Un outil pour la sélection et la visualisation de flux : le package flows

Laurent Beauguitte, Timothée Giraud, Marianne Guérois

Netcom
Réseaux, communication et territoires

29-3/4 | 2015
Visualisation des réseaux, de l'information et de l'espace

Un outil pour la sélection et la visualisation de flux : le *package* flows

Laurent Beauguitte, Timothée Giraud et Marianne Guerois

Édition électronique
URL : https://journals.openedition.org/netcom/2134
DOI : 10.4000/netcom.2134
ISSN : 2431-210X

Éditeur
Netcom Association

Édition imprimée
Date de publication : 16 décembre 2015
Pagination : 399-408
ISSN : 0987-6014

Référence électronique
Laurent Beauguitte, Timothée Giraud et Marianne Guerois, « Un outil pour la sélection et la visualisation de flux : le *package* flows », *Netcom* [En ligne], 29-3/4 | 2015, mis en ligne le 23 mai 2016, consulté le 18 mars 2022. URL : http://journals.openedition.org/netcom/2134 ; DOI : https://doi.org/10.4000/netcom.2134

Networks and Communication Studies,
NETCOM, vol. 29 (2015), n° 3-4
pp. 399-408

NOTES DE RECHERCHE

UN OUTIL POUR LA SELECTION ET LA VISUALISATION DE FLUX : LE *PACKAGE* FLOWS

BEAUGUITTE LAURENT[1], GIRAUD TIMOTHEE[2], GUEROIS MARIANNE[3]

Résumé *- Cette note présente un* package R *permettant de sélectionner, analyser et visualiser les flux. Après avoir brièvement rappelé les enjeux et principes des méthodes mobilisées (flux principaux, flux dominants et flux majeurs), nous présentons les trois principales fonctionnalités mises en œuvre : sélection des flux, production d'indicateurs statistiques décrivant les flux sélectionnés et renseignant sur le volume d'information perdue, cartographie des flux dominants. Enfin, cet outil de sélection raisonnée des flux est appliqué au cas des navettes domicile-travail entre villes du Grand-Est français, afin d'illustrer les effets du choix d'une variante de la méthode des flux dominants de Nystuen et Dacey sur la hiérarchie des villes et la représentation de leurs aires d'influence.*

Mots-clés *- flux dominants, flux majeurs, matrice,* R*, visualisation.*

Abstract *- This note presents an* R package *that allows to select, analyse and visualise flows. After briefly recalling the issues and principles underlying the methods used by this tool (main flows, dominant flows and major flows), we describe the three main functions that are implemented: flows selection, indicators production about selected flows and lost information, cartography of dominant flows. Lastly, this selecting tool is applied to interurban commuters flows in the « Grand-Est » French region. This example shows the effect of choosing a variant of Nystuen and Dacey method on the resulting urban hierarchy and areas of influence.*

Key-words *- dominant flows, major flows, matrix,* R*, visualization.*

[1] CNRS, UMR IDEES -beauguittelaurent@hotmail.com
[2] CNRS, UMS RIATE - timothee.giraud@ums-riate.fr
[3] Université Paris 7, UMR Géographie-cités et UMS RIATE – marianne.guerois@univ-paris-diderot.fr

INTRODUCTION

L'analyse et la représentation des flux constituent un champ de recherche encore insuffisamment développé en géographie, alors même que les processus de mondialisation contemporains rendent leur compréhension de plus en plus cruciale. Si l'analyse de réseaux, qu'ils soient sociaux (Wasserman et Faust, 1994 ; Hennig *et al.*, 2012) ou complexes (Newman, 2010), a connu de nombreux développements ces dernières décennies, tant méthodologiques que conceptuels, l'étude des flux a connu une extension plus thématique que méthodologique (voir cependant la thèse récente de Françoise Bahoken, 2016). De nombreux flux liés à internet sont aujourd'hui couramment analysés (liens entre sites, tweets, flux RSS, etc.) mais le manque d'outils dédiés se fait sentir, excepté pour les applications liées à la logistique et aux transports. L'objet de cette note scientifique est de présenter un outil[4] de sélection, d'analyse et de visualisation des flux développé au sein de l'UMS RIATE.

