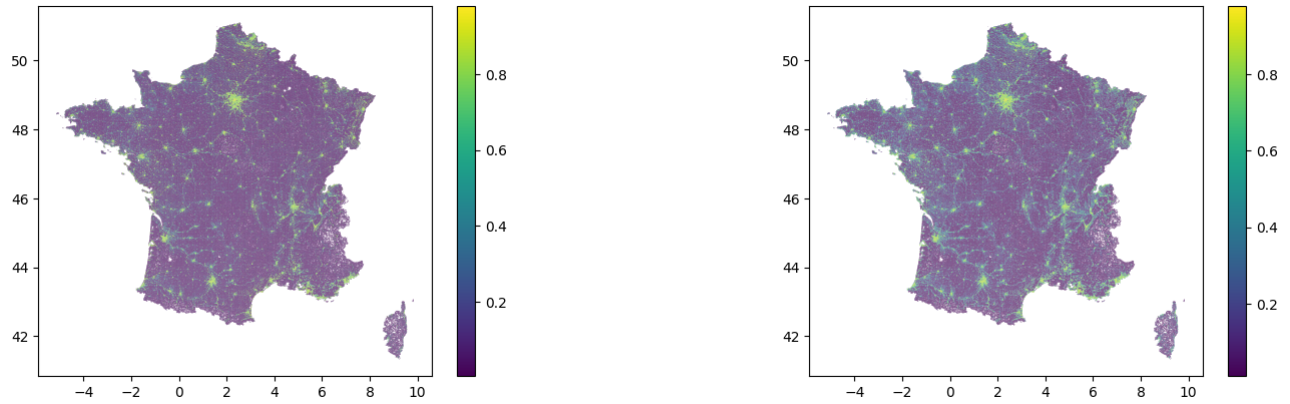
FIGURE 3 – Distinction des lissages

de la tuile comme expliquée dans la section 2.2. En regardant la carte des tuiles correspondantes on peut voir qu'on a ici l'impression de se concentrer sur des tuiles avec un faible taux d'urbanisme.

Par ailleurs le cas où la valeur lissée tire vers le bas le taux d'urbanisme initial se comporte différemment dès le graphe de Moran, on y observe la courbe proposée par la régression linéaire avec une pente plus faible et une densité importante de tuiles dont la valeur lissée est significativement plus faible que le taux initial. Visuellement on lit que c'est dans ce cas là que les tuiles urbaines semblent se ranger bien que bon nombre de tuiles rurales s'y trouve aussi.

En définitive, on ne contredit pas l'impression que le lissage est porteur d'une information censée quand il augmente le taux d'urbanisme et qu'autrement le comportement est plus hasardeux. On peut lire ici que certaines zones sont sûrement moins corrélées spatialement.

Ainsi on s'intéressera maintenant à l'urbanisme corrigée où le taux d'urbanisme n'a que pu être augmenté par le lissage.



(a) Carte choroplèthe de l'encodage

(b) Carte choroplèthe lissé ajusté

FIGURE 4 – Lissage spatial ajusté

3.3 Clustering avec DBSCAN

Avant de voir les résultats du clustering, on observe la répartition des modalités d'urbanisme parmi les tuiles dont le taux d'urbanisme est supérieur à 50%. On voit que les tuiles rurales y sont majoritaire, qu'il y a une part importante de tuiles urbaines et que les tuiles urbaines denses représentent moins d'un 1% des tuiles observées. Voyons comment ces proportions varient dans les clusters selon le paramétrage de l'algorithme.

