Séminaire interne IREDU du 20 janvier 2004 JB/v0,1

Méthodes de classification hiérarchique et de segmentation, quelques applications autour du « classement » des universités en France

Cette note n'est en rien un papier de recherche ; elle a simplement pour but d'illustrer l'utilisation de techniques d'analyse des données en présentant, de manière ouverte, quelques possibilités pour l'évaluation des systèmes éducatifs.

Un séminaire précédent a évoqué la question des analyses factorielles[1], ces analyses comme la régression et les autres méthodes réductives s'adaptent parfaitement aux sciences sociales par le fait qu'elles réalisent des réductions ; c'est à dire qu'elles permettent au chercheur de dominer un univers complet et complexe, en présentant un ensemble réduit et ordonné de manière causale, hélas au prix d'une certaine erreur plus souvent pudiquement nommée variance résiduelle.

Comme le chercheur s'attache à décrire des mondes complexes sur le terrain de l'éducation, il devrait disposer d'une représentation de ces mondes elle aussi complexe et diversifiée. Il est en effet clair de constater que les systèmes éducatifs sont de grands producteurs de données. Ceci n'a pas toujours été le cas et beaucoup d'analyses ont été impactées par la non disponibilité des données, ces dernières ayant alors conditionné les méthodes. On retrouve la question de l'homme qui recherche ses clefs sous le réverbère par facilité, y trouvant de la lumière, quoique n'étant pas certain de les avoir égarées là. Ceci est une question profonde que nous devrons garder à l'esprit.

[1] Séminaire du mardi 16 décembre 2003,– Quelques applications d'analyses factorielles dans le domaine de l'éducation et de la formation / Marc Demeuse, Université de Liège – IREDU – voir l'hyperlien : http://www.u-bourgogne.fr/IREDU/sem16123.htm

Dans les temps présents cette fourniture des données influence largement des méthodes. Ceci vient du fait que les moyens informatiques peuvent fournir des données en quantité considérable. Ainsi on peut citer sans exclusive :

- la classification des banques de données génétiques,
- les systèmes d'information géographiques alimentés par la télédétection et les images satellitaires, associant à des données des coordonnées spatiales,
- l'utilisation directe des données d'activité commerciales, en flux continu, dans la recherche en gestion.

1. Classification et segmentation

En zoologie, le nombre de vertèbres suffit á distinguer les oiseaux des mammifères. La condition idéale est ainsi que les critères typologiques seraient pleinement fondés pour classer sans ambiguïté. Dans les sciences sociales de nombreux critères de description sont concurrents pour un même objet et le risque entropique est constant. Il convient de se demander pourquoi un critère et un seul peut faire la distinction. Les opérateurs de classification permettent de segmenter toutes données en classes selon différents critères d'homogénéité. Les opérateurs de segmentation permettent de construire une carte des classes obtenues des données à partir d'un critère et d'en estimer l'imperfection.

En logique donc avec sa dénomination, la classification sert à définir des classes entre les variables qualitatives ou quantitatives qui caractérisent des individus, d'une part, et d'autre part à transposer cette question en classant des individus en fonction des variables qualitatives ou qualitatives qui les caractérisent.

Les données peuvent se présenter sous différentes formes ; elles concernent n individus supposés affectés, pour simplifier, du même poids :

i. un tableau de distances (ou mesure de dissemblance) (n × n) des individus 2 à 2 ;

ii. les observations de p variables quantitatives sur ces n individus ;

iii. l'observation de variables qualitatives ou d'un mélange de variables quantitatives et qualitatives.

Il s'agit, d'une manière ou d'une autre, de se ramener au premier cas de la connaissance de distances 2 à 2 entre les individus. Le choix d'une matrice de produit scalaire permet de prendre en compte un ensemble de variables quantitatives.

D'autres méthodes sont ensuite traditionnellement incluses dans la chaîne de traitement : ACP avec représentation des classes et de leur enveloppe convexe, pour apprécier la qualité de la classification, AFD (analyse factorielle discriminante) et/ou arbre de classification afin d'aider à l'interprétation de chacune des classes de la typologie par les variables initiales, AFCM dans le cas de variables qualitatives.

1.1 Combinaison entre similitudes

La classification des individus répond donc à une logique économique pure. Comment regrouper des individus et considérer en fonction des variables qui les caractérisent, non plus les individus mais des classes d'individus auxquels on exprimera des actions identiques suivant les classes ou dont on attend un comportement identique suivant leurs classes d'appartenance. Par ailleurs ces méthodes doivent renseigner sur le risque du choix, c'est-à-dire l'imprécision accrue ou au contraire le gain de précision que l'on obtient en choisissant de réduire ou d'augmenter d'une classe marginale le nombre de classe.

Dans bien des études, l'objet essentiel de l'analyse est le degré de ressemblance ou, au contraire, de dissemblance, qui existe entre les objets individus ou variables. Les indices de proximités, déjà évoqués au chapitre 4, sont des valeurs qui mesurent et résument le degré de ressemblance entre deux objets définis par un ensemble commun de variables. Ces indices, que l'on peut classer en deux grands groupes : indices de similarité (s_{ij}) et ceux de dissimilitudes (d_{ij}), se caractérisent par un certain nombre de propriétés mathématiques.

- Propriété 1 : normalisation, cette condition permet de distinguer les indices de similarité de ceux de dissimilitude. On impose au premier d'atteindre son maximum en 1 (lorsque l'objet est comparé á lui-même) et au second d'atteindre son minimum en zéro.

- Propriété 2 : non-négativité : la première condition pour qu'un indice de similarité, dissimilitude soit une proximité est que cet indice soit positif ou nul. $s_{ij} >= 0$ ou dij >_ 0 $\forall$ i,j $\in$ X et $\forall$ i $\in$ X d(i, i) = 0.
- Propriété 3 : Symétrie: la proximité ne dépend pas de l'ordre de présentation: $\forall$ i,j $\in$ X dij = dji

- Propriété 4 : d(i, j) <= d(i, k) + d(j, k); $\forall$(i, j, k) $\in R^3$.

La masse des distances mesurées permet de préciser l'inertie d'un tableau de données, qui a été définie comme la *mesure de dispersion multidimensionnelle*. Ainsi si nous disposons d'un tableau de présentation classique de p variables identifiant n individus, la disposition du tableau est la suivante :

$$y_{ij} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ y_{2,1} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ y_{n,1} & \cdots & \cdots & y_{n,p} \end{bmatrix}$$

$$g = \left\{ \begin{matrix} \bar{y}_1, & \bar{y}_2, & \ldots\ldots, & \bar{y}_n \end{matrix} \right\}$$

Le vecteur g aura comme coordonnées les moyennes de chaque série, aussi l'inertie peut se définir par

$$I(y) = 1/n \sum_{j=1} \left\| y_j - g_j \right\|^2 \text{ où } \left\| y_j - g_j \right\| = \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2$$

1.2 Un critère de qualité d'une partition : le critère d'inertie inter-classes

On se donne une partition de l'ensemble des individus en *k* classes, avec évidemment k<p, par l'intermédiaire des ensembles d'indices des groupes. On note *Is*, ensemble d'indices du groupe *s*, ns_s son effectif, et g son barycentre ou point moyen du groupe. Ce point moyen a encore pour coordonnées les valeurs moyennes des variables dans le groupe. On définit alors :

