# Structure géométrique des distances hiérarchiques

Casanova-del-Angel, F.
Section d'Études et de la Recherche de l'École Supérieure d'Ingénierie et d'Architecture,
Unité Professionnelle « Adolfo López Mateos » de l'Institut Polytechnique National, Mexico, Mexique.
fcasanova@ipn.mx et fcasanova49@prodigy.net.mx

**Résumé**
La structure théorique de la classification hiérarchique nécessaire à la construction de l'arbre ou du dendrogramme est la base pour montrer la relation théorique des distances hiérarchiques géométriques pour une séquence de hiérarchies partielles où, lorsque lors de la suivante élection de classes à ajouter deux hiérarchies partielles égales existent, alors la hiérarchie partielle à ajouter dépendra des distances géométriques présentant les hiérarchies partielles se rapportant à celles de la troisième classe. Le développement théorique est illustré par le biais d'applications avec des données sur l'effet de la corrosion atmosphérique sur l'acier structurel dont est constituée l'infrastructure civile dans la Ville de Mexico, ainsi que sur l'évaluation du rendement des enseignants de troisième cycle au Mexique.

**Mots clés :** dendrogramme, distances hiérarchiques, relations triangulaires : équilatérales, isocèles et scalènes.

**Abstract**
Based on the theoretical structure of hierarchical classification to build the tree or dendrogram, is shown the theoretical relationship of geometrical hierarchical distances for a sequence of partial hierarchies where two partial and equal hierarchies exist in the election of classes to be added, then the partial hierarchy to be added depends on geometric distances shown by partial hierarchies regarding the third class. Theoretical development is exemplified through applications with data from the effect of atmospheric corrosion of structural steel in civil infrastructure in Mexico City and the assessment of teaching performance for postgraduate studies in Mexico.

**Key words:** dendrogram, hierarchical distances, triangular relationships: equilateral, isosceles and scalene.

## Introduction

Ci-dessus sont présentées, comme point de départ de la description théorique de la structure géométrique des distances hiérarchiques, quelques commentaires sur les techniques ayant pour but la recherche de caractéristiques similaires au sein d'un ensemble de données afin d'identifier des groupes, des clusters ou des rassemblements d'un ensemble de valeurs caractérisés comme individuelles ou variables (voir [8]). Ces techniques exploratoires impliquent deux types d'algorithmes: les hiérarchiques et les non hiérarchiques ou de partition. Chacun de ceux-ci a différents niveaux de différenciation, c'est à dire, il est possible de distinguer deux grands types de méthodes de classification:

i. les méthodes non hiérarchiques qui produisent directement une partition en un nombre fixe de classes. Un dendrogramme non hiérarchique se présente sous la forme d'un ensemble de points, où certains éléments sont reliés par des arêtes qui donnent à l'ensemble des individus des propriétés particulières dues aux données. L'un des dendrogrammes les plus intéressants est celui de longueur minimale, lequel est équivalent à une méthode de construction d'une hiérarchie de classes, et
ii. les méthodes hiérarchiques qui produisent une succession de partitions en classes chaque fois similaires à l'image des célèbres classifications des espèces, genres, familles, ordres, etc.

Il est important de se souvenir qu'avant l'application de n'importe quel algorithme de classification, il est d'abord nécessaire d'effectuer une sélection appropriée des variables, l'élection d'une mesure indiquant une similarité entre les objets, c'est à dire, la distance et la méthode de classification (hiérarchique ou pas). La seconde étape est l'application de l'algorithme de classification et l'analyse des résultats à partir de leur description, où les nœuds ou les agrégats qui contiennent les aînés et les benjamins du nœud, son poids, la valeur de l'indice du niveau hiérarchique et le

dendrogramme, lequel est interprété à partir des contributions des variables à chaque nœud, dans la coupe choisie. La description de chaque partition ou coupe se fait partir des aides pour son interprétation, appelées contributions des variables, et contiennent des pourcentages de la variance totale entre classes, ainsi que les contributions par classe ou partition et le calcul des centres de gravité. À partir de la lecture et l'interprétation du dendrogramme, il peut être nécessaire de fragmenter les variables et redémarrer le processus, mais cette fois-ci avec la définition des classes par variable (voir [1 ; 10 ; 11 et 4]), dans le but d'obtenir une interprétation représentative, ou validation, de la structure de l'information analysée. Il me semble nécessaire de rappeler que, bien que cela soit bien connu, certains des algorithmes de classification ont été développés par certains arrangements tabulaires de données similaires ou miscibles à une correspondance, comme par exemple, la relation théorique existant entre l'analyse factorielle des correspondances et une classification hiérarchique établie à partir de la métrique $\chi^2$ (voir [1 ; 9]).

La description précédente fait partie de ce que l'école française connaît sous le nom de la méthode d'analyse de classification, et dont l'objectif est de classifier des unités de la matrice originale en groupes ou *clusters* les plus homogènes possibles à l'intérieur et le plus hétérogène entre eux. La méthode appartient à l'analyse des données. Dans l'école anglaise, le concept est plus limitatif, car parfois il fait référence exclusivement aux analyses univariables ou bivariables. L'école espagnole considère que l'analyse multivariable fait partie de l'analyse des données, c'est à dire, qu'il s'agit de méthodes statistiques. L'école mexicaine suit la nord-américaine.

Mais comment parvenir à l'interprétation représentative ou validation de la structure de données, si l'objectif principal des algorithmes de classification est d'obtenir des groupes d'éléments à partir d'une matrice de distance sans examiner toutes et chacune des combinaisons possibles d'agroupement, ainsi que toutes et chacune des méthodes de classification possibles applicables aux données analysées ? Les diversités d'agroupement des variables ou des classes sont nombreuses, parmi lesquelles : simples, avec ou sans chaînage, où sont détectés les *clusters* non clairement séparés, complètes, moyenne (pondérées ou non pondérées), centroïd et médiane.

