Développement d'un algorithme de détection de communautés spatiales

Serge Lhomme

Développement d'un algorithme de détection de communautés spatiales

Lhomme Serge[1]

1. Lab'Urba (EA 3482), Université Paris-Est Créteil
61 avenue du Général de Gaulle, Créteil, 94000, France
serge.lhomme@u-pec.fr

RESUME. Les méthodes de détection de communautés sont des outils potentiellement puissants pour analyser des réseaux spatiaux puisqu'elles permettent d'identifier des lieux qui interagissent de manière préférentielle. Néanmoins, ces méthodes souffrent d'un biais important : elles ne tiennent pas compte des distances qui séparent les lieux. Une fois cartographiées, les communautés détectées apparaissent généralement très contraintes spatialement et contribuent à mettre simplement en évidence l'importance de la distance dans l'intensité des relations étudiées. Cette contrainte étant déjà bien connue, les résultats obtenus peuvent alors apparaitre triviaux. Pour faire face à ce biais et faire apparaitre des phénomènes plus intéressants, il est proposé d'adapter les algorithmes de détection de communautés à des données spatialisées en y incorporant des modèles de type gravitaire à l'instar de travaux déjà existants. Ce nouvel algorithme se distingue néanmoins des précédents en ayant recours de manière explicite à des modèles gravitaires couramment utilisés en géographie, ce qui doit permettre de simplifier l'interprétation des résultats et d'éviter de possibles erreurs. L'algorithme sera présenté dans une première partie, puis appliqué dans une deuxième partie à un cas d'étude : l'analyse des flux domicile-travail en Ile-de-France. Ce papier confirme que l'ajout de modèles gravitaires permet de dépasser les résultats triviaux obtenus avec des méthodes de détection de communautés plus classiques. Les résultats restent cependant toujours difficiles à interpréter.

ABSTRACT. Community detection methods are powerful tools for analyzing spatial networks since they may highlight preferential interactions between territories. Nevertheless, these methods are biased because they overlook spatial nature of interactions. In most case, these methods produce communities which are strongly determined by spatial factors. Thus, they provide poor or well-known information. To tackle this issue, in the continuity of previous works, we proposed to build community detection algorithm which embedded gravity models. First a new algorithm will be presented and then it will be implemented for analyzing professional mobility in Ile-de-France. Using well-known gravity models, this new algorithm aim to simplify results interpretation. This new algorithm allows to go beyond trivial results obtained with classical community detection algorithms (like in previous works) but results are still difficult to use.

MOTS-CLES : Détection de communautés ; Modèle gravitaire ; Analyse de réseaux.

KEYWORDS: Community detection; Gravity model; Network analysis.

1. Contexte de la recherche

L'intérêt des géographes pour les méthodes de partitionnement et de détection de communautés, issues de l'analyse de réseaux, est compréhensible. En effet, si les données utilisées peuvent être formalisées sous la forme d'un graphe, ces méthodes semblent pertinentes pour effectuer des régionalisations à partir de données économiques (Grasland, 2011 ; Beauguitte, 2011), étudier des systèmes urbains à l'aide des flux domicile-travail (Drevelle, 2015), analyser la géographie des collaborations scientifiques (Maisonobe, 2015)… Cet intérêt est aussi révélé de manière indirecte par l'utilisation croissante des méthodes de visualisation de graphes, qui en rapprochant les lieux les plus « interconnectés » permettent notamment de faire apparaitre visuellement des communautés pouvant alors être interprétées comme des « sous-systèmes » territoriaux (Di Lello & Rozenblat, 2014).

Néanmoins, les méthodes de détection de communautés souffrent d'un biais important, elles ne tiennent pas compte des distances qui séparent les lieux. Elles sont donc aspatiales, alors même que la distance peut influer considérablement sur l'intensité des relations étudiées. Dans un grand nombre de cas, les communautés détectées apparaissent très contraintes spatialement (Ducruet et al., 2011), les résultats obtenus se révélant alors insatisfaisants puisqu'ils ne permettent pas de faire émerger des phénomènes nouveaux ou des structures plus complexes. Pour contourner ce problème et aller au-delà d'une simple mise en évidence de l'importance de la distance dans les relations étudiées, certaines méthodes de classification, voire de détection de communautés, ont alors été appliquées sur les résidus d'un modèle gravitaire (Grasland, 2011). Cette recherche propose plutôt un algorithme dit de détection de « communautés spatiales » en adaptant les algorithmes de détection de communautés à des données spatialisées. L'algorithme sera présenté dans une première partie, puis appliqué dans une deuxième partie à un cas d'étude afin d'évaluer sa pertinence.

