ClustGeo : Classification Ascendante Hiérarchique (CAH) avec contraintes de proximité géographique

4èmes Rencontres R, Grenoble

Marie Chavent [2], Vanessa Kuentz-Simonet [1], Amaury Labenne [1,2] & Jérôme Saracco [2]

[1] *IRSTEA, UR ETBX, 33612 Cestas Cedex, France.*
[2] *INRIA, CQFD, F-33400 Talence, France.*

PLAN

INTRODUCTION

Problématique et objectifs

- Classification d'individus spatialisés : communes, lieux, sections de fleuve, ...
- Méthodes classiques (CAH, k-means) $\Longrightarrow$ forte dispersion géographique des classes obtenues.
- Nécessité d'intégrer une information spatiale dans le clustering :
  - **Q** : Matrice de voisinage entre les individus ($Q_{ij} = 1$ si i et j sont voisins, 0 sinon),
  - $\mathbf{D}_2$: Matrice de distances géographiques entre individus.

Quelques approches existantes :

Avec Q

- [1] Legendre et Legendre, 2009
- [2] Guo, 2009
- [3] Chavent et al., 2009
- [4] Ambroise et al., 1997

Avec D2

- [5] Webster, 1977
- [6] M. A. Oliver, 1988

CAH avec critère additif d'hétérogénéité 1/2

On va définir ici un critère additif d'hétérogénéité $\mathcal{H}(\mathcal{P}_K)$ pour une partition $\mathcal{P}_K$ en K clusters.

Critère d'hétérogénéité de partition

Ce critère est défini comme la somme des critères d'hétérogénéité $H(\mathcal{C}_k)$ associés aux clusters $\mathcal{C}_k$, $k = 1, \ldots, K$.

$$\mathcal{H}(\mathcal{P}_K) = \sum_{k=1}^{K} H(\mathcal{C}_k). \tag{1}$$

L'algorithme de CAH consiste à minimiser ce critère $\mathcal{H}(\mathcal{P}_K)$ à chaque itération.
En effet on rassemble les deux clusters tels que l'augmentation de l'hétérogénéité de la partition soit minimale.
Cela conduit donc à définir une mesure d'agrégation entre clusters comme la différence entre l'hétérogénéité de la partition avant rassemblement et l'hétérogénéité de la partition issue de l'agrégation de 2 clusters.

CAH avec critère additif d'hétérogénéité 2/2

Mesure d'agrégation entre clusters

La mesure d'agrégation $\delta(\mathcal{C}_l, \mathcal{C}_m)$ entre deux clusters d'individus $\mathcal{C}_l$ et $\mathcal{C}_m$ est définie de la manière suivante :

$$\delta(\mathcal{C}_l, \mathcal{C}_m) = \mathcal{H}(\mathcal{P}_{K-1}) - \mathcal{H}(\mathcal{P}_K) = H(\mathcal{C}_l \cup \mathcal{C}_m) - H(\mathcal{C}_l) - H(\mathcal{C}_m). \quad (2)$$

Critère de qualité d'une partition

Un critère de qualité d'une partition $\mathcal{P}_K$ en K clusters peut s'écrire de la manière suivante :

$$1 - \frac{\mathcal{H}(\mathcal{P}_K)}{\mathcal{H}(\mathcal{P}_1)}, \quad (3)$$

où $\mathcal{P}_1$ est la partition en un seul cluster contenant les n individus.
Ce critère varie entre zéro et un. Il vaut 1 pour la partition en n clusters de singletons et zéro pour la partition $\mathcal{P}_1$.

Exemple de la CAH avec critère de Ward 1/2

Critère de Ward d'hétérogénéité de partition

$$\mathcal{H}(\mathcal{P}_K) = \sum_{k=1}^{K} H(\mathcal{C}_k) = \sum_{k=1}^{K} I(\mathcal{C}_k, \mathbf{D}) = W(\mathcal{P}_K, \mathbf{D}). \quad (4)$$

où $I(\mathcal{C}_k, \mathbf{D})$ est l'inertie du nuage de point $\mathcal{C}_k$ et $W(\mathcal{P}_K, \mathbf{D})$ est l'inertie intra-classe de la partition $\mathcal{P}_K$ calculée à partir de la matrice de distance $\mathbf{D}$.

