Impact du changement d'échelle sur l'étude des causes des feux de forêts du sud-est de la France

Romain Louvet [1], **Didier Josselin** [1,2],
Cyrille Genre-Grandpierre [1], **Jagannath Aryal** [3]

1. *UMR ESPACE 7300 CNRS, Université d'Avignon*
   romain.louvet@alumni.univ-avignon.fr
2. *Laboratoire d'Informatique d'Avignon, Université d'Avignon*
   didier.josselin@univ-avignon.fr
3. *University of Tasmania*
   jagannath.aryal@utas.edu.au

RÉSUMÉ. *Le support spatial des données a potentiellement une forte influence sur le traitement statistique des observations. Cette problématique est connue en géographie sous le nom de Modifiable Areal Unit Problem (MAUP). Celle-ci survient lorsque différentes unités surfaciques peuvent être utilisées et que le résultat varie en fonction de ce choix. Dans cet article, nous présentons un état de l'art de ce problème. Considérant un des aspects du MAUP, à savoir l'influence du changement de niveau d'échelle, nous développons une méthode de visualisation de la sensibilité des statistiques à ce problème. Cette méthode est testée sur l'étude des feux de forêt du sud-est de la France, avec des données issues de la base Prométhée, à partir desquelles nous recherchons des variables explicatives. Nos résultats montrent des variations des coefficients de corrélation en fonction des niveaux d'échelle et la possibilité de sélectionner les variables et les niveaux d'échelle en fonction de cette variabilité. Nous proposons deux méthodes : (i) utiliser la visualisation de ces variations afin d'améliorer la robustesse de l'analyse de corrélation en sélectionnant les informations pertinentes selon leur sensibilité au MAUP, (ii) sélectionner un niveau d'échelle pour lequel le résultat est le plus différent possible d'une redistribution spatiale aléatoire de la variable dépendante.*

ABSTRACT. *The Modifiable Areal Unit Problem (MAUP) is a well-known issue related to the influence of the spatial support on statistical observations. It occurs when different spatial units making different spatial partitions are used and when the resulting measures vary according to those partitions. In this paper, we first draw a state of the art. Considering the particular problem of (up)scaling, we propose a method to visualize the sensitivity of the spatial statistics to the support. We test this method on forest fires in Southern France, handling a sample from the Promethée database.*

From these data, we try to find the key explanatory variables. The results show that the correlation coefficient varies significantly, depending on scale, and that we can select variables and scales based on this variability. Then we propose two different ways to deal with the MAUP: (i) by using geovisualization to assess and to improve the robustness of the correlation analysis and to choose the pertinent information that allows to minimize the sensitivity, (ii) by considering as pertinent the spatial partition which is the farthest one from a random spatial distribution of the independent variable.

MOTS-CLÉS : *Modifiable Areal Unit Problem (MAUP), Change Of Support Problem (COSP), feux de forêt, base de données Prométhée, R*

KEYWORDS: *Modifiable Areal Unit Problem (MAUP), Change Of Support Problem (COSP), forest fire, Prométhée database, R*

1. Introduction

Un des principaux problèmes actuels de l'analyse des données géographiques est le biais statistique induit par l'utilisation d'unités surfaciques. Ce problème est connu en géographie sous le nom de MAUP (*Modifiable Areal Unit Problem*, ou problème d'unité spatiale modifiable). Il est défini par le fait que la manière d'agréger les données sous la forme d'unités spatiales a un impact significatif sur le résultat, en particulier sur la recherche de facteurs explicatifs à l'aide de la corrélation (Openshaw, 1984). En l'absence de « règle pour l'agrégation des unités spatiales surfaciques », les découpages administratifs sont massivement utilisés. Puisque le choix du découpage peut avoir un impact sur le résultat, il est particulièrement problématique qu'un découpage soit préféré à un autre, qui plus est s'il ne possède pas nécessairement de rapport avec le phénomène.

Le sud-est de la France, tout comme les autres régions méditerranéennes en Europe, est régulièrement et fortement affecté par les feux de forêts. Il s'agit d'un processus complexe du fait de son statut de catastrophe « naturelle » dont les causes sont en fait principalement humaines (Ganteaume, Jappiot, 2013). Les feux de forêt ne sont *a priori* pas contraints par les découpages territoriaux. Pourtant, certaines études utilisent des limites administratives et un seul niveau d'échelle, ignorant de fait le problème soulevé par l'utilisation d'unités spatiales étant par nature le résultat d'une agrégation, selon des limites plus ou moins arbitraires et possédant des tailles et des formes hétérogènes.

Pour illustrer ce type d'étude, citons Ganteaume et Jappiot qui ont obtenu des résultats intéressants sur les causes des feux de grande taille dans le sud-est de la France (Ganteaume, Jappiot, 2013). Cependant, ces auteurs ont choisi de travailler à l'échelle des départements, alors que des niveaux d'échelle beaucoup plus fins étaient disponibles dans la base sur les feux de forêts utilisée (la base Prométhée). Ce choix du département est justifiable puisque les politiques de

lutte anti-incendie doivent être mises en place à cette échelle. Toutefois, choisir comme unité d'analyse des unités spatiales qui semblent être délimitées de manière arbitraire, dont les limites seraient donc modifiables, est un problème méthodologique important indépendamment de considérations administratives et opérationnelles.