- D'une part, l'*inertie inter-groupe* (ou inertie *entre* les groupes ou inertie « between »), mesurant l'écartement des groupes, par :
  $I_{nb}(y) = \sum_{s=1 à k} n_k / n \left\| g_k - g \right\|^2$ c'est l'inertie du nuage des barycentres des groupes.
- D'autre part, l'*inertie intra-groupe* (ou inertie *à l'intérieur*, within, des groupes) par :

$$I_{nw}(y) = 1/n \sum_{s=1 à k} \sum_{i \in Is} \left\| y_i - g_s \right\|^2$$

On a alors le résultat suivant : **Inertie totale** = **inertie inter groupe** + **inertie intra** groupe

Ainsi pour ce critère, chercher une partition qui maximise l'inertie inter (donc qui tend à disperser au mieux les groupes), revient à chercher une partition minimisant l'inertie intra (donc qui tend à obtenir des groupes les plus compacts). Mais le maximum du critère est obtenu pour la partition discrète, ayant comme classes les singletons (chaque classe est un individu). On est donc amené à maximiser l'inertie inter *pour un nombre de classes fixé a priori.*

On se rend compte que la classification peut s'assimiler dans sa démarche à une approche combinatoire (voir annexe), ce qui dans la pratique se transforme dans des choix entre des algorithmes diversifiés. Les méthodes essaient de réduire le nombre de partitions, à comparer à chaque étape, afin de se diriger vers une partition acceptable. Les méthodes se limitent à l'exécution d'un algorithme itératif convergeant vers une"bonne" partition qui peut correspondre `a un optimum local. Aussi, même si le besoin de classer des objets est très ancien, seule la généralisation des outils informatiques en a permis l'automatisation dans les années 70. Les méthodes de classification hiérarchique peuvent être appliquées à des tableaux de données de taille modeste plus facilement. Toutefois leur application à des grandes bases de données peut souvent être envisagée par la réduction de la taille du tableau initial. Une stratégie souvent adoptée est de réduire d 'abord le nombre de colonnes par une analyse factorielle, puis le nombre de lignes par une méthode rapide de classification par partitionnement. On en déduit ainsi un tableau réduit qui peut être traité par une méthode de classification hiérarchique, mais ses lignes doivent être considérées comme des unités statistiques complexes. ; Celeux et alii (1989) décrivent en détail ces algorithmes. Différents choix sont laissés à l'initiative de l'utilisateur :

- • une mesure d'éloignement, de dissemblance ou de distance entre individus ;
- • le critère d'homogénéité des classes à optimiser : il est, dans le cas de variables quantitatives, généralement défini à partir des traces des matrices de variances inter (ou inertie) ou intra de la partition ;
- • la méthode : la classification ascendante hiérarchique ou celle par réallocation dynamique sont les plus utilisées, seules ou combinées,
- • le nombre de classes ; ceci reste un point délicat et entaché de subjectif

1.3 Les démarches de classification

Les méthodes de classification sont tout aussi variées que les données à analyser, si l'on fait l'exception de méthodes particulièrement utilisées pour la classification génétique où la détection spatiale, on note sept groupes de méthodes :

1. Classification Ascendante Hiérarchique
2. Classification non hiérarchique ou partitionnement
3. Méthodes mixtes
4. Méthodes basées sur le modèle de mélange de Gaussiens
5. Classification « floues »
6. Cartes (SOM) de Kohonen
7. Méthodes basées sur la théorie des graphes

On détaille la question de la CAH, les six autres méthodes n'étant analysées qu'en différentiel par rapport à cette première méthode dans un second temps.

1.3.1. Classification Ascendante Hiérarchique (CAH)

Cette méthode nécessite la définition d'une mesure de similarité ou de distance entre les objets à classifier ou échantillons. Ceci conduit à définir un critère d'agrégation des classes qui peut être défini comme une mesure de similarité ou de distance entre les classes d'objets. Le critère d'agrégation est pris souvent à travers des projections factorielles[2]. La méthode produit une suite de partitions emboîtées de l'ensemble des objets à classifier. Cette méthode classique dans son utilisation est donc proche néanmoins des méthodes d'analyse factorielle.

l'algorithme de Lance et Williams
Au départ, on a partition maximale en n classes, chaque classe étant composée d'un seul objet (partition la plus fine). On agrège, itérativement, à chaque pas, les deux objets les plus ressemblants au sens de la similarité, ici la distance, où on identifie deux classes d'objets optimisant le critère d'agrégation jusqu'à obtenir une seule classe composée de tous les objets (la partition la moins fine)

coupure de l'arbre
Dans la suite, Les partitions sont représentées par un arbre de classification ou un « dendrogramme »[3]. L'ensemble des noeuds définit une «hiérarchie» sur l'ensemble d'objets. Le nombre de classes obtenues dépend du niveau de l'arbre choisi pour la coupe. Il existe des «indices de niveau» pour évaluer la qualité d'une partition et le gain ou la perte de similitude au niveau de chaque nœud.

Les mesures de distances
Les distances les plus «populaires» en sciences sociales sont pour p variables et entre des objets i et j :
La distance euclidienne, celle-ci s'exprime par

$$d^2(x_i, x_k) = \sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2$$

La distance déduite du coefficient de corrélation de Pearson

$$d(x_i, x_k) = 1 - r_{(x_i, x_k)}$$

D'autres mesures de distance alternatives sont utilisées :

- Distance de «Manhattan», distance identique à l'euclidienne mais non quadratique
- Distance *I-norme*, dont la distance euclidienne et la distance de Manhattan sont des cas particuliers et qui regroupe l'ensemble des possibilités d'écarts terme à terme.
- La distance de Malanobis, chère aux analyses factorielles, décrit la distance entre classes de points et non entre points. La confusion vient du fait que quand on sait calculer la distance entre deux points on peut s'en servir pour calculer la distance entre les centres de gravité de deux sous nuages de points et on dit alors qu'on a une distance entre classes.
- Distance du *Khi-deux*, elle recherche une distance, élément par élément, entre le profil de la distribution observée et celui d'une distribution théorique

Les critères d'agrégation, après le calcul des distances, sont eux utilisés pour la détermination des classes et reviennent à rassembler les éléments candidats suivant les références :

- Le lien minimum (single-linkage) : méthode du plus proche voisin,
- Le lien maximum (complete-linkage) méthode du voisin le plus éloigné (ou du diamètre),

[2] Ceci est d'autant plus important que souvent seuls les 2 premiers axes factoriels sont utilisés. Ceci est très stimulant dans la mesure où la représentation fournie dans ce premier plan principal permet de donner directement les distances entre objets. Ceci est très frustrant dans la mesure où la qualité de classification dépend de la valeur des axes factoriels donc des valeurs caractéristiques de la matrice de covariance ; à l'identique des méthodes de régression dont on veut se différencier.

[3] On passe du niveau le plus fin à un niveau immédiatement supérieur (n-1) en agrégeant les deux objets les lus similaires.

- Le lien moyen (average-linkage), connu sous le nom de UPGMA,
- Le critère de Ward (minimisation de la variance intra-classe).

1.3.2 Les méthodes alternatives

Méthode De K-means (ou de centres mobiles ou aussi « nuées dynamiques »)
Cette méthode répond à un algorithme précis:
1. On choisit (éventuellement tiré au hasard) K «noyaux» ou individus de référence,
2. On affecte les objets (individus ou échantillons) aux noyaux dont ils sont les plus proches,
3. On recalcule les noyaux (centres) des classes,
4. On répète 2 et 3 jusqu'à convergence
Au final, on a une partition de l'ensemble d'objets en au plus K classes. Le résultat est fortement dépendant du choix initial de noyaux et de la distance retenue.