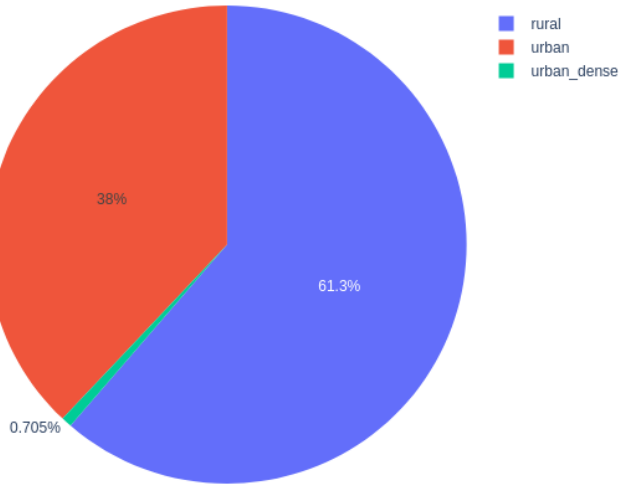
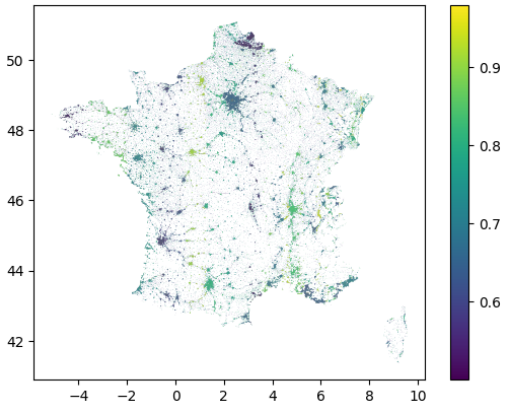


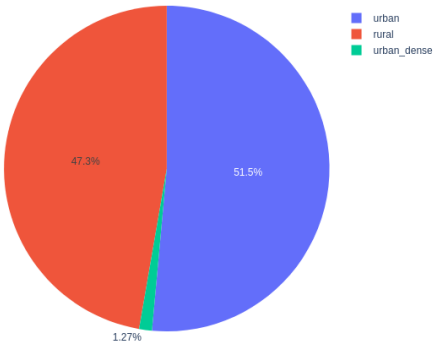
FIGURE 5 – Taux d'occupation chaque modalité d'urbanisme

Le premier clustering est celui dont les paramètres sont ϵ à 1 et le nombre minimum de voisins à 4. Ce clustering étant à une échelle plus micro, c'est celui pour lequel on retrouve le plus de cluster : 3135. Aussi on voit que la population des clusters diffère de la population générale : la tuiles les plus représentée sont maintenant les tuiles urbaines légèrement devant les tuiles rurales. Les deux autres clusterings ont des échelles plus large et cherchent tous les deux 45 voisins dans l'entourage. Le clustering (2) cherche ces voisins avec un rayon équivalent valant 6 tuiles. On y retrouve des

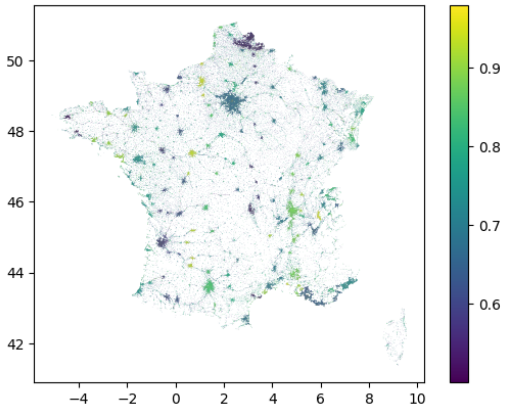
proportions proches de celles du clustering (1) mais avec beaucoup moins de groupes : 163. Enfin le clustering (3) qui a $\epsilon = 4$ présente seulement 73 clusters et leur composition est très intéressante avec plus de deux tiers de tuiles urbaines, et une augmentation significative du nombre de tuiles urbaines denses. Il semble donc que ces méthodes de clusterings des tuiles avec un certain taux d'urbanisme soit pertinentes en vu de prédire leurs classes puisqu'on y trouve des proportions des modalités d'urbanisme significativement différentes. Dans le suite nous étudierons les clusters du paramétrage (3) pour lequel il semblerait qu'on arrive à extraire les grandes villes / métropoles du territoire. Cette propriété est intéressante en vu d'étudier des territoire pour lesquels nous aurions peu de connaissances.