Certains des travaux représentatifs développés dans le but de la validation de la structure des données, du point de vue de la sociologie et de la psychologie, sont ceux menés par *Blashfield et Morey (1980),* lesquels à partir de processus Monte Carlo générant des ensembles de données afin que ceux-ci ressemblent à des tests psychologiques ou névrotiques sur les troubles de la personnalité, ceux de *Bayne, Beauchamp, Begovich et Kane (1980)* qui suggèrent la supériorité de certaines techniques no hiérarchiques due à leur robustesse utilisant des méthodes de Monte Carlo pour estimer les pourcentages d'erreurs de classification de 13 méthodes de hiérarchisation avec six types de paramétrages de deux populations normales bivariées, et ceux de *Scheibler et Schneider (1985),* qui suggèrent une certaine comparabilité avec les techniques hiérarchiques au moment de comparer neuf algorithmes hiérarchiques et quatre non hiérarchiques sur l'habilité de déterminer 200 mélanges normaux multivariés. D'autres travaux de la même nature sont le développement d'un algorithme pour générer des ensembles de données contenant différents *clusters* non superposés (voir [12]).

Le présent développement essaye d'apporter une nouvelle perspective sur la validation de la structure (connue ou inconnue) des données analysées, à partir de la relation géométrique formant les hiérarchies partielles.

**Le dendrogramme hiérarchique**

Partons du fait qu'une classification sur un ensemble fini de variables aléatoires $X_1,...,X_n$ est une partition, c'est à dire, une partie d'un certain nombre de parts vides deux à deux et à l'intersection vide, généralement une hiérarchie de classes emboîtées. L'ensemble fini de variables aléatoires se divise en un nombre fini de classes, se divisant à leur tour en un autre nombre fini de classes ou des

sous-classes. Afin d'obtenir une représentation des relations hiérarchiques entre variables aléatoires, il est nécessaire de définir une structure métrique. Soit $\Re^p$ l'espace réel de $\alpha$-facteurs, obtenus à partir d'une analyse factorielle appliquée à un ensemble fini de variables aléatoires. Au sein de l'espace de probabilité fini ($\Omega$, $P(\Omega)$, P) est définie une distance qui met en relation les facteurs de la classe nommée $\Im^2$ :

$$\Im^2(x_j, x_{j'}) = \sum^p_{\alpha=1} (fj/f_{j'})[F_\alpha(x_j) - F_\alpha(x_{j'})]^2$$

distance factorielle de classes pondérée entre les variables aléatoires qui vérifie la distance de distribution entre lois de fréquence de classes et permet l'invariance entre distance factorielle, où $f_j$ et $f_{j'}$ sont les fréquences des classes $j$ et $j'$ (les fréquences interviennent seulement dans les algorithmes de type agglomératif), et $F_\alpha$ sont les valeurs factorielles des classes $x_j$ et $x_{j'}$. La construction de l'algorithme se base dans l'établissement d'une séquence de hiérarchies partielles, que nous nommerons $C(\alpha) = C_0, C_1,.., C_h ,..., C_{\alpha-1}$, où une hiérarchie partielle est l'union de deux classes. La distance $\Im^2(x_j, x_{j'})$ est calculée sur l'ensemble F de dimension $\alpha$ à partir de l'arrangement tabulaire $F_{JQ}$ tel que :

$C_0 = C_0(\alpha) = \text{Term}[C(\alpha)] = \{x_j\} \ \forall \ j \in F_\alpha(j)$, avec Term l'ensemble des classes terminales de la hiérarchie $C(\alpha)$,
$\text{Ver}[C_0] = C_0(\alpha) = \text{Term}[C(\alpha)] = \{x_j\} \ \forall \ j \in F_\alpha(j)$, avec Ver le sommet des classes terminales de la hiérarchie $C(\alpha)$, et
$\nu(\{x_j\}) = 0 \ \forall \ j \in F_\alpha(j)$ est l'indice du niveau de la classe, $f(\{x_j\})$ est la fréquence liée à $x_j$ &
$\Im^2(x_j, x_{j'}) = \delta(\{x_j\}, \{x_{j'}\}) \ \forall \ x_j, x_{j'} \in F_\alpha(j)$ est la distance entre classes.

Pour l'itération du rang $h$=1, le minimum de $\delta$ sur le sommet de la hiérarchie $C_0$; $\text{Ver}[C_0]$ est recherché. Soit $(\{x_j\}, \{x_{j'}\})$ une paire de classes d'un élément qui satisfait le minimum. Le premier mode ainsi obtenu prend la forme du numéro $\text{Card}(\alpha)+1$. Pour ce que pour $N=\text{Card}(\alpha)+1$ et h=1 nous avons une nouvelle classe que nous nommerons $c$ faite par les variables ou classes initiales $x_j$, $x_{j'}$, est $c_1 = \{x_j, x_{j'}\}$ et l'ensemble de classes placées immédiatement sous la classe $c_1$ de $C(\alpha)$ est :

$$\text{Successeur}\ (c_1, C(\alpha)) = \{x_j, x_{j'}\} \ \forall \in \text{Nœud}(C(\alpha))$$

raison pour laquelle la première hiérarchie partielle est : $C_1 = C_\nu(c_1) = C_1(\alpha) = C_0 \cup c_1$ ,ainsi comme le nouveau sommet est $\text{Ver}[C_1] = \text{Ver}[C_0] \cup c_1 - \{x_j\} - \{x_{j'}\}$, le nouvel indice de niveau de la classe est $\nu(c_1) = \inf \{\delta^0(\{x_j\}, \{x_{j'}\})\} \ \forall \ x_j \neq x_{j'}$ avec $x_j, x_{j'} \in \text{Ver}[C_0]$. La cardinalité de la classe $c_1$ est maintenant de 2 et la fréquence de la nouvelle classe est $f(c_1) = f(\{x_j\}) + f(\{x_{j'}\})$. Selon la nomenclature existante dans la théorie de la classification, $x_j$ sera nommé l'aîné du Nœud de numéro $\text{Card}(\alpha) + 1$ et $x_{j'}$ le benjamin, avec la même cardinalité (voir [4]).

