

Cybergeog : European Journal of Geography

Cartographie, Imagerie, SIG

2021

972

Apport de la classification floue c-means spatiale en géographie : essai de taxinomie socio-résidentielle et environnementale à Lyon

Contribution of the spatial c-means fuzzy classification in geography: a socio-residential and environmental taxonomy in Lyon

Contribución a la clasificación espacial difusa c-means en geografía: una taxonomía socio-residencial y ambiental en Lyon



PPE APPARICIO

14

Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Tout accepter

Tout refuser

Personnaliser

[Politique de confidentialité](#)

Supervisées sont largement utilisées en géographie, même si les classifications ascendantes hiérarchiques (CAH) et *k-means* des données spatialisées. Or, des développements récents en géographie ont introduit de nouveaux algorithmes intégrant l'espace dans le processus de classification. La méthode *ClustGeo* est une extension spatiale de la CAH. Cette méthode est basée sur les méthodes de classification spatiales en logique floue et permet de présenter la méthode SFCM, soit une extension de la CAH. Pour ce faire, nous l'appliquons à des données socio-résidentielles lyonnaises. Aussi, nous la comparons à sa contrepartie non spatiale, la méthode *ClustGeo*. Les résultats soulignent que la méthode SFCM

bénéficie à la fois des avantages de l'approche floue et de l'approche spatiale, facilitant l'interprétation des résultats.

Unsupervised classification methods are common in geography, but the most widely used such as the Hierarchical Clustering Algorithm (HCA) and the k-means are not designed to work with spatial data. New developments in spatial statistics have brought new algorithms that take space into account. For example, the *ClustGeo* method is the spatial extension of the classical HCA. At the same time, little attention was given to spatial fuzzy classification method in geography. The paper aims to present the SFCM method, which is a spatial extension of the fuzzy c-means. To do so, we applied this method to socio-environmental data of the agglomeration of Lyon in France. We thus compared the results to its non-spatial counterpart (FCM), to the HAC and the *ClustGeo* method. The results showed that the SFCM combined both advantages of the fuzzy and spatial approach, making the interpretation of the results easier.

Los métodos de clasificación no supervisada se utilizan regularmente en geografía, incluso si los más empleados como la clasificación ascendente jerárquica (CAH) y k-medias, son poco adaptados a los datos espacializados. Recientes contribuciones en la estadística espacial han dado lugar a nuevos algoritmos que integran el espacio en tal proceso de clasificación. Por ejemplo, el método ClustGeo el cual se constituye como una extensión espacial de la CAH. Sin embargo, una baja atención se ha prestado a los métodos de clasificación espacial de lógica difusa en geografía. Así, el objetivo de este artículo es presentar el método SFCM, es decir una clasificación espacial difusa c-means y para ello, lo aplicamos a datos socioambientales de la aglomeración de Lyon, comparando, además, con su contraparte no espacial (FCM), CAH y el método ClustGeo. Los resultados expresan que el método SFCM se beneficia tanto de las ventajas del enfoque difuso como del espacial, facilitando la interpretación de los resultados.

Entrées d'index

Mots-clés : classification, classification automatique, logique floue, k-means, analyse spatiale, Lyon, équité

Keywords : clustering, automatic classification, k-means, fuzzy logic, spatial analysis, Lyon, equity

Palabras claves : clasificación, clasificación automática, lógica difusa, k-medias, análisis espacial, Lyon, equidad

Texte intégral

Cette recherche a bénéficié du soutien financier de la Chaire de recherche du Canada sur l'équité environnementale et la ville (950-230813).



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Tout accepter

Tout refuser

[Personnaliser](#)

[Politique de confidentialité](#)

odes de classification visant à regrouper les observations classes sur la base de plusieurs variables (Theodoridis *et* ent une distinction entre les méthodes de classification s supervisées (figure 1). Les premières relèvent de la ensionnelle et permettent de classifier automatiquement e a priori sur la nature des classes présentes dans le jeu ; les plus connues sont sans conteste les algorithmes de hique (CAH) et *k-means* (k-moyennes). Quant aux ecter des observations à partir d'un échantillon déjà ses d'entraînement ; les plus connues sont les forêts x de neurones artificiels ou encore l'analyse factorielle . distingue habituellement les méthodes dites strictes (ou floues (*hard* versus *soft clustering* en anglais) (figure 1). ation stricte, chaque observation appartient à une seule

spatiale) comparativement aux méthodes CAH (stricte, a-spatiale), *c-means* (floue, a-spatiale) et *ClustGeo* (stricte, spatiale), en termes de qualité de classification sur les plans sémantique, spatial et de l'interprétation ? Pour ce faire, l'article est organisé comme suit. Dans un premier temps, nous ferons un bref retour sur les principales méthodes de classification non supervisées dites classiques, qu'elles soient strictes (CAH et *k-means*) ou floues (*fuzzy c-means*). Dans un deuxième temps, nous décrirons deux principales méthodes non supervisées intégrant une dimension spatiale (*ClustGeo* et SFCM), en prenant soin de les distinguer préalablement des méthodes de classification par agrégation spatiale intégrant une contrainte spatiale (par exemple AZP, SKATER et AMOEBa) (Aldstadt, Getis, 2006 ; Assunção *et al.*, 2006 ; Openshaw, 1977). Puis, afin d'illustrer l'apport de cette méthode en géographie, nous proposons de comparer les résultats de quatre méthodes de classification non supervisées : une CAH classique (a-spatiale et stricte), un FCM classique (a-spatial et flou), la méthode *ClustGeo* (spatiale et stricte) et la méthode SFCM (spatiale et floue), et ce, en mobilisant un jeu de données sur des variables socio-économiques et environnementales de l'agglomération de Lyon au niveau des îlots regroupés pour l'information statistique (IRIS). De cette façon, il sera possible de distinguer les apports respectifs des méthodes de classification floue/exclusive et spatiale/a-spatiale.

Figure 1 : Synthèse des principales méthodes de classification

	Classifications non supervisées	Classifications supervisées
Classifications strictes	k-moyennes (<i>k-means</i>) k-médianes (<i>k-medians</i>) k-médoïdes (<i>k-medoids</i>) Isodata CAH ¹ Classification mixte BIRCH ² , DBSCAN ³ , OPTICS ⁴ Partitionnement spectral	Forêts d'arbres décisionnels Méthode des <i>k</i> plus proches voisins Machines à vecteurs de support Analyse factorielle discriminante Réseaux de neurones artificiels
Classifications floues	Classification k-moyennes floue (<i>Fuzzy c-means -FCM</i>) Classification k-médianes floue (<i>Fuzzy c-Least Medians clustering</i>) Modèles à mélanges finis	Régression multinomiale Modèle de Markov caché Classification naïve bayésienne

¹ Classification ascendante hiérarchique.

² *Balanced Iterative Reducing and Clustering using Hierarchies*.

³ *Density-Based Spatial Clustering of Applications with Noise*.

⁴ *Ordering Points To Identify the Clustering Structure*.

Conception et réalisation : Jérôme Golb et Philippe Apparicio, 2021.



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Tout accepter

Tout refuser

Personnaliser

[Politique de confidentialité](#)

ur les principales assification non ssiques

érarchique (CAH) est sans aucun doute la méthode de plus utilisée en géographie. À la fin des années 1980, de cette méthode réalisé par Lena Sanders (1989), dans *des données en géographie*, a certainement contribué à CAH auprès des géographes francophones durant les éthode repose en effet sur un principe très simple : elle ent les observations en fonction de leur degré de d'une seule classe. Pour n individus, on obtient ainsi une mière partition avec $n-1$ classes, puis $n-2$ classes et ainsi

définition de zones homogènes pour la marchabilité à Montréal (Riva *et al.*, 2008) et sur la géographie du don de sang au Québec (Cloutier *et al.*, 2011). Quant à la classification mixte, elle a aussi été utilisée pour qualifier les types d'espaces ruraux en Europe méditerranéenne (Chevalier *et al.*, 2010) ou classifier des images Landsat (Apparicio, 2002).

- 8 Le *fuzzy c-means* (FCM), aussi appelé *fuzzy k-means*, a été originalement proposé par Dunn (1973). Il s'agit d'une méthode en logique floue, c'est-à-dire que pour chaque observation, la probabilité d'appartenance aux c groupes est évaluée et présentée sous forme d'une matrice d'appartenance u de taille n par c , où u_{ik} est la probabilité de l'observation k d'appartenir au groupe i . La somme de chaque ligne u est égale à 1. Comme pour le *k-means*, le FCM nécessite de fixer au préalable le nombre de classes et le nombre d'itérations ; les centres de classes initiaux sont aussi soit tirés au hasard, soit spécifiés par l'utilisateur. Ensuite, de façon itérative, le centre des classes (v) est mis à jour avec la matrice u , puis les probabilités d'appartenance sont réévaluées avec les nouveaux centres de classes. Ce processus est répété jusqu'à ce que le nombre d'itérations fixé soit atteint ou jusqu'à ce qu'un critère de convergence soit satisfait. La mise à jour des classes et de u est calculée avec la formule suivante :

$$u_{ik} = \frac{(\|x_k - v_i\|^2)^{-1/(m-1)}}{\sum_{j=1}^c (\|x_k - v_j\|^2)^{-1/(m-1)}} \text{ et } v_i = \frac{\sum_{k=1}^N u_{ik}^m(x_k)}{\sum_{k=1}^N u_{ik}^m}$$

où k et i sont respectivement une observation et une classe.

- 9 Le paramètre m est le principal ajout à la méthode FCM. Il permet de contrôler le degré de flou de la classification. Si $m=1$, la classification obtenue sera stricte (la matrice u sera uniquement remplie de 0 et de 1) ; avec un m plus grand, les valeurs de la matrice u diminuent jusqu'à être parfaitement homogènes. Comme pour le *k-means* classique, différents types de distances peuvent être utilisés (euclidienne, khi2, de Mahalanobis, etc.) et le choix des centres de classes initiaux tout comme le nombre d'itérations peuvent potentiellement affecter les résultats de la classification. Il est habituel de tester plusieurs valeurs pour le nombre de classes jusqu'à trouver la combinaison idéale entre le c et le m . La méthode FCM a été largement utilisée pour l'identification des utilisations du sol à partir d'images satellitaires (Fisher, Pathirana, 1990), mais aussi dans l'analyse du marché immobilier (Hwang, Thill, 2009), ou encore des activités quotidiennes des citoyens (Hafezi *et al.*, 2017). Une solution de rechange aux approches de logique floue a également été proposée : les approches par recouvrement, lesquelles autorisent des observations à appartenir à plusieurs classes dans une classification stricte (Baadel *et al.*, 2016).



