Chapitre 13 : Antécédents des phénomènes ENSO et IOD sur l'évènement 2016-2017

Cette approche constitue le deuxième objectif principal de l'étude, la première étant la description de l'évènement extrême de l'année 2016-2017. Nous utiliserons plusieurs variables caractérisant lesdites phénomènes pour déceler s'il existe un éventuel lien ou pas sur l'évènement extrême de l'année 2016-2017 avec ces variabilités climatiques. Deux variables caractéristiques sont proposées pour l'exécution de cette analyse : les variables quantitatives (les indices) et les variables qualitatives (la nature de l'intensité des phénomènes).

I. Comparaison avec les indices normalisés

Brièvement décrite dans le Chapitre concernant les données et matériels, ces indices représentent la totalité des caractéristiques des deux phénomènes. Rappelons que nous avons en tout 7 indices ENSO et 4 indices IOD, chacun à leur propre signification pour décrire respectivement les deux phénomènes, de cette propriété, on observer le(s)quel(s) de ces indices influent sur les précipitations de notre zone d'étude. Effectivement, la normalisation de ces indices est requise pour l'homogénéité et la concordance des analyses (cette procédure se réfère au centrage et réduction expliqué au chapitre 12-II). Ceci est dit, on va rendre toutes les variables à comparer dans une même dimension (sans unité). On les utilisera de deux façons :

- Comparaison par la méthode graphique
- Analyse de liaisons avec l'analyse en composante principale.

II. Comparaisons avec les Intensités des phénomènes

Dans ce paragraphe, on va caractériser qualitativement les deux phénomènes. Cela ne veut pas dire qu'observe juste le phénomène physiquement et en interpréter ce qui rendra l'étude complètement banale, mais l'utilisation d'un autre indice sera à la page. L'indice ONI, déjà défini dans le chapitre 1-section 1, est utilisé par la NOAA comme facteur standard en vue d'identifier l'intensité des phases chaudes et froides du phénomène ENSO. Quant à l'IOD, l'indice DMI est la plus commode et le seul de nos jours (Yamagata et al, 1999). On utilisera cette méthode comme complémentaire à l'ACP dans l'interprétation en caractérisant les périodes à analyser par l'observation des cartes composite OLR et anomalies SST.

Chapitre 14 : Analyse statistique exploratoire multidimensionnelle

Cette méthode est une branche de l'analyse de données qui utilise l'analyse factorielle et la classification. Etant donnée, l'importance du nombre de données dans notre étude, l'utilisation d'une méthode multivariée qui peut traiter de grands ensembles de données est indispensable. En effet, l'analyse statistique exploratoire multidimensionnelle permet de faire une étude descriptive de la variabilité des données sur plusieurs variables à la fois dans un tableau unique. Plusieurs méthodes sont proposées en analyse factorielle, cela dépend de la nature des données a exploitée. Rappelons que notre but est de trouver une relation quelconque entre les anomalies de précipitations et les variabilités climatiques. Pour cela, on choisira l'Analyse en Composantes Principales associée à une classification Ascendante Hiérarchique, notre choix s'appuie sur les critères suivants :

❖ **La typologie des données** : il s'avère que les données utilisées sont de nature quantitative continue (indices).

❖ **La représentation spatiale** : l'analyse se fera sur plusieurs variables, d'où le terme multidimensionnel (en K dimensions). En tout, nous avons 12 indices à comparer sur 65 ans et 6 mois. Dès que K > 3, K = 12 pour notre cas, l'étude directe du nuage N_I est impossible du fait de la limitation à 3 dimension de notre sens visuel. D'où l'intérêt des méthodes factorielles en général, et dans ce cas particulier de l'ACP, qui fournissent des images planes approchant le mieux possible un nuage de points situé dans un espace de grande dimension.

❖ **Les résultats attendus** : en se référant au deuxième objectif de la présente étude et l'importance de la quantité et de la nature des données traités, cette méthode est la plus commode en raison de sa souplesse et ses caractéristiques (présentées plus bas dans l'objectif) sur l'étude des individus (périodes) et des variables quantitatives (indices) en particulier.

Ici, le résultat principal que nous attendrons et la mise en évidence de la variabilité des précipitations face aux impacts des variabilités climatiques proposées sur la zone d'étude. Et rappelons aussi que nous traiterons la méthode à l'aide de SPSS (Annexe XVIII).

Section 1 Analyse en Composante Principale (ACP) normée

[3] [4] [20] [23] [64] En météorologie et climatologie, on utilise l'appellation courante « analyse des fonctions orthogonales empiriques » ou « EOF ». Leur principe se base sur l'analyse de la variabilité d'une seule catégorie de variables, similaire à ce que les praticiens font habituellement avec l'ACP simple. Dans ces dispositions, non seulement tous les individus ont le même poids dans l'analyse, mais aussi, toutes les variables sont traitées de façon symétrique (on leur fait jouer le même rôle) et les nouveaux axes sont issus de la matrice de covariance empirique des variables. En outre, notre situation se diffère de ceux-ci, tout d'abord, notre analyse se base sur plusieurs champs de variables de différentes unités avec une importance sur l'écart des variabilités entre-elles. En effet, cela pose des problèmes au

niveau de la précision des résultats, d'où l'emploi de l'ACP normée en procédant au centrage et réduction des variables déjà présenter précédemment. Deux améliorations seront obtenues à partir de cette approche :

- Elimination de l'hétérogénéité des variables.
- Augmentation de la contribution des variables en vue de la construction du minimum d'axes, c'est-à-dire, compresser le plus les informations.

Dans la suite du travail, on expliquera les procédures générales en ACP simple, en introduisant les modifications possibles avec une ACP normée, tout en se référant sur les procédures implémentées dans SPSS (présenté dans l'annexe). Notons aussi que toutes les valeurs des variables k pour l'individu i sont « Standardisées ».