Une fois que cette première itération qui construit une nouvelle partition de $\alpha$ a pris fin, il est nécessaire de recalculer les distances entre toutes les classes de la partition dénotée $\text{Ver}[C_1]$. Comme celle-ci se déduit ou s'obtient de $\text{Ver}[C_0]$ remplaçant deux classes par son union, le recalcul de la distance entre parts d'ensembles qui permettent de calculer la distance entre la nouvelle classe crée et les autres classes de $\text{Ver}[C_1]$ à l'exception des deux classes qui ont réalisé la fusion est : $\delta^0(c_1, r) \ \forall \ r \in \text{Ver}[C_1]$ avec $r \neq x_j$ & $r \neq x_{j'}$

Voyons maintenant la h-nième itération. À ce moment nous connaissons déjà les hiérarchies partielles $C_0, C_1, .., C_{h-1}$ . Soit $t_h$ et $t'_h$, deux classes de $\text{Ver}[C_{h-1}]$ qui font minimale la distance $\delta$ calculée sur le $\text{Ver}[C_{h-1}]$, raison pour laquelle la cardinalité est égale a $\text{Card}(\alpha)+h = N$, la $h$-nième classe est $c_h = t_h \cup t'_h$. Le h-nième successeur est : Successeur $(c_h, C(\alpha)) = \{t_h, t'_h\}$ où $t_h$ et $t'_h$ son

respectivement l'aîné et le benjamin de la classe $c_h$. La $h$-nième hiérarchie partielle est $C_h(\alpha) = C_h = C_\nu(c_h) = C_{h-1}(\alpha) \cup c_h = C_{h-1}(\alpha) \cup \{t_h \text{ U } t'_h\}$. Le sommet de la $h$-nième hiérarchie est $\text{Ver}[C_h(\alpha)] = \text{Ver}[C_{h-1}(\alpha)] \cup \{c_h\} - \{t_h\} - \{t'_h\}$. L'indice de la $h$-nième classe est $\nu(c_h) = \inf\{\delta^{h-1}(t, t')\} \ \forall \ t \neq t'$ avec $t, t' \in \text{Ver}[C_{h-1}]$. La cardinalité de la $h$-nième classe est $\text{Card}(c_h) = \text{Card}(t_h) + \text{Card}(t'_h)$ et la fréquence de la $h$-nième classe est $f(c_h) = f(t_h) + f(t'_h)$.

La distance $\delta^h(r, c_h)$ sur le $h$-nième sommet $\text{Ver}[C_h] \ \forall \ r \in \text{Ver}[C_h]$ avec $r \neq t_h$ et $r \neq t'_h$ est recalculé. Cette formule de récurrence est une formulation des paramètres : $\delta^{h-1}(r, h_h)$, $\delta^{h-1}(r, h'_h)$, $\delta^{h-1}(t_h, t_h)$, $f(t_h)$, $f(t'_h)$ $\nu(t_h)$, $\nu(t'_h)$ ainsi que des cardinalités de $t_h$ et $t'_h$. Pour finir, l'itération de rang $\text{Card}(\alpha)$-1. Dans ce cas il ne reste que deux classes à rajouter dont l'union est la conjonction de toutes les classes $\alpha$, c'est à dire, la hiérarchie $C_{\alpha-1}$. Ici, la valeur de la cardinalité du nombre d'éléments est $2*\text{Card}(\alpha)$-1, la classe est $c_h = t_{\text{Card}(\alpha)-1} \cup t'_{\text{Card}(\alpha)-1}$, la hiérarchie partielle est $C_h = C(\alpha) = C_{\text{Card}(\alpha)-1}$, le sommet de cette hiérarchie partielle est $\text{Ver}[C_h] = \text{Ver}[C_{\text{Card}(\alpha)-1}] = \{\alpha\}$, la cardinalité de la classe $c_h = \text{Card } c_{\text{Card}(\alpha)-1} = \text{Card}(\alpha)$, la fréquence de la classe est $f(c_h) = f(\alpha)$, et l'indice de niveau $\nu(c_h) = \nu(\alpha) = \delta^{h-1}(k_{\text{Card}\alpha-1}, k'_{\text{Card}\alpha-1})$.

Les développements théoriques des classifications hiérarchiques posent le problème du choix de classes lorsqu'au moins deux paires de sous-classes présentent le cas d'une égalité de la distance $\delta$ sur $\text{Ver}[C_h]$. La forme non mathématique de résoudre ceci est en choisissant arbitrairement la paire de sous-classes à rajouter, la première lue de forme implicite dans les algorithmes de l'ordinateur. La forme mathématique dépend d'un critère de distance minimale et elle est présentée ci-dessous :

*Théorème des distances hiérarchiques géométriques*. Étant donné une séquence de hiérarchies partielles de $C(\alpha)$, si lors de l'élection suivante de classes à rajouter il existe deux hiérarchies partielles $C_h(\alpha)$ et $C_k(\alpha)$ telles que $C_h(\alpha) = C_k(\alpha) \ \forall \ h \neq k$ et qu'elles présentent la même distance minimale : $\delta(\{x_h\}, \{x_{h'}\}) = \delta(\{x_k\}, \{x_{k'}\})$ en relation avec la classe $C_r(\alpha)$, la hiérarchie partielle à rajouter à cette dernière dépend des distances géométriques $\delta$ que présentent les classes $C_h(\alpha)$ et $C_k(\alpha)$ en relation avec $C_r$.

*Démonstration*. Soit $C(\alpha) = C_0, C_1, ...., C_h, ..., C_{\alpha-1}$ une séquence de hiérarchies partielles. Si deux hiérarchies $C_h(\alpha)$ et $C_k(\alpha)$ telles que $C_h(\alpha) = C_k(\alpha) \ \forall \ h \neq k$ existent et qu'elles remplissent l'égalité de distance $\delta^{h-1}(\{x_h\},\{x_{h'}\}) = \delta^{k-1}(\{x_k\},\{x_{k'}\})$, où $\text{Ver}[C_h] = \text{Ver}[C_k]$ de sorte que les indices de niveau des classes sont égaux, c'est à dire, $\nu(C_h) = \nu(C_k)$, alors :

$$\inf\{\delta^{h-1}(C_h)\} = \inf\{\delta^{k-1}(C_k)\} \quad \text{et} \quad \inf\{\delta^{h-1}(\{x_h\}, \{x_{h'}\})\} = \inf\{\delta^{k-1}(\{x_k\}, \{x_{k'}\})\}$$

