Detecting spatial clusters in functional data: new scan statistic approaches

Camille Frévent [1] & Mohamed-Salem Ahmed [1] & Matthieu Marbac [2] & Michaël Genin [1]

[1] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France.* camille.frevent@univ-lille.fr, mohamed-salem.ahmed@univ-lille.fr, michael.genin@univ-lille.fr

[2] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France.* matthieu.marbac-lourdelle2@ensai.fr

Résumé. Nous avons développé deux statistiques de scan pour détecter des clusters sur des données fonctionnelles indexées dans l'espace. La première méthode est basée sur une adaptation de l'ANOVA fonctionnelle et la deuxième est basée sur une statistique de scan spatiale *distribution-free* pour données univariées. Dans une simulation, la deuxième méthode présente toujours de meilleures performances qu'une statistique de scan non paramétrique pour données fonctionnelles, et l'adaptation de l'ANOVA présente également de meilleures performances pour des données normales. Nos approches détectent de plus petits clusters que la méthode non paramétrique.

Mots-clés. Détection de clusters, données fonctionnelles, statistiques de scan spatiales

Abstract. We have developed two scan statistics for detecting clusters of functional data indexed in space. The first method is based on an adaptation of a functional analysis of variance and the second one is based on a distribution-free spatial scan statistic for univariate data. In a simulation study, the distribution-free method always performed better than a nonparametric functional scan statistic, and the adaptation of the ANOVA also performed better for data with a normal distribution. Our methods can detect smaller spatial clusters than the nonparametric method.

Keywords. Cluster detection, functional data, spatial scan statistics

1 Introduction

Les méthodes de détection de clusters spatiaux ont été longuement étudiées. Le but est de développer des outils capables de détecter l'aggrégation de sites qui se comportent différemment. Les statistiques de scan spatiales permettent de détecter des clusters spatiaux sans information *a priori* quant à leur localisation. Elles ont été principalement proposées par Kulldorff et Nagarwalla (1995) et Kulldorff (1997) dans le cas de modèles de Poisson et de Bernouilli. D'autres statistiques de scan ont ensuite été proposées pour d'autres modèles de distribution (Kulldorff *et al.* (2009), Cucala *et al.* (2017)).

Aujourd'hui les données sont de plus en plus mesurées en temps continu ou quasi-continu. Cela a conduit au développement de méthodes d'analyse de données fonctionnelles (Ramsay et Silverman (2005)). Dans le domaine des statistiques de scan spatiales, Smida *et al.* (2020) a récemment développé une méthode non paramétrique basée sur un test de Wilcoxon-Mann-Whitney fonctionnel. Cependant aucune statistique de scan spatial paramétrique n'a été développée pour des données fonctionnelles. Puisqu'un test d'ANOVA pour données fonctionnelles a été décrit par Cuevas *et al.* (2004), nous allons développer une statistique de scan basée sur ce test. Ensuite puisque ces dernières années des tests statistiques ont été développés pour des données en grande dimension, en résumant l'information après le calcul d'une statistique pour chaque composante (Lin *et al.* (2021)), nous allons proposer une statistique de scan basée sur la combinaison de cette approche et de la statistique de scan *distribution-free* pour données univariées proposée par Cucala (2014).

2 Méthodologie

Soit $s_1, \ldots, s_n$, n sites d'un domaine d'observation $S \subset \mathbb{R}^2$ et $X_1, \ldots, X_n$ les observations de X dans $s_1, \ldots, s_n$. Ici, $\{X(t),\ t \in \mathcal{T}\}$ est un processus stochastique à valeurs réelles, avec $\mathcal{T}$ un intervalle de $\mathbb{R}$. A partir de maintenant les observations sont considérées indépendantes, ce qui est une hypothèse classique en statistique de scan.
Le but est de détecter des clusters spatiaux et de tester leur significativité. On cherche donc à tester $\mathcal{H}_0$ (l'absence de cluster) contre une hypothèse alternative $\mathcal{H}_1$ (la présence d'au moins un cluster $w \subset S$ de valeurs anormales pour X).