Mesure de Ward d'agrégation entre clusters

La mesure d'agrégation $\delta(\mathcal{C}_l, \mathcal{C}_m)$ entre deux clusters d'individus $\mathcal{C}_l$ et $\mathcal{C}_m$ est alors :

$$\begin{aligned} \delta_{Ward}(\mathcal{C}_l, \mathcal{C}_m) &= \mathcal{H}(\mathcal{P}_{K-1}) - \mathcal{H}(\mathcal{P}_K) = H(\mathcal{C}_l \cup \mathcal{C}_m) - H(\mathcal{C}_l) - H(\mathcal{C}_m) \\ &= \frac{1}{n} \frac{n_l \times n_m}{n_l + n_m} d^2(g_l, g_m). \end{aligned} \quad (5)$$

Exemple de la CAH avec critère de Ward 2/2

Critère de qualité d'une partition obtenue avec Ward

Le critère de qualité d'une partition $\mathcal{P}_K$ obtenue avec critère de Ward peut être écrit à partir de l'Equation (3). Il est égal au pourcentage d'inertie expliqué:

$$1 - \frac{W(\mathcal{P}_K, \mathbf{D})}{I(\mathcal{P}_1, \mathbf{D})} \tag{6}$$

Ce critère varie entre 0 et 1. Il vaut 1 pour la partition $\mathcal{P}_n$ en n singletons et 0 pour la partition $\mathcal{P}_1$ en un seul cluster.

Nous allons nous baser sur la CAH de Ward pour développer la méthode hclustgeo de CAH intégrant des contraintes géographiques.

La méthode hclustgeo 1/4

Le but de la méthode est d'intégrer dans la CAH une matrice $\mathbf{D}_2$ de dimension $n \times n$ contenant les distances géographiques entre les individus.
On utilisera également la matrice $\mathbf{D}_1$ de même dimension qui est la matrice de distances euclidiennes entre les individus calculée à partir de la matrice de données $\mathbf{X}$ de dimension $n \times p$ où n individus sont décrits par p variables quantitatives.

Critère d'hétérogénéité de cluster de la méthode hclustgeo

Soit $\alpha \in [0, 1]$. On définit le critère $H(\mathcal{C}_k)$ comme suit :

$$H(\mathcal{C}_k) = \alpha I(\mathcal{C}_k, \mathbf{D}_1) + (1 - \alpha) I(\mathcal{C}_k, \mathbf{D}_2), \tag{7}$$

où $I(\mathcal{C}_k, \mathbf{D}_1)$ (resp. $I(\mathcal{C}_k, \mathbf{D}_2)$) est l'inertie des observations du cluster $\mathcal{C}_k$ calculée à partir de la matrice de distances euclidiennes $\mathbf{D}_1$ (resp. $\mathbf{D}_2$).

Ce critère d'hétérogénéité permet de donner plus ou moins de poids, par le biais du choix du paramètre α, à la matrice de distances $\mathbf{D}_1$ ou à la matrice de distances $\mathbf{D}_2$.

Rq : Le critère d'hétérogénéité d'une partition est toujours défini de la même façon : $\mathcal{H}(\mathcal{P}_K) = \sum_{k=1}^{K} H(\mathcal{C}_k)$.

La méthode hclustgeo 2/4

A partir du critère d'hétérogénéité de clusters défini précedemment, on va définir une mesure d'agrégation entre clusters.

Mesure d'agrégation entre clusters

La mesure d'agrégation correspond en fait à la somme pondérée (par α et $(1-\alpha)$) de 2 mesures de Ward calculées avec 2 matrices de distances différentes : $\delta_1(\mathcal{C}_l, \mathcal{C}_m) = \frac{1}{n} \frac{n_l\ n_m}{n_l+n_m} d_1^2(g_l^1, g_m^1)$ et $\delta_2(\mathcal{C}_l, \mathcal{C}_m) = \frac{1}{n} \frac{n_l\ n_m}{n_l+n_m} d_2^2(g_l^2, g_m^2)$.

On obtient ainsi la mesure d'agrégation globale :

$$\delta(\mathcal{C}_l, \mathcal{C}_m) = \alpha\delta_1(\mathcal{C}_l, \mathcal{C}_m) + (1-\alpha)\delta_2(\mathcal{C}_l, \mathcal{C}_m) \quad (8)$$

Ainsi, lorsque $\alpha = 0$, la méthode hclustgeo est une CAH de Ward effectuée uniquement sur la matrice de distances $\mathbf{D}_2$. Inversement, lorsque $\alpha = 1$, la méthode hclustgeo est une CAH de Ward effectuée sur la matrice de distances $\mathbf{D}_1$.