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

le classification non es spatiales

présentées ci-dessus n'ont pas été élaborées pour des on spatiale. En effet, la ressemblance des observations istiques sémantiques. Dans ce contexte, les appliquer à ose deux problèmes majeurs.

e l'information propre aux données, à savoir leur apte dans le processus de classification. Or, la dimension ructurante ; par conséquent, l'occulter revient à perdre formation. Il convient toutefois de nuancer quelque peu ygraphie des méthodes de classification non supervisées évèlent des effets de voisinage, d'autant plus que les ification sont fortement autocorrélées positivement.

- 12 Deuxièmement, dans un contexte d'autocorrélation spatiale positive, des observations proches spatialement devraient plus vraisemblablement appartenir au même groupe. Avec les méthodes de classification a-spatiales, il est fréquent d'observer des phénomènes de mitage, c'est-à-dire des observations appartenant à un groupe b et isolées au milieu d'un ensemble d'observations appartenant au groupe a . Ce phénomène peut s'expliquer par la présence ici et là d'autocorrélation spatiale locale négative, c'est-à-dire des observations dont les caractéristiques sémantiques diffèrent de leurs voisines. Souvent, la dissimilarité sémantique entre ces observations est négligeable et ne justifie pas cette rupture spatiale. Notons que ce problème se pose moins pour les méthodes de classification floues, puisque l'analyse des probabilités d'appartenance à chacune des classes permet justement de limiter ces problèmes de mitage.
- 13 Le besoin de méthodes de classification produisant des résultats plus cohérents spatialement n'est pas nouveau : il a émergé dès les années 1980, avec les travaux précurseurs en économie régionale de Fischer (1980) sur l'identification de régions homogènes, soit des ensembles d'entités spatiales contiguës ayant un fort degré de similarité par rapport à un ensemble d'attributs. Cet exercice a une application directe dans diverses disciplines : géographie, économie régionale, écologie, études urbaines, transport, etc.
- 14 Les premières approches proposées historiquement imposent une contrainte spatiale de contiguïté. Nous les qualifions ici de méthodes par agrégation spatiale. Elles se distinguent par leur façon de construire des groupes : k entités spatiales sont d'abord sélectionnées au hasard ou par l'utilisateur, puis leurs voisines respectives sont agrégées à ces premières entités pour former des régions, en cherchant à maximiser la variance interrégionale et à minimiser la variance intrarégionale. Le résultat de ce processus est la création de régions très cohérentes spatialement, c'est-à-dire avec une absence de mitage. Autrement dit, avec ces méthodes, il n'est pas possible d'identifier de groupes qui seraient spatialement discontinus, c'est-à-dire composés de plusieurs ensembles régionaux séparés. L'impossibilité d'obtenir du mitage au sein des différentes régions peut masquer la présence de valeurs fortement dissemblables localement, malgré la prise en compte de l'espace. Or, ces observations systématiquement différentes de leurs voisines doivent faire l'objet d'une attention particulière dans les exercices de classification intégrant l'espace, ce que ne permettent pas ces méthodes d'agrégation spatiale. Parmi elles, les plus connues sont l'AZP (Openshaw, 1977 ; Openshaw, Rao, 1995), la méthode SKATER (Assunção *et al.*, 2006) et l'AMOEBa (Aldstadt, Getis, 2006). Les limites de ces méthodes, particulièrement celles relatives au mitage, ont conduit plus récemment à



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Tout accepter

Tout refuser

Personnaliser

[Politique de confidentialité](#)

des incluant l'espace dans le processus de classification, contiguïté. Plus spécifiquement, ces nouvelles méthodes rythmes classiques, tels que la CAH ou le FCM, pour parallèle à la dimension sémantique des données. En s'intégré comme une contrainte dans les algorithmes de une donnée supplémentaire.

Idante hiérarchique spatiale : la

récentement, constitue une version modifiée de la CAH sur des données (Chavent *et al.*, 2018). Elle est basée sur le α et introduit deux matrices de dissimilarité entre les variables sémantique (D_0) et une matrice spatiale (D_1). La première concerne les observations pour les valeurs des variables, tandis que la seconde concerne les observations géographiques entre les observations. Un paramètre α

permet de définir le poids de D_1 comparativement à D_0 : avec $\alpha=0$, aucun poids n'est accordé à D_1 et l'on obtient alors une CAH classique ; avec $\alpha=1$, tout le poids est accordé à D_1 et produit une classification purement spatiale. L'enjeu principal est donc de fixer la valeur du paramètre α , considérant qu'une augmentation de α revient à améliorer l'inertie expliquée de la matrice spatiale, au détriment d'une perte de l'inertie expliquée sur le plan sémantique. Puisque la méthode *ClustGeo* est une extension de la CAH, il s'agit également d'une méthode de classification stricte. Bien que récente, la méthode a déjà été utilisée par exemple à Bandung (Indonésie), pour l'analyse de cas de diarrhée (Jaya *et al.*, 2019), et dans l'Indiana (États-Unis) pour la délimitation de régions de gestion de la population des cerfs de Virginie (Swihart *et al.*, 2020).

Spatial fuzzy c-means (SFCEM)

16 Jusqu'à présent, les méthodes de classification floues incluant une dimension spatiale ont peu fait l'objet d'applications empiriques en géographie, malgré leur intérêt non négligeable face aux méthodes strictes. Pourtant, ces méthodes existent et sont notamment mises à profit dans le domaine de l'imagerie par résonance magnétique (IRM). En effet, l'utilisation de méthodes de classification floues, avec notamment le FCM, est une pratique courante pour segmenter des images médicales, notamment des IRM du cerveau. Cependant, la présence de bruit dans ces images rend le processus de classification plus compliqué, ce qui a conduit à l'introduction d'une dimension spatiale dans l'algorithme classique FCM (Adhikari *et al.*, 2015 ; Cai *et al.*, 2007 ; Chuang *et al.*, 2006). Plus spécifiquement, nous avons retenu ici la méthode SFCEM (Cai *et al.*, 2007) pour sa rapidité d'exécution et parce qu'elle n'introduit qu'un seul paramètre supplémentaire dans l'algorithme. Comme pour la méthode *ClustGeo*, ce paramètre α vient contrôler le poids de la dimension spatiale par rapport à la dimension sémantique.

17 Comparativement au FCM classique, le SFCEM introduit dans son calcul, en plus du jeu de données original (D_0), une version spatialement décalée (D_s) de ce dernier (Cai *et al.*, 2007). En analyse d'image, cela revient à calculer D_s en appliquant un filtre moyen ou médian à D_0 (la médiane étant moins sensible aux valeurs extrêmes locales). Ce processus peut facilement s'appliquer à des entités géographiques vectorielles, en créant une matrice de pondération spatiale W_{kl} (l étant les voisins de k et la diagonale de cette matrice valant 1) (Getis, 2009) et en utilisant les poids de cette matrice dans le calcul d'une moyenne ou

à fonction objectif du SFCEM est la suivante :

$$\sum_{i=1}^c \|v_i\|^2 + \alpha \sum_{i=1}^c \sum_{k=1}^N u_{ik}^m \|x_k - v_i\|^2 \text{ où}$$



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Tout accepter

Tout refuser

Personnaliser

[Politique de confidentialité](#)

res de gravité des groupes, c le nombre de groupes, m le degré de gravité ; k une observation et N le nombre d'observations ; $\|x_k - v_i\|^2$ la distance euclidienne entre l'observation k et le groupe v_i ; $\|x_k - v_i\|^2$ la distance euclidienne entre la version spatialement décalée de k et le groupe v_i ;

u_{ik} la probabilité pour l'observation k d'appartenir au groupe v_i ;

réduite à la version spatialement décalée du jeu de données, alors le SFCEM est réduit à un simple FCM. Si $\alpha=2$, alors la version spatialement décalée aura deux fois plus de poids dans la classification spatiale.

être calculés grâce aux équations suivantes :

$$u_{ik} = \frac{(\|x_k - v_i\|^2 + \alpha \|x_k - v_i\|^2)^{-1/(m-1)}}{\sum_{j=1}^c (\|x_k - v_j\|^2 + \alpha \|x_k - v_j\|^2)^{-1/(m-1)}} \text{ et } v_i = \frac{\sum_{k=1}^N u_{ik}^m (x_k + \alpha x_k)}{(1+\alpha) \sum_{k=1}^N u_{ik}^m}$$

- 19 La méthode SFCM présente des similarités avec la méthode *ClustGeo*, notamment la recherche d'un compromis entre la dimension sémantique du jeu de données et la dimension spatiale avec le paramètre α . On distingue toutefois deux différences notables. Dans le SFCM, l'espace est introduit au moyen d'une version spatialement décalée du jeu de données initial, alors que dans *ClustGeo*, il s'agit simplement d'une matrice de distances. Autrement dit, *ClustGeo* inclut l'agencement spatial des observations tandis que le SFCM inclut une information sémantique localisée dans l'environnement immédiat. Deuxièmement, pour le SFCM, α correspond au poids relatif de la dimension spatiale (valeurs comprises entre 0 et $+\infty$) alors que pour *ClustGeo*, il s'agit de la proportion du poids total à répartir entre la dimension spatiale et la dimension sémantique (valeurs comprises entre 0 et 1).
- 20 L'intérêt d'utiliser une matrice spatiale W dans le SFCM est de laisser à l'utilisateur la possibilité de définir le degré de lissage spatial introduit dans la classification. En effet, une matrice donnant du poids à des observations lointaines contribuera à produire des résultats plus lissés spatialement qu'une matrice plus restrictive. Notons ici un autre développement intéressant proposé par Guo *et al.* (2015), à savoir un FCM spatial basé non pas sur les valeurs de voisins, mais sur la distance spatiale entre les observations.