avec $x_h$, $x_{h'}$ éléments aînés et $x_k$, $x_{k'}$ éléments benjamins des classes $C_h$ et $C_k$ respectives. Considérons une troisième hiérarchie partielle $C_r(\alpha) \subset C(\alpha)$ et $r < h$, $r < k$ à laquelle sera rajoutée l'une des deux hiérarchies partielles d'ordre $h$ ou $k$, utilisant les propriétés de distance ultra métrique. La forma géométrique que peuvent construire ces trois hiérarchies partielles sont : équilatérales, isocèles et scalènes (dans ( [1]) § 4.1 il n'est fait mention que des seules relations triangulaires équilatérales et isocèles et non des relations scalènes). Selon la géométrie formée par les classes hiérarchiques $C_r(\alpha)$, $C_h(\alpha)$ et $C_k(\alpha)$ :

$$\delta^{r-1}(C_r, C_h) \leq \sup\{\delta^{r-1}(C_r, C_k), \delta^{r-1}(C_k, C_h)\} \quad \text{et} \quad \delta^{r-1}(C_r, C_k) \leq \sup\{\delta^{r-1}(C_r, C_h), \delta^{r-1}(C_h, C_k)\}$$

remplissant les inégalités :

$$\delta^{r-1}(C_r, C_h) \leq \delta^{r-1}(C_r, C_k) + \delta^{r-1}(C_k, C_h) \quad \text{et} \quad \delta^{r-1}(C_r, C_k) \leq \delta^{r-1}(C_r, C_h) + \delta^{r-1}(C_h, C_k)$$

ce qui signifie qu'il n'est pas important que le triangle formé par les classes hiérarchiques $C_r(\alpha)$, $C_h(\alpha)$ et $C_k(\alpha)$ soit équilatéral ou isocèle, choisissant arbitrairement la hiérarchie partielle à rajouter. Cependant, si la relation triangulaire est de type scalène, l'une des deux distances à la classe $r$ est plus petite, c'est à dire, si $\delta^{r-1}(C_r, C_k) < \delta^{r-1}(C_r, C_h)$. Ce qui est facile de prouver, de par les propriétés de l'inégalité du triangle.

$$\delta^{r-1}(C_r, C_k) + \delta^{r-1}(C_k, C_h) < \delta^{r-1}(C_r, C_h) + \delta^{r-1}(C_k, C_h)$$

**CQFD**

La séquence de classifications hiérarchiques appliquées dans ce travail de recherche à partir de l'algorithme décrit est montrée dans la Figure 1 : quatre méthodes de classification hiérarchique, trois distances et quatre critères d'agrégation. En particulier, 66 dendrogrammes ont été construit pour la première application et 18 pour la deuxième.

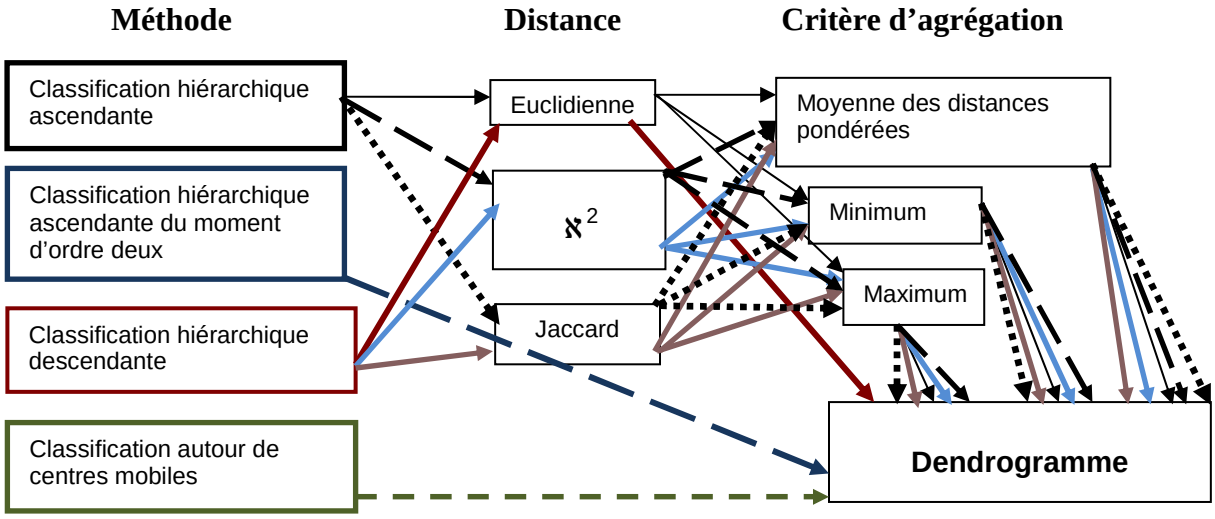


Figure 1. Structure séquentielle des types de classification hiérarchique appliqués.

Les critères employés ou les stratégies d'agrégation partent du fait que lorsque les distances ne sont pas euclidiennes, ce qui se passe lorsque l'inégalité triangulaire $d(a, b) \leq d(a, c) + d(b, c)$ n'est pas vérifiée pour certains points. On parle alors plus d'inégalités que de distance. La notion d'inertie n'a plus de sens et il n'y a plus de critère objectif pour calculer la distance entre deux classes. Pour cette raison, il est possible d'imaginer de nombreuses solutions plus ou moins arbitraires. Des formules existantes de distance entre deux parts, les trois les plus couramment utilisées sont :

i. la distance du saut minimal ou du saut infime (inf) telle que $d(A, B) = \inf [d(e_i, e_j) \ \forall \ e_i \in A \text{ et } e_j \in B]$ qui tend à favoriser le regroupement de deux classes, qu'elles possèdent des points très proches, avec le risque de rencontrer dans une même classe des points très éloignés. Mais même ainsi elle est très utilisée de par ses propriétés mathématiques,

ii. la distance du diamètre ou du supérieur (sup) telle que $d(A, B) = \sup [d(e_i, e_j) \ \forall \ e_i \in A \text{ et } e_j \in B]$ qui remédie au manque de méthode du saut minimum, parce qu'elle exige que les points les plus éloignés soient proches ;et

iii. la distance moyenne telle que $d(A, B) = (1/P_A P_B) \sum_i \sum_j d(e_i, e_j)$ propose un compromis entre les deux impliqués.