Modèle de Mélanges Gaussiens (Mixture Modelling)
Ce sont des méthodes basées sur des modèles ex ante (Model Based) qui évidemment supposent la validité de la distribution gaussienne. Chaque classe est représentée par le modèle de densité répondant à la distribution normale :

$$\phi_k(x|\mu_k,\Sigma_k) = (2\pi)^{\frac{-p}{2}}|\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right\}$$

Où x représente les données; les classes sont ellipsoïdales de centres μk, la matrice de covariances est Σk dans chaque classe k. Les matrices de covariances sont paramétrées par leurs décompositions en valeurs singulières Dk, Ak et λk qui sont respectivement les paramètres d'orientation (orientation des composantes principales des matrices Σk), la forme du contour de la densité et le volume de l'ellipsoïde. Ces paramètres de la distribution sont estimés à partir des données par l'algorithme EM (Expectation Maximization). Cette méthode EM distingue de manière éponyme les étapes E et M qui sont itérées jusqu'à la convergence ; l'étape E calcule une matrice Z dont l'élément z_{ik} est une estimation de la probabilité conditionnelle que l'objet j soit dans la classe k sachant les paramètres de distribution. L'étape M effectue l'estimation des paramètres par la méthode du maximum de vraisemblance sachant Z.

SOM (self Organizing Maps)
Le SOM de Kohonen est une méthode d'apprentissage non supervisée, utilisant le réseau de neurones artificiels, qui semble intéressant pour analyser et visualiser les données lexicales qualitatives. On considère une couche des données (vecteurs dans un espace) et une couche de neurones. Les neurones son liés entre eux (par une fonction de voisinage) et avec les vecteurs de données. Un réseau de neurones est un graphe orienté dont les éléments sont un ensemble de neurones connectés les uns aux autres par des liaisons synaptiques entre axone et dendrite. On dit que le réseau de neurones passe d'un état à un autre lorsque tous ses neurones ont recalculé leur état interne, en fonction de leurs entrées. Ce processus itératif peut être effectué en séquentiel ou en parallèle.
Dans le SOM de Kohonen, la représentation logique de la couche des neurones est un treillis (souvent représenté par une grille dans un plan)[4]. Au début, un certain nombre de vecteurs sont associés à chaque noeud au hasard. Pendant le processus d'apprentissage les vecteurs s'ajustent progressivement pour couvrir l'espace de façon à réduire la complexité. Au final, on obtient un certain nombre de groupes d'objets ordonnancés.

Méthode K-means «Floue»
C'est une combinaison de la méthode K-means et la méthode de classification floue ; chaque objet appartient aux différentes classes avec

[4] La question des réseaux de neurones n'est jamais simple à saisir dans les sciences sociales où le fond de commerce des théories est déjà un réseau de neurones a lui seul. Il semble plus direct de donner une image, un réseau de neurone est une méta connaissance transformée dans une relation rigide (cablée dirait-on), on se retrouve comme dans un modèle statistique dans un ensemble de relations structurée sauf qu'ici suite au processus d'apprentissage l'on passe d'un état probable à un état de certitude. Un exemple dans l'analyse de l'école pourrait distinguer, dans un processus d'évaluation des élèves, la couche statistique du résultat des items et des caractéristiques de classe des variables d'environnement (structure pédagogique, carte scolaire…) qui seraient du domaine du réseau neuronal.

une certaine probabilité cette dernière exprimant l'incertitude

1. Les noyaux initiaux sont déterminés par les vecteurs propres (les axes principaux d'inertie) issus de l'ACP
2. Pour chaque noyaux, on détermine le degré d'appartenance de chaque élément à la classe à partir des similarités ou des distances entre objets.
3. Les nouveaux noyaux sont calculés par les moyennes pondérées de tous les profils d'expression individus, pondérées par leurs degrés d'appartenance à chaque centre,
4. A la convergence, on réévalue les degrés d'appartenance des gènes à chaque classe, suivant la formule suivante où Σk , Dk, Ak et λk conservent les m^mes significations que pour les modèles gaussiens.

$$\sum_k = \lambda_k D_k A_k D_k^T$$

1.4 De la classification vers la segmentation

Les méthodes de classification consistent à créer des arbres qui retracent l'affectation des objets et individus entre classe. Ces méthodes sont uniquement descriptives ; une valeur ajoutée à l'analyse, revient à tester la sensibilité de ce partage entre classes et dans quelle mesure une variable Y peut être reliée à cette typologie de classes. Si l'on part de l'arbre binaire de classification au niveau le pus détaillé, une méthode de segmentation fournit, à partir de l'arbre complet, la séquence des sous-arbres obtenue en utilisant une procédure d'*élagage* basée sur la suppression successive des branches les moins informatives en termes d'explication de la variable réponse Y. Dans cette séquence d'*élagage*, elle sélectionne un sous-arbre "optimal" à l'aide de *l'échantillon de base* (en se basant sur l'estimation de la variance résiduelle des différents segments terminaux). L'idée fondamentale est de sélectionner chaque division d'un sous-ensemble (ou segment) de telle sorte que la variance de la variable à expliquer Y dans les segments descendants soit plus faible que la variance de Y dans le segment parent. Plus précisément, pour toute division d'un segment, on calcule la moyenne pondérée de la variance de Y dans ses segments descendants (variance *intra*). La meilleure division est, parmi toutes les divisions possibles à l'aide des variables explicatives, celle qui minimise cette variance *intra.* On retient ainsi la division qui conduit à deux segments descendants aussi homogènes que possible en Y. Un problème de régression multiple se pose quand on est en présence d'un tableau de données contenant une variable privilégiée (variable réponse), Y continue **à** expliquer par les autres variables du tableau X1, X2 ... Xp dites variables explicatives. Il s'agit alors d'une part, de sélectionner parmi ces variables celles qui expliquent significativement le phénomène Y, et d'autre part, de construire une règle de prédiction de la valeur de Y pour un nouvel individu.

Il existe deux grandes familles d'analyse de la segmentation suivant que l'on s'attache à une variable catégorielle, à lecture qualitative, ou une variable continue ou d'ordonnancement.

2. Mise en pratique : les données

Dans le cas présenté nous prenons l'ensemble statistique formé par les 85 universités françaises avec des variables explicatives X qui les caractérisent. A cet ensemble nous soustrayons les 3 universités technologiques et celles des DOM et TOM pour des raisons qui tiennent à la fois de l'homogénéité des données et de la dissemblance des observations. Les variables décrivent la situation de l'année 2001.

Les cas analysés sont donc les suivants :

- le nom de l'établissement,
- le nom cours utilisé dans les résultats,
- la taille en nombre d'étudiants.