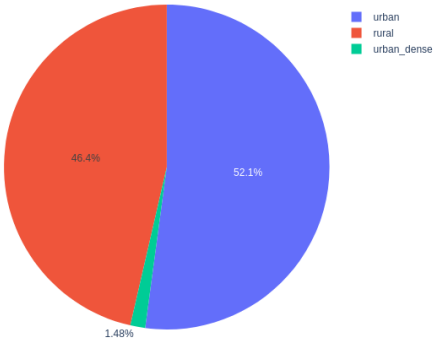
(a) Carte choroplèthe du clustering (1)



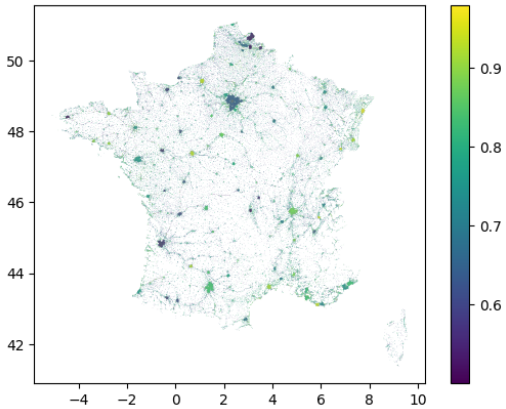
(b) Occupation des clusters (1)



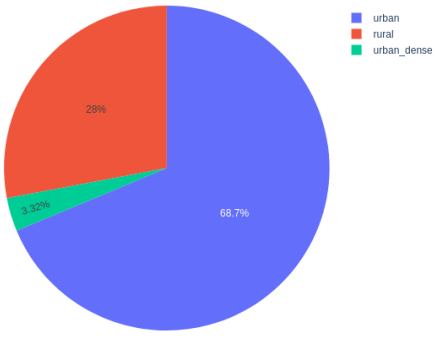
(c) Carte choroplèthe du clustering (2)



(d) Occupation des clusters (2)

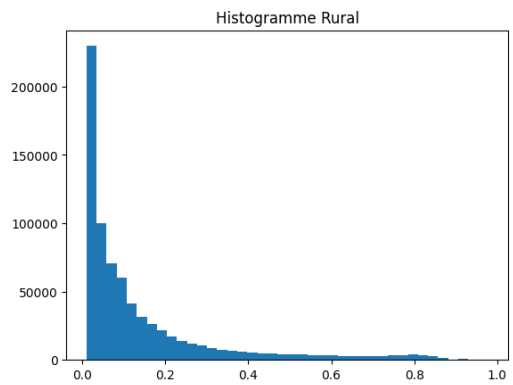


(e) Carte choroplète du clustering (3)

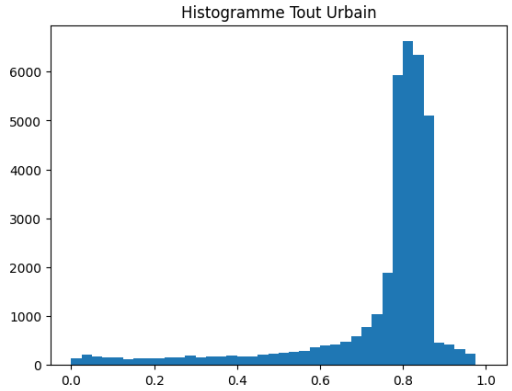


(f) Occupation des clusters (3)

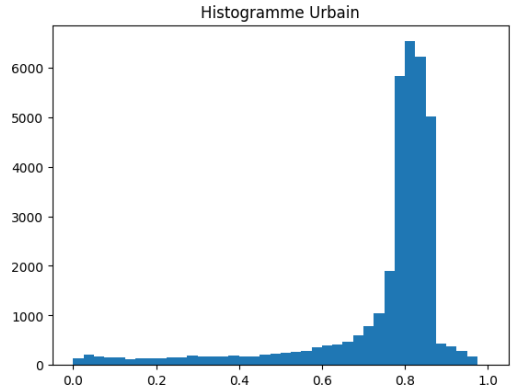
FIGURE 6 – Création des Clusters



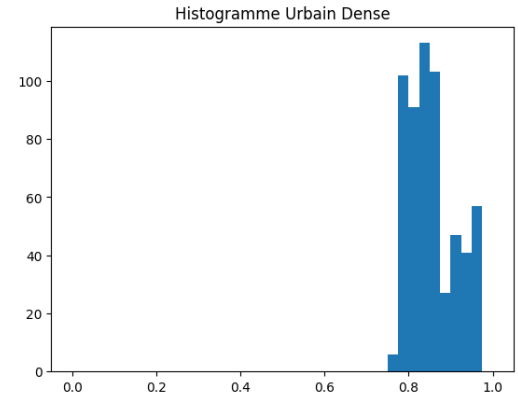
(a) Histogramme des tuiles rurales sur le taux d'urbanisme