Une première partie rappelle les principales méthodes d'analyse de flux (flux dominants, flux majeurs) ; la seconde partie présente les potentialités de l'outil. Enfin, une troisième partie expose les conséquences thématiques des choix possibles *via* une courte étude sur les navettes domicile-travail dans le Grand Est français.

L'ANALYSE DES FLUX GEOGRAPHIQUES : ENJEUX ET METHODES

S'intéresser aux flux suppose de mettre l'accent sur les relations entre lieux plutôt que sur les caractéristiques propres de chacun des lieux, ce qui est le fondement même de l'analyse spatiale. Mais l'analyse et surtout la représentation des flux supposent le plus souvent une sélection pour faciliter la lecture et l'interprétation[5]. Représenter l'ensemble des flux crée des cartes dites en oursins d'une lisibilité toute relative. De nombreuses méthodes ont été mises au point par les géographes au cours des dernières décennies et l'objectif n'est pas ici de proposer une étude exhaustive mais plutôt de donner quelques points de repère bibliographiques et méthodologiques.

L'une des premières méthodes mises au point fut sans doute celle dite des flux dominants (ou des régions nodales) proposée par Nystuen et Dacey en 1961. Travaillant sur les flux téléphoniques entre villes de la région de Seattle, ils cherchaient à mettre en évidence des phénomènes de hiérarchie entre lieux. Selon les auteurs, un lieu *i* est dominé par un lieu *j* si les deux conditions suivantes sont réunies :

- (1) le flux le plus important de *i* est émis en direction de *j* ;
- (2) la somme des flux reçus par *j* est supérieure à la somme des flux reçus par *i*.

[4] Le terme outil est préféré à celui d'application dans la mesure où le package flows nécessite l'environnement R pour fonctionner.

[5] Sur la réduction de la complexité de la carte de flux, voir W. Tobler (1970, 1987) et F. Bahoken (2016) pour un panorama complet.

Cette méthode crée ce qui est appelé en théorie des graphes un arbre (graphe acyclique) ou une forêt (ensemble d'arbres non connexes) où trois catégories de lieu apparaissent : des lieux dominants, des lieux dominés et des lieux intermédiaires. Si la méthode crée une hiérarchie spatiale et fonctionnelle souvent lisible (Potrykowska, 1993 ; Karjalainen, 1993), son inconvénient majeur est de mal prendre en compte la hiérarchie entre flux : il est en effet fréquent qu'une poignée de lieux génère la très grande majorité des flux, or la méthode n'en garde par définition qu'un seul et unique par lieu (Beauguitte, 2014).

Diverses options ont par la suite été proposées pour mieux prendre en compte cette intensité, l'une des plus fréquemment employée étant celle dite des flux majeurs : on ne sélectionne que les flux les plus importants, soit localement, soit globalement, que cette importance soit absolue ou relative. Dans le cas des navettes domicile-travail entre communes par exemple, on pourra par exemple choisir d'analyser :

- les flux supérieurs à 100 ;
- les 50 premiers flux (critère global) ;
- les 10 premiers flux émis par chaque commune (critère local).

Ces critères peuvent aussi être exprimés sous forme relative :

- les flux représentant plus de 10 % de la population active d'une commune (critère local) ;
- les flux prenant en compte 80 % du total des navetteurs (critère global).

Ces critères ont notamment été utilisés pour étudier les flux aériens et ferroviaires en Europe de l'ouest (Cattan, 1995), les échanges commerciaux à l'échelle mondiale (Grasland et Van Hamme, 2011) et les systèmes urbains français (Berroir *et al.*, 2012).

Ces méthodes permettent le plus souvent de mettre en évidence les hiérarchies entre lieux mais la perte d'informations créée par la sélection n'est que rarement questionnée. Si la sélection est nécessaire pour permettre tant l'interprétation que la représentation, il nous semble également utile de proposer des indicateurs statistiques permettant simultanément de connaître le volume d'information perdue et les caractéristiques des flux sélectionnés. L'outil développé au sein de l'UMS RIATE vise autant à proposer des méthodes de sélection qu'à permettre de les justifier d'un point de vue statistique.