2. Formalisation d'un algorithme de détection de communautés spatiales

Les plus anciennes définitions du concept de communauté sont à chercher dans la littérature relative à l'analyse des réseaux sociaux, où dans la continuité de la notion de clique les sociologues ont cherché à identifier des « *cohesive subgroups* » (Wasserman et Faust, 1994 ; Moody et White, 2003). Les premières définitions insistent alors sur la « cohésion » des structures détectées et une communauté peut se définir ainsi : une communauté regroupe des éléments fortement connectés entre eux et peu connectés aux autres communautés. Dans ce cadre, la qualité des partitions obtenues peut se mesurer à l'aide d'indices issus de l'informatique comme celui de Mancoridis (Mancoridis et al., 1998) [1]. Une bonne méthode de détection de communautés tendra à maximiser ces indices.

$$MQ = \frac{1}{p}\sum_{i=1}^{p} S(C_i, C_i) - \frac{2}{p(p-1)}\sum_{i>j}^{p} S(C_i, C_j) \quad (1)$$

p est le nombre de communautés, $S(C_i,C_i)$ est la densité de la communauté Ci, $S(C_iC_j)$ est la densité entre la communauté C_i et la communauté C_j.

Néanmoins, cette définition est considérée aujourd'hui dans certaines recherches comme trop restrictive et surtout comme soulevant certaines apories (Fortunato et Hric, 2016). Ainsi, même si elle ne fait pas consensus, la définition la plus communément partagée s'exprime aujourd'hui davantage en des termes probabilistes. Dans ce cadre, ce sont clairement les travaux de Newman qui font référence avec comme notion centrale la modularité (Newman et Girvan, 2004) [2]. Il convient alors de comparer les valeurs observées (A_{ij}) pour chaque relation avec des valeurs estimées (P_{ij}). Une communauté ne se distingue plus obligatoirement par une structure « cohésive forte », mais par le fait qu'elle regroupe des éléments qui échangent (partagent) plus que ce qui pouvait être attendu. En effet, tout système relationnel peut faire l'objet de prévisions en ayant simplement recours à un modèle de réseau théorique. En l'occurrence, le modèle de base permettant de déterminer des communautés « classiques » est un modèle aléatoire (un réseau aléatoire) [3].

$$Q = \frac{1}{2m}\sum_{C\in P}\sum_{i,j\in C}[A_{ij} - P_{ij}] \quad (2)$$

Q est la modularité, m est la somme totale des valeurs des liens, A_{ij} est la valeur de la relation entre i et j, P_{ij} est l'estimation de la valeur de la relation entre i et j, C est une communauté de la partition P.

$$P_{ij} = k_i k_j / 2m \quad (3)$$

P_{ij} est l'estimation de la valeur de la relation entre i et j, k_i est le degré pondéré du sommet i, m est la somme totale des valeurs des liens.

Dans le cadre de cette définition probabiliste, il est alors possible de remplacer le modèle aléatoire servant de référence pour estimer l'intensité des relations sur le territoire étudié par un modèle de type gravitaire [4], puis de chercher à maximiser la « modularité spatiale » [5] (Expert et al., 2011 ; Cazabet et al., 2017). Les premiers travaux, qui ont mis en œuvre cette approche, se sont appuyés sur un choix de modèle de type gravitaire très pragmatique et adapté aux données utilisées. En effet, les modèles utilisés sont strictement statistiques et ne font appel à aucun modèle théorique de référence.