Tableau 1 les établissements

NOM	NOMC	TAILLE
Aix-Marseille 1	Aix1	25 231
Aix-Marseille 2	Aix2	19 347
Aix-Marseille 3	Aix3	21 648
Amiens	Amiens	20 125
Angers	Angers	15 742
Artois	Artois	11 038
Avignon	Avignon	7 103
Besançon	Besançon	20 750
Bordeaux I	Bdx1	10 722
Bordeaux II	Bdx2	15 175

Bordeaux III	Bdx3	14 847
Bordeaux IV	Bdx4	12 681
Bourgogne	Dijon	24 948
Brest	Brest	16 252
Bretagne-Sud	BretaS	6 341
Caen	Caen	25 900
Cergy-Pont	Cergy	10 261
Chambéry	Chambéry	12 101
Clermont I	Cle1	11 698
Clermont II	Cle2	15 182
Corse	Corse	3 509
Evry-Val-d'es.	Evry-Val-d	9 117
Grenoble I	Gre1	17 471
Grenoble II	Gre2	18 600
Grenoble III	Gre3	6 946
La Réunion	Réunion	10 218
La Rochelle	La Rochell	6 243
Le Havre	Havre	7 111
Le Mans	Le Mans	8 062
Lille I	Lil1	20 442
Lille II	Lil2	19 902
Lille III	Lil3	21 055
Limoges	Limoges	13 677
Littoral	Littoral	10 883
Lyon I	Ly1	27 437
Lyon II	Ly2	25 188
Lyon III	Ly3	19 046
Marne-la-V	M-la-V	8 833
Metz	Metz	15 898
Montpellier I	Montp1	12 708
Montpellier II	Montp2	21 124
Montpellier III	Montp3	19 808
Mulhouse	Mulhouse	7 552
Nancy I	Nan1	15 725
Nancy II	Nan2	18 957
Nantes	Nantes	32 819
Nice	Nice	26 399
Orléans	Orléans	16 120
Paris I	Par1	35 950
Paris II	Par2	17 319
Paris III	Par3	17 675
Paris IV	Par4	23 124
Paris IX	Par9	7 178
Paris V	Par5	27 665
Paris VI	Par6	29 594
Paris VII	Par6	24 744
Paris VIII	Par8	26 804
Paris X	Par10	33 631
Paris XI	Par11	26 488
Paris XII	Par12	23 588
Paris XIII	Par13	19 347
Pau	Pau	13 128
Perpignan	Perpignan	8 554
Poitiers	Poitiers	24 096
Reims	Reims	22 284
Rennes I	Ren1	24 298
Rennes II	Ren2	20 174
Rouen	Rouen	23 988
St-Etienne	St-Etienne	13 159
Strasbourg I	Stb1	16 458
Strasbourg II	Stb2	13 206
Strasbourg III	Stb3	8 478
Toulon	Toulon	9 652
Toulouse I	Tou1	16 804
Toulouse II	Tou2	27 000
Toulouse III	Tou3	28 184
Tours	Tours	22 769
Valenciennes	Valenc	11 102
Versailles	Versail	10 579

Le tableau 2 donne les noms et sources des variables.

Tableau 2, noms et sources des variables, les variables dont le nom est en italiques ne sont pas directement utilisées en fonction des biais de structure induits.

Nom	Intitulé	Source
ACADEMIES	nom de l'académie	MEN DEP
REG	code région	INSEE
Nomreg	Nom région	
DEP	Code département	
NOMDEP	Nom département	
NOM	Nom université	
CLASS	Classement mars 2003	Classement général du nouvel observateur, mars 2003
PartDnat	Part de la population du département dans l'ensemble national	INSEE
PIBD	PIB département frs 2000	INSEE
POPDR	Part du département d'implantation dans la population régionale	
Ptaille	Part de l'établissement dans la population étudiante de la région	
PIBDT	PIBD / tête, pour 2000 en francs courants	
TXEE	Taux d'encadrement (enseignants/étudiants)	
TXPR	Part professeurs	MEN DEP/DES
TXMCF	Part MCF	MEN DEP/DES
TXAUT	Part autres enseignants	MEN
UMRT	Nombre UMR/ Nombre étudiants	Annuaires 2001 – CNRS et INRA
D5A	Deug en 5 ans %	MEN DEP
D3A	Deug en 5 ans %	MEN DEP
D2A	Deug en 5 ans %	MEN DEP
TAILLE	Nombre étudiants	MEN DEP
TAILLR	Part de l'U. dans le total des étudiants	
EVOL99	Tx évolution étudiants 99/2002 (%)	MEN DEP
PREMINSC	Premières inscriptions par an Moyenne 1999/2002	MEN DEP
TPREMCYCL	Taux étudiant premier cycle	MEN DEP
TFC	Taux étudiants formation Continue	MEN DEP
FCTETUD	fonctionnement par étudiant frs 2000 AMU et CNE	
INVETUD	investissement par étudiant moyenne 99/2000	
PARTMED	Part des études médicales	MEN DEP
PARTSCI	Part études scientifiques	MEN DEP
POP99	Population région 99	INSEE
PEMPDREG	Part de l'emploi du département d'implantation dans la région 1999	INSEE

3. Résultats

Trois familles d'analyse vont être mises en œuvre sur ce fichier : la classification hiérarchique, les nuées dynamiques et la segmentation de l'arbre. Les deux premières sous XLSTAT, la troisième sous SPAD.

3.1 classification hiérarchique

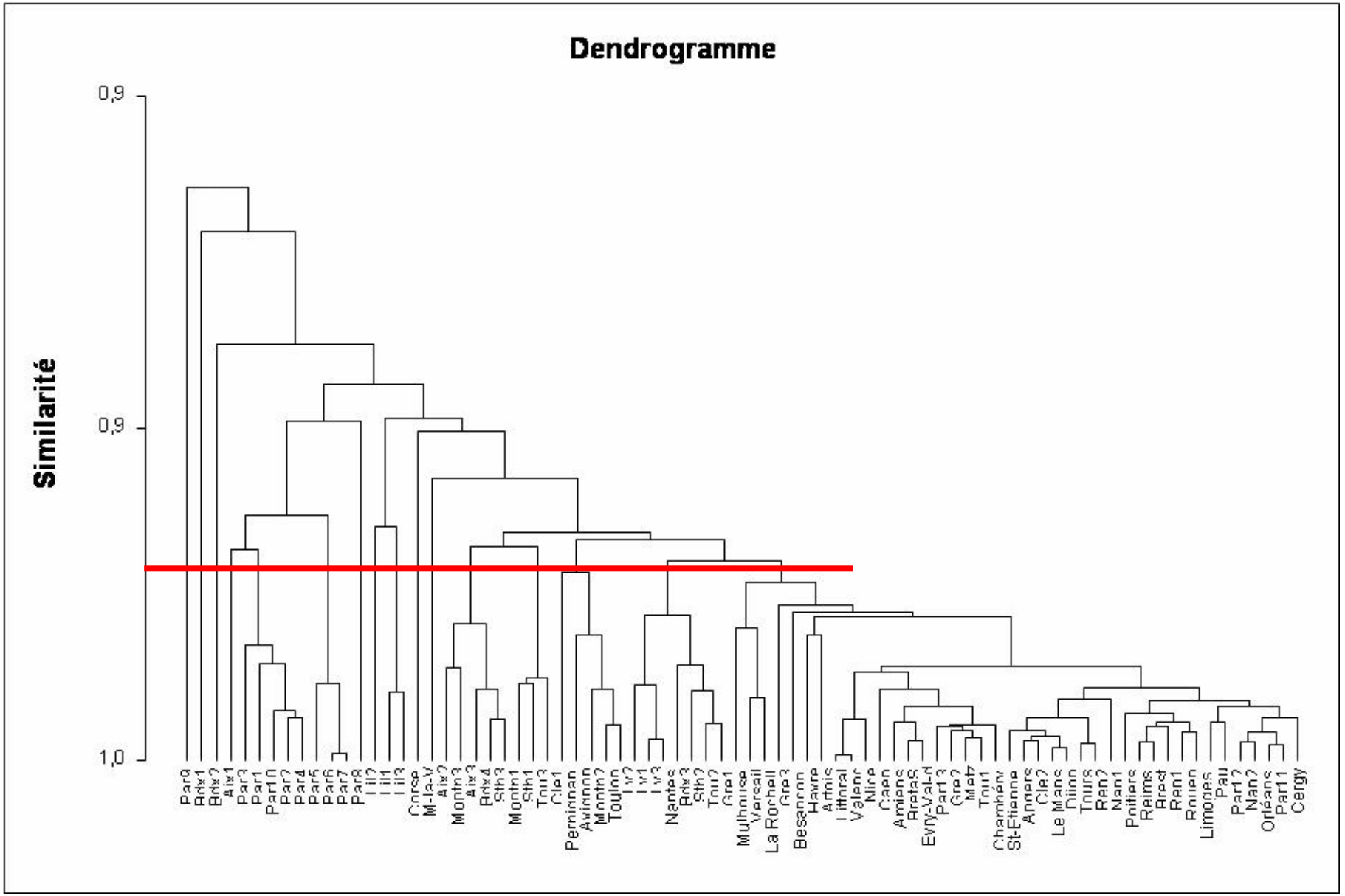
Les premiers tests sont ceux d'une classification hiérarchique ; trois classifications sont réalisées à partir des variables du tableau 2 :