I. Schéma des Données

L'Analyse en Composantes Principales (ACP) s'applique à des tableaux croisant des individus et des variables quantitatives, appelés de façon concise tableaux $Individus \times Variables\ quantitatives$ Selon un usage bien établi, les lignes du tableau représentent les individus et les colonnes représentent les variables. A l'intersection de la ligne *i* et de la colonne *k* se trouve la valeur de la variable *k* pour l'individu *i*. La figure 25 illustre ces notions et complète les notations

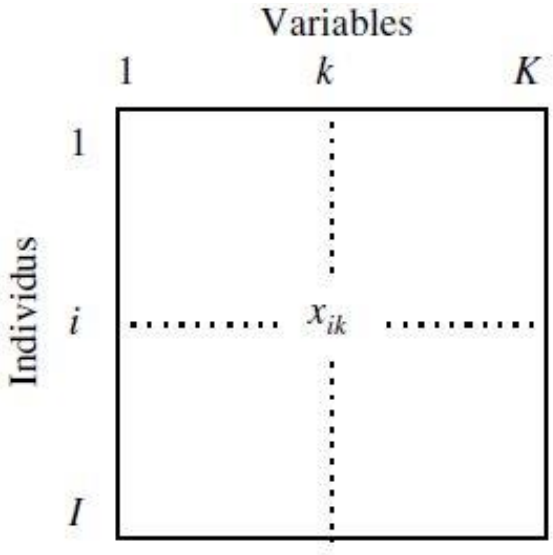

Figure 25 : Tableau des données en ACP.

Avec :

X_{ik} : valeur de la variable *k* pour l'individu i.

I : nombre d'individus et ensemble des individus.

K : nombre et ensemble des variables quantitatives.

Dans notre analyse, nos individus seront les différents mois et saison de la période de référence. Nos variables sont les indices SPI, ENSO et IOD et les x_{ik} sont les valeurs des indices sur chaque mois de la période.

II. Objectifs

Les buts que l'on se pose sur les individus et celles que l'on se pose sur les variables ne sont pas les mêmes. De ce fait, deux cas se présentent : l'étude des individus et l'étude des variables.

1. Etude des individus

L'objectif est d'évaluer la ressemblance entre individus à partir de leur profil de réponse. Pour cela, on cherche à mettre en évidence des groupes homogènes de périodes dans le cadre d'une **typologie des périodes**. Selon un autre point de vue, on cherche les **principales dimensions de variabilité** des périodes. C'est-à-dire, l'analyse des périodes présentant les mêmes variabilités et ceux qui s'opposent. L'approche générale est la recherche d'une projection pour la minimisation de la distance euclidienne entre les individus. En ACP, la distance d(*i* ,*l*) entre deux individus *i* et *l* est définie par :

$$d^2(i,l) = \sum_{k \in K} (x_{ik} - x_l)^2 \qquad (33)$$

Le critère est la suivante : plus cette distance est moindre, plus on observe les périodes qui se ressemblent, qui ont les mêmes caractéristiques climatiques au niveau des précipitations.

2. Etude des variables

❖ **Liaisons entre variables** : L'objectif est de déterminer les liaisons entre les variables en étudiant les liaisons linéaires et en établissant la matrices des corrélations. Cela consiste à résumer l'ensemble des variables par un petit nombre de **variables synthétiques** appelées ici **composantes principales**. Ce point de vue est très lié au précédent : une composante principale peut être considérée comme le représentant (la synthèse) d'un groupe de variables liées entre elles.

❖ **Liaisons linéaires** : En ACP, la liaison entre deux variables est mesurée par le coefficient de corrélation linéaire (dans de rares situations, on utilise la covariance), noté usuellement r présenté au chapitre concernant les statistiques descriptives. Pour l'ACP, la liaison entre les variables k et h est définie par :

$$r(k,h) = \frac{1}{I} \sum_{i \in I} \left(\frac{x_{ik} - \overline{x}_k}{\sigma_k} \right) \left(\frac{x_{ih} - \overline{x}_h}{\sigma_h} \right) \qquad (34)$$

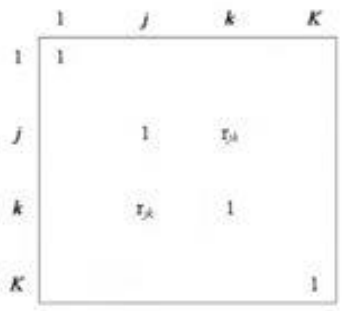

Figure 26 : Matrice des corrélations

Avec en ligne et colonne les variables et r_{jk}, coefficient de corrélation entres les variables j et k, Figure 26. Le critère général est de maximiser r entre les variables, en trouvant la projection idéale ne déformant pas les informations. Dans la présente étude, ceci se traduit par l'observation des indices climatiques qui corrèlent le mieux avec les SPI, chercher une certaine caractéristique d'influence (combinée ou individuelle).

❖ **Indicateur synthétique** : Au cas où le nombre de variables est très considérable, indétectable visuellement ; on a recouru à l'indicateur synthétique qui a pour rôle de grouper les variables présentant les mêmes caractéristiques de liaisons. Nos conditions ne requièrent guère cette procédure puisque notre matrice aura la dimension 12 x 12, détectable visuellement.

III. Respect des postulats

Généralement, toute analyse requiert des conditions initiales pour l'efficacité de la méthode employée et la validité des résultats. En fait, le but est de tester l'intérêt de l'ACP en vérifiant s'il est possible de compresser efficacement l'information disponible ; ainsi que de mesurer le degré de redondance des données. Analyser les résultats d'une ACP, c'est répondre à trois conditions :

1. La matrice des corrélations

Pour vérifier cette condition, dans un premier temps, il convient d'observer la matrice des corrélations (« Correlation Matrix »). Si plusieurs variables sont corrélées (> 0.5), la factorisation est possible. Si non, la factorisation n'a pas de sens et n'est donc pas conseillée.

