Classification ascendante hiérarchique avec contraintes de proximité géographique

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

[a] IRSTEA, UR ETBX, centre de Bordeaux, F-33612 Gazinet Cestas, France.
amaury.labenne@irstea.fr
[b] Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France.
[c] INRIA, CQFD, F-33400 Talence, France.

Résumé. La Classification Ascendante Hiérarchique (CAH) est une méthode bien connue de classification d'individus décrits par différentes variables. Cette méthode vise à rassembler dans une même classe les individus qui se ressemblent du point de vue des variables. Cependant lorsque les individus dont on dispose sont des territoires géographiques, on souhaite parfois que des individus proches géographiquement se retrouvent dans la même classe sans que cela ne nuise trop à la qualité de la partition. La méthode `ClustGeo` que nous avons développée permet d'intégrer des contraintes de proximité géographique au sein d'une CAH, pour cela on utilise le critère d'homogénéité de Ward sur deux matrices différentes de distances.

Mots-clés. CAH, contraintes géographiques, critère de Ward

Abstract. Hierarchical Ascendant Clustering (HAC) is a well-known method of individual clustering. This method aims to bring together individuals who are similar regarding to variables which describe them. However, when individuals are geographical units, we may wish that individuals which are geographically close are put in same clusters without deteriorate too much the quality of the partition. Method `ClustGeo` allows to take into account geographical constraints of proximity within the HAC, to do that we use the Ward homogeneity criterion on two different matrices of distances.

Keywords. HCA, geographical constraints, Ward criterion

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. La création de ces indices synthétiques à partir des données socio-économiques peut être faite de différentes manières. Une des approches retenue est la méthode `ClustOfVar` [1], cette méthode permet de créer des variables synthétiques (indices) fortement corrélées aux variables initiales. 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 indices 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 hiérarchique intégrant des contraintes de voisinages existent déjà [2], [3]. D'autres méthodes basées sur des modèles probabilistes permettent également de prendre en compte ces contraintes de voisinage, en utilisant par exemple un algorithme EM modifié [4]. 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 distances géographiques entre individus. Cette méthode sera appliquée sur un échantillon de communes du Sud-Ouest de la France.

2 Notations, définitions et CAH de Ward

On dispose d'une matrice de données $\mathbf{X}$, de dimension $(n \times p)$, mesurant la qualité de vie (p variables) de n communes. A partir de cette matrice d'observations, on construit la matrice $\mathbf{D}_1$ de distances euclidiennes entre les communes mesurées sur les p variables de qualité de vie. On dispose également d'une matrice $\mathbf{D}_2$ de distances géographiques (en mètres) entre les communes.

Soit ω_i le poids attribué à la commune 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 communes. 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 qualité de vie de la commune 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 communes 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 :

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

Choix du paramètre α. Dans un premier temps, on choisit le nombre K de classes en fonction du dendrogramme issue de la CAH de Ward (équivalent au cas où $\alpha = 1$). Une fois le nombre K de classes fixé, on choisit le paramètre α le plus petit possible tel que la qualité de la partition en K classes soit la moins dégradée possible. Pour cela, on pourra se fixer un seuil de qualité à ne pas dépasser (par exemple 90% de la qualité de la partition d'origine issue de la CAH de Ward basée sur $\mathbf{D}_1$). On rappelle que la qualité d'une partition est définie de la manière suivante : $\left(1 - \frac{W_1}{T_1}\right) \times 100$, qui correspond au pourcentage d'inertie expliquée.

4 Application de la méthode `ClustGeo` sur données réelles

Présentation des données. On dispose d'une matrice de données $\mathbf{X}$ contenant $n = 303$ communes décrites par $p = 5$ variables quantitatives. Ces variables sont issues de la méthode `ClustOfVar` réalisées sur 30 variables relatives aux conditions de vie au sein de chaque commune. Afin de mieux comprendre les variables synthétiques, nous indiquons dans la Table 1 quelles variables d'origine sont le plus liées aux variables synthétiques.