L'utilisation des découpages administratifs à des fins d'étude d'un phénomène complexe tel que les feux de forêts peut s'expliquer par la difficulté à disposer de données à l'échelle individuelle ou selon d'autres types de découpages. Néanmoins, si nous laissons de côté la question des délimitations, pourquoi serait-il plus pertinent d'étudier les causes des feux de forêt en France à l'échelle des départements plutôt qu'à l'échelle des communes ? Car si ce choix influence fortement nos résultats, quel serait le résultat pertinent ? Existe-t-il un "bon" résultat *in fine* ? Autrement dit, puisque le MAUP soulève des questions sur la certitude des résultats statistiques obtenus à partir des données spatiales (Fotheringham *et al.*, 2000), comment prendre en compte ce problème dans l'analyse pour que les résultats soient plus robustes ? C'est ce que nous souhaitons développer dans cet article.

À partir de données de la base Prométhée sur les feux de forêt du sud-est de la France et de variables explicatives agrégées selon les différents niveaux des découpages territoriaux français, nous développons une méthode de visualisation du coefficient de corrélation en fonction des découpages utilisés. Puis, nous proposons d'utiliser la visualisation des variations en fonction de ces niveaux de découpages afin d'améliorer la robustesse de l'analyse en sélectionnant les informations pertinentes selon leur sensibilité au MAUP d'après deux principes :

– la sélection d'un résultat selon sa relative stabilité à travers les échelles ;

– la sélection d'un niveau d'échelle dont le résultat possède la plus grande différence par rapport à un résultat aléatoire.

Nous commençons par présenter un bref état de l'art sur la question du MAUP. Nous développons sa définition, l'effet d'échelle et l'effet de zonage, et nous replaçons ce problème par rapport à ses problèmes connexes, tels que le biais d'inférence écologique, le paradoxe de Simpson, et le COSP (Change of Support Problem). Puis nous décrivons trois approches possibles pour tenter de résoudre le MAUP :

– utiliser des données individuelles ;

– adapter les méthodes statistiques au MAUP ou l'utiliser pour optimiser les résultats ;

– évaluer la sensibilité au MAUP pour choisir le bon découpage ou uniquement les résultats les moins sensibles au changement de partition spatiale.

C'est cette troisième approche que nous avons choisie d'implémenter sous R et que nous décrivons plus en détails ensuite, avec les données utilisées et les

traitements réalisés. Enfin, nous terminons par la présentation des résultats obtenus et leur discussion.

2. État de l'art

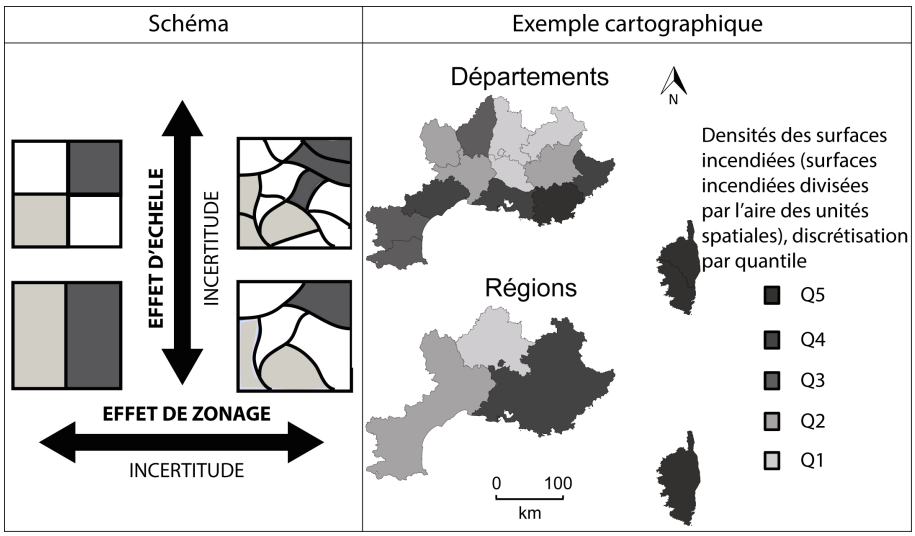

Figure 1. Deux aspect du MAUP, effets de zonage et d'échelle, du fait de l'agrégation de données géographiques et illustration par la cartographie des feux de forêt dans le sud-est de la France à deux niveaux d'échelle

Le MAUP revêt deux aspects, définis par deux effets propres à l'utilisation d'unités spatiales surfaciques. Ces effets sont l'échelle et la délimitation (Figure 1). L'effet de l'échelle est défini par la modification d'un résultat à partir des mêmes données de départ, selon la manière dont ces données sont agrégées à différents niveaux de précision en fonction du nombre et de la taille des unités spatiales. L'effet de la délimitation, également connu sous le nom d'effet du zonage, est dû à l'impact sur les résultats à partir des mêmes données de départ et à une même échelle (même nombre d'unités spatiales) selon différents découpages.

Le MAUP est un problème qui n'est pas restreint à la géographie. Il s'agit d'abord d'un biais statistique commun à toutes disciplines utilisant des agrégats, voir d'une erreur de raisonnement dénoncé y compris en philosophie (Figure 2). Il est proche du problème d'inférence écologique, énoncé par Robinson (W. Robinson, 1950) pour dénoncer l'utilisation par la sociologie de résultats statistiques sur des groupes pour inférer un comportement individuel. En économie et en médecine, le paradoxe de Simpson a été clairement établi comme un biais d'échantillonnage, qui, selon les regroupements d'individus statistiques recensés, amène à inférer des conclusions opposées. Cet effet a été identifié par

(Simpson, 1951) à partir de l'analyse de tableaux de contingence. Sous cette forme, le problème est aspatial. Il devient spatial dès que l'échantillon concerne des données localisées.