Cressie, N. (1977) définit une statistique de scan spatial comme le maximum d'un indice de concentration sur un ensemble de clusters potentiels $\mathcal{W}$. Dans la suite sans perte de généralité, $\mathcal{W}$ est un ensemble de clusters circulaires contenant entre 1 et 50% des sites.

2.1 Statistique de scan paramétrique pour données fonctionnelles

On suppose que le processus X est à valeurs dans l'espace $L^2(\mathcal{T}, \mathbb{R})$ des fonctions réelles de carré intégrable sur $\mathcal{T}$.
Cuevas *et al.* (2004) et Górecki et Smaga (2015) ont adapté la F-statistique de l'ANOVA pour les processus L^2. En considérant deux échantillons indépendants issus de deux processus X_{g_1} et X_{g_2} dans deux groupes g_1 et g_2, le test compare les fonctions moyennes μ_{g_1} et μ_{g_2}.
Pour la détection de cluster, $\mathcal{H}_0$ peut être définie par : $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^c} = \mu_S$, où μ_w, μ_{w^c} et μ_S sont les fonctions moyennes dans w, à l'extérieur de w et dans S. L'hypothèse alternative associée au cluster potentiel w $\mathcal{H}_1^{(w)}$ se réécrit : $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$.

L'ANOVA fonctionnelle peut être utilisée pour comparer les fonctions moyennes dans w et dans w^{c} en utilisant la statistique

$$F_n^{(w)} = \frac{|w|\ ||\bar{X}_w - \bar{X}||_2^2 + |w^{\mathsf{c}}|\ ||\bar{X}_{w^{\mathsf{c}}} - \bar{X}||_2^2}{\frac{1}{n-2}\left[\sum_{j,s_j \in w} ||X_j - \bar{X}_w||_2^2 + \sum_{j,s_j \in w^{\mathsf{c}}} ||X_j - \bar{X}_{w^{\mathsf{c}}}||_2^2\right]}. \tag{1}$$

Alors, $F_n^{(w)}$ peut être considérée comme un indice de concentration et maximisée sur l'ensemble des clusters potentiels $\mathcal{W}$, ce qui amène à la définition suivante de la statistique de scan spatial paramétrique pour données fonctionnelles (PFSS) : $\Lambda_{\text{PFSS}} = \max_{w \in \mathcal{W}} F_n^{(w)}$. Le cluster potentiel pour lequel ce maximum est atteint est appelé "cluster le plus probable" (*most likely cluster* (MLC)) est donc $\text{MLC} = \arg\max_{w \in \mathcal{W}} F_n^{(w)}$.

2.2 Statistique de scan *distribution-free* pour données fonctionnelles

Ici nous proposons de combiner la statistique de scan *distribution-free* pour données univariées proposée par Cucala (2014) et la statistique "max" de Lin *et al.* (2021). Brièvement ces derniers proposent une nouvelle approche au problème de l'ANOVA fonctionnelle en maximisant une statistique au cours du temps.

Nous supposons que pour chaque temps t, $\mathbb{V}[X_i(t)] = \sigma^2(t)\ \ \forall i \in [\![1; n]\!]$. Alors pour chaque t, l'indice de concentration proposé par Cucala (2014) pour tester $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w(t) = \mu_{w^{\mathsf{c}}}(t) = \mu_S(t)$ est

$$I^{(w)}(t) = \frac{|\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t)|}{\sqrt{\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t)]}},\ \text{où}\ \hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t)] = \widehat{\sigma^2}(t)\left[\frac{1}{|w|} + \frac{1}{|w^{\mathsf{c}}|}\right],$$

$$\text{avec}\ \widehat{\sigma^2}(t) = \frac{1}{n-2}\left[\sum_{i,s_i \in w}(X_i(t) - \bar{X}_w(t))^2 + \sum_{i,s_i \in w^{\mathsf{c}}}(X_i(t) - \bar{X}_{w^{\mathsf{c}}}(t))^2\right].$$

Maintenant l'idée est de globaliser l'information en maximisant la quantité précédente au cours du temps pour chaque cluster potentiel w (Lin *et al.* (2021)) : $I^{(w)} = \sup_{t \in \mathcal{T}} I^{(w)}(t)$. Pour la détection de cluster, comme pour le PFSS, $\mathcal{H}_0$ peut être définie par $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^{\mathsf{c}}} = \mu_S$. Et l'hypothèse alternative $\mathcal{H}_1^{(w)}$ associée au cluster potentiel w se réécrit : $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^{\mathsf{c}}}$.