$$P_{SPA}(i,j) = k_i k_j f(d_{ij}) \quad (4)$$

$P_{SPA}(ij)$ est l'estimation de la valeur de la relation entre i et j en utilisant un modèle gravitaire, k_i est le degré pondéré du sommet i, $f(d_{ij})$ est une fonction de dissuasion liée à la distance d_{ij} séparant les lieux i et j.

$$Q_{SPA} = \sum_{C \in P} \sum_{i,j \in C} \left[\frac{A_{ij}}{m} - \frac{P_{SPA}(i,j)}{m_{SPA}} \right] \quad (5)$$

Q_{SPA} est la modularité spatiale, m est la somme totale des valeurs des liens, m_{SPA} est la somme totale des valeurs prédites, A_{ij} est la valeur de la relation entre i et j, P_{ij} est l'estimation de la valeur de la relation entre i et j, C est une communauté de la partition P.

L'algorithme proposé dans cette recherche va quant à lui privilégier l'utilisation de modèles théoriques de référence, avant éventuellement de chercher (si les modèles théoriques se révèlent inappropriés) à coller au mieux aux données à l'aide d'une approche purement statistique. L'algorithme s'appuie pour cela sur deux modèles théoriques régulièrement utilisés en géographie et dans le domaine des transports : une loi de puissance et une loi exponentielle. Ce choix se justifie simplement par la conviction que la maîtrise des modèles de référence utilisés favorise l'interprétation des résultats obtenus et permet ainsi d'éviter de possibles erreurs. Par exemple, avec une approche purement statistique, si la distance ne joue aucun rôle sur l'intensité des relations étudiées, le modèle de référence s'adaptera et produira un résultat identique à un algorithme classique de détection de communautés (Expert et al., 2011) alors même que l'utilisateur pensera avoir pris en compte la dimension spatiale lors de l'interprétation des résultats.

Il convient de préciser ici que les distances calculées sont euclidiennes et non réelles (comme par exemple des distances calculées sur un réseau de transport), puisque l'objectif n'est pas d'avoir un modèle de référence qui soit le plus proche possible des données réelles, mais plutôt d'avoir un modèle produisant des écarts qui soient dans un premier temps facilement interprétables et significatifs.

3. Application de l'algorithme développé aux flux domicile-travail en Ile-de-France

L'algorithme développé a été appliqué afin d'étudier les flux domicile-travail en Ile-de-France. En effet, la multiplicité des flux échangés sur ce territoire questionne la manière dont il est possible de les représenter et ce que l'on peut en extraire (Lhomme, 2016). Plus précisément, nous nous sommes focalisés ici sur les volumes totaux échangés. Ces volumes ont été obtenus à partir des données librement diffusées par l'INSEE (https://www.insee.fr/fr/statistiques/2022109), les entités géographiques étant les communes. Les recherches ont alors été conduites de manière exploratoire pour la région entière, puis se sont focalisées sur la petite couronne et enfin ont été conduites par département. Ces changements dans la zone d'étude s'expliquent par la complexité des résultats obtenus avec l'algorithme développé, nécessitant alors de manipuler un nombre relativement faible d'unités spatiales pour être en mesure de donner du sens à ces résultats.

L'utilisation des algorithmes de communautés classiques peut apparaitre comme un moyen relativement simple de synthétiser les informations contenues au niveau des liens au niveau des sommets (les liens étant beaucoup plus nombreux que les

nœuds). Ces algorithmes permettent de produire des régionalisations qui simplifieront davantage l'interprétation des cartes produites. Néanmoins, comme il transparait de la littérature abordée plus haut, l'application de ces algorithmes simplifie énormément la complexité des échanges se produisant au sein des territoires en ne faisant finalement apparaitre que le caractère contraignant de la distance sur l'intensité des relations avec la création de communautés très contraintes spatialement (Fig. 1). A titre d'exemple, pour le Val-de-Marne, une communauté s'organise au centre du département autour de Créteil, une autre au nord-est principalement autour de Champigny et une dernière à l'ouest principalement autour de Vitry-sur-Seine et Villejuif. Cette représentation des mobilités si elle n'est pas inintéressante, ne fait pas émerger de phénomènes spatiaux saillants et complexes. Elle peut alors apparaitre triviale.