La question de l'inertie sémantique et de la cohérence spatiale

- 21 Tel que souligné précédemment, l'enjeu d'une classification spatiale est de parvenir à trouver un équilibre entre la dimension spatiale et la dimension sémantique des données. Ainsi, la modification d'ordre spatial de la fonction objectif des algorithmes FCM et CAH s'accompagne à la fois d'un inconvénient et d'un avantage. D'un côté, la prise en compte de la dimension spatiale génère inévitablement une perte de la qualité d'ajustement du point de vue sémantique. Autrement dit, l'inertie expliquée par les méthodes de classification spatiales sera inévitablement moins grande que celle expliquée par des méthodes a-spatiales. En revanche, elle permet un gain de cohérence spatiale, en augmentant la probabilité que des observations voisines ou proches les unes des autres appartiennent au même groupe.



ement spatiale

Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

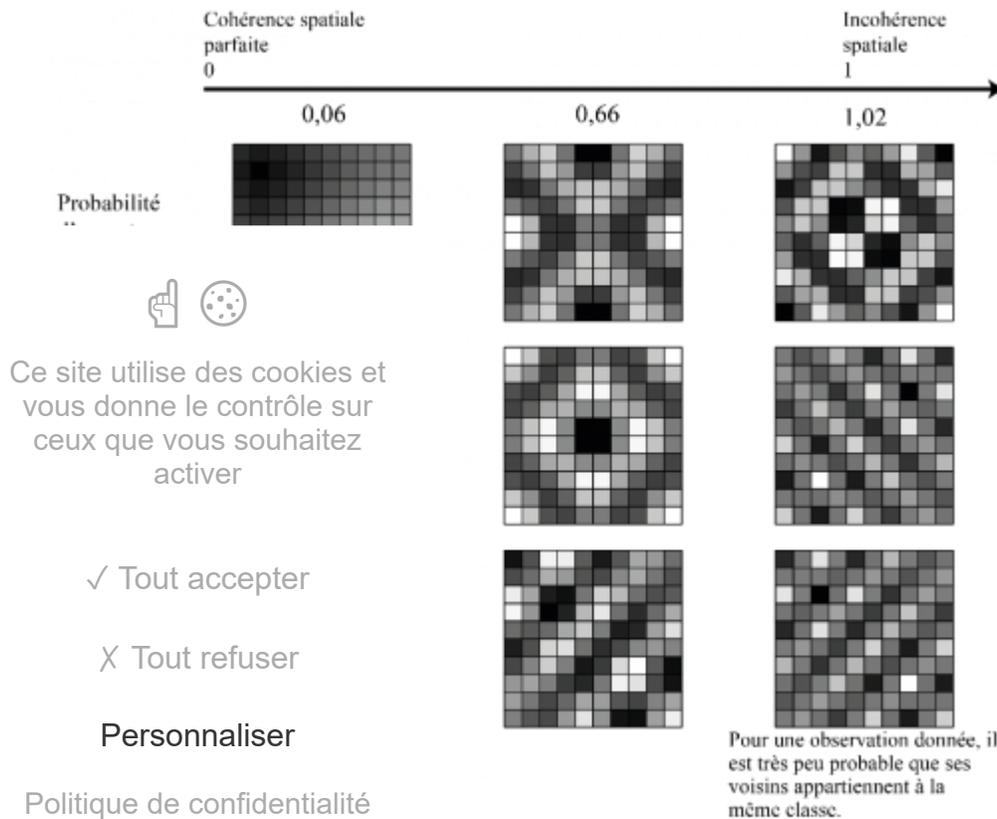
Geo, il est possible de calculer la part de l'inertie spatiale pose sur une matrice de distance spatiale en plus de la Par contre, cette méthode ne peut pas être appliquée au posons de quantifier le degré d'incohérence spatiale (le observations proches) comme étant la somme des écarts t leurs voisins sur la matrice d'appartenance u . Cette ne une quantité au même titre que l'inertie : elle est ie classification, chaque observation appartient fférent de ses voisins (autocorrélation spatiale négative ju'au contraire, les observations appartiennent pe (autocorrélation spatiale positive parfaite). Nommons iale S_{cr} , calculée comme suit :

$$\|u_k - u_l\|^2$$

où W_{kl} est la pondération spatiale entre les observations voisines k et l , et $\|u_k - u_l\|^2$ est la distance euclidienne entre les probabilités d'appartenance sur la matrice u entre l'observation k et sa voisine l .

23 Pour standardiser cette somme, il serait possible de la diviser par la plus grande incohérence spatiale possible produite par une classification spatiale, avec une autocorrélation spatiale négative parfaite. Cependant, cette approche rendrait la comparaison de deux matrices u plus difficile, car le dénominateur serait arbitrairement trop grand. En effet, une classification spatiale tend à produire des résultats spatialement autocorrélés. Il serait donc plus judicieux de comparer la valeur de S_{cr} à l'incohérence spatiale obtenue sous l'hypothèse d'une répartition spatiale aléatoire des données (S_{ca}). Pour estimer S_{ca} , il est possible de recourir aux méthodes de Monte-Carlo, c'est-à-dire de permuter les observations dans la matrice u_{ik} un grand nombre de fois (habituellement 999), afin de simuler la distribution de S_{ca} . Nous pouvons ensuite comparer la valeur de S_{cr} avec la moyenne de S_{ca} , reporter un intervalle de confiance à 95 %, ou encore considérer la distribution du ratio S_{cr}/S_{ca} dans son ensemble. Il est donc attendu que le ratio S_{cr}/S_{ca} soit compris entre 0 (cohérence spatiale parfaite, atteinte dans le cas où toutes les observations appartiennent au même groupe) et 1 (incohérence spatiale équivalente à celle attendue si les observations étaient aléatoirement réparties dans l'espace). Il est mathématiquement possible que ce ratio soit supérieur à 1, ce qui signifierait que la classification est marquée par une forme d'autocorrélation spatiale négative. Ce cas de figure est cependant peu probable, considérant que l'objectif même du SFCM est de produire une classification marquée par de l'autocorrélation spatiale positive. La figure 2 présente les valeurs de l'indicateur d'incohérence spatiale obtenues pour des résultats fictifs de trois classifications. Sans surprise, plus l'autocorrélation spatiale positive est importante pour les probabilités d'appartenance de trois classes, plus la valeur de l'indicateur d'incohérence spatiale sera proche de 0.

Figure 2 : Illustration graphique des valeurs prises par l'indicateur d'incohérence spatiale



cio, 2021.

- 24 Il est intéressant de noter que cette approche peut également être appliquée aux méthodes de classification strictes, à la seule différence que ces dernières ont des matrices d'appartenance u binaires (uniquement avec des valeurs de 0 et de 1). D'autres indicateurs peuvent aider à quantifier le degré de cohérence spatiale, comme l'indice de Moran, calculé sur chacune des colonnes de la matrice d'appartenance u , ou encore les *joint count statistics*, calculées sur les classes de la classification après avoir définitivement arrêté les classes de chaque observation.

Degré d'incohérence sémantico-spatiale

- 25 L'indice S_{cr} permet de quantifier le degré d'incohérence spatiale d'une classification floue. Il est calculé uniquement à partir de la matrice d'appartenance et ne dépend donc pas des données originalement utilisées dans la classification. Une de ses caractéristiques est d'augmenter avec le nombre de groupes, car un plus grand nombre de frontières diminue la cohérence spatiale. Cette caractéristique n'est pas nécessairement désirable ; en effet, une frontière nette entre des observations très différentes sur le plan sémantique ne devrait pas pénaliser l'indicateur final. Pour cela, nous proposons un second indicateur d'incohérence sémantico-spatiale S_{mcr} , pondérant les différences observées entre voisins dans la matrice d'appartenance u par l'inverse de la distance sémantique entre ces observations. La formule est identique à l'équation [3], la seule différence résidant dans la matrice spatiale W , calculée comme suit :

$$W_{kl} = \frac{1}{\|k-l\|^2}$$

où W_{kl} est la pondération entre l'observation k et sa voisine l , et $\|k-l\|^2$ est la distance euclidienne sur le plan sémantique entre k et l .

- 26 Cette matrice W est standardisée globalement, et non en ligne, afin de donner plus de poids aux écarts sémantiques importants dans toute la zone d'étude, et pas seulement localement. Les deux indicateurs S_{cr} et S_{mcr} peuvent être utilisés conjointement pour déterminer le meilleur équilibre entre l'inertie sémantique expliquée par la classification et sa cohérence spatiale.

Territoire d'étude, entités spatiales et les



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Tout accepter

Tout refuser

Personnaliser

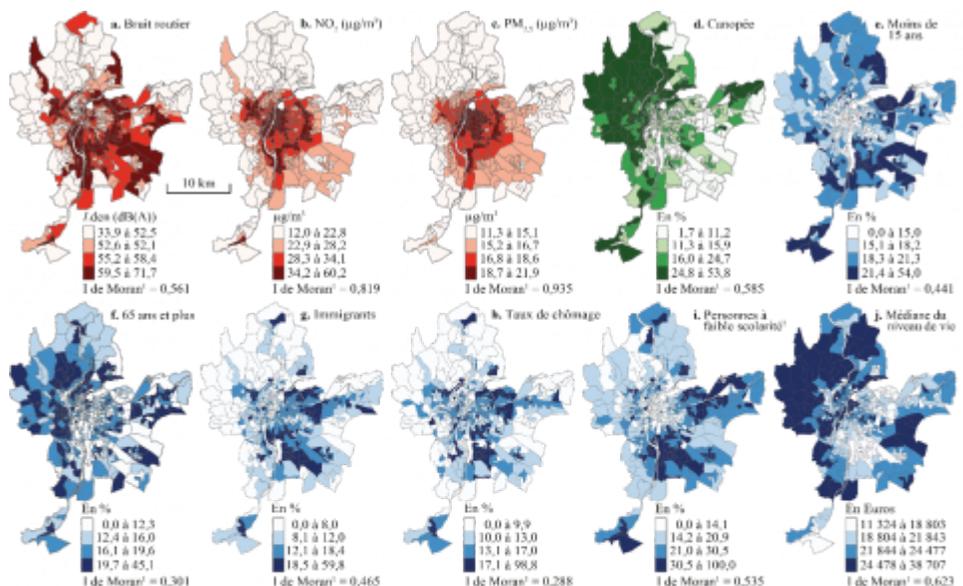
[Politique de confidentialité](#)

sons ici est une analyse exploratoire d'un ensemble de vironnementales de l'agglomération lyonnaise au niveau 800 et 500 habitants dans le cas des IRIS d'habitation s comprend dix variables (figure 3). Les trois premières s nuisances urbaines, à savoir les pollutions sonores et moyen durant la journée (L_{den} en 2010) ainsi que les concentration de dioxyde d'azote (NO_2 en $\mu g/m^3$ en 2015) en $\mu g/m^3$ en 2015). Quant à la quatrième variable à un élément positif du cadre de vie urbain, soit la le pourcentage de la surface de chaque IRIS couverte par plus de trois mètres, selon des données LiDAR de 2015, uivent ensuite six variables socio-économiques (en bleu zes de jeunes de moins de 15 ans, de personnes de 65 ans : chômage, le pourcentage de personnes de 15 ans et plus

faiblement scolarisées (ayant obtenu au plus le brevet des collèges) et finalement, la médiane du niveau de vie (en €), soit le revenu disponible du ménage divisé par le nombre d'unités de consommation. Les données socio-économiques proviennent du recensement de 2015. Les données environnementales sont accessibles sur le site des données ouvertes du Grand Lyon (Métropole de Lyon) et de l'Atmo Auvergne-Rhône-Alpes (2019).