Une des formes spatiales de ce paradoxe est le COSP, ou Change Of Support Problem. En statistiques spatiales, il regroupe un ensemble de problèmes liés au changement du support des données (points vers surfaces, surfaces vers points). Il s'agit par exemple d'un problème d'interpolation spatiale lorsque deux supports, en général des unités surfaciques, ne sont pas calés. Lorsque le processus de traitement consiste à désagréger l'information spatialisée, on parle de *downscaling*, qui inclut un problème de précision des données. Cela est très proche du problème d'inférence écologique (King *et al.*, 2004), qui est décrit en écologie ou dans les sciences de l'environnement. À l'opposé, le MAUP est un problème d'*upscaling*, puisqu'on agrège l'information spatiale en changeant les partitions. En un sens, le COSP couvre l'ensemble des problèmes dus aux procédures d'agrégation et de désagrégation spatiales et peut ainsi être considéré comme l'extension spatiale du paradoxe de Simpson, des paramètres de proximité entre individus statistiques ou d'autocorrélation spatiale intervenant alors dans le (ré-)échantillonnage (Figure 2). On entend ici par échantillonnage le processus qui consiste, dans une population ou un échantillon connu, à regrouper d'une certaine façon et selon certains critères les individus en sous-échantillons.

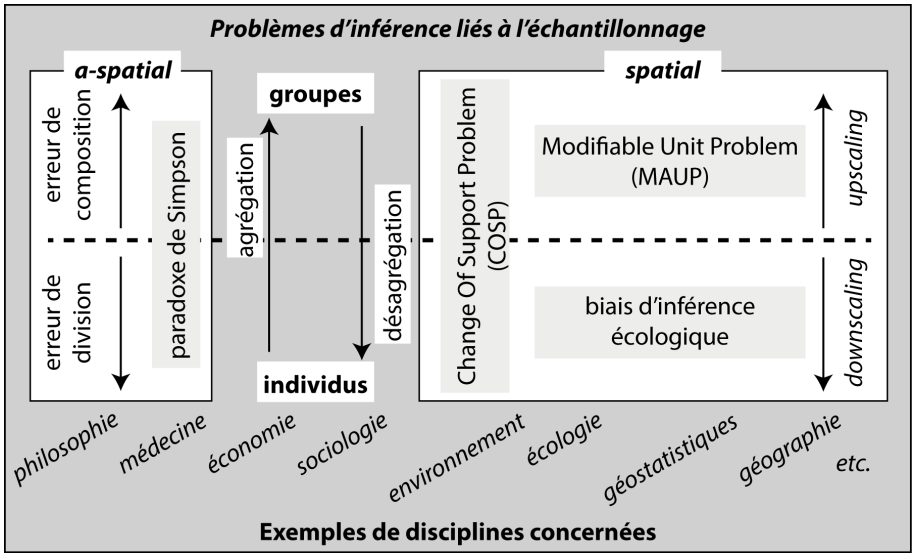

Figure 2. Les différentes formes de problèmes liées à l'échantillonnage spatial ou a-spatial

Depuis sa découverte attribuée à Gehlke et Biehl (Gehlke, Biehl, 1934) ce problème a été décrit par des travaux de référence tels que ceux de Robinson

(W. Robinson, 1950) et d'Openshaw et Taylor (Openshaw, Taylor, 1979). Bien que de nombreuses études fassent état de ce problème, peu de solutions efficaces ont été appliquées et aucune solution globale ne fait consensus (Swift *et al.*, 2008 ; Arsenault *et al.*, 2013). Nous définissons trois approches dans les solutions proposées. La première, la plus simple, serait d'utiliser uniquement des données désagrégées, au niveau individuel. Malheureusement, ce type de données est rarement disponible, souvent par nécessité de maintenir le secret statistique. Par ailleurs, ces données possèdent un faible pouvoir de communication en comparaison avec leur équivalent cartographié selon des limites spatiales bien connues. Elles peuvent également souffrir d'un problème d'atomisme ou de la tentation d'ignorer la dimension spatiale des données pour produire des résultats (Swift *et al.*, 2008 ; Arsenault *et al.*, 2013).

Les données individuelles n'étant pas moins problématique, la seconde approche consiste à s'adapter au MAUP. Une première solution consiste à utiliser des formules statistiques le prenant en compte, comme par exemple une corrélation pondérée par la taille des unités spatiales (A. Robinson, 1956). Une seconde solution adaptative est fondée sur le fait de considérer le MAUP comme un outil plutôt que comme un problème, ce qui se justifie par le fait qu'il est directement lié à la structure spatiale de la variance (Swift *et al.*, 2008). À partir de là, il est possible de préconiser le choix d'un découpage pertinent à partir de l'optimisation du résultat recherché (Openshaw, 1984). Plusieurs méthodes peuvent être employée, comme l'autocorrélation ou la régression géographiquement pondérée (GWR) (Charleux, 2005), afin de créer le découpage qui montrera le plus d'information sur la structure spatiale (King, 1997). Toutefois, ces méthodes d'optimisation du découpage sont problématiques du point de vue de la définition classique de l'objectivité scientifique.