**Application 1**

L'objectif de cette application a été d'analyser les effets de la corrosion atmosphérique sur l'acier structurel, un constituant essentiel de l'infrastructure civile de la Ville de Mexico, où se trouvent quatre parcs industriels enclavés dans la zone nord de la zone métropolitaine : Vallejo, Azcapotzalco, Xalostoc et Tultitlán. L'information analysée et utilisée est l'union de trois arrangements tabulaires de données. Un arrangement tabulaire de l'information météorologique *I*x*J*1 des valeurs moyennes hebdomadaires obtenues de la Station Météorologique Expérimentale de l'Institut Polytechnique National et qui se compose de cinq variables : vitesse du vent, température de l'air, humidité relative, radiation solaire et précipitation pluviale. Un arrangement tabulaire d'information *I*x*J*2, avec des données moyennes hebdomadaires de deux polluants : le bioxyde de soufre, mesuré dans 4 stations du Réseau Automatique de Surveillance Atmosphérique de la Ville de Mexico (Vallejo, La Villa, Azcapotzalco et Xalostoc), et les chlorures ou la concentration d'anions pour des valeurs maximales par semaine à la station Xalostoc, ainsi que des valeurs de *pH* ou de potentiel d'hydrogène de l'eau de pluie. Le troisième arrangement tabulaire d'information *I*x*J*3 contient les valeurs de la vitesse de corrosion de deux dépositions d'éléments structurels. La première donnée correspond à la semaine du 2 au 9 juin 2002 et la dernière à la semaine du 13 au 18 novembre 2005, c'est à dire, trois ans et cinq mois d'information moyenne hebdomadaire pour un total de 181 observations.

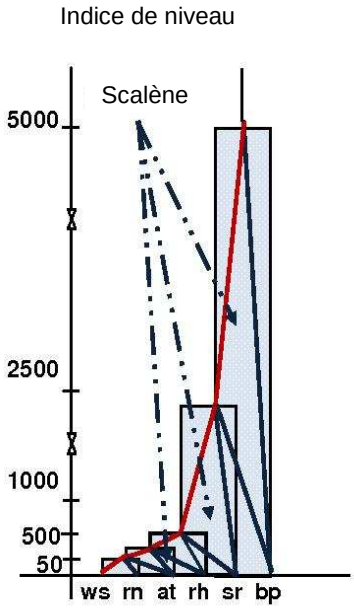


Figure 2.a. Dendrogramme à partir de l'arrangement *I*x*J*1.

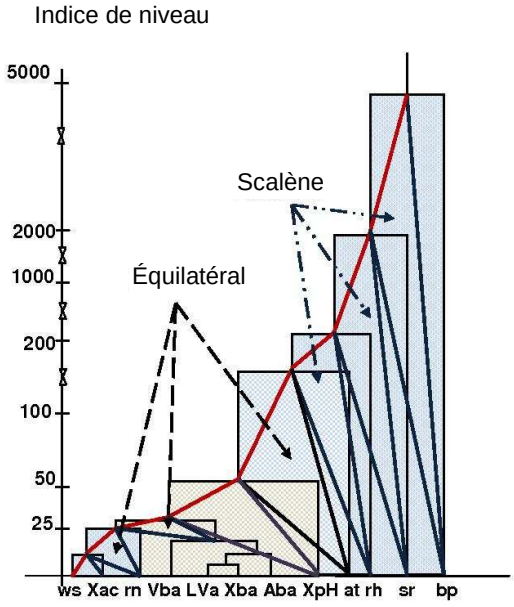


Figure 2.b. Dendrogramme à partir de l'arrangement (*I*x*J*1) ∪ ( *I*x*J*2).

**Analyse factorielle des données brutes et leur classification hiérarchique**

La méthode factorielle choisie pour l'étude et la description des données en étude est l'Analyse Factorielle des Correspondances. Le premier plan factoriel 1-2 n'a pas de forme définie, car la totalité des observations se trouvent centrées sur l'origine. Pour cette raison, une coupe en classes des variables en étude a été réalisée et un arrangement tabulaire de Burt a été construit (voir [14]).

Une classification hiérarchique ascendante avec distance euclidienne et critère d'agrégation minimal a été construite à partir de l'arrangement *IxJ1*. Son dendrogramme hiérarchique est présenté dans la Figure 2a. La structure géométrique formée ici par les classes hiérarchiques $C_r(\alpha)$, $C_h(\alpha)$ et $C_k(\alpha)$ est scalène. À l'arrangement antérieur ont été rajouté les variables de contamination atmosphérique, ce qui a crée l'arrangement (*I*x*J*1) ∪ (*I*x*J*2) et construit une classification hiérarchique ascendante avec distance euclidienne sous le critère d'agrégation minimale. Son

dendrogramme hiérarchique est présenté dans la Figure 2b, et sa structure géométrique est de deux types : équilatérale et scalène. Enfin, en joignant l'arrangement tabulaire *I*x*J*3 qui contient les valeurs de vitesse de corrosion des deux dépositions d'éléments structurels et, nous construisons pour l'arrangement complet [(*I*x*J*1)∪(*I*x*J*2)]∪*I*x*J*3 le même type de classification hiérarchique ascendante avec distance euclidienne et critère d'agrégation minimal en Figura 3. Sa structure géométrique est conservée.