2. L'indice KMO (Kaiser-Meyer-Olkin)

Ou encore "Measure of sampling adequacy ou MSA". Dans un deuxième temps, l'idée est de confronter la matrice des corrélations brutes avec la matrice des corrélations partielles. Si MSA ≈ 1, l'ACP peut agir efficacement parce que corrélations partielles sont quasi-nulles (en valeur absolue) ; si KMO << 1, problème car pas de redondance entre les variables (**Erreur ! Source du renvoi introuvable.**). C'est plutôt un indice de compressibilité de l'information qu'un indicateur de l'intérêt d'une ACP pour un fichier de donnée. L'indice KMO est défini par la relation :

$$Indice\ KMO = \frac{\sum_j \sum_{m \neq j} \left(r_{jm}\right)^2}{\sum_j \sum_{m \neq j} \left(r_{jm}\right)^2 + \sum_j \sum_{m \neq j} \left(r^*_{jm}\right)^2} \qquad (35)$$

Où :

j et m : respectivement les variables disposées en lignes et colonnes.

r_{jm} : Corrélations brutes entre les variables prises 2 à 2.

r^*_{jm} : Corrélations partielles c.-à-d. après avoir retranché l'influence des (p-2) autres.

Tableau 14 : Interprétation de l'indice KMO selon Kaiser, Meyer & Olkin (1974)

Valeurs d l'indice KMO	Signification
0,90 à 1,00	Merveilleux
0,80 à 0,89	Excellent
0,70 à 0,79	Moyen
0,60 à 0,69	Médiocre
0,50 à 0,59	Misérable
0,00 à 0,49	Ne pas factoriser

3. Le Test de sphéricité de Bartlett

La dernière condition, ce test est basé sur l'analyse de la matrice des corrélations R. L'idée est de procéder à un test d'hypothèse, en supposant comme H_0 : les variables sont deux à deux indépendantes (R = matrice unité). Deux situations extrêmes se présente, notons que |R| = produit des valeurs propres de R.

- |R| = 1, les variables sont deux à deux orthogonales, l'ACP est inutile, puisqu'il est impossible de résumer l'information.
- |R| = 0, il y a une colinéarité parfaite (le 1er facteur explique 100% de l'inertie totale).

Donc, en conclusion, on effectue le rejet de H_0. Il est possible de compresser l'information avec l'ACP. Efficacement même si l'on en juge la valeur de |R|.

IV. Nuage des individus

Sur la Figure 27, à chaque individu est associée une suite de *K* nombres. Selon ce point de vue, un individu peut être représenté comme un point de l'espace vectoriel à *K* dimensions, noté $\mathbb{R}^K$, dont chaque dimension représente une variable. L'ensemble des individus constitue le nuage N_I dont le centre de gravité G est confondu avec l'origine O des axes du fait du centrage ; G représente l'individu moyen précédemment cité.

Dans l'espace $\mathbb{R}^K$, la notion de ressemblance entre deux individus introduite section 1.1 n'est autre que la distance euclidienne usuelle. Donc, l'ensemble des distances inter-individus constitue ce que l'on appelle la forme du nuage N_I. Un de nos objectifs en ACP est de réaliser un bilan de ces distances qui revient à étudier la forme du nuage N_I, c'est-à-dire à y déceler une partition des individus (la

typologie mentionnée dans les objectifs) ou des directions d'allongement remarquables (les principales dimensions de variabilité).

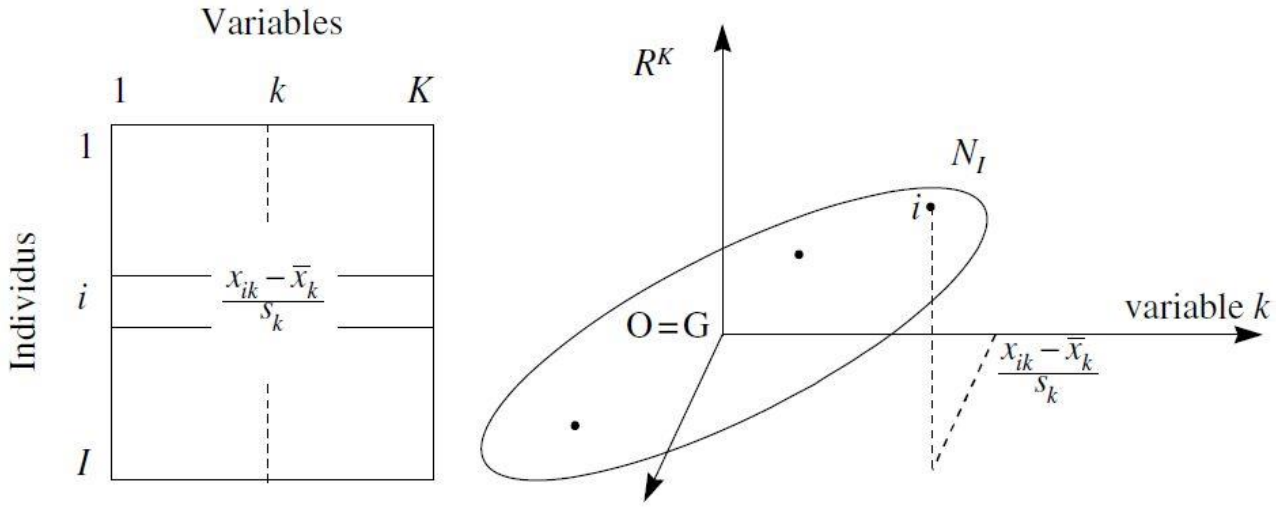

Figure 27 : Tableau des données et nuages des individus associé dans l'espace $\mathbb{R}^K$

V. Ajustement du nuage des individus

L'objectif est de fournir des images planes approchées du nuage N_I situé dans l'espace $\mathbb{R}^K$. Premièrement, on va essayer d'avoir une image approchée dans un espace à une dimension,c'est à dire, retrouver la meilleure représentation axiale des individus en déformant le moins possible le nuage, c'est l'axe d'inertie maximum. Après, on Cherche une deuxième représentation axiale, orthogonale à la première qui satisfait les mêmes conditions, c'est le plan d'inertie maximum et ainsi de suite.