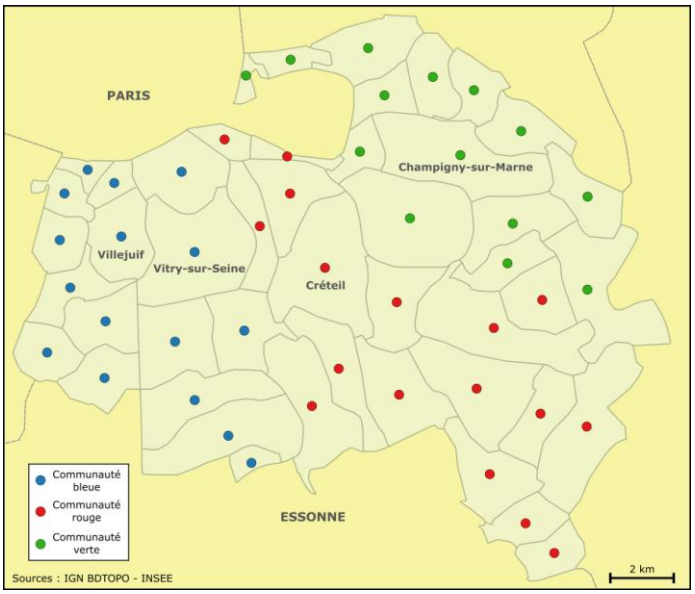

FIGURE 1. Application de l'algorithme de Louvain (algorithme classique de détection de communautés) sur les flux-domicile au sein du Val-de-Marne.

L'utilisation de l'algorithme de détection de communautés spatiales développé va complétement renverser cette perspective. En effet, la plupart des résultats obtenus se révèlent complexes à interpréter, les communautés n'étant plus que contraintes spatialement (Fig. 2). Dans les faits, il apparait nécessaire de maitriser parfaitement les logiques existantes sur le territoire étudié afin d'être en mesure d'exploiter les résultats obtenus. A titre d'exemple, pour le Val-de-Marne (Fig. 2), la communauté représentée en rouge ne s'articule pas simplement autour de Champigny-sur-Marne, mais souligne le caractère singulier de certaines communes comme Rungis ou Boissy-Saint-Léger. Ainsi, la commune de Rungis (et son marché d'intérêt national) va effectivement avoir un pouvoir attractif relativement important

sur des communes intermédiaires comme Alfortville et ce sur des distances plus importantes que la normale. Tandis que la commune de Boissy-Saint-Léger représente un bon exemple des liens particuliers que certaines communes peuvent avoir avec des pôles plus centraux en présence d'une infrastructure de transport transformant les rapports à l'espace (lien direct et efficace entre Champigny-sur-Marne et Boissy-Saint-Léger par le RER A).

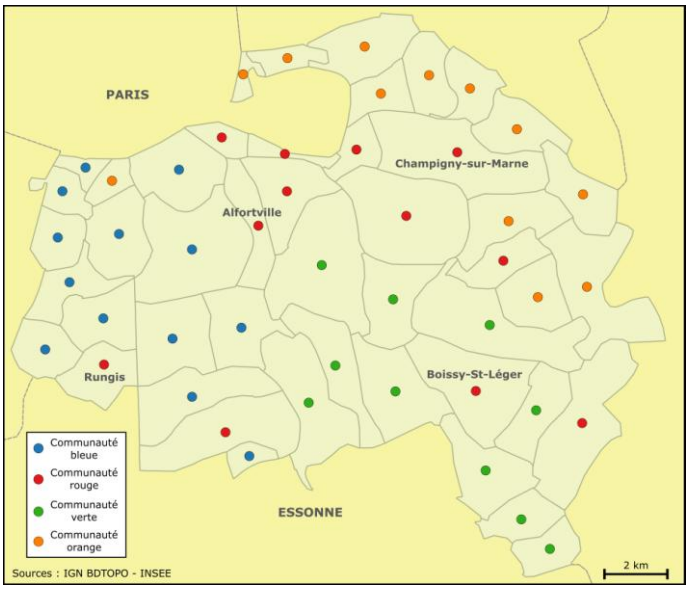

FIGURE 2. Application de l'algorithme de détection de communautés spatiales développé sur les flux-domicile au sein du Val-de-Marne.