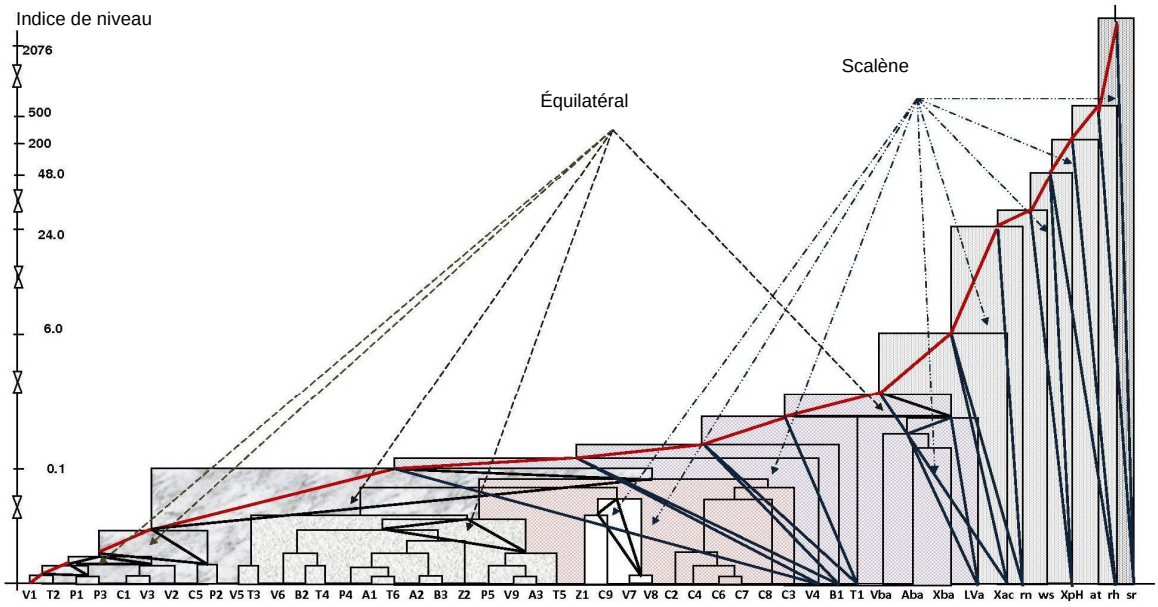


Figure 3. Dendrogramme de l'arrangement tabulaire complet.

On peut noter que la structure graphique des dendrogrammes ne change pas, ils sont tous du type croissant avec une structure équilatérale et scalène, entre des séquences de hiérarchies partielles (Figures 2.a, 2.b et 3). L'information fondamentale que contient l'arrangement final de l'information est cachée et les données de météorologie dominent, bien que dans la Figure 3 la relation entre la corrosion et la pollution montre déjà d'autres relations géométriques entre hiérarchies partielles.

La Figura 4 contient la hiérarchie de la relation existant entre les classes construites de la météorologie avec les concentrations maximales de polluants qui agirent pendant la période d'étude et les vitesses de corrosion des profils structuraux étudiés. La lecture et l'interprétation se sont faites sur la base de la valeur de l'indice de niveau hiérarchique, montré sur la gauche du dendrogramme, entendant par cela l'ordre successif des valeurs que donne le produit du poids de la classe analysée et son diamètre (la distance $d(i, i')$ est le diamètre de la partie la plus petite d'une hiérarchie contenant à la fois $i$ et $i'$) (voir [6]). Le dendrogramme hiérarchique construit est formé par cinq branches, dont l'interprétation est complètement congruente avec ce qui est connu sur le sujet. Ici, la structure géométrique des classes est variée. La géométrie équilatérale agroupe les phases dénommées stabilisation de l'effet de corrosion et transition de l'effet de corrosion. La géométrie scalène agroupe la phase d'excès d'humidité et la phase de météorologie. Et, enfin, la géométrie isocèle se manifeste avec la phase initiale de la corrosion.

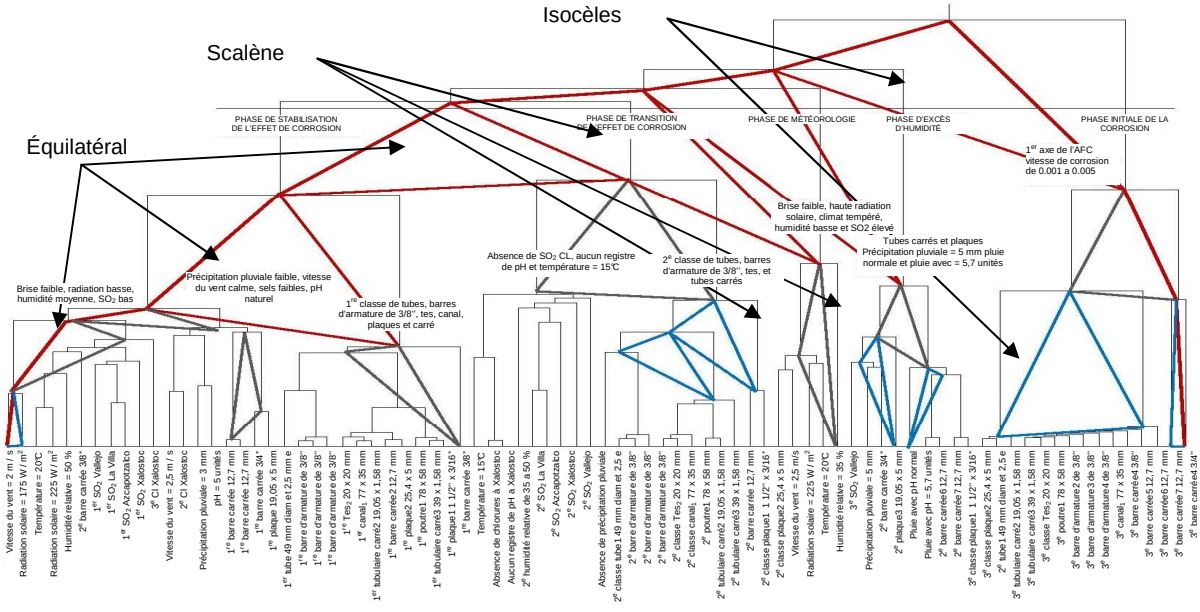


Figure 4. Dendrogramme de la corrosion des profils structuraux à partir d'un arrangement tabulaire de Burt.

**Application 2**

À partir du développement d'une méthodologie d'évaluation du rendement des enseignants pour les études de troisième cycle au Mexique, et avec la conception d'un questionnaire ou instrument d'évaluation appliqué aux élèves de niveau Maîtrise en Sciences et divisé en quatre sections (voir [7]). Les élèves ont répondu à 543 questionnaires, un par matière étudiée pendant les années 2003 et 2004. Une étude statistique a été faite matière par matière et, une analyse aussi bien globale que par matière a été faite du point de vue éducatif et psychologique. L'objectif principal de cette recherche a été de développer une méthodologie propre permettant d'évaluer le rendement académique du personnel enseignant de troisième cycle au Mexique (voir [5]). La population sur laquelle a été appliqué cet instrument sont les élèves d'un cursus de troisième cycle appartenant à l'Institut Polytechnique National du Mexique. Dans celui-ci sont enseignées 22 matières, dont 3 sont propédeutiques, 8 obligatoires, 3 des séminaires et 8 optionnelles parmi lesquelles l'élève choisit 2. Parmi les élèves interrogés, 19 étaient en cours d'écriture du mémoire et le reste suivait les cours obligatoires ou optionnels. Les 18 élèves qui suivaient les cours propédeutiques ne furent pas interrogés. Chaque élève a répondu à un questionnaire par matière, accumulant un total de 543 questionnaires répondus pendant 2003.