LE *PACKAGE* FLOWS

L'outil développé est un *package* R. Rappelons que ce logiciel libre et multi-plateforme de statistiques est devenu ces dernières années un logiciel de traitements de données au sens large, autorisant tant la cartographie que l'analyse textuelle ou l'analyse de graphes (Groupe ElémentR, 2014). La structure de R est modulaire : une base est commune à tous les utilisateurs, puis chacun installe les modules (*packages*) nécessaires aux traitements à réaliser. L'ensemble du code est libre et rigoureusement documenté, ce qui contribue à la diffusion de l'outil et assure la reproductibilité des résultats obtenus.

La chaîne de traitement proposée par le *package flows* (Giraud *et al.*, 2015) est la suivante :

- préparation des données ;
- sélection des flux ;
- données statistiques sur la sélection réalisée ;
- représentation sous forme de graphe ou de carte (flux dominants).

Un texte accessible en ligne, ou en local une fois le *package* installé, détaille les différents principes et étapes du traitement (https://cran.r-project.org/web/packages/flows/vignettes/flows.html).

Les données en entrée peuvent être de format *i-j-fij*, c'est-à-dire origine – destination – intensité du flux. Une première fonction **preflows** permet alors de passer de cette liste de liens (appelée aussi matrice en format *long*) à une matrice carrée. Cette étape est facultative si les données sont déjà en format matriciel (format *wide*), la diagonale pouvant être pleine ou vide. Par contre, la matrice doit être carrée (liste identique d'origines et de destinations).

Les méthodes de sélection basées sur l'origine *i* des flux sont accessibles *via* la fonction **firstflows** et sont au nombre de trois :

- sélection de *k* premiers flux provenant de tout *i* (méthode *nfirst*) ;
- sélection des flux *fij* supérieurs à une intensité *k* donnée (méthode *xfirst*) ;
- sélection des flux *fij* tels que la somme de ces flux permet d'atteindre un seuil *k* (absolu ou relatif) donné (méthode *xsumfirst*).

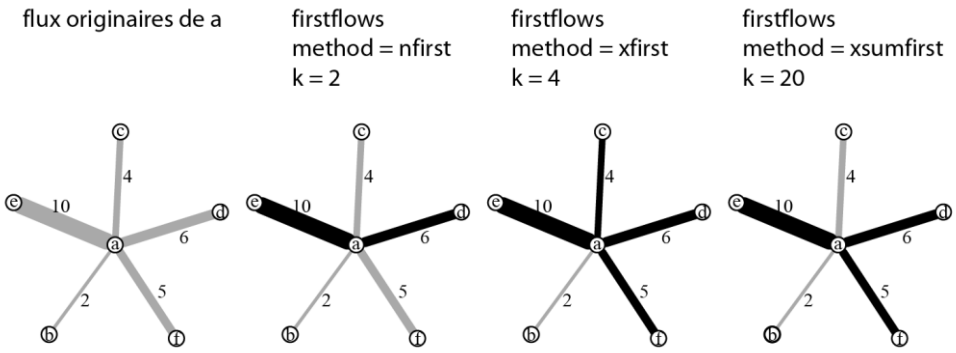

Figure 1** : les trois méthodes de la fonction **firstflows.
En noir apparaissent les liens sélectionnés en fonction de la méthode et du seuil k *choisi.*

En fonction des objectifs de la recherche, il sera par exemple possible de sélectionner les deux premiers flux émis par chaque lieu, les flux supérieurs à un seuil donné ou encore les flux représentant 50 % du volume total des flux pour tout *i* (figure 1).

Les méthodes de sélection prenant en compte le volume total des flux sont implémentées dans la fonction **firstflowsg** et sont identiques aux trois options détaillées ci-dessus : sélection des *k* premiers flux, sélection des flux supérieurs à une intensité donnée et enfin sélection des flux tels que la somme permet d'atteindre un seuil donné (absolu ou relatif).