Variable synthétique	Corrélation positive	Corrélation négative
V1 : Accès aux services	Nombreux services.	Peu de services.
V2 : Conditions de logement	Maisons, Peu de bâtiments, Propriétaires, Peu d'HLM, Faible densité.	Appartements, Beaucoup de bâtiments, locataires, Présence d'HLM, Forte densité.
V3 : Diplômes et catégories socio-professionnelles	RNI important, Population diplômée, Emplois qualifiés.	RNI faible, Population peu diplômée, Emplois non qualifiés.
V4 : Situations familiales et emplois	Retraités, Emploi sur la commune, Peu de résidences principales, Faible taux d'emploi.	Familles avec enfants, Emplois sur le département, Part élevée de résidences principales, Taux d'emploi élevé.
V5 : Environnement	Beaucoup de végétation, Peu d'agriculture.	Peu de végétation, Beaucoup d'agriculture.

TABLE 1 – Description des variables synthétiques

A partir de cette matrice de données $\mathbf{X}$ on calcule la matrice de distances euclidiennes $\mathbf{D}_1$. On dispose également de la matrice $\mathbf{D}_2$ de distances géographiques entre les $n = 303$ communes.

CAH de Ward sur la matrice de distance $\mathbf{D}_1$. Dans un premier temps on réalise une CAH de Ward sur la matrice de distance $\mathbf{D}_1$. Au vu du dendrogramme (non représenté ici), on décide de retenir la partition en $K = 4$ classes de communes. À partir de cette typologie, on peut représenter sur une carte les communes en fonction de leur classe, voir Figure 2(a).

Interprétation des classes issues de la typologie de Ward par les variables synthétiques. L'interprétation des classes de la typologie peut se faire en regardant les valeurs moyennes que prennent les variables synthétiques dans les différentes classes. Ainsi la classe 1 représente les communes les plus urbaines, la classe 2 représente principalement les commune péri-urbaines, la classe 3 regroupe majoritairement les communes rurales et littorales, alors que la classe 4 contient des communes rurales isolées avec une proportion importante de territoires agricoles.

Méthode `ClustGeo` sur les données qualité de vie et choix du paramètre α. Nous avons appliqué la méthode `ClustGeo` sur les données qualité de vie. Pour cela, on

applique la méthode pour différentes valeurs de $\alpha \in [0, 1]$ afin de choisir le meilleur α pour la partition retenue en $K = 4$ classes. Le choix du paramètre est fait à l'aide de la Figure 1. On retient la valeur $\alpha = 0.5$. En effet on voit que la perte de qualité de la partition correspondante est inférieure à 10% pour $K = 4$ classes. De plus, comme on peut le voir sur la Figure 2(b) la partition obtenue avec cette méthode est plus "compacte" au sens géographique.

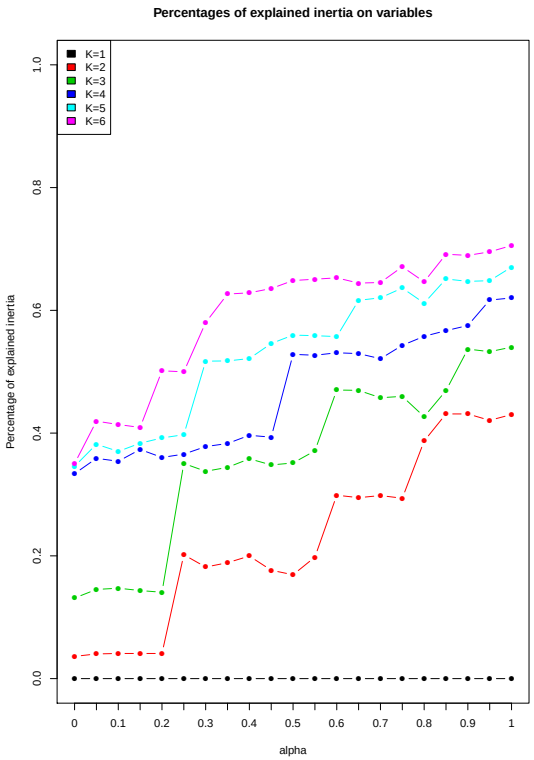

FIGURE 1 – Graphique des qualités de partitions en fonction de K et α.

