Choix d'un polygone régulier comme unité géographique pour l'agrégation de données en statistique spatiale

Sylvain Coly [1] & Benoît Coly [2] & Anne-Françoise Yao-Lafourcade [3] & David Abrial [4] & Myriam Charras-Garrido [5]

[1,3] *Université Blaise Pascal - Laboratoire de Mathématiques*
UMR 6620 - CNRS
Campus des Cézeaux
BP 80026
63171 Aubière CEDEX
France

[1] *sylvain.coly@clermont.inra.fr*
[3] *Anne-francoise.Yao@math.univ-bpclermont.fr*

[1,4,5] *Centre INRA Auvergne/Rhône-Alpes*
Unité d'Épidémiologie Animale
Route de Theix
63122 Saint-Genès-Champanelle

[4] *david.abrial@clermont.inra.fr*
[5] *myriam.charras-garrido@clermont.inra.fr*

[2] *Lycée Montdory*
7 ter Avenue Jean Jaurès
63304 Thiers

[2] *benoit.coly@gmail.com*

Résumé. En statistique spatiale, on peut étudier des données ponctuelles ou agrégées. Selon le contexte et les objectifs, il peut être pertinent d'agréger des données ponctuelles. Les unités administratives sont alors souvent utilisées, mais elles présentent plusieurs inconvénients (grande hétérogénéité de taille, population, longueur de frontières, ...) et peuvent ainsi introduire un biais dans les analyses. Si l'on souhaite utiliser un polygone régulier comme unité géographique, il existe trois possibilités : le triangle équilatéral, le carré et l'hexagone régulier. Si le carré est satisfaisant, l'hexagone est la meilleure forme pour agréger des données, car la plus proche du cercle, c'est-à-dire de la situation où tous les points situés à la frontière sont équidistants du centre de gravité. L'hexagone présente par ailleurs de meilleures propriétés de stabilité par des isométries : l'orientation du pavage est donc moins influente. À partir de chacun de ces pavages, on peut construire différentes relations de voisinage pour prendre en compte les corrélations spatiales. La stratégie la plus cohérente avec la distance euclidienne est de considérer un quadrillage ou

un pavage triangulaire munis d'une relation de voisinage où les voisinages par les sommets sont pris en compte et pénalisés par une pondération. Il faut noter, *a contrario*, que les voisinages simples (côté commun) induisent une complexité plus faible, et donc des temps de calcul plus courts. En conclusion, l'hexagone est un bon compromis global, même si, selon le contexte, on pourra privilégier différentes stratégies de pavage.

Mots-clés. Statistiques spatiales, Données agrégées, Unités géographiques, Polygones réguliers, Qualité d'agrégation, Corrélations spatiales, Complexité de modèle.

Abstract. In spatial statistics, one can study point data or aggregated data. Depending on the context and objectives, it may be suitable to aggregate point data. The administrative units are often used, but they have many drawbacks (very different shapes, size, population size, length of borders, ...) that can introduce bias into the statistical analyses. It could be relevant to use a regular polygon as the geographical unit; the equilateral triangle, the square and the regular hexagon are the only three possibilities. If the square is appropriate to synthetise information at a point, the hexagon is the best; in fact it is the closest of the circle shape, for which all points on the border are equidistant from the center of gravity. The hexagon is also the most stable (by isometries) polygon: the orientation of the tiling ilmplies a lower bias. For each of these polygons, we can build different neighbor relations to take into account the spatial correlations. The strategy which is the most congruent with the Euclidean distance is to consider square or triangular tilings with neighbor relations which allow the adjacency by the edge and include weight coefficients. It should be noted that simple neighbor relations (edge in common) imply more simple models and shorter computation durations. As a conclusion, the hexagon is a good compromise, even if one should rather consider other strategies to tile the plan depending on the context of its study.

Keywords. Spatial statistics, Aggregated data, Geographical units, Regular polygons, Aggregation quality, Spatial correlations, Model complexity.

