Le package `ClustGeo` :
Classification ascendante hiérarchique avec contraintes de proximité géographique

A. Labenne[a], **M. Chavent**[b], **V. Kuentz-Simonet**[a] and **J. Saracco**[b]

[a]IRSTEA, UR ETBX, 33612 Cestas Cedex, France
amaury.labenne@irstea.fr

[b]Univ. Bordeaux, IMB & INRIA Bordeaux Sud-Ouest, CQFD, F-33400 Talence

Mots clefs : CAH, contraintes géographiques, critère de Ward

1 Introduction

Ce travail s'inscrit dans le cadre du projet ANR ADAPT'EAU dont une des tâches vise à établir un diagnostic socio-économique des territoires à l'échelle communale. Pour cela, l'approche par la qualité de vie a été retenue. Le but de cette tâche est de créer des indicateurs synthétiques des conditions de vie des communes afin d'évaluer leur vulnérabilité. Par la suite, afin de mieux comprendre la ressemblance entre les différents territoires, il est pertinent d'effectuer une classification (CAH de Ward par exemple) de ces communes en fonction des valeurs des indicateurs qui leur sont associées. Cette classification permet d'aboutir à différentes typologies et ainsi de représenter les communes sur une carte en fonction de leur classe d'appartenance. Cependant, afin de faciliter l'interprétation nous avons souhaité obtenir une typologie qui soit plus "compacte" géographiquement. En effet il semble naturel que deux communes proches géographiquement présentent des caractéristiques similaires de conditions de vie et se retrouvent donc dans la même classe. Certaines méthodes de classification intégrant des contraintes de voisinages existent déjà [1], [2]. Nous développons ici une nouvelle méthode de classification ascendante hiérarchique, appelée `ClustGeo`, qui ne se base pas sur des contraintes de voisinage mais sur des contraintes de distance géographique entre individus. Cette méthode sera appliquée sur un échantillon de communes du Sud-Ouest de la France et les fonctions du package R associé seront présentées.

2 Notations, définitions et CAH de Ward

On dispose d'une matrice de données $\mathbf{X}$, de dimension $(n \times p)$, mesurant les caractéristiques (p variables) de n individus. A partir de cette matrice d'observations, on construit la matrice $\mathbf{D}_1$ de distances euclidiennes entre les individus mesurées sur les p variables. On dispose également d'une matrice $\mathbf{D}_2$ de distances géographiques (en mètres) entre les individus. Soit ω_i le poids attribué à l'individu i. Soient $\mu_k = \sum_{i \in C_k} \omega_i$, le poids de la classe k et $g_k \in \mathbb{R}^p$, le centre de gravité de la classe k. Soit $g \in \mathbb{R}^p$ le centre de gravité de l'ensemble des individus. On note par $T = \sum_{i=1}^{n} \omega_i d^2(x_i, g) = \sum_{i=1}^{n} \sum_{i'=1}^{n} \frac{\omega_i \omega'_i}{2\mu_k} d^2(x_i, x'_i)$ l'inertie totale et $W = \sum_{i \in Ck} \omega_i d^2(x_i, g_k) = \sum_{i \in Ck} \sum_{i' \in Ck} \frac{\omega_i \omega'_i}{2\mu_k} d^2(x_i, x'_i)$ l'inertie intra-classe, où $x_i \in \mathbb{R}^p$ est le vecteur des p variables de l'individu i.

Les mesures T et W pourront être indexées par 1 ou 2 en fonction de la matrice de distances ($\mathbf{D}_1$ ou $\mathbf{D}_2$) à partir de laquelle elles ont été calculées.

Partitions et homogénéité. Soit $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ une partition des n individus en K classes. On définit un critère d'homogénéité de partition $H(\mathcal{P}_K)$ que l'on cherche à minimiser. Pour cela on note $H(\mathcal{C}_k)$ l'homogénéité d'une classe $\mathcal{C}_k$ et on définit l'homogénéité de la partition $\mathcal{P}_K$ comme $H(\mathcal{P}_K) = \sum_{k=1}^{K} H(\mathcal{C}_k)$.

Exemple de la CAH avec critère de Ward. La CAH avec critère de Ward consiste à minimiser le critère d'homogénéité de partition $H(\mathcal{P}_K)$ en prenant $H(\mathcal{C}_k) = W_1(\mathcal{C}_k)$, où $W_1(\mathcal{C}_k)$ est l'inertie intra-classe calculée à partir de la matrice de distance $\mathbf{D}_1$.
La mesure d'agrégation entre deux classes $\mathcal{C}_l$ et $\mathcal{C}_m$ est alors égale à :

$$\mathcal{D}(\mathcal{C}_l, \mathcal{C}_m) = H(\mathcal{P}_{K-1}) - H(\mathcal{P}_K) = \frac{\mu_l\ \mu_m}{\mu_l + \mu_m}\ d_1^2(g_l, g_m).$$

3 La méthode `ClustGeo`

Le but ici est d'intégrer une matrice de distances géographiques. Pour cela, on va définir un nouveau critère d'homogénéité de classe, le calcul du critère d'homogénéité de partition reste, quant à lui, le même ($H(\mathcal{P}_K) = \sum_{k=1}^{K} H(\mathcal{C}_k)$). Cela va nous conduire à une nouvelle mesure d'agrégation entre classes.

Homogénéité de classe. Soit $\alpha \in [0, 1]$. On considère ici :

$$H(\mathcal{C}_k) = \alpha W_1(\mathcal{C}_k) + (1 - \alpha) W_2(\mathcal{C}_k). \tag{1}$$

Où $W_1(\mathcal{C}_k)$ (resp. $W_2(\mathcal{C}_k)$) est l'inertie intra-classe calculée à partir de la matrice de distance $\mathbf{D}_1$ (resp. $\mathbf{D}_2$).

Mesure d'agrégation $\mathcal{D}$ entre deux classes. Cette mesure d'association correspond aux distances de Ward calculées sur deux matrices de distances différentes ($\mathbf{D}_1$ et $\mathbf{D}_2$) et pondérées respectivement par α et $(1 - \alpha)$. Ainsi si $\alpha = 1$ cette méthode revient a effectuer une CAH de Ward sur la matrice de distances $\mathbf{D}_1$. Inversement, si $\alpha = 0$, cette méthode revient à effectuer une CAH de Ward basée uniquement sur la matrice $\mathbf{D}_2$. A partir du critère d'homogénéité de classe défini ci-dessus on obtient la mesure d'agrégation entre classes suivante :

$$\mathcal{D}(\mathcal{C}_l, \mathcal{C}_m) = \alpha \frac{\mu_l\ \mu_m}{\mu_l + \mu_m}\ d_1^2(g_l, g_m) + (1 - \alpha) \frac{\mu_l\ \mu_m}{\mu_l + \mu_m}\ d_2^2(g_l, g_m). \tag{2}$$

Références

[1] Marie Chavent, Yves Lechevallier, Françoise Vernier, and Kevin Petit. Monothetic divisive clustering with geographical constraints. In Paula Brito, editor, *COMPSTAT 2008*, pages 67–76. Physica-Verlag HD, 2008.

[2] Pierre Legendre and Louis Legendre. Chapter 12 - ecological data series. In Pierre Legendre and Louis Legendre, editor, *Developments in Environmental Modelling*, volume 24 of *Numerical Ecology*, pages 711–783. Elsevier, 2012.