Interprétation des classes de la typologie par les variables synthétiques. Comme précédemment, on peut interpréter les classes de communes en regardant les valeurs moyennes que prennent les variables synthétiques dans les différentes classes. La classe 1 issue de `ClustGeo` contient les communes les plus urbaines, elle correspond à la fusion des classes 1 et 2 de la typologie de Ward. La classe 2 est une nouvelle classe principalement issue de la classe 4 de la typologie de Ward, elle contient des communes à dominante forestière. Les classes 3 et 4 sont relativement semblables aux classes 3 et 4 de la typologie de Ward.

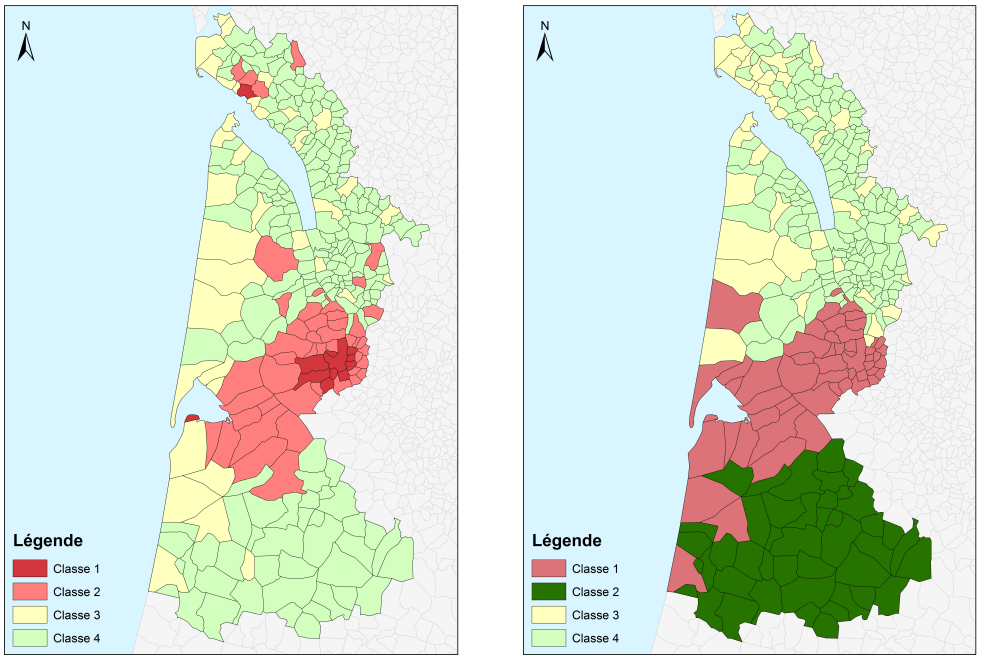

FIGURE 2 – (a) Carte de la typologie issue de la CAH de Ward (à gauche) et (b) Carte de la typologie issue de la méthode `ClustGeo` avec $\alpha = 0.5$ (à droite).

Références

[1] M. Chavent, V. Kuentz Simonet, B. Liquet, and J. Saracco. ClustOfVar : An R Package for the Clustering of Variables. *Journal of Statistical Software*, 50(13) :1–16, 2012.

[2] 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.

[3] 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.

[4] C. Ambroise, M. Dang, and G. Govaert. Clustering of Spatial Data by the EM Algorithm. In Amílcar Soares, Jaime Gómez-Hernandez, and Roland Froidevaux, editors, *geoENV I — Geostatistics for Environmental Applications*, number 9 in Quantitative Geology and Geostatistics, pages 493–504. Springer Netherlands, 1997.