$I^{(w)}$ peut être considérée comme un indice de concentration et maximisée sur l'ensemble des clusters potentiels $\mathcal{W}$ pour obtenir la statistique de scan spatial *distribution-free* pour données fonctionnelles (DFFSS) : $\Lambda_{\text{DFFSS}} = \max_{w \in \mathcal{W}} I^{(w)}$. Et pour cette méthode le *most likely cluster* est défini par $\text{MLC} = \arg\max_{w \in \mathcal{W}} I^{(w)}$.

2.3 Significativité du MLC

Une fois le MLC détecté, sa significativité doit être évaluée. Pour cela nous générons des données par permutation des données de départ (*random labelling*) et nous en déduisons une estimation de la p-valeur (Dwass (1957)). Enfin le MLC est considéré significatif si la p-valeur associée est inférieure à l'erreur de type I.

3 Etude d'une simulation

Dans une étude de simulation nous avons comparé (i) la statistique de scan spatial paramétrique pour données fonctionnelles (PFSS) Λ_{PFSS}, (ii) la statistique de scan spatial *distribution-free* pour données fonctionnelles (DFFSS) Λ_{DFFSS} et (iii) la statistique de scan spatial non paramétrique pour données fonctionnelles (NPFSS) Λ_{NPFSS} développée par Smida *et al.* (2020). Cette simulation a montré que le DFFSS présente de meilleures performances que les deux autres méthodes, surtout dans le cas d'un shift local dans le temps. Le PFSS et le NPFSS présentent des performances similaires dans le cas gaussien mais les performances du PFSS diminuent quand on s'éloigne de la normalité. Le NPFSS a tendance à détecter des clusters plus grands que le PFSS et le DFFSS.

Bibliographie

Cressie, N. (1977). On Some Properties of the Scan Statistic on the Circle and the Line, *Journal of Applied Probability*, 14, pp. 272-283.

Cucala, L. (2014). A distribution-free spatial scan statistic for marked point processes, *Spatial Statistics*, 10, pp. 117–125.

Cucala, L., Genin, M., Lanier, C. et Occelli, F. (2017). A Multivariate Gaussian scan statistic for spatial data, *Spatial Statistics*, 21, pp. 66-74.

Cuevas, A., Febrero-Bande, M. et Fraiman, R. (2004). An ANOVA test for functional data, *Computational Statistics & Data Analysis*, 47, pp. 111-122.

Dwass, M. (1957). Modified Randomization Tests for Nonparametric Hypotheses, *Annals of Mathematical Statistics*, 28, pp. 181-187.

Górecki, T. et Smaga, Ł. (2015). A comparison of tests for the one-way ANOVA problem for functional data, *Computational Statistics*, 30, pp. 987-1010.

Kulldorff, M. et Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference, *Statistics in Medicine*, 14(8), pp. 799-810.

Kulldorff, M. (1997). A Spatial Scan Statistic, *Communications in Statistics - Theory and Methods*, 26, pp. 1481-1496.

Kulldorff, M. et Huang, L. and Konty, K. (2009). A scan statistic for continuous data based on the normal probability model, *Int J Health Geogr*, 8(58).

Lin, Z., M. E. Lopes et H.-G. Müller (2021). High-dimensional manova via bootstrapping

and its application to functional and sparse count data.
Ramsay, JO et Silverman, BW (2005). Functional Data Analysis *Springer*.
Smida, Z., Cucala, L. et Gannoun, A. (2020). A nonparametric spatial scan statistic for functional data, *hal-02908496*.