Détection d'agrégats spatiaux pour données ponctuelles

Lionel Cucala

Détection d'agrégats spatiaux pour données ponctuelles

Lionel Cucala

Institut de Mathématiques et de Modélisation de Montpellier
Université Montpellier 2
Place Eugène Bataillon
34095 Montpellier

Résumé

Nous proposons une méthode originale de détection d'agrégats pour données ponctuelles dans le cadre spatial. Cette problématique se retrouve notamment dans le domaine de l'épidémiologie, lorsque l'on souhaite identifier les régions dans lesquelles le nombre de cas d'une maladie est anormalement élevé. La procédure classique consiste à calculer la statistique de balayage ("scan statistic") de Kulldorff (1997): l'indice de concentration utilisé est basé sur un rapport de vraisemblance et cette concentration est maximisée sur une collection de cercles, voire d'ellipses, répartis sur la zone d'observation. La première nouveauté consiste à s'appuyer sur une collection d'agrégats possibles de taille limitée et dont les éléments sont de formes très diverses. Pour celà, nous utilisons la méthode de Bar-Hen *et al.* (2007) qui consiste à associer au processus ponctuel une collection de graphes associés et nous limitons les agrégats potentiels à l'ensemble des composantes connexes de ces graphes. La seconde nouveauté est l'introduction d'un nouvel indice de concentration spatiale adapté de celui introduit par Cucala (2008) dans un cadre temporel. L'application de ces méthodes à des jeux de données épidémiologiques, ainsi qu'une étude de simulation, montrent que ces deux méthodes ont des performances similaires à la statistique de balayage classique alors qu'elles sont beaucoup plus rapides. On s'aperçoit également que l'indice de concentration spatiale introduit est sensiblement meilleur que celui utilisé habituellement.

Mots-clés: Epidémiologie, détection d'agrégats, données ponctuelles, théorie des graphes.

Abstract

We propose an original method for detecting clusters in spatial point processes. This problem is common in an epidemiological context, when one wants to identify the regions where the rate of a certain disease is abnormally high. The classical method is the computation of the spatial scan statistic introduced by Kulldorff (1997): the concentration

index is based on a likelihood ratio and this concentration is maximised over plenty of circular or elliptic windows over the whole area. Our first idea consists in introducing a small family of data-based possible clusters. For this, we rely on the method introduced by Bar-Hen *et al.* (2007) which associates to the point process a collection of graphs and we consider only the connected components of these associated graphs as potential clusters. The second idea is the use of a new spatial concentration index which is a spatial version of the one introduced by Cucala (2008) in the temporal setting. These methods perform well when applied to epidemiological data sets and a simulation study shows that they are almost as powerful as the classical scan statistic but much less computationnally demanding. It is also obvious that the new spatial concentration index is more efficient than the classical one.

Keywords: Epidemiology, cluster detection, point processes, graphs theory.

1 Introduction

De plus en plus de jeux de données, notamment en épidémiologie, se présentent sous la forme de processus ponctuels observés sur un domaine $D \subset \mathbb{R}^d$, avec $d = 2$ ou 3 généralement: on associe à chaque événement observé (cas de maladie par exemple) sa localisation géographique. Afin d'analyser le processus aléatoire d'apparition de ces événements, on cherche généralement les zones spatiales, appelées agrégats, dans lesquelles la densité de cas est anormalement élevée, en prenant en compte la densité de population sous-jacente $h(x), x \in D$, supposée connue. Pour celà, on introduit tout d'abord l'hypothèse nulle H_0 selon laquelle les localisations spatiales des événements $X_1, \cdots, X_n$ sont indépendentes et réparties selon la densité de population $h(.)$ sur le domaine d'observation D.

2 Une collection d'agrégats potentiels basés sur les données

Nous proposons ici comme agrégats potentiels un ensemble de fenêtres basées uniquement sur les données, de forme non contrainte et dont le nombre est équivalent au nombre d'évènements recensés.