28 Sans surprise, l'analyse cartographique de ces variables et de leurs valeurs respectives de l'indice de Moran révèle une forte autocorrélation spatiale positive des indicateurs environnementaux, surtout ceux de la pollution de l'air. En effet, les IRIS localisés dans les arrondissements de la Ville de Lyon présentent des niveaux de pollution bien plus élevés comparativement à ceux des communes périphériques de l'agglomération. Notons aussi la forte autocorrélation spatiale de la médiane du niveau de vie : les IRIS localisés dans les 4^e, 2^e et 6^e arrondissements de la Ville de Lyon et dans les communes à l'ouest de la Saône présentent les valeurs les plus fortes, tandis que ceux du 9^e arrondissement et des communes de Vaulx-en-Velin, Villeurbanne, Saint-Fons, Vénissieux et Givors affichent les revenus médians les plus faibles. Il est à noter que toutes ces variables sont préalablement centrées réduites (moyenne = 0 et écart-type = 1), afin de réaliser les classifications.

Figure 3 : Cartographie des différents indicateurs



- 12 Découpage administratif
- 17 Feytaud
- 18 Fleury-sur-Saône
- 19 Fontaines-Saint-Martin
- 20 Fontaines-sur-Saône
- 21 Francheville
- 22 Genay
- 23 Givors
- 24 Grigny
- 25 Irigny
- 26 Jonage
- 27 La Mulotière
- 28 La Tour-de-Salvagny
- 29 Lirioisnet
- 30 Lissieu
- 31 Mancy-Fénelin
- 32 Meyzieu
- 33 Mions
- 34 Montanay
- 35 Neuville-sur-Saône
- 36 Oullins
- 37 Pierre-Bénite
- 38 Poleymieux-au-Mont-d'Or
- 39 Rillieux-la-Pape
- 40 Rochetaillière-sur-Saône
- 41 Saint-Cyr-au-Mont-d'Or
- 42 Saint-Denis-sur-Meurville
- 43 Sainte-Poye-Lyon
- 44 Saint-Fons
- 45 Saint-Genis-Laval
- 46 Saint-Germain-lès-Ollières
- 47 Saint-Germain-au-Mont-d'Or
- 48 Saint-Priest
- 49 Saint-Romain-au-Mont-d'Or
- 50 Sathonay-Camp
- 51 Sathonay-Village
- 52 Solaise
- 53 Tassin-la-Demi-Lane
- 54 Vaulx-en-Velin
- 55 Vénissieux
- 56 Verançon
- 57 Villeurbanne

Elle est calculée avec une matrice de contiguïté (selon le partage d'un nœud) standardisée en ligne. Toutes les significations avec P < 0,001. * Population de plus de 15 ans ayant au plus le brevet des collèges. Grand Lyon, 2010. In et e. Atmo Auvergne-Rhône-Alpes, 2015. Base ouverte des données publiques françaises, 2015; F. J. INSEE, 2015. Source : Jérémy Gelb et Philippe Apparcia, 2021.

Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

Application empirique au de Lyon

netres k et m pour les méthodes

29 Pour déterminer le nombre optimal de classes à retenir (k) ainsi que le paramètre m de logique floue, nous réalisons dans un premier temps un FCM classique. Pour ce FCM, nous avons calculé pour chaque valeur de k entre 2 et 9 et chaque valeur de m entre 1 et 3 (avec un saut de 0,1) cinq mesures permettant d'évaluer la qualité des classifications.

30 La première, sans aucun doute la plus utilisée, est la part de l'inertie expliquée variant de 0 à 100 %. La deuxième, l'indice de Xie et Beni (1991), correspond au ratio de la compacité et de la séparation d'une classification floue : une plus grande séparation entre les groupes formés est souhaitable, au même titre qu'une petite compacité ; une valeur faible de l'indice de Xie et Beni indique donc une meilleure classification. La troisième, le critère de silhouette ajusté aux classifications floues (Campello, Hruschka, 2006), permet de mesurer la qualité d'une classification en comparant les distances entre chaque observation et leurs deux centres de groupes les plus proches, pondérées par les probabilités d'appartenance à ces groupes : une valeur plus élevée indique une meilleure classification, c'est-à-dire une faible superposition des groupes. La quatrième, le coefficient de partition (Oliver, Webster, 1989), permet de quantifier l'intensité du floutage d'une classification floue comparativement à une classification stricte, mais ne rend pas compte de la qualité de la prédiction en tant que telle : ce coefficient varie de 0 (complètement flou) à 1 (séparation stricte). La cinquième, l'indice de Fukuyama et Sugeno (1989), combine à la fois la compacité des groupes, la séparation entre les groupes et le degré de flou dans la classification : une augmentation de la compacité et une augmentation de la séparation entraînent une diminution de l'indice, une valeur plus faible indiquant donc une meilleure classification.

31 Les résultats de ces indicateurs sont présentés à la figure 4. Sans surprise, plus grand est le nombre de groupes, plus importante est la valeur de l'inertie expliquée ; k semble optimal avec des valeurs de 3 ou 4 (le gain d'inertie étant limité au-delà de 4, figure 4a). Concernant le paramètre m (degré de flou), l'indice de silhouette est maximal quand m est proche de 1,5 (figure 4c). L'intensité du floutage augmente rapidement avec m (figure 4b), mais n'atteint que 60 % lorsque $m=1,5$ et $k=4$, indiquant ainsi une distinction encore relativement nette entre les groupes. Enfin, les indices de Fukuyama et Sugeno et de Xie et Beni indiquent tous les deux une classification moins efficace, avec l'augmentation de m . Au-delà de $m=1,5$, la dégradation de la classification est particulièrement marquée. Il apparaît donc que $m=1,5$ donne les meilleurs résultats, considérant l'ensemble des indicateurs. Au final, avec $k=4$ et $m=1,5$, la classification obtenue permet d'expliquer 30 % de l'inertie totale. À des fins de comparaison, la CAH classique a également été calculée avec quatre classes ($k=4$) une métrique euclidienne au carré et le critère d'agrégation de



les classifications FCM de 2 à 9 classes

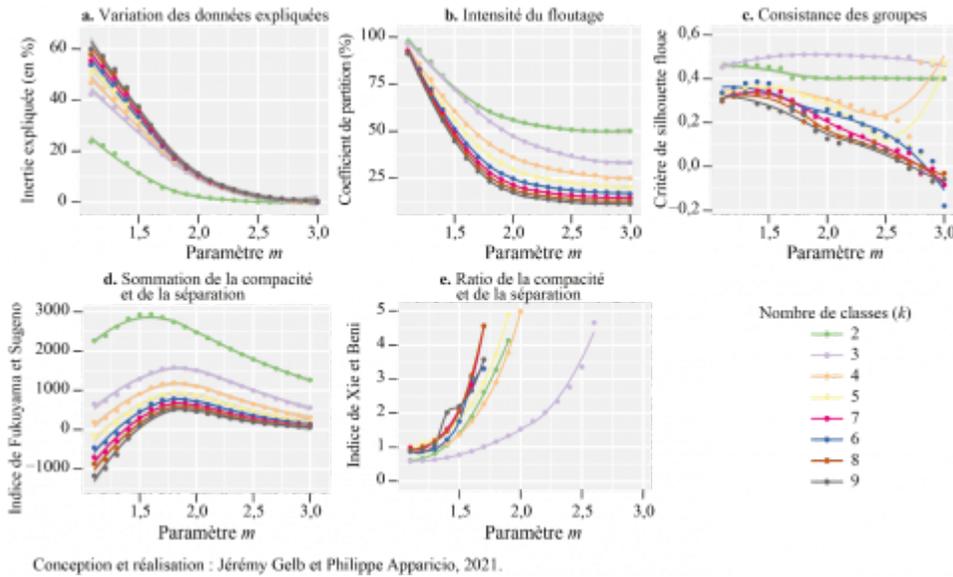
Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

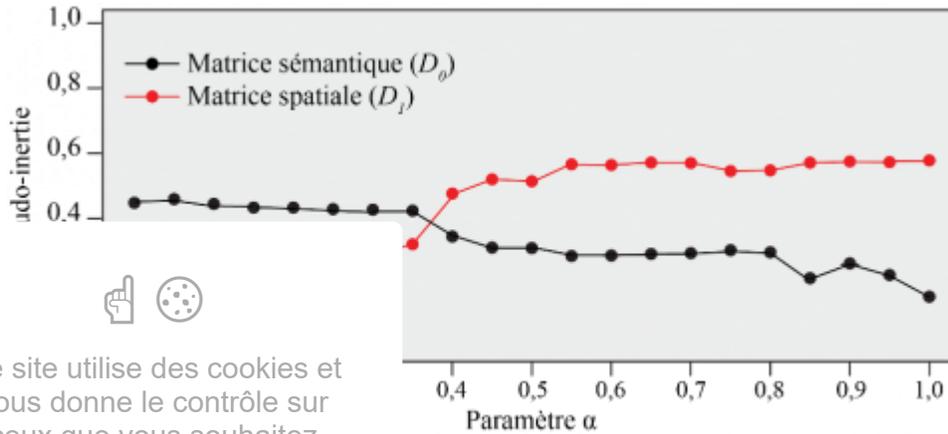
Politique de confidentialité



Sélection des paramètres alpha pour les méthodes spatiales

32 Pour la méthode *ClustGeo* (avec $k=4$), nous évaluons l'impact du paramètre α pour des valeurs de 0 à 1, avec un saut de 0,05. À la lecture de la figure 5, on observe qu'avec une pondération spatiale fixée à $\alpha=0,35$, l'inertie de la matrice spatiale expliquée augmente de plus de 25 % pour une perte d'inertie expliquée dans la matrice de distance sémantique de seulement 2 %. Avec $\alpha=0,4$, on observe une chute importante de l'inertie sémantique expliquée. Nous retenons donc une valeur de $\alpha=0,35$ pour la méthode *ClustGeo*.