Enfin, la fonction **domflows** permet de sélectionner les flux en fonction d'un critère de domination. Cette fonction permet notamment la sélection des flux obéissant au deuxième critère de la méthode de Nystuen et Dacey.

Toutes ces fonctions prennent en entrée une matrice de flux carrée et génèrent des matrices binaires de même taille : les flux sélectionnés par la méthode sont représentés par des 1, les autres par des 0. Il est donc possible de combiner les fonctions de sélection pour obtenir des matrices de flux multi-critères (figure 2). Il est également possible de récupérer l'intensité des liens sélectionnés en opérant une multiplication terme à terme, dite aussi produit matriciel d'Hamamard, des matrices.

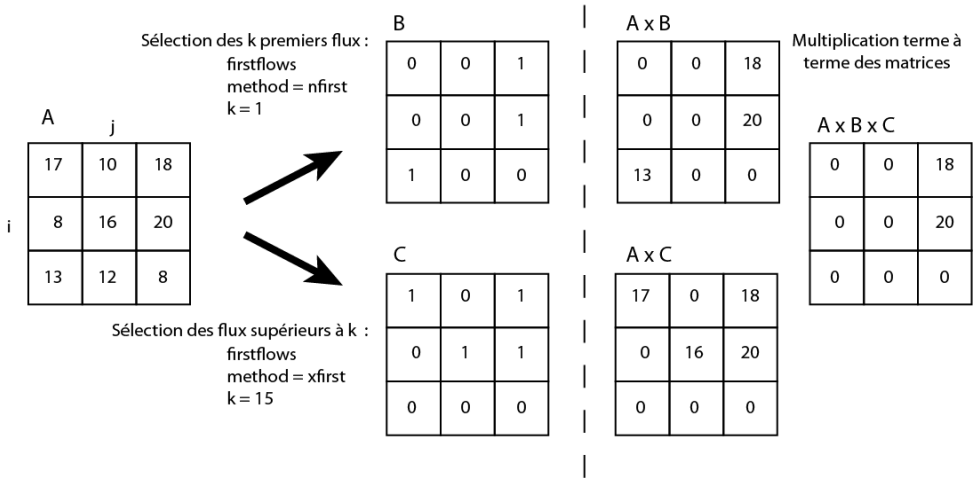

***Figure 2** : Sélection de flux et combinaison de matrices.*

Il est ensuite possible de calculer un certain nombre d'indicateurs à l'aide de la fonction **statmat**. Dans la littérature, l'un des rares indicateurs mobilisé compare le nombre de liens conservé et le volume de flux qu'ils prennent en charge (cf par exemple Drevelle, 2012). Il nous a semblé pertinent de compléter cette information à l'aide d'indicateurs issus entre autres de la théorie des graphes. Cette fonction fournit notamment la densité de flux (nombre de flux présents divisé par le nombre de flux possibles), le nombre, la taille et la composition des composantes connexes, le volume total de flux, les quartiles et la moyenne de l'intensité des flux. Il est donc possible de faire une sélection raisonnée et justifiée statistiquement. Cette fonction peut également générer quatre sorties graphiques : la distribution des degrés (par défaut, nombre de

liens sortants) brute et pondérée par l'intensité des liens ; une courbe de Lorenz et une boîte à moustaches relatives à l'intensité des flux.

En ce qui concerne la visualisation, la fonction **plotDomFlows** produit un graphe où la taille et la couleur des sommets sont dépendants de leur position dans le graphe (sommet dominant, intermédiaire ou dominé) et où l'épaisseur des liens varie selon l'intensité du flux. La fonction **plotMapDomFlows** permet une représentation cartographique des flux sélectionnés selon le même principe. Ces deux fonctions ne s'appliquent qu'à des sélections de type flux dominants, différents *packages* permettant déjà la visualisation des sélections autres.

L'outil a été conçu de manière à être aussi générique et souple que possible : il s'adapte donc à toute matrice, qu'elle concerne ou non des unités spatiales.