La troisième approche développée pour résoudre le MAUP est l'évaluation de la sensibilité des résultats au problème. L'analyse de sensibilité peut d'abord être conçue pour comparer les résultats des différents découpages à des variables statistiques connues au niveau individuel afin de sélectionner le meilleur découpage (Arsenault *et al.*, 2013). Un exemple de ce type de solution, proche des méthodes statistiques de downscaling, consiste à utiliser un ensemble de variables connues au niveau individuel, puis d'ajuster une matrice de variance-covariance des données agrégées afin de sélectionner uniquement le découpage montrant la plus grande similarité avec le niveau individuel de variance (Steel, Holt, 1996 ; Holt *et al.*, 1996 ; Tranmer, Steel, 1998). Cette solution pose toutefois un problème de taille : elle nécessite la connaissance de variables au niveau individuel. Les statistiques bayésiennes peuvent être employées pour pallier à ce problème en mesurant la sensibilité au MAUP à partir d'un estimateur calculé selon des données individuelles générées aléatoirement (Hui, 2009). Ces solutions sont fondées sur l'hypothèse que le MAUP n'affecte que les données non aléatoires (Openshaw, 1984). Une troisième solution d'analyse de sensibilité consiste à re-échantillonner aléatoirement les données observées pour éliminer l'effet du support du calcul d'indice (Mahfoud *et al.*, 2007 ; Josselin *et al.*, 2008 ;

Mahfoud *et al.*, 2009), en faisant l'hypothèse que le MAUP a le même effet dans les deux organisations spatiales des données (observées *versus* aléatoires). Dans cette approche, les auteurs recherchent ce qu'ils appellent «l'échelle pertinente» de mesure d'un indice statistique, qui est celle qui maximise cet indice, une fois l'effet du support spatial supposé éliminé par l'analyse des échelles, c'est à dire celle qui montre le plus grand écart à une distribution de ces données qui serait aléatoire. Enfin, l'analyse de sensibilité peut être également conçue comme la sélection non pas d'un découpage mais d'undifférents découpages, si un résultat est stable malgré les changements de limites ou d'échelle, il est possible d'affirmer que ce résultat est plus robuste et plus pertinent qu'un résultat sensible à la manière dont les données sont agrégées (Fotheringham *et al.*, 2000).

Ce papier propose d'appliquer l'analyse de sensibilité, en particulier :

– la sélection du résultat selon sa variabilité en fonction des découpages (Figure 4) ;

– la sélection d'un découpage dont les résultats sont les plus différents des données re-échantillonées aléatoirement (Figure 5).

Il s'agit uniquement d'une analyse de sensibilité au problème spécifique qu'est le MAUP, et plus particulièrement à son effet d'échelle (Figure 1). La question de la qualité des données de départ n'est pas prise en compte dans cette méthode, ni à proprement parler la proposition d'un modèle explicatif et reproductible des feux de forêt.

3. Données et méthode

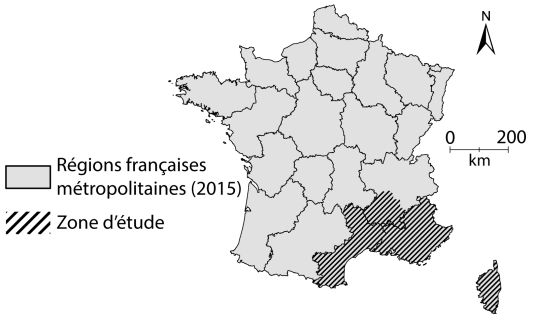

Figure 3. Localisation de la zone d'étude

Nos données [1] (variables dépendantes et explicatives) ont d'abord été obtenues par communes, puis ont été agrégées. Nous avons utilisés les découpages territoriaux français sélectionnés dans la base de données Géofla 2014 de l'IGN

1. Les scripts et les données utilisées pour obtenir nos résultats sont accessibles à l'adresse suivante : https://github.com/romain-louvet/sageo_rig2015.git

à partir de la zone couverte par la base Prométhée (régions Corse, Languedoc-Roussillon, Paca, et départements de l'Ardèche et de la Drôme, voir Figure 3). Les codes d'identification des niveaux de découpages proviennent de Géofla, hormis les EPCI (établissement public de coopération intercommunale) de 2014 qui ont été extraits du site *collectivites-locales.gouv.fr*. Grace à ces codes, nous avons généré six niveaux d'échelle, sous la forme de *shapefiles* distincts. La zone d'étude a ainsi été découpée en 3571 communes (LAU2), 476 cantons (LAU1), 244 EPCI, 43 arrondissements, 15 départements (NUTS3) et 4 régions (NUTS1), dont les superficies moyennes calculées sont respectivement de 22, 169, 1 868, 5 356, et 20 084 km^2.

La base de données de départ est un extrait de Prométhée, base officielle d'enregistrement des incendies dans la zone méditerranéenne française, pour la période de 1997 à 2013. Cette base de données fut créée suite à la décision en 1973 de l'État français de se doter d'un outil de recensement des feux de forêts du sud-est de la France. Chaque incendie y est enregistré individuellement. Son point d'éclosion est localisé à l'échelle communale et au carreau DFCI (carreaux de 2 km de côté). Ces données sont librement téléchargeables, toutefois le carroyage DFCI n'est accessible que par demande [2]. C'est à partir de cette base que nous avons calculé nos variables dépendantes par unité spatiale : nombre de feux, surfaces incendiées, taille moyenne des feux (surface divisée par le nombre), densité du nombre des feux (nombre par km^2), et densité des surfaces incendiées (m^2 de surfaces incendiées par km^2). Enfin, les variables dépendantes ont été log transformées afin de suivre un modèle paramétrique (Ganteaume, Jappiot, 2013). Nous avons utilisé 14 variables explicatives : la densité de population par km^2, la densité routière par km^2, la densité ferroviaire par km^2, le taux de chômage, le nombre de lits touristiques par habitants, le nombre de lits touristiques par km^2, le taux d'évolution du cheptel et de la surface agricole utile, et six variables d'occupation du sol (taux d'occupation sur l'aire totale). Ces variables ont été choisies car elles sont souvent mentionnées comme facteurs potentiellement déclenchant des feux de forêts pour la zone d'étude (Ganteaume, Jappiot, 2013).