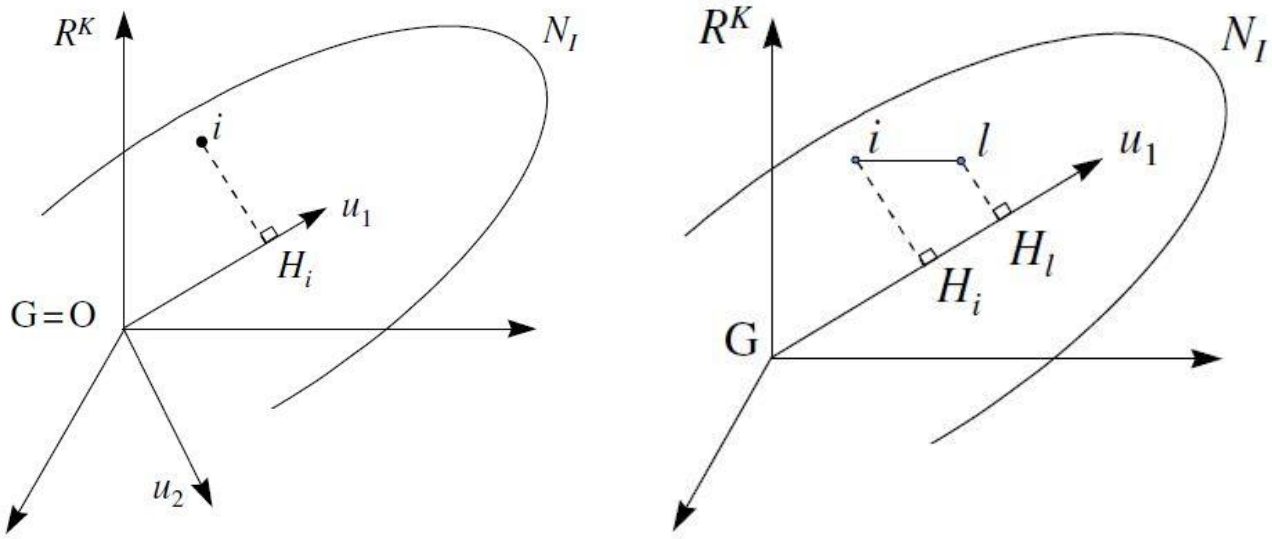

Figure 28 : Ajustement du nuage des individus (à gauche) ; représentation des distances inter-individuelles (à droite).

1. Axe d'inertie maximum

C'est le critère à une dimension. On recherche un axe de variabilité maximum $\vec{u}_1$ sur lequel on va projeter les individus i en H_i et qui rend maximum $\sum_i OH_i^2$.

2. Plan d'inertie maximum

Le critère à deux dimensions. On cherche $\vec{u}_2$, orthogonal à $\vec{u}_1$, qui satisfait le même critère de façon à obtenir un plan pour visualiser les distances inter-individuelles. C'est-à-dire est tel que $\sum_i \sum_l d^2\,(H_i - H_l)$ est le plus proche possible de $\sum_i \sum_l d^2\,(i,l)$.

3. Vecteurs et valeurs propres

Plus généralement, on procède à un emboîtement de solution. C'est-à-dire, après la recherche du premier et du deuxième axe, on réeffectue la procédure pour trouver une suite d'axes orthogonaux d'inertie maximum à condition qu'ils sont orthogonaux aux axes de rang inférieur. Tout ceci peut se formuler comme suit :

- $\vec{u}_s$ un vecteur unitaire de l'axe de rang s, les vecteurs propres ou les axes d'inertie maximum.
- H_i^s la projection de M_i sur $\vec{u}_s$.
- Trouver $\vec{u}_s$ tel que $\lambda_s = \sum_i \frac{1}{I}(OH_i^s)^2$, les valeurs propres ou inerties projetées soient maximum avec la contrainte : $\vec{u}_s \perp \vec{u}_t \; pour \; t < s$.

Au final, $\vec{u}_s$ est un vecteur propre unitaire de la matrice des corrélations associé à la valeur propre de rang s (λ_s). Cette dernière constitue la contribution des individus aux axes factoriels.