Nous décrivons tout d'abord la technique de Bar-Hen *et al.* (2007) qui associe une collection de graphes au processus ponctuel original. Posons $X_1, \cdots, X_n$ comme dans l'Introduction. Pour tout $\delta \in \mathbb{R}^+$, un graphe, noté $\mathcal{G}(\delta)$, est défini: l'ensemble de ses sommets est $\{1, \cdots, n\}$ et l'ensemble de ses arêtes est $\{(i,j) : d(X_i, X_j) \leq \delta, 1 \leq i \leq n, 1 \leq j \leq n\}$, où $d(.,.)$ est la distance euclidienne. La composante connexe du sommet i dans ce graphe est notée $\mathcal{N}_i(\delta)$. Soit $A_i(\delta) = \{x \in A : \exists j \in \mathcal{N}_i(\delta), d(x, X_j) \leq \delta\}$ le δ-voisinage du sommet i. Bar-Hen *et al.* (2007) cherchent principalement à tester

l'hypothèse d'homogénéité spatiale contre un phénomène d'agrégation globale en s'appuyant sur le nombre de composantes connexes en fonction de δ. Nous préférons nous concentrer sur la recherche d'un agrégat local et pour celà nous décidons d'analyser toutes les composantes connexes issues des graphes $\{\mathcal{G}(\delta), \delta \in \mathbb{R}^+\}$.

Puisqu'un agrégat contient généralement des évènements qui sont proches d'au moins un autre évènement de l'agrégat, il paraît logique de considérer les δ-voisinages comme agrégats potentiels. Nous nous intéressons donc aux zones

$$\{A_i(\delta) : 1 \leq i \leq n,\, \delta \in \mathbb{R}^+\}.$$

A première vue, le nombre de ces zones peut paraître élevé mais il peut être fortement réduit. D'abord, les distances δ à analyser sont juste les distances $d_{i,j} = d(X_i, X_j)$, puisque le graphe $G(\delta)$ reste le même entre deux $d_{i,j}$ consecutifs. Ensuite, une nouvelle arête est ajoutée au graphe $G(\delta)$ lorsque δ atteint $d_{i,j}$ mais les composantes connexes du graphe peuvent rester les mêmes: dans ce cas, la concentration ne peut pas être maximisée pour $\delta = d_{i,j}$. Enfin, seul l'agrégat potentiel $A_i(d_{i,j}) = A_j(d_{i,j})$ doit être analysé lorsque δ atteint $d_{i,j}$ puisque c'est le seul voisinage dans lequel le nombre d'évènements augmente. Soit $\mathcal{G}^-(\delta)$ le graphe dont les sommets sont $\{1, \cdots, n\}$ et les arêtes $\{(i,j) : d(X_i, X_j) < \delta, 1 \leq i \leq n, 1 \leq j \leq n\}$. La composante connexe du sommet i dans ce graphe est notée $\mathcal{N}_i^-(\delta)$. L'ensemble final des agrégats potentiels est donc

$$\mathcal{C} = \{A_i(d_{i,j}) : 1 \leq i < n,\, i < j \leq n,\, \mathcal{N}_i^-(d_{i,j}) \neq \mathcal{N}_i(d_{i,j})\}.$$

On remarque que le nombre d'agrégats potentiels $\mathcal{C}$ est exactement le nombre d'arêtes nécessaires pour lier les n sommets d'un graphe, soit $n-1$.

3 Deux statistiques de balayage spatial basées sur les graphes

Maintenant que l'ensemble des agrégats potentiels est défini, nous devons choisir un indice de concentration pour les comparer. Rappelons les indices utilisés dans le cadre unidimensionnel, c'est-à-dire pour détecter un agrégat d'évenements sur le segment $[0,1]$. Le premier fut introduit par Nagarwalla (1996) et est basé sur le rapport de vraisemblance entre l'hypothèse uniforme et une densité alternative constante par morceaux. Avec cet indice, la concentration d'un intervalle de longueur d contenant m évènements est

$$I^{1D}(d,m) = \Big(\frac{m}{nd}\Big)^m \Big(\frac{n-m}{n(1-d)}\Big)^{n-m} \mathbb{1}\Big(\frac{m}{n} > d\Big),$$

où n est le nombre total d'évènements. Récemment, un autre "hypothesis-free" (HF) indice de concentration a été défini par Cucala (2008), basé sur la distribution nulle de

tous les m-espacements issus du processus ponctuel. Avec cet indice, la concentration d'un intervalle de longueur d contenant m évènements est

$$I_{HF}^{1D}(d,m) = 1/B_{inc}(d, m-1, n+2-m),$$

où $B_{inc}(.,a,b)$ est la fonction Beta incomplete, c'est-à-dire la fonction de répartition de la loi Beta de paramètres a et b.