UN EXEMPLE D'APPLICATION : LES NAVETTES DOMICILE-TRAVAIL DANS LE GRAND EST FRANÇAIS

A titre d'illustration, nous présentons un aperçu rapide des conséquences induites par ces choix de méthodes et de paramètres à partir de l'exemple des flux de navetteurs entre les villes du Grand Est français[6]. Ces données de déplacements fréquents et réguliers sont de plus en plus mobilisées pour analyser la mise en réseau des villes à des échelles non seulement locales (aires urbaines voisines) mais aussi plus vastes (voir par exemple Brutel, 2011 ; Gingembre et Baude, 2014). L'élargissement de la portée des déplacements contribue en effet à redessiner la géographie des territoires. De plus, les navettes à longue distance, si elles restent minoritaires (moins de 1% des actifs ayant un emploi en 2008[7]), sont en progression régulière (+2,1%/an entre 1994 et 2008). Le traitement de ce jeu de données à l'aide du package ***flows*** doit permettre d'apporter différents éclairages complémentaires sur la manière dont les échanges tissés entre aires urbaines structurent ce territoire du Grand Est[8].

---

[6] Les données de flux traitées dans cette partie sont issues des fichiers détail du Recensement Général de la Population de l'INSEE (2011) (source : http://www.insee.fr/fr/themes/detail.asp?reg_id=99&ref_id=mobilite-professionnelle-11). L'espace d'étude comprend cinq régions administratives de l'Est de la France, dont les regroupements dans le contexte de la Réforme territoriale ont été particulièrement discutés : la Champagne-Ardenne, la Lorraine, l'Alsace, la Bourgogne et la Franche-Comté. Les villes correspondent aux aires urbaines dans leur délimitation de 2010 (source : http://professionnels.ign.fr/geofla#tab-3).

[7] Sources : SOeS (Service de l'Observation et des Statistiques du Ministère de l'Ecologie, du Développement durable et de l'Energie), INSEE, INRETS, enquête nationale transport 2008. Les navettes à longue distance sont dans ce cas définies comme les navettes dont la destination se situe à plus de 80 km à vol d'oiseau du domicile.

[8] L'ensemble des données utilisées dans cette partie sera inclus dans le *package*.

Le tableau 1 montre les statistiques correspondant à plusieurs critères de sélection absolus : on garde l'ensemble des flux supérieurs à un seuil donné. On constate par exemple que garder les seuls liens supérieurs à 500, soit moins de 5 % du nombre total de liens, permet de conserver plus de 60 % du volume total des flux.

	Matrice complète	Flux > 100	Flux > 500	Flux > 1000
Volume total	313298	270705 (86%)	193196 (62%)	145365 (46%)
Nombre de liens	3191	484	137	62
Intensité moyenne	98	559	1410	2345
Nb de composantes connexes (> 1)*	1	3	10	7

***Tableau 1** : Effectifs et intensité des liens selon différents seuils absolus de sélection des flux.*
**Les composantes connexes ne comprennent pas les villes dites isolées (non reliées aux autres villes).*

L'identification des flux les plus structurants est réalisée à l'aide de la fonction **domflows**. Pour ces mobilités de proximité, le choix a été fait de modifier la première règle de la méthode classique de Nystuen et Dacey, afin d'élargir la sélection, de conserver des cas de domination multiple et de faire émerger des bassins d'influence moins fragmentés (figure 3).

Ainsi, en retenant tous les liens supérieurs à 20% des flux d'actifs sortants au départ de chaque ville (et pas seulement le premier flux), tout en conservant la deuxième règle de domination, on aboutit à une structure composée par 6% de l'ensemble des liens et 39% du volume des flux. La hiérarchie des « têtes de réseau » évaluée à partir de la sélection de flux dominants et du nombre d'actifs entrants fait nettement ressortir, par ordre décroissant, la domination de Nancy, Strasbourg, Mulhouse et Metz, qui captent chacune plus de 10 000 actifs. Cette hiérarchie est très proche de celle qui ressort de l'application de la méthode classique.