VI. Représentation des variables associée à la représentation des individus

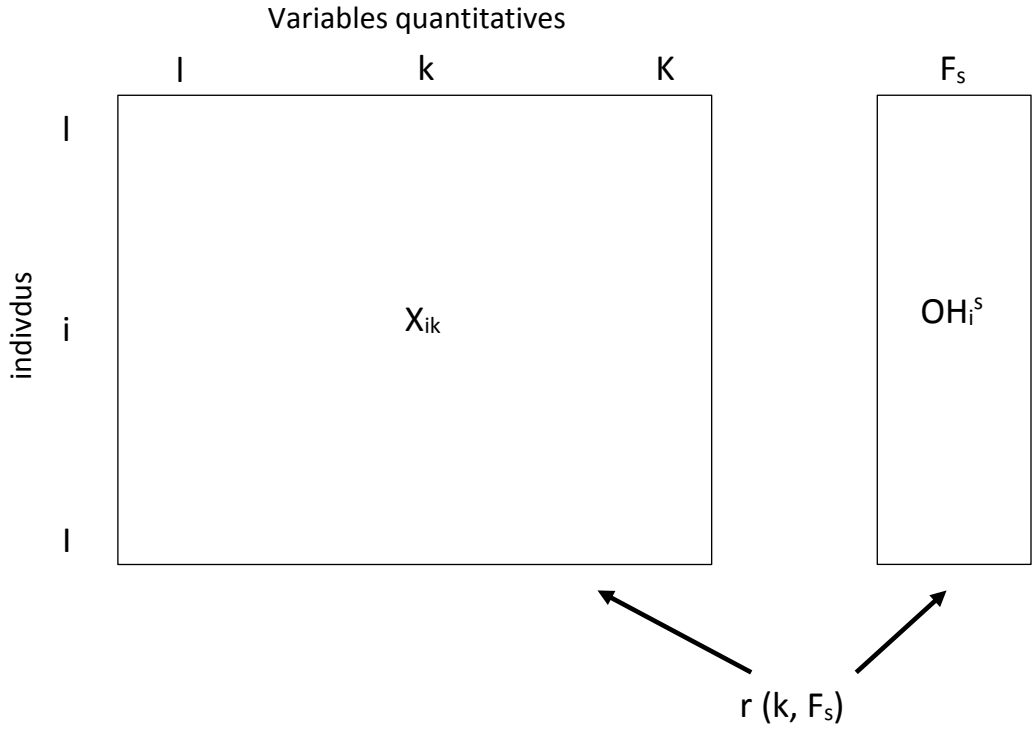

Figure 29 : Représentation des variables associée à la représentation des individus

Pour interpréter la représentation des individus sur les axes, on va calculer le coefficient de corrélation entre les variables classiques et d'autres variables artificielles qui sont les coordonnées sur

un axe. Ainsi, on peut expliquer, les principales dimensions de variabilité d'un ou des groupes individus sur les axes. Le principe est de prendre le tableau de données et de la comparer avec les valeurs propres de chaque axe factoriel (Figure 29).

A partir de cette représentation, on déduit le cercle des corrélations qui représente la contribution des variables aux axes factoriels.

VII. Nuage des variables

Sur la Figure 30, à chaque variable, est associée une suite de *I* nombres. Selon ce point de vue, une variable peut être représentée comme un vecteur de l'espace vectoriel à *I* dimensions, noté $\mathbb{R}^I$, dont chaque dimension représente un individu. L'ensemble des extrémités des vecteurs représentant les variables constitue le nuage N_K.

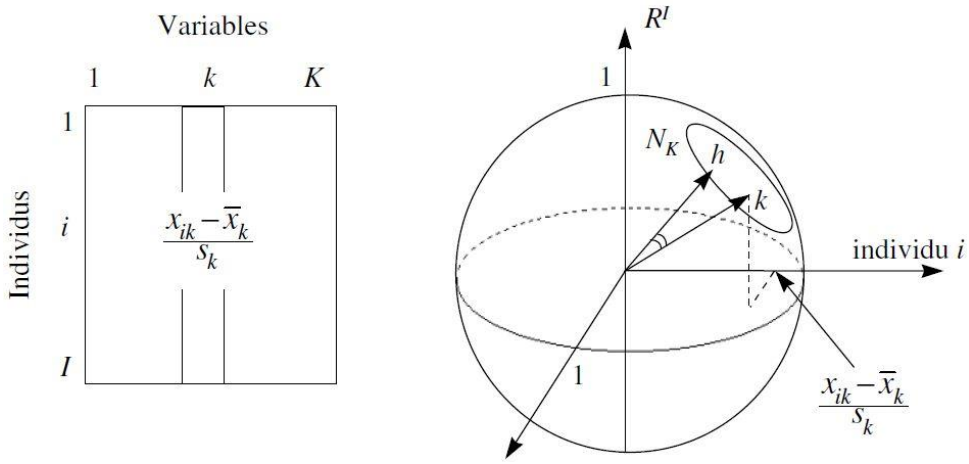

Figure 30 : Tableau des données et nuages des variables associé dans l'espace $\mathbb{R}^I$

. Les variables ont pour coordonnées les valeurs $\frac{x_{ik}-\bar{x}_k}{\sigma_k}$ (centrées-réduites) rangées en colonnes, Figure 30(à gauche).

Dans la figure 30 (à droite), les variables sont représentées par des vecteurs partant de l'origine $\overrightarrow{Oh}\ et\ \overrightarrow{Ok}$, des vecteurs à I coordonnées. Par rapport à l'étude des individus, ici, on étudie la liaison entre les variables en déterminant le cosinus l'angle $\theta_{hk} = \left(\overrightarrow{Oh}, \overrightarrow{Ok}\right)$ que fait les deux vecteurs. Cet angle peut s'exprimé par le produit scalaire de la variable $x_{ik}\ et\ x_{ih}$ divisé par la norme des deux, qui s'écrit :

$$\cos\left(\overrightarrow{Oh}, \overrightarrow{Ok}\right) = \frac{\langle x_k, x_h \rangle}{\|x_k\|\|x_h\|} = \frac{\sum_{i=1}^{I} x_{ik} x_{ih}}{\sqrt{\sum_{i=1}^{I} x_{ik}^2}\sqrt{\sum_{i=1}^{I} x_{ih}^2}} \qquad (36)$$

Comme les valeurs sont centrées, le coefficient de corrélation est :

$$cos\left(\overrightarrow{Oh}, \overrightarrow{Ok}\right) = r(x_h, x_l) \qquad (37)$$

Et du fait de la réduction, la longueur des vecteurs vaut :

$$\|Variable\ K\|^2 = 1 \qquad (38)$$

VIII. Ajustement du nuage des variables

L'objectif est d'obtenir une suite de *S* variables synthétiques $\{v_s; s = 1, \dots, S\}$ et une représentation approchée des corrélations entre les variables, l'ACP applique au nuage N_K des variables la même démarche qu'au nuage des individus (Figure 31).