Figure 5 : Impact du paramètre α (pondération spatiale) dans la méthode *ClustGeo*



et Philippe Apparicio, 2021.

,5), nous avons calculé toutes les valeurs du paramètre α 0,5. Lorsque α augmente, la part de l'inertie sémantique en revanche, le critère de silhouette augmente avec α , 1 (figure 6.c). Par contre, l'indice de silhouette et le une situation optimale quand α varie entre 0,45 et 0,55 yama et Sugeno et de Xie et Beni (figures 6.d et 6.e) lpha, mais cette augmentation est minimale lorsqu'on la s graphiques des figures 4.d et 4.e.

fication finale, nous avons fixé $\alpha=0,7$, soit une valeur r équilibre entre les différents indicateurs. Nous avons

Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Tout accepter

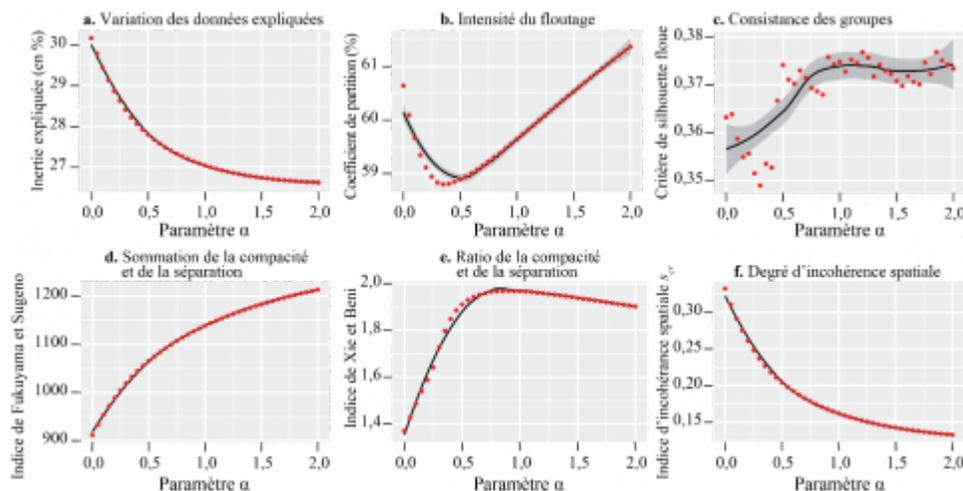
Tout refuser

Personnaliser

[Politique de confidentialité](#)

aussi comparé les résultats obtenus si la valeur médiane des voisins était retenue au détriment de la valeur moyenne. L'utilisation de la médiane a été écartée, car si les solutions obtenues améliorent très légèrement la cohérence spatiale, c'est au détriment d'une forte perte de l'inertie sémantique expliquée. De même, nous avons comparé les résultats obtenus en utilisant plusieurs matrices de pondération spatiale, à savoir des matrices de contiguïté selon le partage d'un nœud ou d'une frontière (*Queen* et *Rook*) et des matrices de plus proches voisins allant de 5 à 10. Les différences étant anecdotiques, nous retenons la matrice classique, c'est-à-dire la plus souvent utilisée, soit la matrice de contiguïté selon le partage d'un nœud standardisée en ligne.

Figure 6 : Impact du paramètre α sur le SFCM



Conception et réalisation : Jérémy Gelb et Philippe Apparicio, 2021.

Comparaison des résultats des quatre classifications

35 Afin de comparer les résultats des quatre classifications, nous avons calculé plusieurs mesures de qualité d'une partition reportées au tableau 1. L'analyse des valeurs de ces indicateurs permet d'avancer trois constats intéressants. Premièrement, les méthodes de classification strictes (CAH et *ClustGeo*) parviennent à expliquer une plus grande partie de l'inertie sémantique que les méthodes de classification floues. Par exemple, le pourcentage



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

pour la CAH contre 0,302 pour le FCM. Deuxièmement, les méthodes floues obtiennent de meilleurs indices de silhouette, de consistance des groupes et de compacité. Troisièmement, les méthodes spatiales obtiennent des indices de cohérence spatiale plus faibles que leur contrepartie a-spatiale, mais un constat qui n'est pas le cas du SFCM.

Les quatre classifications

	A-spatiale		Spatiale	
	Ha	FCMb	<i>ClustGeo</i> c	SFCMd
Inertie sémantique expliquée (%)	0,8	0,363	0,261	0,370
Inertie spatiale expliquée (%)	0,0	0,606	1,000	0,592
Indice de silhouette	0,47	0,302	0,422	0,274
Indice de cohérence spatiale s_r	2,45	911,55	694,45	1100,08

^a Classification ascendante hiérarchique classique. ^b Classification *c-means* classique. ^c Classification ascendante hiérarchique spatiale. ^d Classification *c-means* spatiale.

Conception et réalisation : Gelb et Apparicio, 2020.

- 36 Pour que ce diagnostic soit complet, il convient d'évaluer le rôle joué par la dimension spatiale dans les quatre classifications (tableau 2). Pour ce faire, nous avons calculé le I de Moran sur les colonnes de la matrice d'appartenance, ainsi que les indices d'incohérence spatiale présentés précédemment (équation 3). Il est intéressant de constater que les méthodes introduisant l'espace parviennent à augmenter significativement l'autocorrélation spatiale dans les résultats et d'améliorer nettement la cohérence spatiale. Le SFCM produit d'ailleurs les résultats les plus cohérents spatialement, et ce, pour les deux indicateurs S_{cr} et S_{mcr} (la moyenne ainsi que le 5^e et 95^e percentiles des 999 permutations sont reportés).

Tableau 2 : Indicateurs de cohérence spatiale des quatre classifications

Classification	I de Moran sur l'appartenance aux groupes ^e				Indice d'incohérence spatiale (S_{cr}) ^f	Indice d'incohérence sémantico-spatiale (S_{mcr}) ^f
	A	B	C	D		
CAHa	0,489	0,699	0,647	0,231	0,415 (0,404-0,429)	0,304 (0,294-0,315)
FCMb	0,548	0,835	0,724	0,388	0,334 (0,325-0,343)	0,229 (0,222-0,236)
ClustGeoc	0,498	0,779	0,809	0,784	0,267 (0,261-0,275)	0,173 (0,168-0,179)
SFCMd	0,749	0,908	0,860	0,645	0,183 (0,178-0,189)	0,129 (0,125-0,134)

^a Classification ascendante hiérarchique classique. ^b Classification *c-means* classique. ^c Classification ascendante hiérarchique spatiale. ^d Classification *c-means* spatiale. ^e I de Moran calculé avec une matrice de contiguïté selon le partage d'un nœud ; plus sa valeur est élevée, plus la classe est cohérente spatialement. ^f À titre de rappel, une valeur proche de 0 traduit une forte cohérence spatiale de la classification, tandis qu'une valeur proche de 1 signale une forte incohérence spatiale.

Conception et réalisation : Gelb et Apparicio, 2020.

- 37 Il apparaît donc clairement que dans le cas qui nous intéresse, les méthodes de classification strictes obtiennent de meilleurs résultats d'un point de vue de l'inertie expliquée. Cependant, l'indicateur de silhouette tend à indiquer que les groupes sont plus homogènes pour les méthodes floues. L'ajout de la dimension spatiale réduit nécessairement la qualité de la classification, mais cette réduction est minime et se fait au point de vue de leur distribution spatiale conséquent.



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

er les quatre groupes obtenus en fonction des quatre point de vue de leur distribution spatiale au sein de premier groupe (A) regroupe des IRIS localisés dans les de Lyon correspondant à des banlieues. Le deuxième s arrondissements de la Ville de Lyon situés au centre de upe (C) fait moins consensus entre les quatre méthodes. secteurs périphériques, alors que le FCM classique et les it essentiellement aux secteurs périphériques à l'ouest. me groupe (D) se situent davantage dans les périphéries t les classifications spatiales. Pour la CAH classique, le teurs entourant directement le centre.

ité des classes – valeurs moyennes des observations) algorithmes pour les dix variables (tableau 3) – nous à D.

pour les quatre algorithmes. Il est caractérisé par des levés (entre 56 et 57 dB) ainsi que par des niveaux de

concentration de NO₂ (26 ug/m³) et de PM_{2,5} relativement faibles (16 ug/m³) comparativement aux autres groupes. Il présente en revanche des valeurs moyennes très faibles pour la médiane du niveau de vie (autour de 16 000 euros) et, inversement, très fortes pour les pourcentages d'immigrants (autour de 26 %), de jeunes de moins de 15 ans (autour de 25 %), de personnes de 15 à 64 ans faiblement scolarisées (ayant obtenu au plus le brevet des collèges, 39 %) et le taux de chômage (autour de 24 %). En croisant ces résultats avec les cartes précédentes, il apparaît clairement que ce groupe décrit les banlieues socialement défavorisées de Lyon (communes de Vaulx-en-Velin, Saint-Fons et Vénissieux, entre autres). Il est intéressant de noter que la défavorisation sociale ne semble pas se coupler avec une surexposition aux pollutions atmosphériques et sonores.

41 À nouveau, le groupe B est très similaire pour les quatre algorithmes. Il est caractérisé par des niveaux de pollutions atmosphériques plus importants que les autres groupes (NO₂ : 35 ug/m³; PM_{2,5} : 18,7 ug/m³) et par des niveaux de bruit moyennement élevés (autour de 56,5 dB). Il s'agit également d'un groupe ayant une médiane du niveau de vie élevée (environ 22 500 euros) ainsi qu'un taux de chômage, un pourcentage d'immigrants et un pourcentage de personnes ayant obtenu au plus le brevet des collèges relativement faibles (13,5 %, 12 % et 15,5 %). Si l'on se reporte aux cartes, ce groupe décrit les quartiers centraux de Lyon, marqués par une forte exposition aux différentes pollutions, mais dont la population est favorisée sur le plan socio-économique.