La densité de population a été calculée à l'aide de la moyenne des données de recensement de l'INSEE de 1999, 1997 et 2012. La densité routière et la densité ferroviaire ont été obtenues par croisement des surfaces des unités spatiales avec les tronçons routiers et ferroviaires de la base Route 500 de l'IGN de 2012. Le taux de chômage est une moyenne des taux de 1999, 2066 et 2011, des données INSEE. Les lits touristiques sont des données INSEE de 2013. Il s'agit d'une capacité d'accueil d'hébergement touristique exprimée en lits selon la méthode de calcul de l'INSEE. Le nombre de lits touristiques par habitants a été estimé avec la moyenne de population de 1999 à 2012. L'évolution du cheptel et de la surface agricole utile sont issues du recensement agricole. Il s'agit du taux de

2. Pour plus d'informations, voir http://www.promethee.com

variation entre 1988 et 2010. Enfin, les données d'occupation du sol sont issues des statistiques de Corine Land Cover à la commune. Il s'agit de moyennes de l'occupation du sol de 2000 révisée et de 2006. Les variables qui ont été utilisées sont la part sur le total de la surface de l'unité spatiale : des terres arables ; des cultures permanentes, cultures annuelles associées et agroforesterie ; des friches agricoles ; des forêts ; des landes, broussailles, végétations sclérophylles et végétations arbustives et/ou en mutation ; des prairies, pelouses et pâturages naturels.

Presque la totalité des traitements a été réalisée sous R, à l'exception du calcul des densités routières et ferroviaires effectué avec le logiciel ArcGIS à l'aide d'un script en Python. R a été choisi comme outil principal pour ses capacités de traitements statistiques et son intégration de fonctions SIG au sein d'une chaîne de traitements unique (Commenges *et al.*, 2014). Ces traitements incluent : l'utilisation de packages spécifiques pour les données spatiales (*rgeos* et *rgdal*), le chargement et le pré-traitement des données à l'échelle des communes, puis l'agrégation des variables, la création d'un nouveau *shapefile* par niveau d'échelle, et d'une liste d'objets spatiaux correspondant aux différents niveaux d'échelle. Ensuite, à partir de cette liste, des corrélations ont été calculées pour l'ensemble des variables à chaque niveau d'échelle. Enfin, le ré-échantillonnage aléatoire des données a été effectué pour les variables explicatives. Plus de 150 000 points aléatoires correspondant au nombre de feux dans la base Prométhée ont été générés 100 fois à l'aide de la fonction *spsample()* et agrégé par unité spatiale grâce à la fonction *over()*. Les corrélations ont ensuite été calculées pour chacun des 100 ré-échantillonnages aléatoires et les résultats enregistrés dans une liste afin d'en extraire le minimum, maximum, et la moyenne par niveau d'échelle et par corrélation. Lorsque plusieurs niveaux d'échelle possèdent un coéfficient de corrélation significatif, nous avons calculé la différence avec le coefficient maximum calculé à partir des données aléatoires de deux manières : une différence simple et une différence relative (différence divisée par le coefficient de corrélation de l'échelle considérée).

4. Résultats et discussion

Si nous analysons la variation générale des résultats de la corrélation à différents niveaux d'échelle, nous observons une relation forte entre le nombre d'unités spatiales par niveau d'échelle, la part des corrélations significatives et l'intensité de la corrélation. Plus le niveau d'échelle est haut (c'est-à-dire moins il y a d'unités spatiales) et, en moyenne, moins les r^2 sont significatifs et plus ils sont grands. A partir des 5 variables dépendantes et des 14 variables explicatives, nous avions 70 corrélations à calculer pour six niveaux d'échelle, soit au total 420 corrélations. Sur 420, 173 corrélations sont significatives (p-value inférieure à 0.05). Parmi les relations significatives, nous avons 44 relations à la commune, 42 au canton, 39 à l'EPCI, 28 à l'arrondissement, 15 au départe-