L'analyse statistique de l'information a été faite matière par matière et de forme groupée. Une classification hiérarchique a été faite à partir de la distance euclidienne à l'arrangement tabulaire des données brutes, dans le but d'obtenir une hiérarchisation ascendante des variables, sous le critère d'agrégation de distance minimale pour les questions faites aux élèves et déterminer ses agroupements primaires. L'arbre a un indice de niveau de 153 unités hiérarchiques (Figure 5). Suivant la valeur de l'indice de niveau optimal par lequel le dendrogramme a été *coupé*, quatre patrons de comportement primaire ont été identifiés et qui paraissent évidents. Le premier patron contient les moyens audiovisuels, le deuxième patron est l'offre d'activités hors cursus et sont regroupés dans l'engagement de l'enseignant face aux élèves. Le quatrième patron est le travail en tant qu'enseignant du professeur.

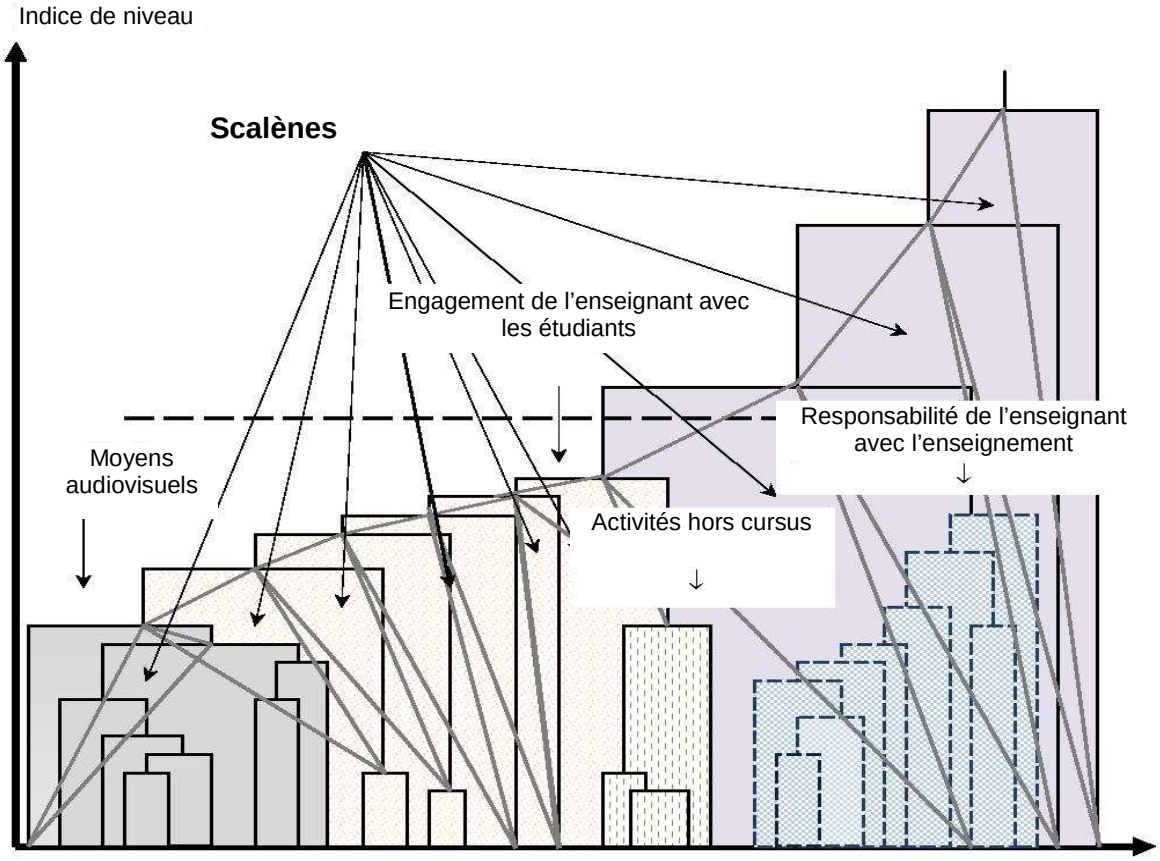


Figure 5. Hiérarchie du rendement des enseignants de troisième cycle au Mexique.

La structure géométrique est d'un seul type, scalène, ce qui paraît logique et concorde avec la théorie exposée sur les distances géométriques.

**Conclusions**

Du point de vue théorique, il est possible de signaler que le développement théorique de la classification hiérarchique pose le problème du choix des classes lorsqu'au moins deux paires de sous-classes présentent le cas d'une égalité minimale dans la distance δ sur Ver[$C_h$], situation qui se présente lorsque l'information est trop fragmentée. La façon traditionnelle de le résoudre a été d'élire arbitrairement la paire de sous-classes à ajouter, la première lue. La façon correcte, dont la structure mathématique dépend d'un critère de distances minimales, a été démontrée avec le théorème nommé *des distances hiérarchiques géométriques*.

Du point de vue dendrogrammatique, la relation géométrique formée par les classes hiérarchiques identifient clairement les patrons de comportement primaire qui se trouvent de forme diffuse dans l'information analysée. C'est de plus une aide à l'interprétation de l'arbre hiérarchique.

**Références**

[1] **Benzécri J. P, Benzécri F, Bellier L, Bénier B, Blaise S, Bourgarit Ch, Briane J. P, Cazes P, Dreux Ph, Escofier B, Fénelon J P, Forcade J, Giudicelli X, Grosmangin A, Guibert B, Hassan A R, Hathout A, Jambu M, Kamal I H, Kerbaol M, Lacoste A, Lacourt P, Laganier J, Lebesux M O, Lechat J, Le Chappelier M, Lenoir P, Leroy P, Mahé J, Mann C, Marano Ph, Masson M, Moitry J, Müller J, Nakhlé F, Piétri M, Richard J F, Rousseau R, Roux G & M, Salem A, Sandor G, Stérpan S, Tabet N, Thauront G, Volle**

**M, Yagolnitzer E y Zloptowicz M**. **1976**. *L´Analyse des Données. 1. La Taxinomie*. Ed. Dunod. París. ISBN: 2-04-003316-5.