La méthode hclustgeo 3/4

Nous allons voir à travers deux exemples de classification de communes à quel point la typologie obtenue est différente en fonction du choix du paramètre α.

Exemples de hclustgeo avec $\alpha = 0$ et $\alpha = 1$

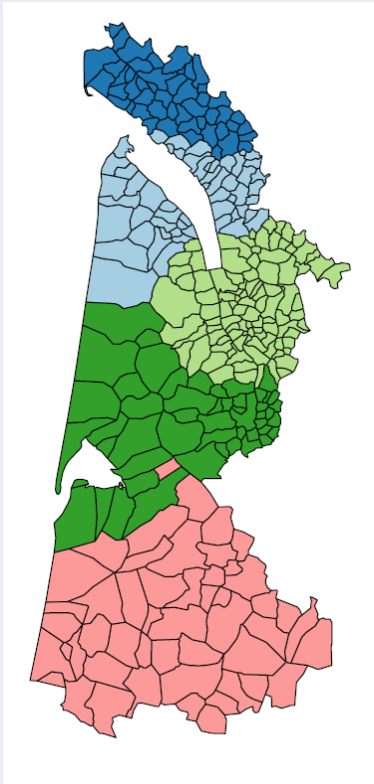


Figure : hclustgeo avec $\alpha = 0$

$\Leftrightarrow$ CAH de Ward sur **D**2

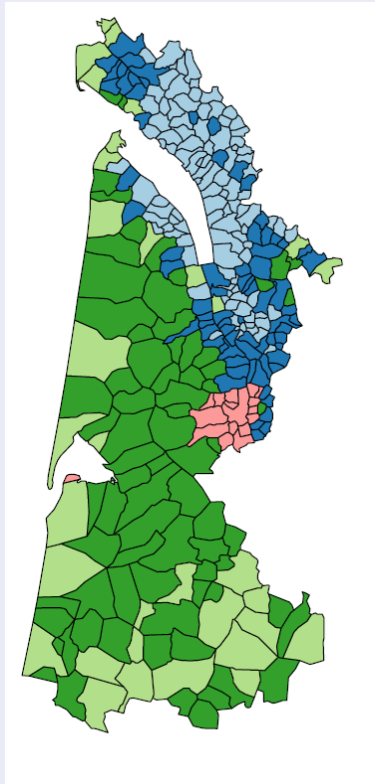


Figure : hclustgeo avec $\alpha = 1$

$\Leftrightarrow$ CAH de Ward sur **D**1

La méthode hclustgeo 4/4

- Le choix du paramètre α est très important.
- 2 critères de qualité de partition peuvent être définis :
  - Qualité de "resemblance des individus" : $Q_1 = 1 - \frac{W(\mathcal{P}_K, \mathbf{D}_1)}{I(\mathcal{P}_1, \mathbf{D}_1)}$
  - Qualité de "rassemblement géographique" : $Q_2 = 1 - \frac{W(\mathcal{P}_K, \mathbf{D}_2)}{I(\mathcal{P}_1, \mathbf{D}_2)}$
- On privilégie Q_1. En effet on cherche avant tout à avoir des individus qui se ressemblent (du point de vue des variables) au sein d'une même classe.
- On choisit comme référence la valeur de Q_1 calculée avec $\alpha = 1$ (CAH de Ward sur $\mathbf{D}_1$). Ensuite, on choisira le plus petit α possible tel que la valeur de Q_1 soit la moins dégradée possible.

Application sur données réelles 1/6

Nous allons illustrer l'utilisation de la fonction `hclustgeo` du R-package **ClustGeo** sur un jeux de données réelles où 303 communes du Sud-Ouest de la France sont décrites par 4 variables quantitatives.

Le package permet également la représentation des typologies sur une carte à l'aide de l'appel au package **rCarto**. Pour cela les shapefiles (.dbf, .shp, .shx) relatifs aux individus sont nécessaires. Les shapefiles relatifs à l'exemple sont disponibles dans le package.

Application sur données réelles 2/6

Nous présentons, étape par étape, la méthodologie utilisée ainsi que des extraits de codes.