Dans le contexte spatial, l'incice de concentration classique est celui de Kulldorff (1997), qui est le pendant spatial de celui de Nagarwalla. Avec cet indice, sous l'hypothèse Poissonnienne, la concentration d'une zone Z est

$$I(Z) = \Big(\frac{n(Z)}{n\int_Z h(x)|dx|}\Big)^{n(Z)} \Big(\frac{n-n(Z)}{n\big(1-\int_Z h(x)|dx|\big)}\Big)^{n-n(Z)} \mathbb{1}\Big(\frac{n(Z)}{n} > \int_Z h(x)|dx|\Big),$$

où $n(Z)$ est le nombre d' évènements dans Z et $|.|$ est la mesure de Lebesgue. Evidemment, on peut introduire le pendant spatial de l'indice de concentration de Cucala en remplaçant la longueur de l'intervalle d par la proportion de la population incluse dans Z, $\int_Z h(x)|dx|$. Avec cet indice, la concentration de la zone Z est

$$I_{HF}(Z) = 1/B_{inc}\Big(\int_Z h(x)|dx|, n(Z)-1, n+2-n(Z)\Big).$$

Avec ces deux indices de concentration et l'ensemble des agrégats potentiels défini précédemment, on introduit les deux statistiques de balayage spatial basées sur les graphes

$$\Lambda_{G,1} = \sup_{Z\in\mathcal{C}} I(Z)$$

et

$$\Lambda_{G,2} = \sup_{Z\in\mathcal{C}} I_{HF}(Z).$$

Une fois ces statistiques calculées, leur p-valeur est estimée par une procedure de Monte Carlo, exactement comme la statistique de balayage spatial classique. En effet, connaître la distribution nulle de ces statistiques semble très compliqué.

4 Application à des données épidémiologiques

On applique ces méthodes à un jeu de données décrivant la distribution spatiale de cancers du larynx ou du poumon recensés entre 1973 et 1984 dans une zone du Lancashire, UK, et publié par Diggle *et al.* (1990). Il y a 917 cas de cancer du poumon et 57 du larynx et le lieu de résidence associé est connu. Les cancers du larynx sont moins fréquents et on les soupconne d'être plus nombreux autour d'un incinérateur industriel désaffecté. Comme les cancers du poumon ne semblent pas dépendants d'un facteur environnemental