1 Agrégation de données ponctuelles et pavage du plan par des polygones réguliers

Données ponctuelles et agrégées en statistique spatiale. En statistique spatiale, on analyse des données ponctuelles ou agrégées. Les données ponctuelles indiquent exactement où s'est produit un événement (survenue d'une avalanche, apparition d'un cas de grippe aviaire, etc.), ou où a été échantillonnée une valeur mesurée (température mesurée en un point, hauteur d'un arbre, etc.). Les données agrégées à des unités géographiques sont en fait des comptages d'événements (nombre de cambriolages par an et par département, cas de grippe par commune, etc.) ou des calculs d'indicateurs

(IDH d'un pays, résultats électoraux pour une municipalité, etc.) associés à un ensemble de zones données. Dans ce cadre, les entités géographiques considérées sont souvent administratives, mais peuvent être aussi environnementales (pré, forêt, etc.) ou populationnelles (bassin d'activité, etc.) par exemple.

Méthodes associées aux données ponctuelles et agrégées. Les analyses statistiques effectuées sur un jeu de données spatiales dépendent de leur quantité, de la problématique d'étude, mais également du type de ces données. Ainsi, les données ponctuelles permettent la recherche d'agrégats, le lissage, le kriegeage ou encore les analyses par graphe. Les données agrégées sont, quant à elles, utilisées en classification, en cartographie du risque, pour le calcul d'indicateurs permettant de qualifier globalement la région d'étude (indice de Moran, indice de Tango, ...) ou encore pour les méthodes de dépassement de seuil (cartes CUSUM, ...). Les objectifs peuvent être identiques (identification de phénomènes anormaux par exemple), mais dans la plupart des cas, les méthodes diffèrent (détection d'agrégats/cartes de contrôle).

Agrégation des données ponctuelles. L'objectif d'une étude peut parfois inciter à transformer les données ponctuelles en données agrégées, auquel cas on peut librement choisir l'unité géographique. On considère souvent des unités administratives, ce qui facilite l'interprétation des résultats. Néanmoins leur irrégularité peut introduire un biais dans l'analyse. En effet, leur forme, leurs dimensions et la longueur de leurs frontières peuvent être très variables; en outre la position de leur "centroïde" (point représentant l'ensemble de la zone) peut également être source de biais dans le prise en compte des dépendances spatiales. Il peut être préférable de considérer des maillages réguliers, constitués de formes géométriques identiques et pouvant paver le plan. On pourra alors choisir leur centre de gravité comme centroïde.

Pavages possibles. Seuls trois polygones réguliers ont la propriété géométrique de paver le plan : le triangle équilatéral, le carré et l'hexagone régulier. Classiquement, le carré est la forme la plus utilisée pour le maillage de zones géographiques. Bien que le quadrillage soit le pavage le plus fréquemment rencontré en statistique spatiale, rien *a priori* ne devrait inciter à privilégier une forme plutôt qu'une autre.

2 Représentativité des centroïdes pour les unités géographiques

Représentativité du centre de gravité pour l'ensemble des points du polygone. L'objectif dans le choix d'un maillage, est que le centre de gravité des polygones considérés soit le plus représentatif possible des données qui vont lui être assimilées. On remarque au

premier abord que les demi-diagonales et demi-médianes ont des longueurs très proches pour l'hexagone. Ainsi les points les plus proches et les plus éloignés du centre ont une très importante différence de distance au centre dans le cas du carré, et surtout du triangle équilatéral par rapport à l'hexagone. Par ailleurs, la forme même de l'hexagone est celle qui est la plus proche de ses cercles inscrits et circonscrits : le rapport entre les aires des cercles inscrits et circonscrits vaut 1/4 dans le cas du triangle équilatéral, 1/2 pour le carré et 3/4 pour l'hexagone. Enfin, si on place un point aléatoirement dans le domaine que constitue un polygone régulier, (pris avec une aire de 1) il sera approximativement à même distance du centre en moyenne pour le cercle, l'hexagone régulier et le carré. En revanche, cette distance est presque 2 fois supérieure pour le triangle équilatéral ; ainsi, les points situés dans les "pointes" du triangle pèsent lourdement dans leur mauvaise représentation par le centre. On eut donc conclure que l'hexagone régulier est le polygone le plus pertinent dans lequel agréger des données, même si le carré reste une option valable.