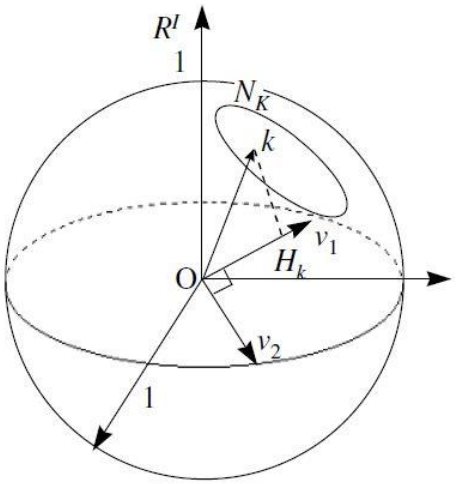

Figure 31 : Ajustement du nuage des variables

Le critère (inertie projetée maximum) satisfait dans le choix des axes est exactement le même que pour le nuage d'individus. Mais il prend une signification différente du fait que le nuage n'est pas centré (son centre de gravité n'est pas à l'origine) et que tous les points sont situés sur la sphère unité : ce sont les angles entre les vecteurs représentant les variables qui sont peu déformés par les projections et non pas les distances entre les points du nuage. En effet, le plan (v_1, v_2), en maximisant l'inertie à l'origine du nuage projeté, rend maximum la somme des cosinus carrés des angles entre les vecteurs et leur projection : il ajuste les vecteurs et déforme donc le moins possible leurs angles.

1. Les Composantes Principales

Le vecteur v_1 qui caractérise la direction d'inertie maximum définit une nouvelle variable. Les variables étudiées étant centrées et réduites, leur projection sur v_1 est égale à leur coefficient de corrélation avec cette variable. De ce fait, rechercher le vecteur v_1 qui rend maximum $\sum_k OH_k^2$ équivaut à rechercher la combinaison linéaire la plus liée à l'ensemble des variables (au sens du critère : somme des carrés des corrélations maximum). En ce sens, v_1 est la variable qui synthétise le mieux l'ensemble des variables initiales. Les axes factoriels étant orthogonaux deux à deux, on met en évidence une suite de variables synthétiques, les composantes principales, non corrélées entre elles, qui résument au mieux l'ensemble des variables initiales.

Plus généralement, on procède à un emboîtement de solution comme pour les vecteurs et valeurs propres. Tout ceci peut être formuler comme suit :

- Rechercher v_s, un vecteur unitaire de la direction de rang s, des variables synthétiques considérée comme axes d'inertie maximum, mais qui ne le sont pas.
- M_k, la représentation de la variable k dans $\mathbb{R}^I$.
- H_k^s, la projection de M_k sur v_s.
- Trouver v_s tel que $\sum_k (OH_k^s)^2$, les composantes principales soient maximums avec la contrainte $v_s \perp v_t$

Au final, on déduit la représentation optimale des variables avec les composantes principales qui s'écrivent :

$$v_s = \sum_k (OH_k^s)^2 = \sum_k (\cos\theta_k^s)^2 = \sum_k \big(r(k, v_s)\big)^2 \qquad (39)$$

2. L'effet taille

Si, dans un jeu de données, les variables sont toutes corrélées positivement deux à deux, alors le nuage N_K est loin de l'origine. Le premier axe factoriel rend alors surtout compte de la position de N_K par rapport à l'origine : parallèlement, la forme du nuage N_K est mal représentée en ce sens que les projections des variables sont proches les unes des autres (Figure 32).

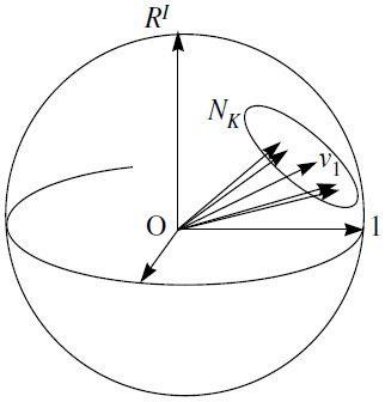

Figure 32 : L'effet taille dans $\mathbb{R}^I$.

Ce cas de figure est couramment appelé « effet taille » : Les variables, étant corrélées positivement deux à deux, forment entre elles des angles aigus. Le nuage N_K est concentré sur un petit secteur de la sphère. La projection des variables sur le premier axe factoriel, défini par v_1, rend compte principalement de la position de N_K par rapport à O. Il existe donc dans ce cas, une structure commune à l'ensemble des variables : c'est ce que traduit la première composante principale.

IX. Extraction des Composantes Principales

SPSS permet de fixer le nombre de composantes voulues. Deux critères reviennent fréquemment pour déterminer le nombre de facteurs à extraire :

❖ **La valeur dite de « eigen » (*eigenvalue*)**, traduite en français par « Valeur propre initiale », est le critère le plus largement utilisé. Plus la valeur propre initiale est élevée, plus le facteur explique

une portion significative de la variance totale. Par convention, tout facteur avec une valeur propre initiale supérieure à 1 est considéré comme facteur significatif. La sortie de résultats SPSS affiche le pourcentage de variance expliquée uniquement pour les facteurs ayant une valeur propre initiale supérieure ou égale à 1.

❖ **Le coude de Cattell** constitue un critère plus sévère pour déterminer le nombre de facteurs. Il est possible de demander dans les options associées à l'ACP de réaliser un graphique à partir des valeurs propres. Tous les points représentent les valeurs propres des composantes. Ils sont reliés par une ligne. On ne retient que les facteurs qui se situent avant le changement abrupt de la pente. Les points qui suivent ce changement, appelé rupture du coude, semblent former une ligne droite horizontale. L'information ajoutée par les facteurs représentés par ces points est peu pertinente

1. Rotation des axes factoriels – Rotation VARIMAX

L'objectif est d'obtenir des composantes plus facilement interprétables. La condition initiale nous dit qu'il se peut très bien que chaque variable présente individuellement une corrélation « moyenne » avec la composante. Dans ce cas, l'interprétation sera difficile.

L'idée est de faire pivoter les facteurs (à nombre de facteurs fixés) de manière à rendre plus tranchées (clarifier) les liaisons ou non-liaisons des variables avec l'un des facteurs (on peut préserver ou non l'orthogonalité) dans le but de faciliter l'interprétation des facteurs, Figure 33.