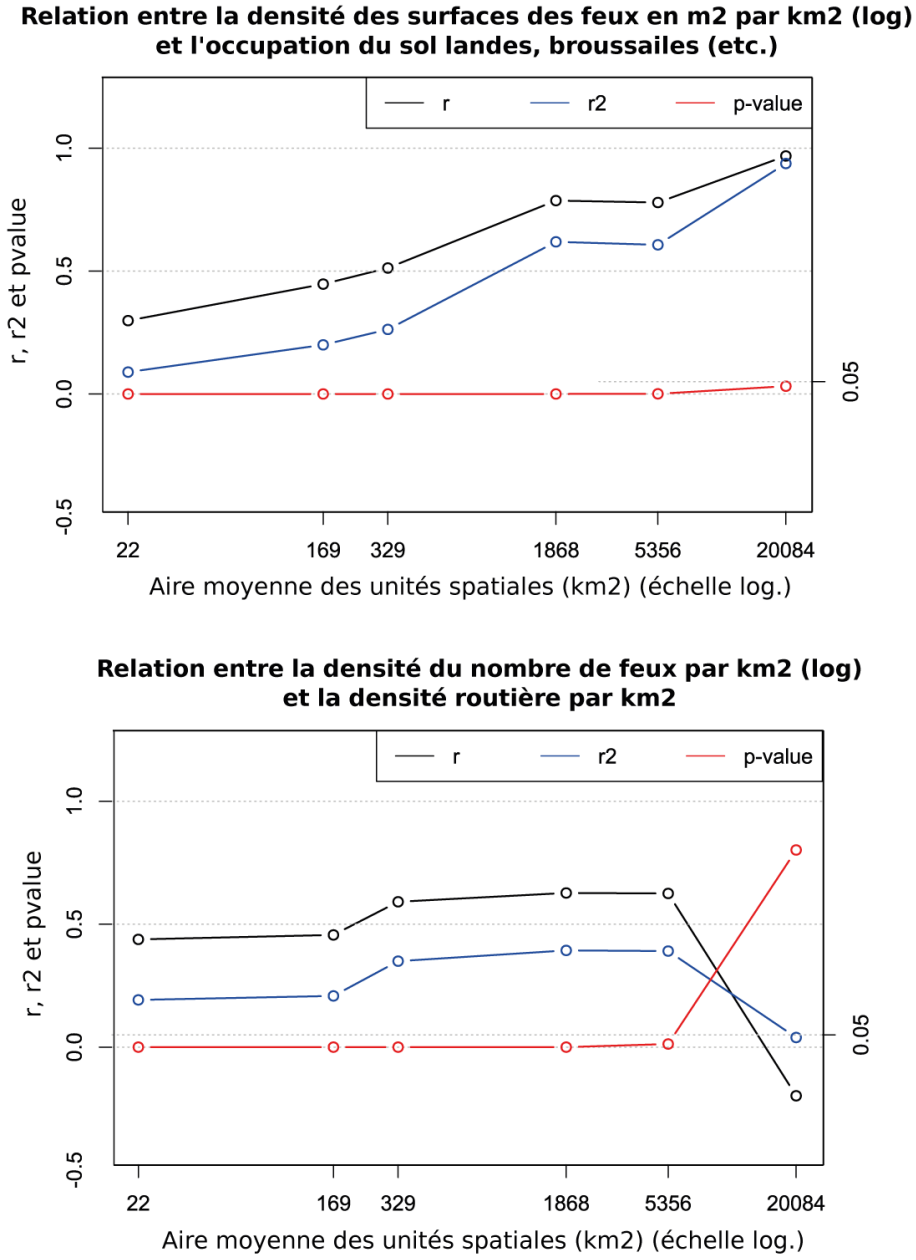

Figure 4. Exemples des résultats de visualisation des calculs de corrélation (r, r^2, p-value) en fonction des niveaux d'échelle (représentés par les moyennes des unités spatiales)

ment, et 5 à la région. Ce qui fait respectivement sur le nombre de relation par niveau d'échelle un pourcentage de 63 %, 60 %, 55 %, 40 %, 21.4 %, et 7 %.

Seuls 37 r^2 sont supérieurs ou égaux à 0.25, dont seulement 1 à la commune, 6 au niveau des EPCI, 10 des arrondissements, 15 des départements et 5 des régions. Seuls 11 r^2 sont supérieurs ou égaux à 0.5, dont 2 arrondissements, 4 départements, et 5 régions. Parmis eux, 5 sont des coefficients de détermination très élevés, supérieurs ou égaux à 0.9, tous au niveau des régions. Le r^2 moyen significatif par échelle est de 0.047, 0.087, 0.135, 0.253, 0.410, et 0.931. Ainsi nous observons que pour un même jeu de données, il y a environ 9 fois plus de chance de trouver une relation à la commune qu'à la région, 3 fois plus de chance à la commune qu'au département. Parallèlement, le r^2 est tendanciellement 20 fois plus grand à la région qu'à la commune et 9 fois plus grand au département qu'à la commune. L'effet d'échelle sur la corrélation est généralement une augmentation du coefficient de corrélation avec l'augmentation de la taille des unités spatiales (Blalock M, 1964), ce que confirme nos résultats.

Après cette vision globale, si nous prenons en compte uniquement les couples de variables dépendantes et explicatives (sans tenir compte des niveaux d'échelle), nous observons 59 relations significatives à au moins un niveau. En visualisant la variation des résultats de la corrélation à différentes échelles (Figure 4), il est possible de détecter des tendances, des anomalies et de sélectionner des corrélations stables à plusieurs échelles. En effet, certaines relations seraient stables alors que d'autres resteraient très sensibles à l'effet d'échelle (Fotheringham, Wong, 1991). Commencer par sélectionner les corrélations significatives à plusieurs niveaux d'échelles permet d'écarter un grand nombre de relations, aux résultats variant trop. Les relations significatives à un seul niveau d'échelle sont en effet les plus fréquentes, avec 18 cas, soit 30 %. Mais arrivent ensuite les relations significatives à 5 niveaux d'échelles qui représentent tout de même 20 % et les relations significatives à deux niveaux d'échelle (19 %). Seulement 3 relations sont significatives à toutes les échelles.