4. Conclusion

L'algorithme développé remplit l'objectif qui lui était assigné en permettant l'émergence de communautés moins contraintes spatialement à l'instar de ce que l'on trouve déjà dans la littérature existante (Expert et al., 2011 ; Cazabet et al., 2017). Il offre donc la possibilité de dépasser les résultats triviaux généralement obtenus avec les algorithmes classiques de détection de communautés. En contrepartie, l'utilisation de cet algorithme requiert une très bonne connaissance des relations étudiées sans laquelle l'exploitation des résultats devient impossible. L'algorithme utilisé ne convient donc pas à une démarche exploratoire de découverte thématique ou territoriale, mais pourra venir enrichir des analyses existantes et offrir des perspectives de réflexion nouvelles. Dans ce cadre, la maîtrise des modèles de référence utilisés par l'algorithme de détection de communautés semble nécessaire et justifie le développement de ce nouvel algorithme par rapport à ceux déjà existants. Enfin, il conviendra à l'avenir d'interroger les différents choix effectués pour évaluer leur impact sur les résultats obtenus : d'un point de vue

technique (les seuils, l'algorithme de Louvain, la méthode de calibration, les modèles théoriques…) et thématique (taille des unités géographiques, mode de transport étudié…).

Références

Beauguitte L., (2011), Graph theory and network analysis, in. *Statistical toolbox for flow and network analysis*, dir. Van Hamme G. & Grasland C., pp. 57-68.

Cazabet R., Borgnat P., Jensen P., (2017), Using Degree Constrained Gravity Null-Models to understand the structure of journeys' networks in Bicycle Sharing Systems, *ESANN 2017 - European Symposium on Artificial Neural Networks*, Computational Intelligence and Machine Learning, Apr 2017, Bruges, Belgium.

Di Lello O., Rozenblat C., (2014), Les réseaux de firmes multinationales dans les villes d'Europe centre-orientale, *Cybergeo : European Journal of Geography*.

Drevelle M., (2015), Structure des navettes domicile-travail et polarités secondaires autour de Montpellier, *M@ppemonde*.

Ducruet C., Ietri D., Rozenblat C., (2011), Cities in Worldwide Air and Sea Flows: A multiple networks analysis, *Cybergeo : European Journal of Geography*.

Expert P., Evans T.S., Blondel V.D., Lambiotte R., (2011), Uncovering space-independent communities in spatial networks, *Proceedings of the National Academy of Sciences of the United States of America*, 108(19), pp.7663–7668.

Fortunato S., Hric D., (2016), Community detection in networks: A user guide, *Physics Reports*, 44p.

Grasland C., (2011), Natural/Residual regions based on spatial interaction models, in. *Statistical toolbox for flow and network analysis*, dir. Van Hamme G. & Grasland C., pp. 21-42.

Lhomme S., (2016), Formalisation d'une méthode de filtrage graphique pour simplifier la cartographie des interactions spatiales, *Conférence SAGEO 2016*, Nice, décembre 2016.

Mancoridis S., Mitchell B.S., Rorres C., Chen Y., Gansner E.R., (1998), Using automatic clustering to produce high-level system organizations of source code, *Proceedings of the 6th International Workshop on Program Comprehension*, IWPC'98, pp. 45–52.

Maisonobe M., (2015), *Étudier la géographie des activités et des collectifs scientifiques dans le monde : de la croissance du système de production contemporain aux dynamiques d'une spécialité, la réparation de l'ADN*, Thèse de Géographie, Université Toulouse le Mirail - Toulouse II, 502p.

Moody J., White D.R., (2003), Structural cohesion and embeddedness: A hierarchical concept of social groups, *American Sociological Review*, 68(1), pp. 103-127.

Newman M.E.J., Girvan M., (2004), Finding and evaluating community structure in networks, *Physical Review E*, 69(2).

Wasserman S., Faust K., (1994), Social Network Analysis: Methods and Applications, *Cambridge University Press*, 866p.