1ère étape : Choix du nombre de classes avec CAH de Ward basée sur $\mathbf{D}_1$ et dendrogramme

```
# load data
> library("ClustGeo")
> data(comm303)
> base <- comm303$data.303
> Dgeo <- comm303$Dgeo.303

# path to shapefiles and identifier of observations (needed for the plot)
> path.303 <- file.path(path.package("ClustGeo"), "shapes/comm303")
> ID.ind <- "INSEE_COM"

# perform hclustgeo for alpha=1
> res.alpha1 <- hclustgeo(data=base, D.geo=Dgeo, alpha=1)
```

Application sur données réelles 3/6

1ère étape : Choix du nombre de classes avec CAH de Ward basée sur $\mathbf{D}_1$ et dendrogramme

```
# plot of the dendrogram
> plot(res.alpha1, choice="dendro")
```

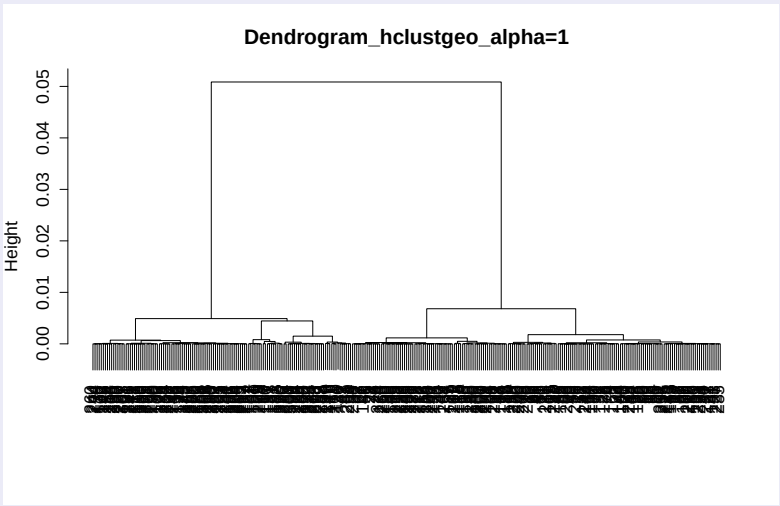


Figure : hclustgeo avec $\alpha = 1$

```
# plot of the map
> plot(res.alpha1, choice="maps",
       K.range=5,  path.shp=path.303,
       name.ind.shp=ID.ind)
```

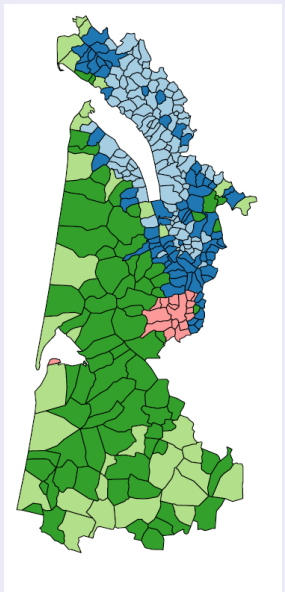
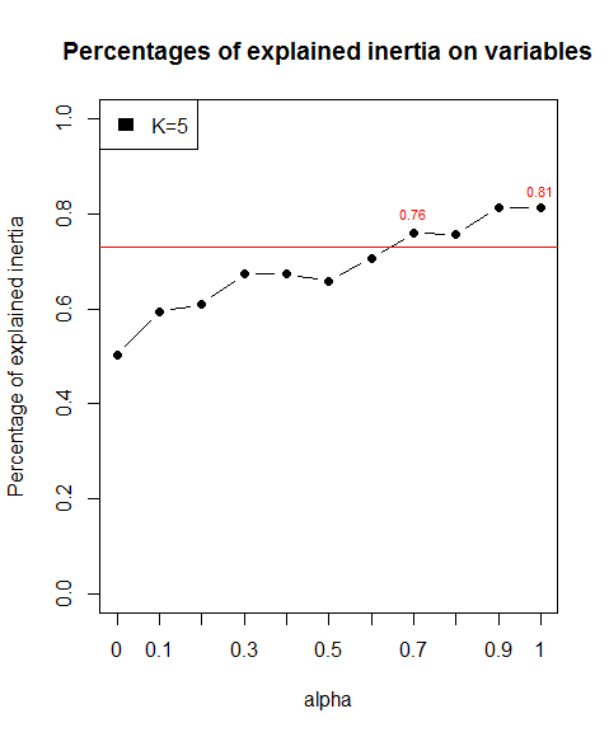