Tableau 3 : Comparaison des centres de gravité des classes pour les quatre classifications

Groupe	Méthode	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	N	
A	CAHa	56,5	25,8	16,2	15,0	24,9	13,9	28,2	25,7	41,4	15 763	97	
	FCMb	57,2	27,6	16,6	16,0	24,2	14,2	26,5	24,3	38,3	16 259	92	
	ClustGeoc	57,2	27,0	16,5	16,0	25,0	13,6	28,0	25,9	41,2	15 862	92	
	SFCMd	56,7	26,6	16,3	15,8	23,7	14,7	24,7	22,5	37,1	16 892	97	
B	CAHa	56,3	34,9	18,8	15,5	14,8	14,0	11,9	13,1	14,7	23 196	185	
	FCMb	56,7	34,8	18,6	15,1	15,0	14,1	12,4	13,5	16,3	22 940	171	
	ClustGeoc	57,0	35,3	18,9	14,4	15,1	13,9	13,1	13,7	16,6	22 371	200	
	SFCMd	56,2	34,7	18,6	15,1	14,9	14,1	12,1	13,5	15,8	23 289	132	
					4,8	24,2	18,5	20,1	8,8	10,3	21,2	24 612	180
					4,9	27,3	18,4	19,4	8,9	10,8	20,1	25 101	117
					5,0	29,3	18,0	20,6	8,6	10,6	19,0	25 464	125
					4,9	26,1	18,7	19,3	9,4	11,1	20,9	24 638	147
					3,0	18,4	19,7	15,2	18,3	17,3	30,2	17 785	44
					3,6	17,3	18,6	17,2	13,6	13,6	24,6	21 399	126
				4,9	16,5	20,0	18,0	12,0	12,0	27,0	21 532	89	
				7,2	17,2	18,2	16,4	14,8	14,6	24,4	20 930	130	



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

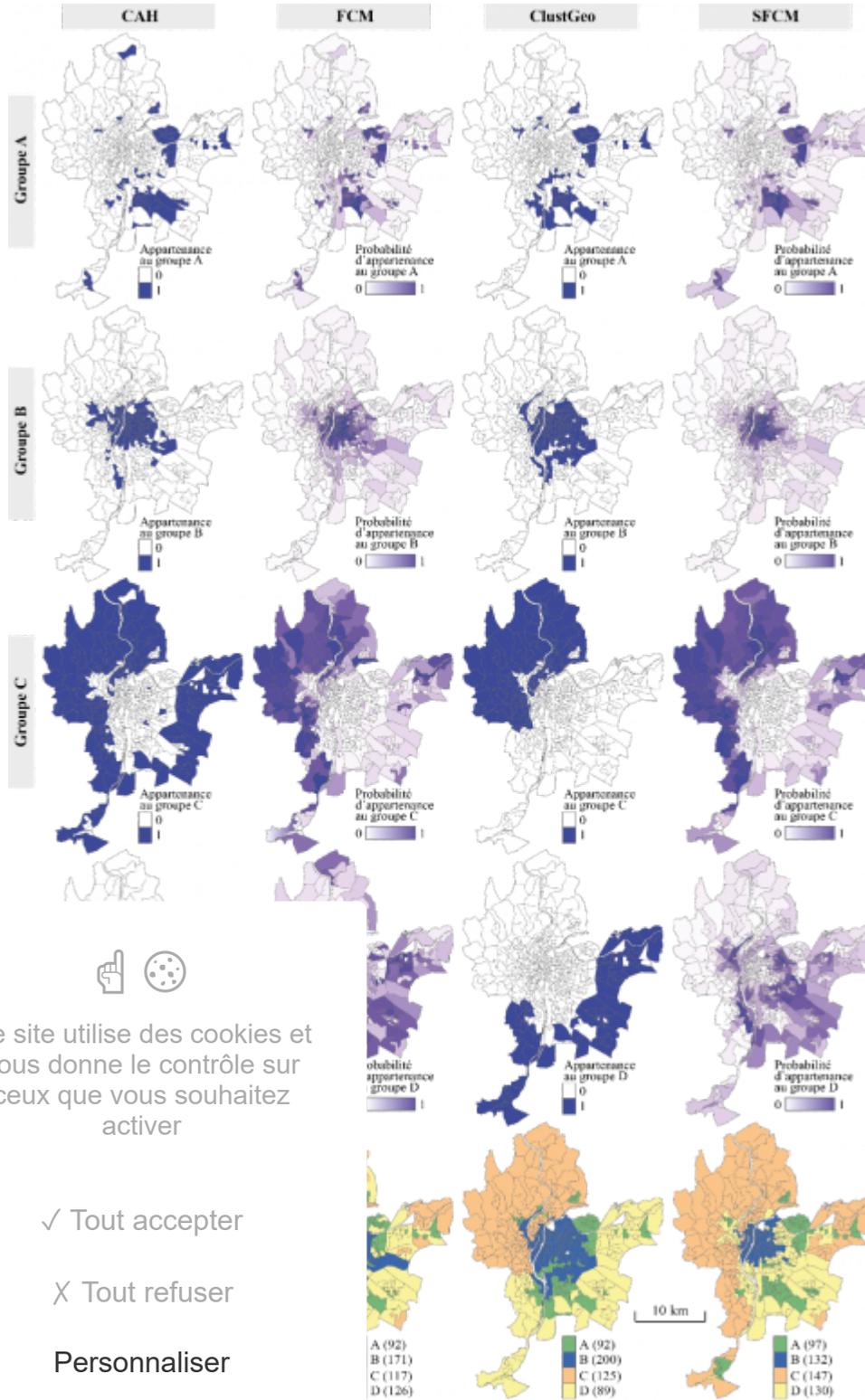
Politique de confidentialité

*L*_{den}). **V2** : NO₂. **V3** : PM_{2,5}. **V4** : Végétation (%). **V5** : Personnes de 15 ans et plus (%). **V7** : Immigrants (%). **V8** : Taux de chômage. **V9** : Personnes ayant obtenu au plus le brevet des collèges (%). **V10** : Médiane du niveau de vie (euros).

io, 2020.

42 Il apparaît donc clairement que dans le cas qui nous intéresse, les méthodes de classification strictes obtiennent de meilleurs résultats d'un point de vue de l'inertie expliquée. Cependant, l'indicateur de silhouette tend à indiquer que les groupes sont plus homogènes pour les méthodes floues. L'ajout de la dimension spatiale réduit nécessairement la qualité de la classification, mais cette réduction est minime et se fait au profit d'un gain de cohérence spatiale conséquent.

Figure 7 : Cartographie des résultats des quatre méthodes de classification



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

aussi très semblable selon les quatre algorithmes, il est médian ainsi qu'un indice de canopée très élevé

comparativement aux autres groupes (respectivement 24 500 euros et 26 %). Il est aussi associé au pourcentage de population de 65 ans et plus le plus élevé (environ 20 %), au pourcentage d'immigrants le plus faible (environ 9 %), au taux de chômage le plus bas (environ 10,5 %) et à un faible pourcentage de personnes ayant obtenu au plus le brevet des collèges (20 %). Ce groupe combine également les niveaux de concentration de NO₂, de PM_{2,5} et de bruit les plus faibles de tous les groupes. Il peut ainsi correspondre aux communes riches et vieillissantes sises dans un environnement de qualité à la périphérie ouest du Grand Lyon.

- 44 Finalement, le quatrième groupe est certainement plus difficile à qualifier, car les algorithmes ont abouti à des solutions légèrement dissemblables. La CAH classique décrit un groupe avec des niveaux de bruit et de NO₂ relativement élevés (61,4 dB et 35,5 ug/m³), un pourcentage d'immigrants et un taux de chômage relativement élevés eux aussi (18,3 % et 17,3 %) ainsi qu'une médiane du niveau de vie plutôt basse (17 700 euros). À l'inverse, les trois autres algorithmes semblent refléter un groupe avec des distinctions bien moins tranchées, se distinguant seulement par sa tendance géographique vers l'est.

Résultats finaux du SFCM et analyse des observations incertaines

- 45 Si l'on conserve uniquement les résultats du SFCM, la figure 8.b permet de résumer les différences entre les groupes. Ce graphique en radar est construit en calculant pour chaque groupe la moyenne se rapportant à chacune des variables originales (pondérée par les probabilités d'appartenance des observations) ; soit les valeurs des centres de gravité pour le SFCM présentées au tableau 3. Pour chaque variable, le groupe ayant obtenu la plus faible valeur est situé à l'origine de l'axe correspondant (au centre du graphique), alors que le groupe ayant obtenu la moyenne la plus forte se trouve à l'extrémité de l'axe. Les autres groupes sont ensuite localisés sur cette échelle linéaire entre la plus forte et la plus faible valeur. Cette figure permet donc d'obtenir un aperçu des différences relatives entre les groupes.

Figure 8 : Résultats finaux de la classification SFCM



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Tout accepter

Tout refuser

Personnaliser

Politique de confidentialité

- 46 L'un des intérêts non négligeables d'une classification en mode flou est la possibilité d'identifier des observations pour lesquelles l'attribution à un groupe est très incertaine. Nous définissons ici ces observations comme étant celles pour lesquelles la probabilité la plus haute d'appartenance à un groupe est inférieure à 40 %. En attribuant ensuite chaque observation à son groupe le plus vraisemblable, la cartographie des classes est réalisée à la figure 8.a. Ainsi, 20 observations ont une affectation incertaine (en noir) ; elles sont principalement localisées à l'est de la zone d'étude et majoritairement au sein du groupe D. On peut donc qualifier ce groupe de territoires interstitiels et hétérogènes séparant le centre, les banlieues et la périphérie riche.
- 47 Il est intéressant de noter que la classification proposée corrobore d'autres travaux décrivant les logiques sociale, urbaine et environnementale de la région du Grand Lyon. Le triptyque centre aisé, banlieue défavorisée et périphérie ouest aisée avait déjà été bien décrit par Ravalet (2007), dans une analyse des caractéristiques socio-économiques et des habitudes de déplacement dans la région. Dans leur monographie de la région du Grand Lyon, Authier *et al.* (2010) avaient également signalé que les populations étrangères et plus socialement défavorisées se concentrent essentiellement dans les grands ensembles d'habitat social édifiés dans les années 1960-1970, au sein des communes de l'Est lyonnais (Vénissieux, Saint-Fons et Vaulx-en-Velin). La commune même de Lyon regroupe une population très diplômée, composée majoritairement de cadres et de personnes exerçant des professions intellectuelles supérieures. L'urbanisation et l'industrialisation ont longtemps été ralenties par le relief à l'ouest de la commune de Lyon, ce qui explique notamment la forte valeur de l'indice de canopée du groupe C et de plus faibles niveaux de pollutions atmosphériques et sonores. Au fil du XXe siècle, les territoires à l'ouest ont été peu à peu développés par l'habitat pavillonnaire, occupé par des populations socialement favorisées. Tous ces points se retrouvent dans la description de la région par Charmes et Rousseau (2014). Ceux-ci soulignent également la grande diversité des couronnes périurbaines de Lyon, ce qui explique le mitage et les observations incertaines que nous avons constatés dans notre classification.