Nous avons retenus deux tendances de relations stables : des relations qui augmentent fortement et des relations qui augmentent légèrement (Figure 4). Parmi les relations qui augmentent fortement, la relation la plus forte est celle entre l'occupation du sol des landes, broussailles (etc.) et les densités de surfaces des feux de forêt : trois niveaux d'échelle dont le r^2 est supérieur à 0.5. La deuxième relation la plus intense qui augmente est celle entre le taux de chômage et le nombre de feux de forêt, avec quatre niveaux d'échelle dont le r^2 est supérieur à 0.25. Bien que ces relations puissent être fortes (r^2 proche de 1 à l'échelle régionale, voir premier cas de la Figure 4), comme nous constatons un accroissement important du coefficient de détermination avec les niveaux d'échelle, nous pouvons conclure à une forte influence de l'effet d'échelle sur ces résultats qu'il faudrait donc écarter au profit des relations n'augmentant que faiblement. Parmi les relations à faible augmentation, nous avons retenu trois relation qui possèdent trois niveaux d'échelles avec un r^2 supérieur à 0.25 :

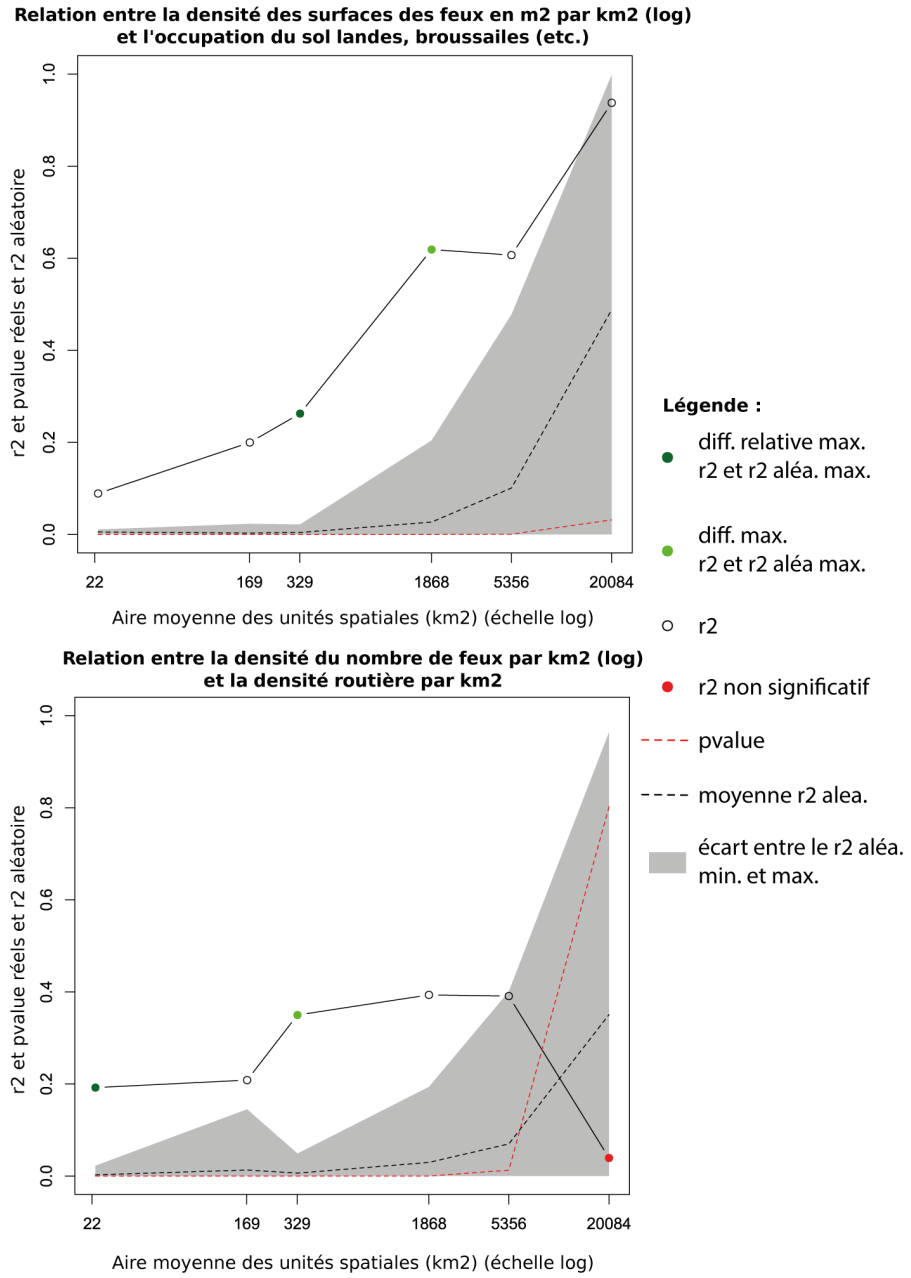

Figure 5. Exemples de sélections d'un niveau d'échelle "pertinent" et de son r^2 en fonction de la distance à un résultat aléatoire

la densité routière et la densité du nombre de feux (Figure 4), les cultures permanentes et la densité du nombre de feux, la part des forêts dans l'occupation du sol et la densité du nombre de feux. Ces relations, bien que moins fortes, devraient être privilégiées dans l'analyse car elles sont moins sensibles à l'effet d'échelle.

Enfin, pour compléter la sélection des résultats en fonction de leur sensibilité au MAUP, nous proposons déterminer si une échelle est plus ou moins pertinente en fonction de sa proximité avec des résultats aléatoires (Figure 5). Ainsi, pour les deux relations que nous avons retenues de notre exemple, nous pouvons nuancer ce que nous avons dit jusqu'à présent. En effet, dans les deux cas, le ou les derniers niveaux d'échelle significatifs sont très proches ou inclus dans l'intervalle gris représentant l'écart entre le maximum et le minimum des coefficients de détermination calculés aléatoirement. Bien que stable, la relation entre la densité routière et la densité du nombre des feux est donc potentiellement similaire à un résultat aléatoire et uniquement le fait de l'échantillonnage spatial à l'échelle des départements. Elle n'est en cela plus différente de la relation augmentant fortement avec les échelles. *A contrario*, cette méthode permet de sélectionner deux niveaux d'échelle (EPCI et arrondissements) pour la relation augmentant fortement (relation entre les landes et la densité des surfaces des feux) grâce au fait que les r^2 de ces niveaux sont particulièrement différents du résultat aléatoire (en différence absolue et relative).