(b) Histogramme des tuiles urbaines et urbaines denses sur le taux d'urbanisme



(c) Histogramme des tuiles urbaines sur le taux d'urbanisme



(d) Histogramme des tuiles urbaines denses sur le taux d'urbanisme

FIGURE 7 – Histogrammes des modalités d'urbanisme dans l'encodage

3.4 Régression

D'abord on voit sur les histogrammes 7 que la répartition des taux d'urbanisme diffère significativement selon la modalité d'urbanisme étudié. Ce résultat était attendu et c'est cela qui nous pousse confirme qu'une régression pour chacune de ces modalité est plus opportune.

On peut voir que les régressions ne performent pas tous aussi bien. La régression pour les tuiles rurales exprime 73% de la variance des individus sur le taux d'urbanisme. Ici et pour les tuiles urbaines denses où R^2 est assez fort on a des régressions qui expliquent une assez bien les données. En revanche, on voit que le modèle a plus de difficultés pour les tuiles urbaines, on peut expliquer cela en regardant l'histogramme dédié où l'on voit que c'est la modalité qui s'étale le plus sur les différentes valeurs du taux d'urbanisme.

Régime	R^2
Rural	0.7341
Urbain	0.3216
Urbain dense	0.6263

TABLE 2 – Score R^2 de le régression sur chaque modalité d'urbanisme

Ensuite, on regarde comment les variables interagissent entre elles et lesquelles impactent signifi-

cativement la régression. Premièrement, on peut constater que la régression se comporte de manière tout à fait particulière avec les variables qui dénombrent les routes. Ces variables ont toutes, quelque soit la modalité d'urbanisme étudiée, un coefficient très fort (une valeur absolue proche de 4 pour les rurales, de 12 pour les urbaines et de 43 pour les urbaines denses). Sachant qu'on veut établir un taux entre 0 et 1 cela est très étonnant, mais il faut aussi savoir que la variable ALL équivaut à la somme de toutes les autres routes ($MINOR$, $PATH$, $PRIMARY$, $SECONDARY$, $SERVICE$, $TERTIARY$, $TRACK$ et $TRUNK$), ainsi on comprend que le coefficient positif de la variable ALL vient compenser d'une certaine manière tous les coefficients négatifs des autres routes. Ces variables ne semblent pas particulièrement porteuses d'information car aucune d'entre elles n'a une probabilité de d'indépendance suffisamment faible, quelque soit la modalité encore une fois. La seule autre variable pour laquelle on ne peut pas rejeter l'hypothèse d'indépendance intervient uniquement dans la régression sur les tuiles urbaines : $BUILDPERCENTAGE$, c'est le taux de la tuile recouvert de bâtiment. Pour les tuiles rurales, on voit que les informations sur le taux et nombre de bâtiments dans une tuile sont les informations les plus dépendantes selon la t-statistic, c'est une caractéristique qui est propre a cette modalité, car ailleurs la t-statistic est bien moindre. Enfin on peut constater que l'information de la variable $clustered$ qui dit si la tuile appartient a un cluster, cette information étant construite à partir du taux d'urbanisme on comprends bien que cela soit porteur d'information.