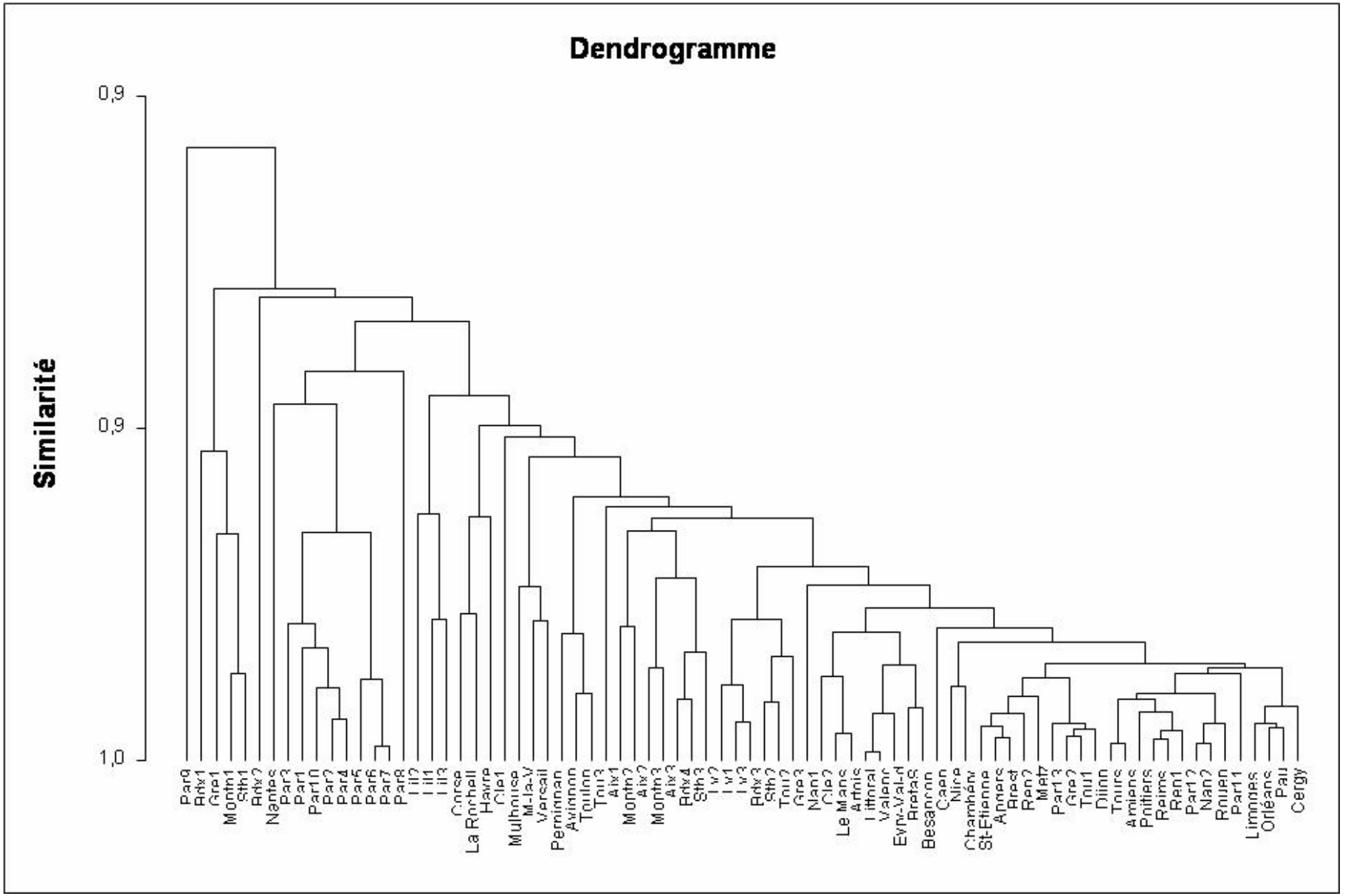
- la description par les variables de caractéristiques de taille, structure de discipline, rendement au DEUG;
- les mêmes variables avec les dépenses récurrentes et d'investissement et de structure du corps enseignant de l'Université ;
- l'ensemble des variables en ajoutant structures de recherche mixtes et les variables socio-économiques.