Figure : hclustgeo avec $\alpha = 1$

Application sur données réelles 4/6

2ème étape : Choix du paramètre α

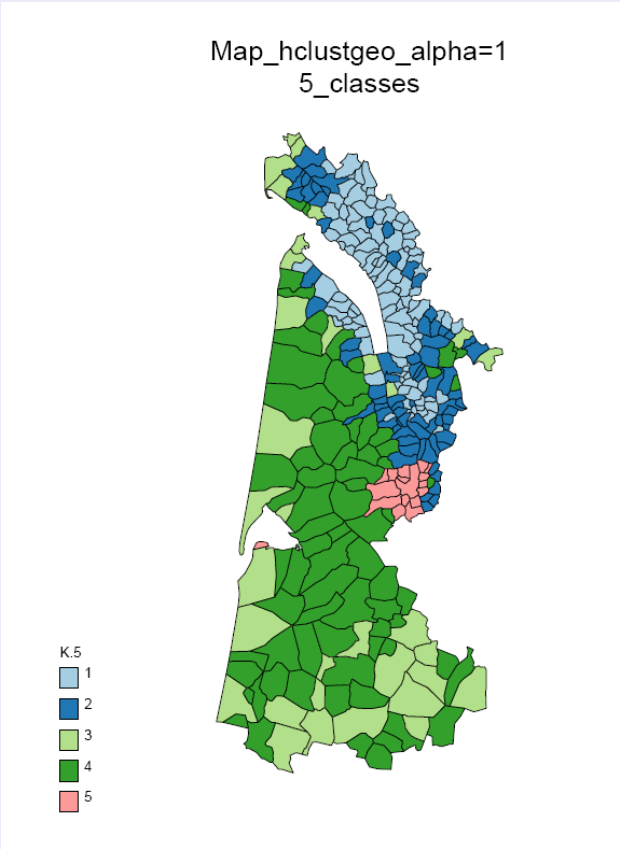
```
# perform hclustgeo for several values of alpha
> multi.alpha <- seq(0, 1, 0.1)
> res.alpha <- hclustgeo(data=base, D.geo=Dgeo, alpha=multi.alpha)

# plot of the qualities
> plot.qual <- plot(res.alpha, choice="quality", K.range=5)
```

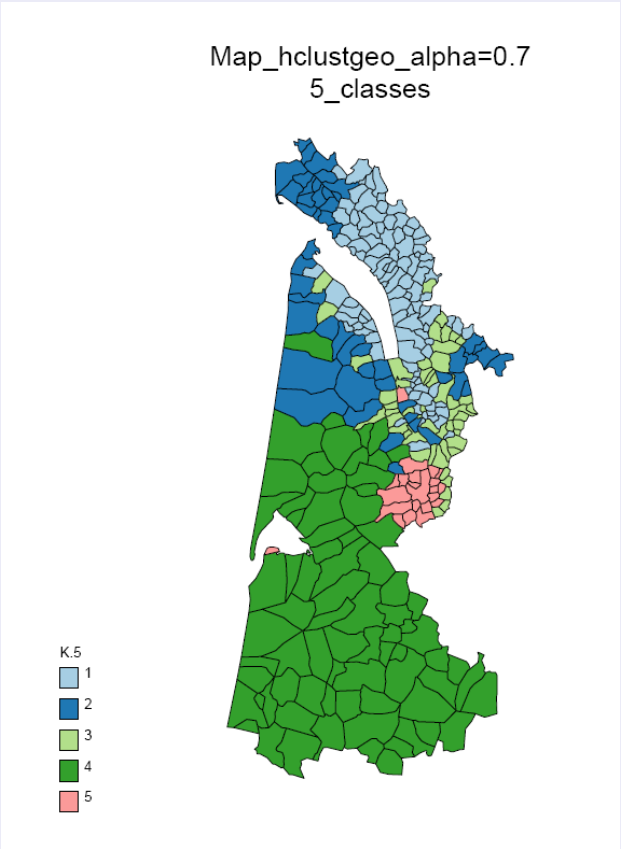


Application sur données réelles 5/6

3ème étape : Classification obtenue pour $\alpha^\star = 0.7$ et description des classes

```
#plot of the two maps
plot(res.alpha, choice="maps", K.range=5, choice.alpha=c(0.7, 1),
     path.shp=path.303, name.ind.shp=ID.ind)
```





Application sur données réelles 6/6

3ème étape : Classification obtenue pour $\alpha^\star = 0.7$ et description des classes

```
#Obtain description of partitions
summ<-summary(res.alpha, K.range=5, choice.alpha=c(0.7, 1))
```