Stabilité des polygones par les isométries du plan. Un autre critère de comparaison est de privilégier une forme ayant les propriétés les plus proches du cercle. Le triangle équilatéral comporte trois axes de symétrie et pas de symétrie centrale, au contraire du carré et de l'hexagone qui présentent par ailleurs respectivement quatre et six symétries axiales. Ces considérations incitent à privilégier l'hexagone. Elles sont importantes dans la mesure où plus un polygone est stable par des isométries du plan, plus le choix de l'orientation du pavage obtenu a une faible influence. Ainsi, l'hexagone engendre moins d'*a priori* sur les résultats des analyses statistiques.

3 Voisinages, distances et corrélations spatiales

Voisinages possibles. En statistique spatiale, la plupart des analyses sur données agrégées les munit d'une structure de voisinage. Cette structure prend en compte les similarités entres unités géographiques proches (environnementales, populationnelles, ...) et l'influence que chacune peut exercer sur ses voisines. Le choix de l'unité géographique doit intégrer cette notion de voisinage. Pour chaque polygone, plusieurs stratégies sont possibles. Pour le triangle équilatéral, on peut considérer que deux cellules sont voisines : **(T1)** si elles ont seulement un côté commun ; **(T2)** si elles ont un côté ou un sommet en commun ; **(T3)** si elles ont un côté ou un sommet en commun mais que l'on pénalise par un coefficient de pondération les relations entre sommets. De même que pour le triangle équilatéral, on considère que deux carrés sont voisins : **(C1)** si elles ont seulement un côté commun ; **(C2)** si elles ont un côté ou un sommet en commun ; **(C3)** si elles ont un côté ou un sommet en commun mais que l'on pondère les relations entre sommets. Concernant l'hexagone, le seul voisinage possible est l'adjacence par le côté **(H)**. On peut associer à un pavage une structure de graphe en considérant chaque polygone comme un

nœud qui sera relié par une arète à chaque polygone dont il est le voisin. On peut alors définir, pour chaque graphe, une distance entre deux centroïdes comme le chemin le plus court les reliant au sens de la théorie des graphes.

Adéquation entre distance euclidienne et distance induite par les relations de voisinages. On peut juger de la bonne qualité d'un pavage muni de sa relation de voisinage au fait que la distance qu'il induit soit cohérente avec la distance euclidienne. Pour cela, on munit le plan des trois pavages (T), (C) et (H) (tels que la distance entre deux centres de polygones adjacents soit de 1), et on simule des couples de points selon une loi uniforme dans un cercle de rayon 10. On mesure alors les écarts entre la distance euclidienne et les 7 distances induites par les structures de voisinage. Les résultats obtenus par simulation montrent que les approches (T1) et (C1) surestiment fortement les distances, respectivement de 27% et 29% en moyenne. *A contrario*, les approches (T2) et (C2) sous-estiment les distances entre les points, de 35% et 10% en moyenne. Les hexagones (H) induisent une surestimation de 12% de la distance en moyenne (avec par ailleurs une variance beaucoup plus faible que pour les 4 modèles précédents). Les approches par pondération améliorent grandement les résultats, avec une surestimation de la distance de seulement 4% pour (C3) et 2% pour (T3). On voit ainsi l'intérêt de considérer des modèles complexes, incluant des pondérations, pour une bonne modélisation de la vraie distance.