La méthode utilisée est la Rotation VARIMAX qui permet de maximiser la variance des carrées des corrélations intra-facteurs (c'est-à-dire, les rendre les plus différents possibles les uns des autres)

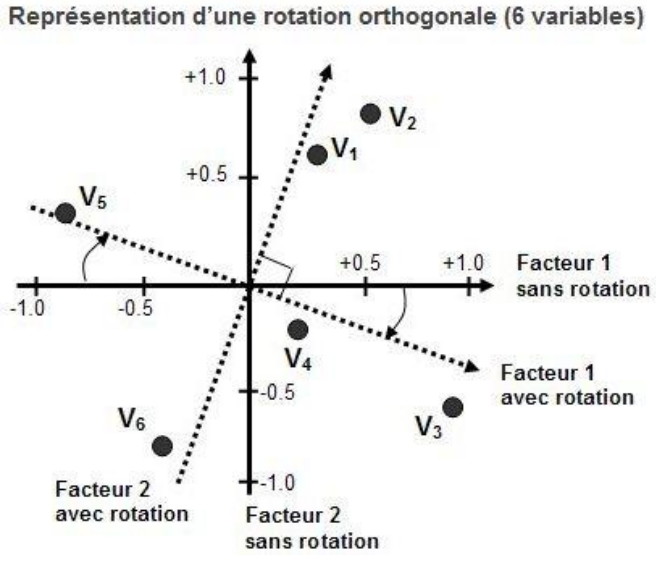

Figure 33 : Représentation d'une rotation des axes factoriels

X. Dualité et formules de transition en ACP normée

Le nuage N_I des individus et le nuage N_K des variables sont deux représentations du même tableau, l'une à travers ses lignes et l'autre à travers ses colonnes. Des relations très fortes, dites relations de dualité lient ces deux nuages. Ceci est une approche pour l'aide à l'interprétation.

1. Inerties

Tout d'abord, leur inertie totale est la même. Elle est égale au nombre de variables en ACP normée :

$$Inertie\ totale\ de\ N_I(ou\ de N_K) = \frac{1}{I}\sum_k \sum_i \left(\frac{x_{ik} - \overline{x}_k}{\sigma_k}\right)^2 = K \qquad (40)$$

La projection de chacun de ces deux nuages sur une suite d'axes orthogonaux correspond à une décomposition de l'inertie totale. On peut montrer que les deux décompositions sont identiques : les inerties des nuages N_I et N_K projetés sur les axes factoriels de même rang sont égales (et notées λ_s). Soit, pour les axes de rang *s* :

$$Inertie(N_I/u_s) = Inertie(N_K/v_s) = \lambda_s \qquad (41)$$

2. Facteurs

L'ensemble des projections de tous les points du nuage d'individus N_I sur le s^e axe factoriel u_s , appelé s^e facteur sur les individus, constitue une nouvelle variable notée F_s . Avec la s^e composante principale v_s, la relation entre le *s*e facteur sur *I* et le s^e axe factoriel de R_I s'écrit donc :

$$v_s = \frac{1}{\sqrt{\lambda_s}} F_s \qquad (42)$$

Sur la Figure 34, les coordonnées de N_I sur u_s (*s*e axe factoriel de N_I) constituent le *s*e facteur sur les individus (noté F_s). Le vecteur F_s dans R^I est colinéaire à v_s (s^e axe factoriel de N_K). Ainsi, les projections planes des individus dans R^K sont des représentations graphiques des couples de variables synthétiques obtenues dans R^I.

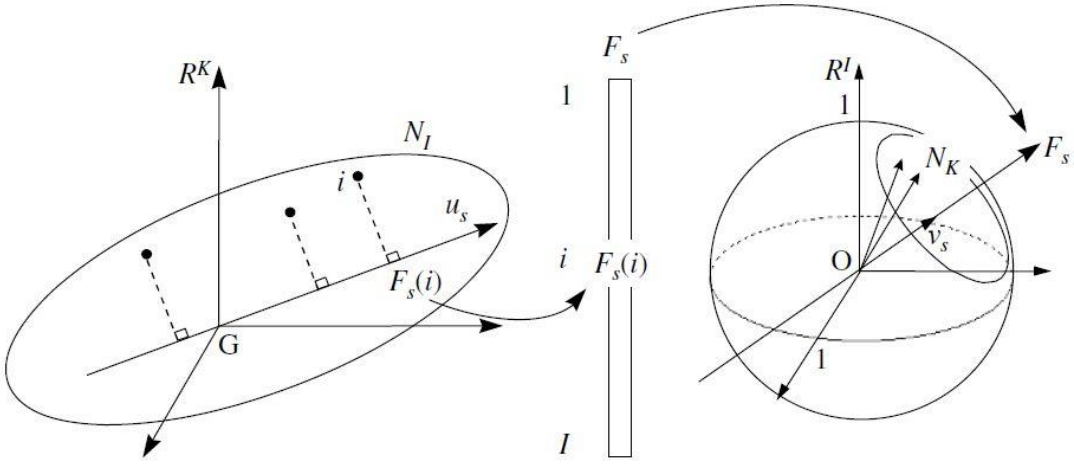

Figure 34 : La première forme de la dualité

Sur la Figure 35, les coordonnées de N_K sur v_s (*s*e axe factoriel de N_K) constituent le s^e facteur sur les variables (noté G_s). Le vecteur G_s dans R^K est colinéaire au s^e axe factoriel *us* de N_I.