Discussion

Apport de la méthode SFCM



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

s donc comparé les résultats produits par quatre trictes), FCM (a-spatiale et floue), *ClustGeo* (spatiale et ie) ; sur un jeu de données combinant des indicateurs rentaux de l'agglomération de Lyon (France) concernant constaté que pour les deux premiers groupes identifiés, es résultats similaires. En revanche, des différences plus e plan spatial que sémantique pour les deux derniers : clairement que les méthodes spatiales parviennent à e spatiale des résultats obtenus, sans pour autant réduire pliquée, facilitant ainsi l'interprétation des résultats. De es méthodes floues donnaient des indices de silhouette : plus homogènes, et ce, alors que les méthodes strictes le l'inertie sémantique du jeu de données. Il est possible une forme de résistance des méthodes floues à la nment" les caractéristiques des classes sans avoir à vations situées en limite de celles-ci. Ainsi, la méthode ntages de l'approche floue sur l'approche stricte et les

avantages de l'approche spatiale sur l'approche a-spatiale. Cependant, comparativement à un FCM classique, deux nouveaux paramètres doivent être fixés : α , représentant le poids accordé à la dimension spatiale, et W , la matrice de pondération spatiale contrôlant l'effet de floutage spatial. Dans le jeu de données étudié, nous avons pu constater que le rôle de W est mineur et qu'il est relativement facile de sélectionner une valeur optimale de α une fois le nombre de classes (k) et la valeur du paramètre de logique floue (m) choisis. Avec un jeu de données de taille modeste ou suffisamment de puissance de calcul, il serait aussi possible de trouver la meilleure combinaison entre α , m , k , et W en testant toutes les combinaisons possibles. Pour répondre à la question posée en introduction, la méthode SFCM permet de bénéficier à la fois des avantages des méthodes de classification floues (analyse des observations incertaines et classes plus homogènes) et de l'apport de la dimension spatiale (réduction de l'incohérence spatiale). Sans surprise, dans l'exemple proposé, les différences entre les résultats des quatre méthodes de classification ne sont pas diamétralement opposées et n'ont pas vocation à l'être (bien que pour les classes les moins homogènes, la figure 7 ait montré des écarts notables). Autrement dit, l'ajout de l'espace dans l'exercice de classification ne devrait pas radicalement changer les résultats et prendre le pas sur la dimension sémantique des données. L'objectif est de lisser spatialement les classes produites, dans une logique de réduction du bruit. La méthode SFCM semble proposer le meilleur compromis entre qualité sémantique de classification, cohérence spatiale et interprétabilité des résultats. En effet, comparativement à la méthode *ClustGeo*, la méthode SFCM présente à la fois de meilleures cohérences spatiale et sémantico-spatiale (voir les résultats présentés au tableau 3) et un meilleur critère de silhouette (homogénéité des classes). Finalement, la cartographie des probabilités d'appartenance des classes permet d'identifier les observations incertaines et affine l'interprétation finale des résultats.

Description succincte du paquet R *geocmeans*

49 L'ensemble des analyses présentées dans cet article peuvent être reproduites à l'aide du paquet R *geocmeans*. En effet, le jeu de données sur la région du Grand Lyon est fourni avec le paquet et une vignette présente la syntaxe R pour les quatre méthodes de classification (CAH, *ClustGeo*, FCM et SFCM). Le paquet peut actuellement être téléchargé en lançant les deux lignes de code suivantes dans R :

```
tools::install_github("JeremyGelb/geocmeans").
```



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

et de réaliser la classification SFCM. La fonction `parer` regroupe les résultats de plusieurs classifications en faisant varier W ; l'utilisateur peut ainsi choisir les valeurs optimales du degré de flou (m) et le nombre de classes (k), mais aussi la matrice de pondération spatiale (W). D'autres fonctions – `calcqualityIndexes` et `qualityIndexes` – permettent de calculer les indicateurs de qualité de la classification, tels que le degré de cohérence spatiale des classes. Enfin, les fonctions `summarizeClusters` et `mapClusters` permettent de visualiser les résultats obtenus.

Cet article évalue la pertinence et l'efficacité de la méthode *spatial c-means* appliquée à des données socio-environnementales. Nous avons pu constater que l'ajout de la dimension spatiale permet de combiner à la fois les avantages des méthodes de classification

floues et spatiales. Ces avantages en font une méthode particulièrement pertinente en géographie, où les phénomènes sont souvent caractérisés par une forte autocorrélation spatiale positive et où une approche stricte tend à masquer la complexité des phénomènes étudiés. Cette méthode présente trois avantages majeurs comparativement aux méthodes strictes a-spatiales (CAH ou *k-means*) : réduire les problèmes de mitage et améliorer à la fois la cohérence sémantico-spatiale et l'homogénéité des classes.

52 Dans les dernières décennies, nous avons assisté au développement de nouvelles méthodes incluant explicitement l'espace dans les méthodes inférentielles (entre autres, les modèles autorégressifs et géographiquement pondérés) et exploratoires (analyses factorielles spatiales). Ces méthodes sont aujourd'hui largement mobilisées par les géographes. Nous espérons que cet exercice contribuera humblement à démocratiser l'intégration de la dimension spatiale dans les méthodes de classification, notamment avec le recours à la méthode SFCM.

Bibliographie

Adhikari S.K., Sing J.K., Basu D.K., Nasipuri M., 2015, "Conditional spatial fuzzy C-means clustering algorithm for segmentation of MRI images", *Applied Soft Computing*, Vol.34, 758-769.

DOI : 10.1016/j.asoc.2015.05.038

Aldstadt J., Getis A., 2006, "Using AMOEBA to create a spatial weights matrix and identify spatial clusters", *Geographical Analysis*, Vol.38, No.4, 327-343.

DOI : 10.1111/j.1538-4632.2006.00689.x

Apparicio P., 2002, *Apport des systèmes d'information géographique à l'étude de l'insertion des HLM dans les quartiers montréalais*, Thèse de doctorat en géographie, Université du Maine.

Apparicio P., Cloutier M.S., Shearmur R., 2007, "The case of Montreal's missing food deserts: evaluation of accessibility to food supermarkets", *International Journal of Health Geographics*, Vol.6, No.1, 4.

Apparicio P., Riva M., Séguin A.M., 2015, "A comparison of two methods for classifying trajectories: a case study on neighborhood poverty at the intra-metropolitan level in Montreal", *Cybergeo: European Journal of Geography*.

DOI : 10.4000/cybergeo.27035

Apparicio P., Séguin A.M., 2006, "L'insertion des HLM montréalaises dans le milieu social environnant", *L'Espace géographique*, Vol.35, No.1, 63-85.

DOI : 10.3917/eg.351.85

Assunção R.M., Neves M.C., Câmara G., da Costa Freitas C., 2006, "Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees", *International Science*, Vol.20, No.7, 797-811.



"Données ouvertes de qualité de l'air", *Atmo Auvergne-Rhône-Alpes*, [donnees-ouvertes-de-qualite-de-lair-o](http://donnees-ouvertes-de-qualite-de-lair-atmo.fr/donnees-ouvertes-de-qualite-de-lair-o)

Vogel M., 2010, *Sociologie de Lyon*, La Découverte.

Overlapping clustering: A review", 233-237 in: *SAI Computing*

analyse typologique en milieu urbain : les aires homogènes géographiques", *Revue géographique de Lyon*, No.4, 267-275.

la sociorésidentielle d'une ville latino-américaine : Puebla au Mexique", *Revue géographique de Lyon*, Vol.50, No.139, 45-63.

données. Tome 1. La taxinomie. Tome 2. L'analyse des

fast and robust fuzzy c-means clustering algorithms incorporating spatial information", *Pattern Recognition*, Vol.40, No.3, 825-838.

"A fuzzy extension of the silhouette width criterion for cluster validation", *Pattern Recognition*, Vol.157, No.21, 2858-2875.

Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

Charmes E., Rousseau M., 2014, "Le pavillon et l'immeuble : géopolitique de la densification dans la région métropolitaine de Lyon", *Géographie, économie, société*, Vol.16, No.2, 155-181.

DOI : 10.3166/ges.16.297-314

Chavent M., Kuentz-Simonet V., Labenne A., Saracco J., 2018, "ClustGeo: an R package for hierarchical clustering with spatial constraints", *Computational Statistics*, Vol.33, No.4, 1799-1822.

DOI : 10.1007/s00180-018-0791-1

Chevalier P., Dedeire M., Ghiotti S., Hirczak M., Razafimahefa L., 2010, "L'espace rural euro-méditerranéen : approches méthodologique et typologique", *Cahiers de géographie du Québec*, Vol.54, No.152, 291-312.

Chuang K.S., Tzeng H.L., Chen S., Wu J., Chen T.J., 2006, "Fuzzy c-means clustering with spatial information for image segmentation", *Computerized Medical Imaging and Graphics*, Vol.30, No.1, 9-15.

DOI : 10.1016/j.compmedimag.2005.10.001

Cloutier M.S., Apparicio P., Charbonneau J., 2011, "La géographie du don de sang au Québec : quelles variations régionales ?", *Cahiers de géographie du Québec*, Vol.55, No.156, 471-490.