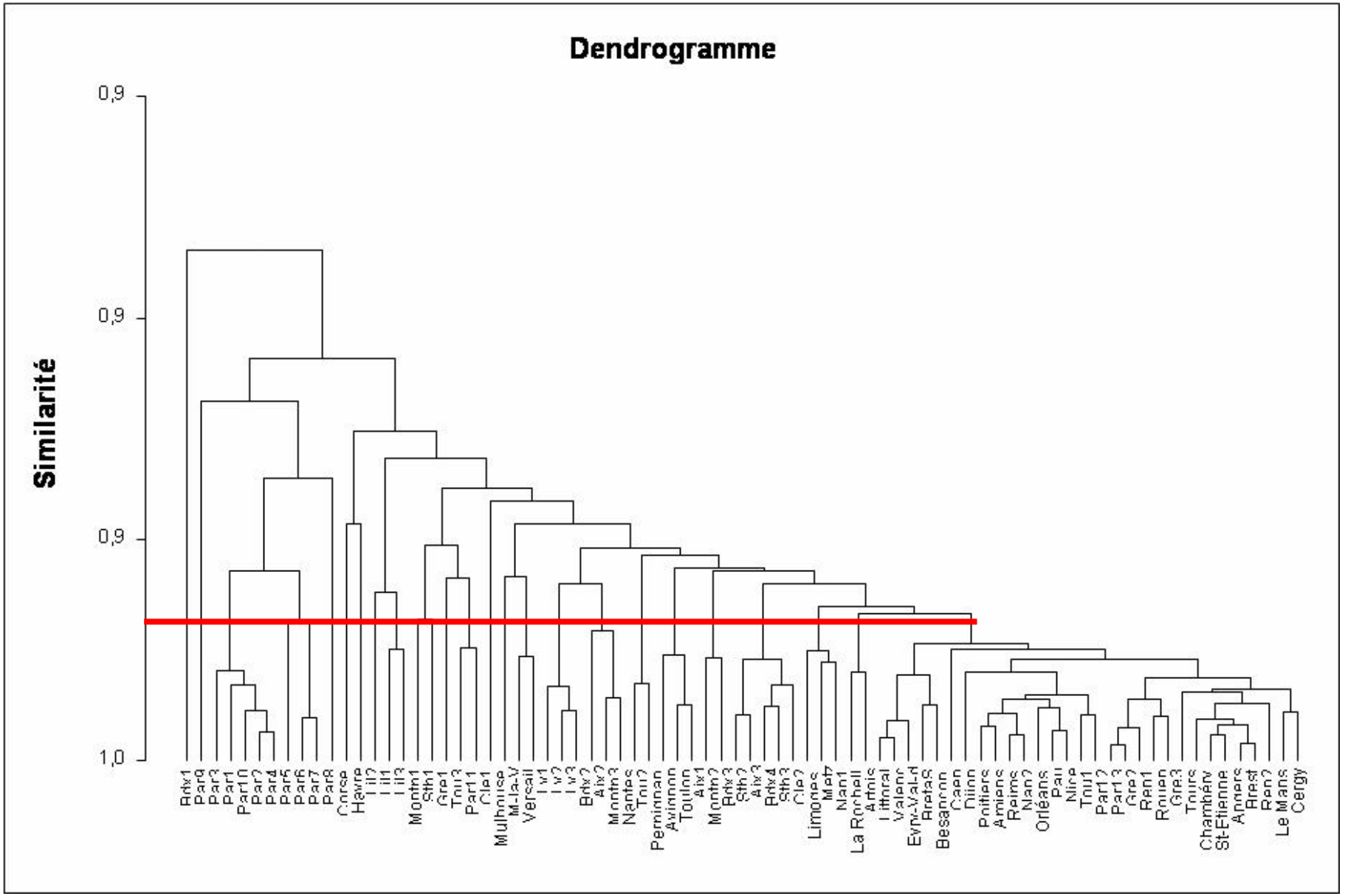
La lecture du *dendogramme* est assez intuitive[5], la lecture verticale est celle de la séparation des variances. Au plus haut du graphique toutes les observations sont agrégées dans un seul tronc ; le graphique ne compte qu'un segment unique. Les segments vont en se séparant, au niveau de *nœuds*, jusqu'au niveau ultime où l'effectif de classe est individuel. La longueur verticale de chacun de ces segments représente donc le gain en variance interclasses venant de la séparation. Plus un segment « père » possède de segment fils, donc de nœuds, moins il rassemble d'unités observations spécifiques. La lecture du dendogramme se réalise donc en coupant l'arbre suivant l'axe vertical ; ainsi la coupure représentée par la droite horizontale, sur le dendogramme de la page 10 permet d'identifier, d'une part, des établissements très spécifiques (Paris 9, Bordeaux 1, Bordeaux 2), un sous ensemble d'établissements parisiens à dominante juridique et scientifique. D'autre part des homogénéités géographiques remarquables (Lyon, Lille), ou en second lieu des ensembles géographiques non homogène mais dont l'ensemble des éléments est assez spécifique (Bordeaux, Strasbourg, Toulouse…). En fin d'arbre, disons à partir de Mulhouse (U. de Hte Alsace) un ensemble d'établissement peu spécifique qui paraissent être plus orienté vers le service public d'enseignement au niveau local, de création plus récente en moyenne et plutôt généralistes. Le fait de retrouver des établissements « périphériques » de l'Ile de France n'est pas anodin.

Le second dendogramme (page 11) introduit les éléments qualitatifs et de financement de l'éducation. Ceci conduit à que des situations spécifiques apparaissent comme Nantes, Montpellier 1, ou se renforcent comme Strasbourg 1. Des établissements récents montrent une spécificité renforcée (Mulhouse par exemple), alors que les blocs géographiquement homogènes comme Lille et Lyon ne paraissent pas remis en cause.
En introduisant les variables de recherche et d'environnement économique, dendogramme de la page 12, des modifications sont notables comme la transition vers plus de caractère spécifique de Paris 11, mais aussi le renforcement du dualisme Paris s'opposant au reste du territoire. On reste aussi assez surpris par la relative modification des homogénéités de groupes géographiques suivant les critères retenus (Bordeaux, Grenoble,…) qui s'oppose à des homogénéités géographiques quelles que soient les variables prises en compte (Lille, Lyon). On remarquera surtout que cette introduction des variables économiques territoriales tend à « casser » davantage les classes. Ainsi, si l'on réalise une coupure d'arbre pour un niveau identique de rapport entre les variances inter et les variances intra, on est confronté à un plus grand nombre de classe. Ce constat courant, même si les variables sont ici sans dimension et normalisés vient de la conjonction, entre ces indicateurs, d'une dimension de variance entre les établissements basée sur l'histoire (donc l'ancienneté de ceux-ci) et l'hétérogénéité des potentiels économiques à travers le territoire.

[5] Ici n'est conservée que la seule information graphique du dendogramme, une analyse exhaustive doit comporter l'information de modification du partage variances inter et intra marginales au niveau de chaque nœud.


Dendrogramme
Similarité
0,9
0,9
1,0
Par9
Rbx1
Rbx2
Aix1
Par3
Par1
Par10
Par2
Par4
Par5
Par6
Par7
Par8
Lil2
Lil1
Lil3
Corse
M-la-V
Aix2
Mont3
Aix3
Rbx4
Sth3
Mont1
Sth1
Tou3
Cle1
Perpignan
Avignon
Mont2
Toulon
Lyo2
Lyo1
Lyo3
Nantes
Rbx3
Sth2
Tou2
Gre1
Mulhouse
Versail
La Rochell
Gre3
Besancon
Havre
Arlnic
Littoral
Valenc
Nice
Caen
Amiens
Bretag
Evry-Val-d
Par13
Gre2
Metz
Tou1
Chambéry
St-Etienne
Angers
Cle2
Le Mans
Dijon
Tours
Ren2
Nan1
Poitiers
Reims
Brest
Ren1
Rouen
Limoges
Pau
Par12
Nan2
Orléans
Par11
Cergy


Dendrogramme
Similarité
0,9
0,9
1,0
Par9
Rdx1
Gre1
Montn1
Sth1
Rdx2
Nantes
Par3
Par1
Par10
Par2
Par4
Par5
Par6
Par7
Par8
Lil2
Lil1
Lil3
Corse
La Rochell
Havre
Cla1
Mulhouse
M-la-V
Versail
Perpignan
Avignon
Toulon
Tou3
Aix1
Montn2
Aix2
Montn3
Aix3
Rdx4
Sth3
Iv2
Iv1
Iv3
Rdx3
Sth2
Tou2
Gre3
Nan1
Cla2
Le Mans
Arlnis
Littoral
Valanc
Evry-Val-N
Bretas
Besancon
Caen
Nice
Chambéry
St-Etienne
Annecy
Brest
Ran2
Met7
Par13
Gre2
Tou1
Dijon
Tours
Amiens
Poitiers
Reims
Ran1
Par12
Nan2
Rouen
Par11
Limoges
Orléans
Pau
Cergy


Dendrogramme
Similarité
0,9
0,9
0,9
1,0
Rdx1
Par9
Par3
Par1
Par10
Par2
Par4
Par5
Par6
Par7
Par8
Corse
Havre
Lil2
Lil1
Lil3
Montn1
Sth1
Gre1
Tour3
Par11
Cle1
Mulhouse
Malakv
Versail
Lyo1
Lyo2
Lyo3
Rdx2
Aix2
Montp3
Nantes
Tour2
Perpignan
Avignon
Toulon
Aix1
Montp2
Rdx3
Sth2
Aix3
Rdx4
Sth3
Cle2
Limoges
Metz
Nan1
La Rochell
Arlnis
Littoral
Valence
Evry-Val-d
Brest3
Besancon
Caen
Poitiers
Amiens
Reims
Nan2
Orléans
Pau
Nice
Tou1
Par12
Par13
Gre2
Ren1
Rouen
Gre3
Tours
Chambéry
St-Etienne
Angers
Brest
Ren2
Le Mans
Cergy

3.2 Nuées dynamiques

Cette méthode, présentée au 1.3, tend donc à constituer des classes à variance intra minimales avec un choix ex ante d'un nombre k de classes. Il s'agit donc ici d'une agglomération autour de « noyaux », alors que la CAH qui précède découle d'une démarche d'individu candidat à la similitude par rapport à une classe donnée.