[2] **Blashfield Roger K y Morey Leslie C. 1980**. A Comparison of Four Clustering Methods Using MMPI Monte Carlo Data. Applied Psychological Measurement. Vol. 4, No. 1, pp. 57-64. DOI: 10.1177/014662168000400107.

[3] **Bayne Charles K, Beauchamp John J, Begovich Connie L and Kane Victor E. 1980**. Monte Carlo comparisons of selected clustering procedures. Pattern Recognition. Vol. 12, issue 2, pp. 51-62. DOI: 10.1016/0031-3203(80)90002-3

[4] **Casanova del Angel F. 2001.** *Análisis multidimensional de datos.* Ed. Logiciels, México. ISBN: 970-92662-1-7

[5] **Casanova del Angel, F. 2006**. Postgraduate Teaching Performance Evaluation System. International Conference on Teaching Statistics. July 2-7, 2006. Salvador Bahía, Brazil.

[6] **Casanova del Angel, F y Toquiantzi Butrón, R. 2008**. Corrosion phases of structural shapes exposed to the atmosphere. Corrosion Science. Vol. 50, issue 8, August (2008) pp. 2288-2295. ISSN: 0010-938X doi: 10.1016/ j.corsi2008.05.015.

[7] **Cattel, J. M., & Farrand, L. 1896**. Physical and Mental Measurments of the Students of Colombia University. *The Psychological Review*, 3(6), pp. 618-648.

[8] **Estarelles R, de la Fuente E. I y Olmedo P. 1992**. Aplicación y valoración de diferentes algoritmos no-jerárquicos en el análisis *clúster* y su representación gráfica. Anuario de Psicología, núm. 55, pp. 63-90. Universitat de Barcelona. España.

[9] **Jambu M. 1978**. *Classification automatique pour l´analyse des deonnés. 1.Méthodes et algorithmes*. Ed. Dunod. ISBN: 2-04-010251-9.

[10] **Jambu M y Lebeaux M O. 1978**. *Classification automatique pour l´analyse des données. 2. Logiciels*. Ed. Dunod. ISBN: 2-04-010451-8.

[11] **Lerman I. C. 1981**. *Classification et analyse ordinale des données*. Ed. Dunod. Paris, ISBN: 2-04-015405-1.

[12] **Milligan Glenn W. 1985**. An algorithm for generating test clusters. Psychometrika. Vol 50, number 1. Pp. 123-127. March. Springer New York. ISSN: 0033-3123 (Print) 1860-0980. Doi: 10.1007/BF02294153.

[13] **Scheibler Dieter y Schneider Wolfgang. 1985**. Monte Carlo Test of the Accuracy of Cluster Analysis Algorithms: A Comparison of Hierarchical and Nonhierarchical Methods. Multivariate Behavioral Research. Vol. 20, issue 3 July, pp. 283-304. Doi: 10.1207/s15327906mbr2003_4.

[14] **Toquiantzi Butrón, R y Casanova del Angel, F. 2007**. Modelo jerárquico multidimensional de las fases de la corrosión de perfiles estructurales expuestos a la atmósfera. El Portulano de la Ciencia. Año VII, Vol. II, Núm. 17, pp. 647-665. Editorial Logiciels. México.

**Références consultée et non referéncées**

[1] **Bragard A, Bonnarens H. 1980**. Atmospheric Conditions and Durability of Weathering Steels. C.R.M. Metall. Rep. no. 57. pp. 15-24.

[2] **Casanova del Angel F. 1997**. Estructura gráfica de las precipitaciones ácidas en el área metropolitana de la Ciudad de México entre 1987 y 1994. pp. 1-6. IPN DEPI 923265. Ed. Instituto Politécnico Nacional. México.

[3] **Casanova del Angel, F. 2003**. Modelo matemático de previsión de fenómenos meteorológicos a pequeña escala con base en predictores cuantitativos. Parte I. El Portulano de la Ciencia. Año III, Vol I, Núm. 9. Editorial Logiciels. México.

[4] **Cole I.S, Neufeld A.K, Kao P, Ganther W.D, Chotimongkol L, Bhamornsut C, Hue y Bernardo N.V. 1999**. Performance Based Tests for Metals in Tropical Countries.

[5] **Da Silva Toribio, J. 1998**. La predicción de la vida útil y la vida residual de las construcciones. Universidad Federal de Uberlandia. Brasil.

[6] **Feliu S, Morcillo M, Feliu S, Jr**. **1993**. Predicción de la corrosión atmosférica mediante parámetros contaminantes y meteorológicos. Centro Nacional de Investigaciones Metalúrgicas. 28040. Madrid España.
[7] **International Organization for Standardization ISO/ DIS 8407. 1983**. Corrosion of Metals and Alloys – Removal of Corrosion Products from Corrosion Test Specimens. Estados Unidos de Norte América.
[8] **Leygraf C, E. Graedel T. 2000**. Atmospheric Corrosion. ISBN: 0-471-37219-6. pp 101-102. Ed. Wiley Interscience. New Jersey.
[9] **Lipfert L. F, Benarie M. M. 1986**. A General Corrosion Function in Terms of Atmospheric Pollution Concentrations and Rain pH. Atmospheric Environment. Vol. 20. no. 10. pp. 1947-1958.
[10] **Norma Mexicana NMX-B-252-ONNCE. 1988**. Requisitos generales para planchas, perfiles, tablaestacas y barras, de acero laminado, para uso estructural.
[11] **Norma Mexicana NMX-C-407-ONNCE. 2001**. Industria de la construcción – varilla corrugada de acero proveniente de lingote y palanquilla para refuerzo de concreto-especificaciones y métodos de prueba.
[12] **Timoshenko P. S, Gere M. J. 1974**. *Mecánica de materiales*. Ed. Hispano- Americana. México.
[13] **Toquiantzi Butrón, R. 2007**. Corrosión de perfiles estructurales expuestos a la atmósfera. Tesis de Posgrado. Instituto Politécnico Nacional. México.
[14] **Tutti. K. 1982**. Corrosion of Steel in Concrete. Swedish Cement and Concrete Institute. Stockholm.