Boules associées aux distances induites par les voisinages. Dans la plupart des analyses en statistiques spatiales, les données sont fortement influencées par un ou quelques point(s). Il est important que les zones d'influence soient les plus proches possibles du cercle afin que certaines directions ne soient pas plus fortement impactées que d'autres, selon l'orientation du pavage. On a donc étudié les boules associées aux distances induites par les voisinages. On remarque que les boules associés (C1) et (C2) sont des carrés, ce qui est logique dans la mesure où leurs distances associées sont celles des normes $\|.\|_1$ et $\|.\|_\infty$. Les boules associées (T1) et (T2) sont des pseudos-hexagones de dimensions peu cohérentes avec la distance euclidienne. *A contrario* l'hexagone et le pseudo-hexagone associés (H) et (T3) respectivement sont de taille cohérente avec le cercle de même rayon. On constate enfin que la boule associée à (C3) est un pseudo-octogone, forme plus proche du cercle que le carré et l'hexagone, ce qui en fait le meilleur modèle pour traiter de données fortement influencées par un point.

4 Complexité et temps de calcul

Considérons des pavages constitués du même nombre de polygones. La complexité algorithmique des méthodes statistiques, et donc la durée de calcul associée, seront alors

fonctions (linéaires, quadratiques, ...) du nombre de voisins que l'on considère pour chaque polygone. Selon les méthodes utilises, minimiser les temps de calcul peut être un enjeu crucial; il est alors important de considérer des modèles les plus simples possibles. Le schéma (T1) est le plus simple, (C1) et (H) aussi ont peu de voisins. (C2) et (T2) sont un peu plus complexes en raison de l'augmentation du nombre de voisins, surtout pour (T2) (12 contre seulement 3 pour (T1)). Les approches (C3) et surtout (T3) sont beaucoup plus complexes, en raison de l'ajout des coefficients de pondération.

Il faut également remarquer que l'ajout de ces coefficients complique (voire rend impossible) la mise en place de certaines méthodes, que ce soit au niveau de leur implémentation informatique, ou de l'applicabilité même de la méthode statistique. Le pavage hexagonal, et surtout le triangulaire, sont très rarement utilisés, et il peut être impossible de les utiliser dans certains cadres. Par ailleurs, le quadrillage est beaucoup plus fréquemment utilisé, et donc beaucoup plus systématiquement implémenté sur les logiciels. Il faut noter que le quadrillage est beaucoup plus intuitif et cohérent avec le système de coordonnées habituellement utilisé pour référencer les points. Il permet donc de simplifier les calculs en général. Il sera donc généralement plus simple d'appliquer des méthodes sur des données agrégées à des unités carrées.

5 Conclusion

Il est préférable d'utiliser le quadrillage lorsque simplicité et praticité d'usage sont les principaux critères de choix ; dans ce cas (C2) est légèrement meilleur que (C1) dans la prise en compte des corrélations spatiales. Cependant, si l'on souhaite utiliser le quadrillage, (C3) est l'idéal, par la très grande cohérence entre la distance qu'elle induit et la distance euclidienne. En outre, l'approche (C3) est la meilleure si on constate (ou suppose) une forte influence d'un (ou quelques) point(s) sur l'ensemble des données; ceci peut notamment survenir quand on considère une capitale dans la région d'étude et que le phénomène d'intérêt est lié aux activités humaines. En revanche, (T3) est le meilleur choix si l'enjeu principal est la prise en compte des distances ; cependant il s'agit de la structure la plus complexe, et l'utilisation d'une grille triangulaire est synonyme de médiocre qualité d'agrégation. Les stratégies (T1) et (T2) à proscrire, sauf volonté de modèle extrêmement simple ; mais on privilégiera alors (C1) en général, puisque le carré est meilleur que le triangle équilatéral pour agréger les données. Le pavage (H) est le meilleur choix si le bon rattachement d'un phénomène à son unité spatiale est le paramètre le plus important. Ainsi, il semble que le choix d'un pavage doive essentiellement dépendre des données et des objectifs de l'étude. Néanmoins, (H) est un compromis intéressant entre simplicité (seulement 6 voisins, pas de coefficient de pondération) et bon ajustement de la distance euclidienne, malgré son assez faible usage dans la littrature.