Le tableau 3 reprend les résultats obtenus pour un nombre de noyaux égale à 5, sachant que l'ensemble des variables caractérisant les établissements est ici introduit d'entrée. Le tableau donne pour chaque groupe la variance intra, le nombre d'éléments groupés et leur identifiant court.

Tableau 3 : Groupement autour de 5 noyaux

Classes	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Inertie intra	1,24E+08	1,45E+08	2,28E+07	6,21E+07	7,32E+07
Effectif	20	18	4	18	17
	Avignon	Bdx2	Nantes	Aix2	Caen
	Bdx1	Bdx3	Par1	Aix3	Par12
	Corse	Bdx4	Par10	Amiens	Dijon
	Evry-Val-d	Cle1	Par6	Besançon	Ly1
	M-la-V	Cle2		Par13	Ly2
	Gre3	Gre1		Gre2	Nice
	Artois	Chambery		Lil1	Tours
	Littoral	Limoges		Lil2	Par4
	Valenc	St-Etienne		Lil3	Par5
	Perpignan	Montp1		Ly3	Par7
	Le Mans	Nan1		Montp2	Par8
	Toulon	Metz		Montp3	Poitiers
	Par9	Angers		Nan2	Ren1
	La Rochell	Orléans		Par2	Rouen
	BretaS	Brest		Par3	Tou2
	Havre	Stb1		Reims	Tou3
	Mulhouse	Stb2		Ren2	Par11
	Stb3	Pau		Tou1	
	Versail				
	Cergy				

Ici des résultats restent communs par rapport à la CAH, opposition Paris-province, relative indépendance avec la dominante scientifique ou littéraire de l'université et l'on retrouve certaines classes d'homogénéité spatiale (Lille). Il est intéressant de voir que la classe 4 qui reste la plus homogène se caractérise par un rendement interne médiocre au DEUG par rapport à la classe 5

3.3 Segmentation par régression non paramétrique d'un arbre

Cette procédure, mise en œuvre ici sous Spad, effectue la construction d'un arbre de décision binaire complet pour la régression non-paramétrique d'une variable à expliquer quantitative (variable réponse) sur un ensemble de variables explicatives qui peuvent être de nature quelconque : continues, ordinales ou nominales.

Dans le cas présenté nous prenons, pour représenter notre variable Y le classement des universités françaises établies par le « Nouvel observateur » en mars 2003[6]. Cet indicateur continu Y indique le rang de classement, est mis en relation avec les variables explicatives X suivantes, observées pour les 78 universités hiérarchisées dans le classement. Sous SPAD, la filière d'études va demander la mise en oeuvre de 2 méthodes : la régression par arbre binaire, l'élagage de l'arbre L'ensemble des indicateurs constitue ici les continues exogènes du modèle. La régression par arbre binaire permet de mesurer l'impact des exogènes sur la variable réponse Y. La procédure d'élagage entraîne la suppression des branches les moins informatives de l'arbre et produit cette séquence de sous arbres (A1, A2,...A78). La démarche d'élagage nécessite un échantillon d'individus de référence et des échantillons tests[7]. Ce tableau fournit pour chaque sous arbre de la séquence : i-son nombre de segments terminaux, ii- l'erreur relative de prédiction. Les éléments du tableau 4 permettent d'obtenir le sous-arbre optimal repéré par un * dans la dernière colonne. Il s'agit du plus petit (en terme de nombre de segments terminaux) sous arbre de la séquence correspondant à la plus petite estimation de l'erreur théorique de prédiction sur Y. Dans cet exemple le sous arbre contient 2 segments terminaux dont l'estimation de l'erreur relative théorique de prédiction est 0,5875. L'erreur relative correspondant à l'échantillon de base représente la variance résiduelle de l'arbre, rapportée à la variance s2 de Y à la racine de l'arbre (ou erreur apparente initiale), c'est-à-dire, la part de la variance non expliquée par l'arbre de régression construit à partir de l'échantillon de base.

Tableau 4 d'élagage - Recherche du sous arbre optimal

Arbre	Nombre de segments terminaux	Erreur relative de prédiction - échantillon test	Ecart type associé	Erreur relative de prédiction - échantillon base	Sous arbre optimal
1	14	1,064	0,347	0,065	
2	13	1,014	0,340	0,067	
3	12	0,969	0,321	0,070	
4	11	0,954	0,321	0,084	
5	10	0,909	0,298	0,100	
6	9	0,911	0,298	0,117	
7	8	0,831	0,284	0,143	
8	7	0,708	0,251	0,192	
9	6	0,714	0,252	0,242	
10	5	0,648	0,237	0,313	
11	4	0,713	0,232	0,394	
12	3	0,639	0,226	0,478	
13	2*	0,587	0,165	0,575	Optimum
14	1	1,000	0,034	1,000	

Le résultat essentiel est ici d'expliquer le rôle des variables dans les positions de Y. Ceci est synthétiser dans le tableau 5. Dans ce tableau le numéro de segment indique la position sur l'arbre dendogramme, plus ce numéro est petit plus le segment est élevé dans l'arbre. Ce tableau contient la description de l'arbre élagué (pour l'échantillon de base) : i- les différentes divisions, ii- l'effectif (et la proportion) des différents segments intermédiaires ou terminaux., iii- la moyenne prédite et l'erreur de prédiction des différents segments. La robustesse du résultat est testée en modifiant les choix de répartition entre tests et bases des observations. Le résultat important, dans ce tableau 5, revient à comparer les moyennes prédites, la variable explicative et la valeur de la coupure, soit les colonnes les

[6] Ce classement sur ses critères peut faire l'objet de débats, même s'il échappent à cette note technique, on peut renvoyer vers l'avis du CNE à
www.cne-evaluation.fr/fr/actualite/reponsenouvelobs.pdf

[7] La répartition de taille entre test et référence est laissée à l'utilisateur de la métode, ici l'on a pris 50-50.

plus à droite. Une variable de coupure est ainsi une variable qui est significative quant à segmenter les établissements dans leur classement Y. La mention *segment terminal* traduira l'impact d'observations très spécifiques qui a contrario ne voient que très mal leur position dans Y expliquée par l'ensemble des indicateurs X pris ici en compte. Des positions de ce type en haut de tableau et plus nombreuses que les variables X de coupure induirait une structure de données avec de faibles covariances. La valeur de coupure identifie pour chaque variable X la valeur pour laquelle son impact marginal est le plus important. Il faut aussi mettre en rapport la moyenne prédite de Y et cette dernière valeur d'impact. Ainsi la première ligne du tableau indique que INVETUD, dotation équipement en francs par étudiant 2000 permet de significativement segmenter l'arbre autour d'une valeur Y prédite de 36 pour une valeur de 1462 francs. L'interprétation directe est ainsi la suivante : une université aura une probabilité réelle d'être dans la première moitié du classement si son investissement par étudiant est proche de 1500 frs. A l'identique si le taux de MCF est supérieur à 50%, ceci explique qu'elle se situe dans les 50 premières. Parmi les variables qui influencent la segmentation c'est-à-dire qui permettent de fournir la coupure qui minimisera la variété (erreur de prédiction) interne à chaque sous-groupe de l'arbre, on retrouve aussi l'évolution des effectifs sur moyen terme, la part des filières scientifiques, la taille relative de l'établissement.