Variable	Coefficient	Std.Error	t-Statistic	Probability
rural CONSTANT	0.0344411	0.0002395	143.7903367	0.0000000
rural BUILD PERCENTAGE	2.4623688	0.0166534	147.8598339	0.0000000
rural BUILD COUNT	0.0015280	0.0000028	553.6484953	0.0000000
rural ALL	3.9184543	22.7278419	0.1724077	0.8631171
rural SHOP	-0.0133469	0.0001939	-68.8414440	0.0000000
rural MINOR	-3.9189833	22.7278418	-0.1724309	0.8630988
rural PATH	-3.9120712	22.7278419	-0.1721268	0.8633379
rural PRIMARY	-3.8531048	22.7278431	-0.1695324	0.8653780
rural SECONDARY	-3.8820327	22.7278427	-0.1708052	0.8643770
rural SERVICE	-3.8938266	22.7278417	-0.1713241	0.8639690
rural TERTIARY	-3.8923477	22.7278422	-0.1712590	0.8640202
rural TRACK	-3.9120375	22.7278416	-0.1721253	0.8633390
rural TRUNK	-3.7998737	22.7278427	-0.1671903	0.8672204
rural clustered	0.0576218	0.0016973	33.9488000	0.0000000

TABLE 3 – Regression sur les tuiles rurales

Variable	Coefficient	Std.Error	t-Statistic	Probability
urban CONSTANT	0.5413380	0.0022526	240.3155617	0.0000000
urban BUILD PERCENTAGE	-0.0025466	0.0205669	-0.1238222	0.9014567
urban BUILD COUNT	0.0000705	0.0000030	23.1455574	0.0000000
urban ALL	12.7243870	35.2312453	0.3611677	0.7179763
urban SHOP	-0.0013613	0.0001109	-12.2760765	0.0000000
urban MINOR	-12.7026465	35.2312427	-0.3605506	0.7184376
urban PATH	-12.7186719	35.2312432	-0.3610055	0.7180975
urban PRIMARY	-12.7113631	35.2312432	-0.3607980	0.7182526
urban SECONDARY	-12.6992309	35.2312400	-0.3604537	0.7185101
urban SERVICE	-12.7252118	35.2312434	-0.3611911	0.7179588
urban TERTIARY	-12.6934786	35.2312427	-0.3602904	0.7186322
urban TRACK	-12.7187988	35.2312415	-0.3610091	0.7180948
urban TRUNK	-12.7036454	35.2312530	-0.3605789	0.7184165
urban clustered	0.0175523	0.0021422	8.1935655	0.0000000

TABLE 4 – Regression sur les tuiles urbaines

Variable	Coefficient	Std.Error	t-Statistic	Probability
urban dense CONSTANT	0.8239992	0.0080092	102.8818682	0.0000000
urban dense BUILD PERCENTAGE	0.0458525	0.0221980	2.0656178	0.0393136
urban dense BUILD COUNT	0.0000062	0.0000024	2.6311483	0.0087387
urban dense ALL	43.1960417	39.0553355	1.1060215	0.2691813
urban dense SHOP	0.0003388	0.0000264	12.8379342	0.0000000
urban dense MINOR	-43.1990388	39.0553111	-1.1060990	0.2691479
urban dense PATH	-43.1941253	39.0553339	-1.1059725	0.2692025
urban dense PRIMARY	-43.2043529	39.0553507	-1.1062339	0.2690895
urban dense SECONDARY	-43.2008448	39.0553467	-1.1061442	0.2691283
urban dense SERVICE	-43.1958281	39.0553454	-1.1060158	0.2691838
urban dense TERTIARY	-43.2004907	39.0552644	-1.1061375	0.2691312
urban dense TRACK	-43.1939145	39.0554017	-1.1059652	0.2692057
urban dense TRUNK	-43.1881276	39.0553750	-1.1058178	0.2692695
urban dense clustered	0.0187453	0.0038059	4.9252938	0.0000011

TABLE 5 – Regression sur les tuiles urbaines denses

Taux obsérvé	R^2
Urbain	-0.4089
Urbain Dense	-0.4123
Moyenne urbanisme	-0.408

TABLE 6 – Score R^2 des autres regressions

Enfin on regarde si la taille du cluster aurait pu avoir un impact sur le taux d'urbain, d'urbain dense et sur la moyenne des taux d'urbanisme et on voit que les courbes n'ont pas de pentes significatives, cela est confirmé par des score R^2 qui sont tous négatifs. Il aurait été intéressant de tester ces statistiques sur des clusters obtenus avec d'autre paramètre, on peut imaginer que les paramètres (1) auraient