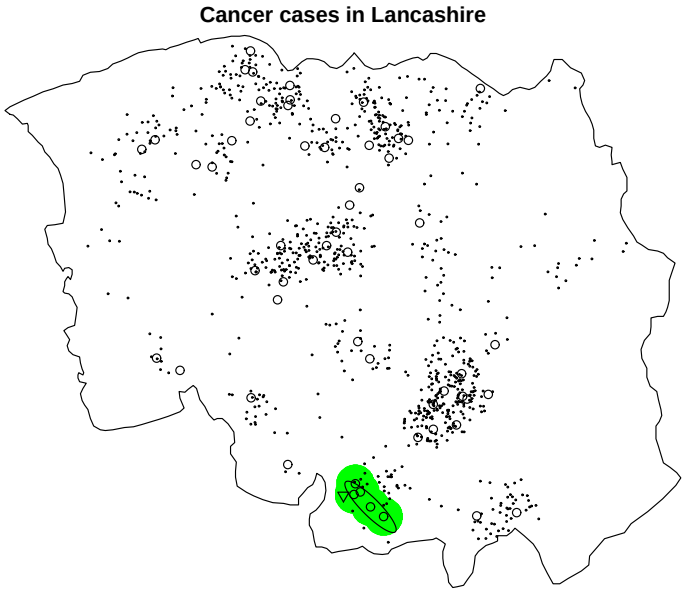

Figure 1: Detection of cancer clusters

mais plutôt distribués selon la densité de population, ils seront utilisés comme données de contrôle et leurs localisations sont notées $\{z_j, j = 1, \cdots, N\}$. Les intégrales introduites précédemment sont estimées en utilisant ces données: pour tout $B \subset A$, l'intégrale $\int_B h(x)|dx|$ est approximée par $\sum_{i=1}^{N} \mathbb{1}(z_i \in B)/N$. On utilise le logiciel SaTScan pour calculer la statistique de balayage elliptique, notée Λ_E. Toutes les p-valeurs sont estimées à partir de 999 simulations. La figure 1 illustre les résultats.

Les cancers du poumon sont représentés par les petits points et ceux du larynx par les plus grands points. Le triangle représente l'incinerateur désaffecté. On observe que beaucoup de cancers du larynx sont concentrés dans des petites zones mais la population à risque est aussi plus importante dans ces zones, sauf dans une zone autour de l'incinérateur contenant cinq cas. L'aire grisée est l'agrégat estimé par les deux statistiques de balayage spatial basées sur les graphes: elle correspond parfaitement à cette zone. Les p-valeurs sont 0.070 pour $\Lambda_{G,1}$ et 0.038 pour $\Lambda_{G,2}$. L'agrégat correspondant à la statistique de balayage elliptique, représenté par l'ellipse contenant les mêmes cinq cas, est très similaire. La p-valeur de Λ_E est 0.012. Ainsi, si l'erreur de première espèce est fixée à $\alpha = 0.05$, les statistiques $\Lambda_{G,2}$ et Λ_E sont toutes deux significatives et les agrégats estimés sont très semblables, alors que $\Lambda_{G,1}$ n'est pas significative. Le test HF basé sur les graphes et le test de balayage elliptique indiquent clairement une concentration anormale de cancers du larynx autour de l'incinérateur désaffecté. Sur un simple ordinateur personnel, les tests basés sur les graphes ont pris environ 141 secondes et le test elliptique a pris 229 secondes. Nous devons préciser que le test de balayage circulaire, beaucoup plus rapide

que l'elliptique car le nombre d'agrégats potentiels est beaucoup plus faible, conclut qu'il n'y a pas d'agrégat spatial significatif. En effet, contrairement à l'ellipse représentée sur la figure 1, un cercle contenant les cinq cas autour de l'incinérateur contient également un nombre important de cancers du poumon.

5 Conclusion

L'application décrite précédemment ainsi qu'une étude de simulation nous donnent à penser que l'utilisation des agrégats potentiels basés sur les graphes donnent des résultats comparables aux méthodes classiques, tout en réduisant énormément le nombre de zones spatiales à étudier, et donc le temps de calcul. Remarquons également que le nombre d'agrégats potentiels reste constant lorsque la dimension des données augmente. De plus, il apparaît que le nouvel indice de concentration spatiale que nous avons introduit conduit à un test plus puissant et qu'il pourrait donc être utilisé quel que soit l'ensemble des agrégats potentiels définis.

Bibliographie

[1] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics. Theory and Methods*, 26, 1481–1496.
[2] Bar-Hen, A., Koskas, M., Picard, N. (2007). Spatial cluster detection using the number of connected components of a graph. Technical report.
[3] Cucala, L. (2008). A hypothesis-free multiple scan statistic with variable window. *Biometrical Journal*, 50, 299–310.
[4] Nagarwalla, N. (1996). A scan statistic with a variable window. *Statistics in Medicine*, 15, 845–850.
[5] Diggle, P.J., Gatrell, A.C. et Lovett, A.A. (1990). Modelling the prevalence of cancer of the larynx in part of Lancashire: a new methodology for spatial epidemiology. In *Spatial epidemiology* (ed. R.W. Thomas), 21. London: Pion.