Les changements introduits par la variante sont beaucoup plus importants en termes de délimitation des zones d'influence : 6 sous-régions peuvent être identifiées, au lieu de 27 d'après la méthode classique (les différentes cartes ne sont pas reproduites ici mais l'ensemble de ces traitements peut être réalisé à l'aide du *package*). La plus vaste correspond au périmètre des actuelles régions d'Alsace et de Lorraine. Un autre sous-ensemble associe les villes de Franche-Comté avec celles de l'est et du nord-ouest de la Bourgogne. Les villes de Champagne-Ardennes ne sont quasiment connectées qu'entre elles. On pourrait facilement, à l'aide d'un tel *package*, réitérer la sélection avec un seuil encore plus élevé ou au contraire plus restreint, afin de repérer les liens les plus robustes, les liens intermittents, et l'évolution du nombre de sous-régions en fonction de ces seuils.

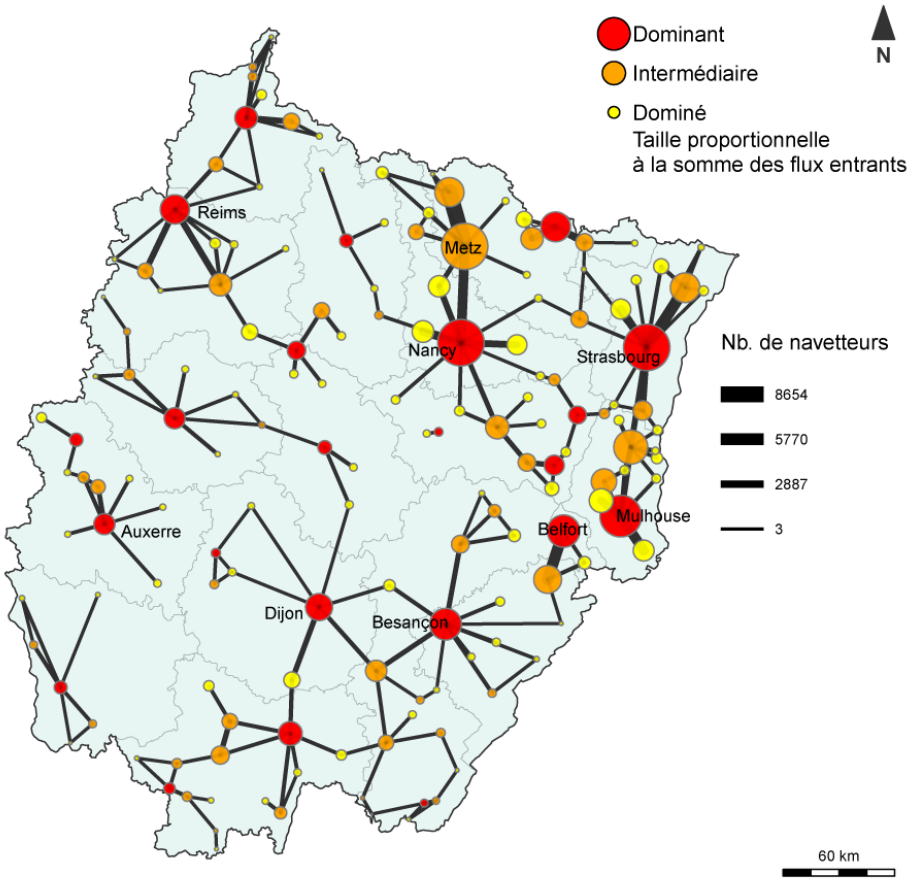

Figure 3 *: Une variante des flux dominants appliquée aux navettes inter-urbaines dans le Grand Est.*

Flux dominants des navettes entre aires urbaines, avec une variante sur le premier critère de Nystuen et Dacey (flux supérieurs à 20% de l'ensemble des actifs sortants d'une ville). Cette figure est la sortie brute du package et nécessite donc un travail de mise en page a posteriori.

La modification voire la suppression de la seconde règle de domination aboutirait à une carte plus complexe de la hiérarchie des flux, les échanges réciproques entre villes pouvant alors être mis en valeur.