`summ$"summary_hclustgeo.alpha=0.7"$desc$"K=5"`

	Variables	V test	Mean in cluster	Overall mean	Sd in cluste	Overall Sd
Classe 1	agri.land	13.83	78.09	44.87	8.08	30.39
	graduate.rate	-3.83	14.55	15.47	3.18	3.02
	housing.appart	-5.30	3.59	8.55	3.75	11.83
Classe 2	employ.rate.city	2.38	31.28	27.39	14.58	12.65
	agri.land	-2.97	33.21	44.87	15.21	30.39
Classe 3	agri.land	2.86	56.11	44.87	7.28	30.39
	graduate.rate	-2.70	14.41	15.47	2.33	3.02
	employ.rate.city	-5.00	19.20	27.39	5.66	12.65
Classe 4	graduate.rate	5.27	17.05	15.47	2.74	3.02
	agri.land	-12.04	8.48	44.87	7.56	30.39
Classe 5	housing.appart	13.56	41.53	8.55	16.17	11.83
	graduate.rate	3.51	17.64	15.47	2.42	3.02
	agri.land	-5.09	13.03	44.87	10.71	30.39

`summ$"summary_hclustgeo.alpha=1"$desc$"K=5"`

	Variables	V test	Mean in cluster	Overall mean	Sd in cluster	Overall Sd
Classe 1	agri.land	13,72	78,32	44,87	8,1	30,39
	graduate.rate	-4,21	14,45	15,47	3,23	3,02
	housing.appart	-5,24	3,58	8,55	3,77	11,83
Classe 2	agri.land	2,6	52,79	44,87	9,95	30,39
	graduate.rate	-2,37	14,75	15,47	2,36	3,02
	employ.rate.city	-4,95	21,1	27,39	5,95	12,65
Classe 3	employ.rate.city	11,35	51,02	27,39	9,43	12,65
	housing.appart	2,44	13,29	8,55	9,55	11,83
	agri.land	-5,27	18,5	44,87	15,3	30,39
Classe 4	graduate.rate	5,06	17	15,47	2,64	3,02
	employ.rate.city	-4,05	22,24	27,39	7,77	12,65
	agri.land	-11,27	10,52	44,87	8,11	30,39
Classe 5	housing.appart	13,52	46,28	8,55	14,92	11,83
	graduate.rate	2,19	17,03	15,47	2,27	3,02
	agri.land	-4,85	10,11	44,87	8,17	30,39

CONCLUSION

- La méthode hclustgeo permet d'intégrer de l'information géographique (par le biais de α) au sein de la CAH de Ward.
- Une méthode graphique permet de choisir $\alpha^\star$ simplement.
- La fonction `plot` permet l'affichage de cartes directement à partir de shapefiles.
- Possibilité de choix de différentes matrices $\mathbf{D}_2$: par exemple la distance par la route peut être plus pertinente.
- Le package **ClustGeo** est disponible sur GitHub et très prochainement sur le CRAN.

```
devtools::install_github("AmauryLabenne/ClustGeo")
# This needs the devtools package to be installed :
# install.packages("devtools")
```

BIBLIOGRAPHIE

- [1] P. Legendre and L. Legendre. Chapter 12 - Ecological data series. In P. L. a. L. Legendre, editor, *Developments in Environmental Modelling*, volume 24 of *Numerical Ecology*, pages 711–783. Elsevier, 2012
- [2] D. Guo. Greedy Optimization for Contiguity-Constrained Hierarchical Clustering. In *IEEE International Conference on Data Mining Workshops, 2009. ICDMW '09*, pages 591–596, Dec. 2009.
- [3] M. Chavent, Y. Lechevallier, F. Vernier, and K. Petit. Monothetic Divisive Clustering with Geographical Constraints. In P. Brito, editor, *COMPSTAT 2008*, pages 67–76. Physica-Verlag HD, 2008.
- [4] C. Ambroise, M. Dang, and G. Govaert. Clustering of Spatial Data by the EM Algorithm. In A. Soares, J. Gómez-Hernandez, and R. Froidevaux, editors, geoENV I —*Geostatistics for Environmental Applications*, number 9 in Quantitative Geology and Geostatistics, pages 493–504. Springer Netherlands, 1997.
- [5] R. Webster. *Quantitative and numerical methods in soil classification and survey*. Clarendon Press, Oxford; New York, 1977.
- [6] R. W. M. A. Oliver. A geostatistical basis for spatial weighting in multivariate classification. 21(1): 15–35, 1988. ISSN 0882-8121. doi: 10.1007/BF00897238.

MERCI DE VOTRE ATTENTION...