5. Conclusion

Nos résultats mettent en évidence l'effet d'échelle du MAUP sur l'étude des causes des feux de forêt. Ils montrent que la corrélation dépend en grande partie du niveau d'échelle et donc que si une analyse est conduite à un seul niveau, celle-ci prend le risque de trouver une relation significative qui ne le serait pas à un autre niveau d'échelle et inversement, ou une relation plus ou moins intense. L'intensité de la relation semble en grande partie dépendre du niveau d'échelle. Toutefois, l'explication tient sans doute au fait que le nombre d'individus a une forte influence sur la résultat de la corrélation. Pour pallier à ce problème de variation du résultat, nous avons ainsi proposé, d'une part, de visualiser la variation de la corrélation en fonction des niveaux échelles (Figure 4), et, d'autre part, de visualiser la différence de ces résultats avec une corrélation calculée à partir de données issues d'un re-échantillonnage aléatoire (Figure 5). Afin de poursuivre ce travail, cette approche pourrait être complétée par un test appliqué à d'autres zones d'étude, afin d'observer si nous obtenons des résultats équivalents avec des variables similaires ou avec d'autres variables. Il serait également nécessaire d'étendre la méthode à l'étude de l'effet de zonage et chercher à implémenter des solutions optimales fondées sur la délimitation de nouveaux découpages en unités spatiales. Par ailleurs, le re-échantillonnage aléatoire, ici effectué uniquement sur les variables dépendantes, pourraient être amélioré en ajoutant les variables explicatives.

Bibliographie

Arsenault J., Michel P., Berke O., Ravel A., Gosselin P. (2013). How to choose geographical units in ecological studies: Proposal and application to campylobacteriosis. *Spatial and Spatio-temporal Epidemiology*, vol. 7, p. 11-24.

Blalock M H. (1964). *Causal inferences on nonexperimental research.* Chapel Hill, NC: University of North Carolina Press.

Charleux L. (2005). GWR, MAUP et lissage par potentiels. *Revue Internationale de Géomatique*, vol. 15-2, p. 195-209.

Commenges H., Beauguitte L., Buard E., Cura R., Le Néchet F., Le Texier M. *et al.* (2014). *R et espace : Traitement de l'information géographique.* Groupe ElementR, Framabook, Paris.

Fotheringham A. S., Brunsdon, C. M., Charlton. (2000). *Quantitative geography: Perspectives on spatial data analysis.* SAGE.

Fotheringham A. S., Wong D. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, vol. 23, p. 1025-1044.

Ganteaume A., Jappiot M. (2013). What causes large fires in southern france. *Forest Ecology and Management*, vol. 294, p. 76-85.

Gehlke C., Biehl H. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, vol. supplement 29, p. 169-170.

Holt D., Steel D., Tranmer M., Wrigley N. (1996). Aggregation and ecological effects in geographically based data. *Geographical Analysis*, vol. 28, p. 244-261.

Hui C. (2009). Foundations of computational intelligence. In A.-E. Hassanien, A. Abraham, F. Herrera (Eds.),, vol. 2, p. 175-196. Springer Berlin Heidelberg.

Josselin D., Mahfoud I., Fady B. (2008). Impact of a change of support on the assessment of biodiversity with shannon entropy. In *Spatial Data Handling, SDH'2008"*, p. 109-131. Montpellier, June, 23-25.

King G. (1997). *A solution to the ecological inference problem. reconstructing individual behaviour from aggregate data.* Princeton University Press.

King G., Rosen O., Tanner A. M. (Eds.). (2004). *Ecological inference. new methodological strategies.* Cambridge University Press.

Mahfoud I., Josselin D., Fady B. (2007). Sensibilité des indices de diversité à l'agrégation. *Revue Internationale de Géomatique*, vol. 3-4, p. 293-308.

Mahfoud I., Josselin D., Fady B. (2009). Analyse exploratoire des effets de support spatial et de robustesse statistique sur la fiabilité de la mesure de la (bio)diversité. *Photo-interprétation / European Journal of Applied Remote Sensing*, vol. 45, p. 3-11;35-41.

Openshaw S. (1984). *The modifiable areal unit problem.* Norwich: Geo Books, CATMOG 38.

Openshaw S., Taylor P. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In N. Wrigley (Ed.),, p. 127-144. Statistical Applications in the Spatial Sciences, London: Pion.

Robinson A. (1956). The necessity of weighting values in correlation analysis of areal data. *Annals of the Association of American Geographers*, vol. 46, p. 233-236.

Robinson W. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, vol. 15, p. 351-357.

Simpson E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society - Series B (Methodological)*, vol. 13-2, p. 238-241.

Steel D., Holt D. (1996). Rules for random aggregation. *Environment and Planning A*, vol. 28, p. 957-978.

Swift A., Liu L., Uber J. (2008). Reducing maup bias of correlation statistics between water quality and gi illness. *Computers, Environment and Urban Systems*, vol. 32, n°2, p. 134-148.

Tranmer M., Steel D. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A*, vol. 30, p. 817-831.