CONCLUSION

Le *package* **flows**, déposé depuis juin 2015 sur le CRAN (dépôt officiel des *packages* R), vise à permettre une sélection raisonnée des flux, tout en laissant à l'utilisateur une grande liberté quant aux critères choisis. Les options de visualisation ne concernent que la famille de méthodes dites des flux dominants dans la mesure où

d'autres outils existent pour assurer la représentation sous forme de graphe (*packages* **igraph** ou **statnet**) ou de cartes de flux (*package* **cartography**). Enfin, nous souhaiterions bénéficier de retours d'utilisateurs afin d'enrichir les possibilités de ce *package*, notamment en ce qui concerne les indicateurs statistiques proposés sur les sélections effectués.

REFERENCES

BAHOKEN F. (2016), *Contribution à la cartographie d'une matrice de flux,* Thèse de doctorat en Géographie, Université Paris 7.

BEAUGUITTE L. (2014), (re)lire les classiques : A graph theory interpretation of nodal flows de Nystuen et Dacey, 1961, Blog du groupe fmr, http://groupefmr.hypotheses.org/3517 (visité le 2 juin 2015).

BERROIR S., CATTAN N., GUÉROIS M., PAULUS F. et VACCHIANI-MARCUZZO C. (2012), Les systèmes urbains français, Synthèse DATAR, Travaux en Ligne 10, http://www.datar.gouv.fr/sites/default/files/travaux_en_l_10_synthese_susm.pdf (visité le 2 juin 2015).

BRUTEL C. (2011), « Un maillage du territoire français. 12 aires métropolitaines, 29 grandes aires urbaines », *INSEE Première*, n°1333.

CATTAN N. (1995), Barrier Effects: The Case of Air and Rail Flows, *International Political Science Review / Revue internationale de science politique*, vol. 16(3), pp. 237-248.

DREVELLE M. (2012), Structure des navettes domicile-travail et polarités secondaires autour de Montpellier, *M@ppemonde*, 107 (visité le 4 février 2016)

GINGEMBRE J., BAUDE J. (2014), Les mobilités domicile-travail dans les réseaux d'agglomérations, *EchoGéo* [En ligne], 27 | 2014, mis en ligne le 23 mai 2014, consulté le 19 juin 2015.

GIRAUD T., BEAUGUITTE L. et GUÉROIS M. (2015), *flows : Flow Selection and Analysis*, CRAN, https://cran.r-project.org/web/packages/flows/ (visité le 04 février 2016).

GROUPE ELEMENTR (2014), *R et espace. Traitement de l'information géographique*, Framabook, http://framabook.org/16-r-et-espace/ (visité le 2 juin 2015).

HENNIG M., BRANDES U., PFEFFER J. et MERGEL I. (2012), *Studying Social Networks. A Guide to Empirical Research*, Frankfort et New York, Campus Verlag.

KARJALAINEN E. (1993), Structure of migration flows in Kainuu, Finland, *Geographia Polonica*, 61, pp. 317-328.

NEWMAN M. (2010), *Networks : An introduction*, Oxford, Oxford University Press.

NYSTUEN J. et DACEY M. (1961), A graph theory interpretation of nodal regions, *Papers and Proceedings of the Regional Science Association*, vol. 7, pp. 29-42.

POTRYKOWSKA A. (1993), Intra-urban migration in the Warsaw urban region, *Geographia Polonica*, 61, pp. 281-291.

TOBLER W. R. (1970), A computer movie simulating urban growth in the Detroit region, *Economic geography*, 46, pp. 234-240.

TOBLER W. R. (1987), Experiments in migration mapping by computer, *The American Cartographer*, 14(2), pp. 155-163.

VAN HAMME G. et GRASLAND C. (2011), Divisions of the world according to flows and networks, *EuroBroadMap working paper*, https://halshs.archives-ouvertes.fr/EUROBROADMAP/ (visité le 2 juin 2015).

WASSERMAN S. et FAUST K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge, Cambridge University Press.