Tableau 5 Nombre de segments terminaux (15)

Numéro du segment	Effectif échant. test	Pourcentage	Moyenne prédite classement	Libellé de la variable de coupure	Valeur de coupure
1	43	100,00	36,302	INVETUD	1462,550
2	25	58,14	49,800	TXMCF	0,505
4	12	27,91	39,917	EVOL99	3,050
8	9	20,93	47,222	PARTSCI	0,190
16	6	13,95	53,500	Ptaille	0,010
32	3	6,98	43,333	Segment terminal	
33	3	6,98	63,667	Segment terminal	
17	3	6,98	34,667	Segment terminal	
9	3	6,98	18,000	Segment terminal	
5	13	30,23	58,923	INVETUD	798,315
10	2	4,65	38,500	Segment terminal	
11	11	25,58	62,636	INVETUD	1184,990
22	7	16,28	67,714	D5A	82,950
44	5	11,63	64,400	PartDnat	0,023
88	3	6,98	60,667	Segment terminal	
89	2	4,65	70,000	Segment terminal	
45	2	4,65	76,000	Segment terminal	
23	4	9,30	53,750	Segment terminal	
3	18	41,86	17,556	INVETUD	2683,860
6	12	27,91	23,583	TXPR	0,260
12	4	9,30	12,750	Segment terminal	
13	8	18,60	29,000	POPDR	0,380
26	6	13,95	32,500	PartDnat	0,010
52	2	4,65	25,000	Segment terminal	
53	4	9,30	36,250	Segment terminal	
27	2	4,65	18,500	Segment terminal	
7	6	13,95	5,500	EVOL99	0,550
14	3	6,98	8,000	Segment terminal	
15	3	6,98	3,000	Segment terminal	

Ainsi en terme de comportement opportuniste un établissement aura d'autant plus intérêt à jouer sur les indicateurs pour lesquels la valeur moyenne (position prédite dans le classement) est la plus faible en terme de rang dans le classement ; ainsi une université voulant jouer le classement aura un taux d'investissement élevé, devra augmenter ses étudiants tout en restant dans une taille modeste privilégier la part des MCF par rapport aux autres statuts et si possible avoir la change d'être situé dans un département de potentiel important.

On notera, point délicat d'interprétation, que la même variable peut jouer différemment à divers niveaux de segmentation de l'arbre. Le tableau 5 montre que ceci est particulièrement vrai pour l'investissement par étudiant. (INVETUD). De fait il suffit de suivre les impacts de haut en bas des seuils de coupure, dans le tableau, soit de plus au moins significatif. Ainsi comme nous l'avons déjà précisé une université sera au moins dans les 36 premières si elle investit autour de 1500 frs. De manière moins significative, il faut qu'au moins elle investisse 800 frs par étudiant pour ne pas être au-delà du rang 58 et, quoique moins significatif, au moins 2700 pour se situer dans les 18 premières. A l'inverse, d'autres interprétations sont plus ambiguës, ainsi en est-il de EVOL99, une dynamique forte sur moyen terme du nombre d'étudiants (+3%) faciliterait un classement dans la première moitié alors qu'une croissance réduite (0,5%) garantirait à un moindre niveau de vraisemblance un classement très favorable. Ceci revient à des questions de structure de covariance et s'assimile aux effets non linéaires des impacts marginaux dans la régression.

Conclusion

Les méthodes de la classification et de l'analyse des segmentations peuvent apporter des compléments intéressants dans la boîte à outils de l'évaluation. On peut en effet les qualifier de méthodes exploratoires dans la mesure où elles permettent de mieux apprécier les interactions de groupes et les sources d'inertie dans un fichier complexe. Ces méthodes possèdent toutefois des limites qui leurs sont propres, d'une part, et communes aussi avec toutes les méthodes d'analyses des données, en particulier sur les structures de covariation pour ces dernières. L'appel à des analyses combinatoires, des procédures itératives et d'échantillonnage, les rend toutefois assez largement dépendantes de la complexité même des données. L'empilement de méthodes proches, utilisées avec des différences sensibles suivant les logiciels fait aussi que ces méthodes peuvent paraître plus ésotériques en rapport à des méthodes beaucoup présentes dans le coutumier de la pratique statistique comme le modèle gaussien autour de la régression et de l'analyse en composantes principales.

Références

Celeux, G., E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy (1989). *Classification automatique des données*. Dunod.
Jobson J.D. (1992). Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 209-278.
Lebart L., A. Morineau & M. Piron (1997). Statistique exploratoire multidimensionnelle. 2ème édition. Dunod, Paris, pp. 251-277.
Tomassone R., M. Danzart, J.J. Daudin & J.P. Masson (1988). Discrimination et classement. Masson, Paris.

Annexe

Dénombrement de l'ensemble des partitions d'un ensemble à n éléments

On note $N_{k,n}$ le nombre de partitions contenant exactement k parties d'un ensemble E_n à n éléments, et N_n le nombre total de ces partitions. On construit l'ensemble des partitions d'un ensemble à n éléments à partir de celui obtenue pour n-1 éléments de la façon suivante.

Soit E n-1={1,2,…..n}l'ensemble à n-1 éléments et n le nom du n^{ieme} élément.

Une partition de E_n:

- soit admet {n}comme singleton,
- soit possède l'élément n dans une partie de plus d'un élément.

Le nombre de partitions contenant k parties du premier ensemble est exactement égal au nombre de partitions contenant k-1 parties de E_{n-1}, soit $N_{n-1,k-1}$. Pour les partitions n'admettant pas {n}comme singleton, on passe d'une partition contenant k parties de E_{n-1} à k différentes partitions (contenant toujours k parties) de E_n en ajoutant l'élément n à chacune des k parties de cette partition. Au total, on arrive à: $N_{n,k} = N_{n-1,k-1} + k\, N_{n-1,k}$
A l'aide d'un tableur, on obtient les résultats suivants :

K / N	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	1	1	1	1	1	1	1	1
2	1	3	7	15	31	63	127	255	511	1023	2047
3		1	6	25	90	301	966	3025	9330	28501	86526
4			1	10	65	350	1701	7770	34105	145750	611501
5				1	15	140	1050	6951	42525	246730	1379400
6					1	21	266	2646	22827	179487	1323652
7						1	28	462	5880	63987	627396
8							1	36	750	11880	159027
9								1	45	1155	22275
10									1	55	1705
11										1	66
12											1
$\sum_{j=1}^{12} Nj$	2	5	15	52	203	877	4140	21147	115975	678570	4213597

On peut remarquer que le facteur multiplicatif permettant de passer de N_{n-1} à N_n est proche d'une fonction linéaire en n. Le nombre N_n augmente donc plus qu'exponentiellement en fonction de n et vaudrait environ 510^{13} pour n=20. Cette brève étude permet de voir qu'il ne sera pas possible, si on a défini un critère de qualité d'une partition, de passer toutes les partitions en revue pour trouver la meilleure (un ensemble d'individus contient en général de 50 à 10 000 individus).
Cette question de combinatoire a souvent conduit à la conclusion que ces méthodes devaient se limiter à des tableaux de taille réduite.