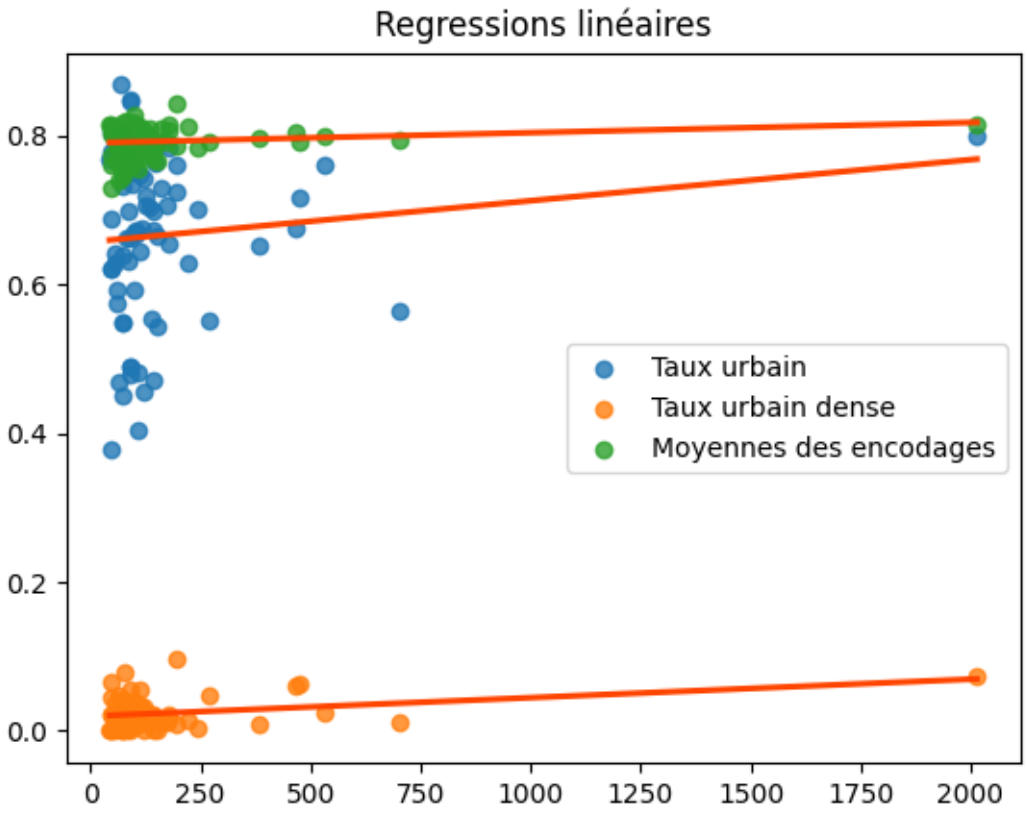


FIGURE 8 – Régressions linéaires de différents taux par rapport à la taille du cluster

rendu des résultats différents avec des clusters plus petits au vu de l'évolution des proportions dans les graphiques 6.

4 Conclusion

Pour conclure, on peut dire que nous avons pu utiliser des méthodes pour ajouter de l'information spatiale au données en appliquant, de manière tronquée, un lissage spatial et en utilisant du clustering. Cela nous a permis d'apporter aux tuiles une information sur le contexte spatial. Ensuite nous avons pu voir comment les variables pouvaient interagir entres elles au sein de chaque modalité d'urbanisme pendant une régression. Néanmoins, comme nous l'avons précédemment dit, l'entreprise n'a pas vraiment d'intérêt clair à prédire cet encodage mais on peut penser à des méthodes spatiales qui copie les mécanismes de méthodes de régressions spatiales avec par exemple une classification qui utilise les informations du voisinage, en plus des informations qu'on a vu ici comment obtenir. En regardant encore plus loin on peut imaginer qu'une fois l'étape de classification accomplie, on pourra utiliser, pour chaque modalité d'urbanisme (à la manière de ce qui a été fait ici) une régression pour évaluer les vitesses qui prend utiliserait les informations du voisinage.