Diday E., 1971, "Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques", *Revue de statistique appliquée*, Vol.19, No.2, 19-33.

Dunn J.C., 1973, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters", *Journal of Cybernetics*, Vol.3, No.3, 32-57.

DOI : 10.1080/01969727308546046

Everitt B.S., Landau S., Leese M., Stahl D., 2011, *Cluster analysis*, John Wiley & Sons.

Fischer M.M., 1980, "Regional taxonomy: a comparison of some hierarchic and non-hierarchic strategies", *Regional Science and Urban Economics*, Vol.10, No.4, 503-537.

Fisher P.F., Pathirana S., 1990, "The evaluation of fuzzy membership of land cover classes in the suburban zone", *Remote Sensing of Environment*, Vol.34, No.2, 121-132.

DOI : 10.1016/0034-4257(90)90103-S

Fleuret S., Apparicio P., 2011, "Essai de typologie des centres de santé et de services sociaux au Québec", *The Canadian Geographer/Le Géographe canadien*, Vol.55, No.2, 143-157.

DOI : 10.1111/j.1541-0064.2010.00318.x

Fukuyama Y., Sugeno M., 1989, "A new method of choosing the number of clusters for the fuzzy c-mean method", 247-250 in: *Proc. 5th Fuzzy Syst. Symp.*

Getis A., 2009, "Spatial weights matrices", *Geographical Analysis*, Vol.41, No.4, 404-410.

Guo Y., Liu K., Wu Q., Hong Q., Zhang H., 2015, "A new spatial fuzzy c-means for spatial clustering", *WSEAS Transactions on Computer*, Vol.14, 369-381.

Hafezi M.H., Liu L., Millward H., 2017, "Identification of Representative Patterns of Time Use Activity Through Fuzzy C-Means Clustering", *Transportation Research Record*, Vol.2668, No.1, 38-50.

DOI : 10.3141/2668-05



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

neating urban housing submarkets with fuzzy clustering", *ing and Design*, Vol.36, No.5, 865-882.

L : <https://www.insee.fr/fr/metadonnees/definition/c1523>

ms for clustering data. Prentice-Hall, Inc.

gata R., 2019, "Clustering with spatial constraints: The case of a", 012068 in: *Journal of Physics: Conference Series*, IOP

2068

ustering by means of medoids. *Statistical Data Analysis based cal data analysis based on the L1-norm and related methods*,

5, *Statistique exploratoire multidimensionnelle*, Paris, Dunod.

for classification and analysis of multivariate observations", *ifth Berkeley symposium on mathematical statistics and*

ouvertes de Lyon", *Données ouvertes de Lyon*. URL :

t différenciation sociospatiale à Dakar (Sénégal)", *Cahiers de* 5, 47-69.

Oliver M., Webster R., 1989, "A geostatistical basis for spatial weighting in multivariate classification", *Mathematical Geology*, Vol.21, No.1, 15-35.

DOI : 10.1007/BF00897238

Openshaw S., 1977, "A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling", *Transactions of the Institute of British Geographers*, Vol.2, No.4, 459-472.

DOI : 10.2307/622300

Openshaw S., Rao L., 1995, "Algorithms for reengineering 1991 Census geography", *Environment and Planning A*, Vol.27, No.3, 425-446.

Pumain D., 1976, "La composition socio-professionnelle des villes françaises : essai de typologie par analyse des correspondances et classification automatique", *L'Espace géographique*, No.4, 227-238.

DOI : 10.3406/spgeo.1976.1663

Ravalet E., 2007, "Modes de vie et modes de ville, activités et déplacements quotidiens à Montréal et Lyon", *Environnement urbain/Urban Environment*, Vol.1, 1-20.

Richer C., Palmier P., 2012, "Mesurer l'accessibilité territoriale par les transports collectifs : proposition méthodologique appliquée aux pôles d'excellence de Lille Métropole", *Cahiers de géographie du Québec*, Vol.56, No.158, 427-461.

Riva M., Apparicio P., Gauvin L., Brodeur J.M., 2008, "Establishing the soundness of administrative spatial units for operationalising the active living potential of residential environments: an exemplar for designing optimal zones", *International Journal of Health Geographics*, Vol.7, No.1, 43.

DOI : 10.1186/1476-072X-7-43

Saad T., 1982, "Typologie d'un espace urbain en pays en voie de développement : l'exemple de Nabeul (Tunisie)", *L'Espace géographique*, No.2, 134-142.

DOI : 10.3406/spgeo.1982.3729

Sahraoui Y., Lunardi N., Antoni J.P., 2014, "Essai de typologie des espaces résidentiels à partir d'indicateurs désagrégés à l'échelle du bâtiment. Application à Besançon et à Metz", *Cybergeo: European Journal of Geography*.

Sanders L., 1989, *L'analyse statistique des données en géographie*, Montpellier, GIP Reclus.

Swihart R.K., Caudell J.N., Brooke J.M., Ma Z., 2020, "A Flexible Model-based Approach to Delineate Wildlife Management Units", *Wildlife Society Bulletin*, Vol.44, No.1, 77-85.

DOI : 10.1002/wsb.1037

Tabard N., 1993, "Des quartiers pauvres aux banlieues aisées : une représentation sociale du territoire", *Economie et Statistique*, Vol.270, 5-22.

DOI : 10.3406/estat.1993.5822

Theodoridis S., Pikrakis A., Koutroumbas K., Cavouras D., 2010, *Introduction to pattern recognition: a matlab approach*, Academic Press.

Ward J.H., 1963, "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association*, Vol.58, No.301, 236-244.

Wang M., 2008, "A hybrid clustering method for identifying high-density clusters", *Journal of the American Statistical Association*, Vol.103, No.380, 841-847.



measure for fuzzy clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.30, No.8, 841-847.

Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

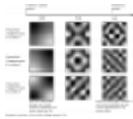
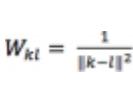
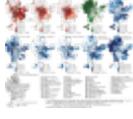
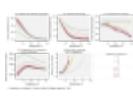
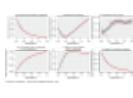
Tout accepter

Tout refuser

Personnaliser

Politique de confidentialité

hèse des principales méthodes de classification
openedition.org/cybergegeo/docannexe/image/36414/img-1.png
5k
openedition.org/cybergegeo/docannexe/image/36414/img-2.png
ik
openedition.org/cybergegeo/docannexe/image/36414/img-3.png
ik
openedition.org/cybergegeo/docannexe/image/36414/img-4.png
k

	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-5.png
Fichier		image/png, 4,5k
	Titre	Figure 2 : Illustration graphique des valeurs prises par l'indicateur d'incohérence spatiale
	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-6.png
Fichier		image/png, 120k
$W_{klt} = \frac{1}{\ k-l\ ^2}$	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-7.png
Fichier		image/png, 2,1k
	Titre	Figure 3 : Cartographie des différents indicateurs
	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-8.png
Fichier		image/png, 862k
	Titre	Figure 4 : Évaluation de la qualité des classifications FCM de 2 à 9 classes
	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-9.png
Fichier		image/png, 292k
	Titre	Figure 5 : Impact du paramètre α (pondération spatiale) dans la méthode <i>ClustGeo</i>
	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-10.png
Fichier		image/png, 58k
	Titre	Figure 6 : Impact du paramètre α sur le SFCM
	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-11.png
Fichier		image/png, 210k
	Titre	Figure 7 : Cartographie des résultats des quatre méthodes de classification
	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-12.png
Fichier		image/png, 1,2M
	Titre	Figure 8 : Résultats finaux de la classification SFCM
	URL	http://journals.openedition.org/cybergegeo/docannexe/image/36414/img-13.png
Fichier		image/png, 479k



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

Apport de la classification floue c-means spatiale en « socio-résidentielle et environnementale à Lyon », *Cybergegeo* : [Géographie, Cartographie, Imagerie, SIG, document 972, mis en ligne le 19 novembre 2021]. URL : <http://journals.openedition.org/cybergegeo/docannexe/image/36414> ; DOI : <https://doi.org/10.4000/cybergegeo.36414>

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité

Centre de Recherche sur les Mobilités (LAEQ), Institut national de la recherche scientifique, 385 rue de la République, 69622 Villeurbanne cedex, France

position individuelle aux pollutions atmosphériques et acoustiques [Texte intégral]

Paru dans *Cybergeo : European Journal of Geography*, GeOpenMod - Modèles et logiciels

Philippe Apparicio

Professeur titulaire

Laboratoire d'équité environnementale (LAEQ), Institut national de la recherche scientifique, 385 rue

Sherbrooke Est, Montréal (QC), Canada, H2X 1E3

Philippe.apparicio@ucs.inrs.ca

Articles du même auteur

Un atlas-web pour comparer l'exposition individuelle aux pollutions atmosphérique et sonore selon le mode de transport [Texte intégral]

Article 903

Paru dans *Cybergeo : European Journal of Geography*, GeOpenMod - Modèles et logiciels

A comparison of two methods for classifying trajectories: a case study on neighborhood poverty at the intra-metropolitan level in Montreal [Texte intégral]

Article 727

Paru dans *Cybergeo : European Journal of Geography*, Espace, Société, Territoire

La répartition de la population âgée dans huit métropoles canadiennes de 1981 à 2006 : un groupe de moins en moins ségrégué [Texte intégral]

Article 639

Paru dans *Cybergeo : European Journal of Geography*, Espace, Société, Territoire

Segregation Analyzer: a C#.Net application for calculating residential segregation indices [Texte intégral]

Article 414

Paru dans *Cybergeo : European Journal of Geography*, Systèmes, Modélisation, Géostatistiques

Déploiement d'atlas interactifs sur Internet : nouvelles avenues avec le Scalable Vector Graphics, le C# et l'ASP.Net [Texte intégral]

Article 340

Paru dans *Cybergeo : European Journal of Geography*, Dossiers

Les indices de ségrégation résidentielle : un outil intégré dans un système d'information géographique [Texte intégral]

Article 134

Paru dans *Cybergeo : European Journal of Geography*, Espace, Société, Territoire

Droits d'auteur



La revue *Cybergeo* est mise à disposition selon les termes de la Licence Creative Commons Attribution 4.0 International



Ce site utilise des cookies et vous donne le contrôle sur ceux que vous souhaitez activer

✓ Tout accepter

X Tout refuser

Personnaliser

Politique de confidentialité