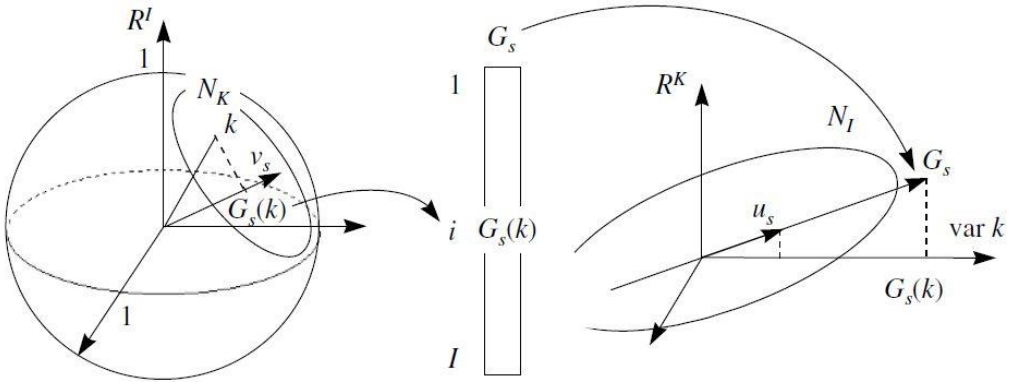

Figure 35 : La deuxième forme de dualité

On montre que le point représentant dans R^K cet individu type est situé sur le s^e axe du nuage des individus. Plus précisément :

$$u_s = \frac{1}{\sqrt{\lambda_s}} G_s \qquad (43)$$

XI. Relation de transition

On appelle relations de transition entre les facteurs de rang *s*, F_s et G_s , l'écriture algébrique des propriétés illustrées par les figures 37 et 38. Ces relations s'écrivent, en notant λ_s l'inertie projetée de N_I (ou de N_K) sur l'axe de rang *s* :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{x_{ik} - \overline{x}_k}{\sigma_k} G_s(k) \qquad (44)$$

$$G_s(k) = \frac{1}{I} \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{x_{ik} - \overline{x}_k}{\sigma_k} F_s(i) \qquad (45)$$

La formule 43 exprime le fait que la projection $F_s(i)$ d'un individu *i*, est une combinaison linéaire des projections $G_s(k)$ de toutes les variables. Dans cette combinaison linéaire, le coefficient d'une variable *k* est positif si la valeur x_{ik} de cette variable pour l'individu *i* dépasse la moyenne $\overline{x}_k$. Dans le cas contraire, ce coefficient est négatif. Ainsi, lorsque l'on regarde simultanément les deux graphiques, un individu est du côté des variables pour lesquelles il a de fortes valeurs et à l'opposé des variables pour lesquelles il a de faibles valeurs. Le graphique des individus est une représentation approchée des distances interindividuelles. Celui des variables peut être considéré en tant qu'élément explicatif de cette représentation : deux individus situés à une même extrémité d'un axe sont proches car ils ont tous deux généralement de fortes valeurs pour les variables situées du même côté qu'eux et de faibles valeurs pour les variables situées à l'opposé.

Réciproquement, le graphique des individus peut intervenir en tant qu'aide à l'interprétation du graphique des variables : si deux variables sont très corrélées positivement, elles sont situées du même côté sur un axe. Sur l'axe correspondant du nuage d'individus, les individus qui ont de fortes valeurs pour ces deux variables se situent du même côté qu'elles et ceux qui ont de faibles valeurs se situent à

l'opposé. Les individus extrêmes pour ces variables sont loin de l'origine. Les éventuels individus particuliers induisant à eux seuls des corrélations fortes sont ainsi repérés facilement.

XII. Aide à l'interprétation

La richesse des interprétations vient de la connaissance des données à analyser. L'intérêt de l'ACP est de fournir des variables synthétiques qui constituent un résumé de l'ensemble des variables initiales et sont la base d'une représentation plane approchée des variables et de leurs angles. La représentation des variables aide à l'interprétation de celle des individus et réciproquement Il en résulte que l'étude d'un plan est toujours réalisée conjointement avec la consultation d'un ensemble d'indicateurs regroupés sous le terme d'aides à l'interprétation.

1. Qualité de représentation d'un élément par un axe

La qualité de représentation de l'élément *i* (individu ou variable) par l'axe *s* est mesurée par le rapport :

$$QLT_s(i) = \frac{[inertie\ de\ la\ projection\ de\ l'élément\ i\ sur\ l'axes]}{[inertie\ totale\ de\ i]} \qquad (46)$$

C'est aussi le cosinus carré de l'angle θ entre O_i et l'axe *s, comme dans la Figure 36.*

$$QLT_s(i) = \frac{(OH_i^s)^2}{(O_i)^2} = cos^2\theta \qquad (47)$$

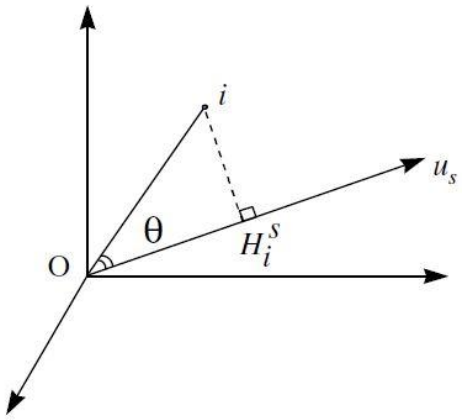

Figure 36 : Qualité de représentation d'un élément par un axe

Si la qualité de représentation d'un point sur un axe ou un plan est proche de 1, ce point est très proche de l'axe ou du plan.

2. Qualité de représentation d'un nuage par un axe

La définition précédente se généralise à l'ensemble d'un nuage par le rapport :

$$\frac{Inertie\ de\ la\ projection\ du\ nuage\ sur\ l'axe}{Inertie\ totale\ du\ nuage} \qquad (48)$$

Cet indicateur, appelé pourcentage d'inertie associé à un axe, mesure en outre « l'importance » relative d'un axe factoriel dans la variabilité des données.

3. Contribution d'un élément à l'inertie d'un axe

Un axe factoriel rend maximum (sous contrainte d'orthogonalité avec les axes précédents) l'inertie projetée d'un nuage. Cette inertie projetée du nuage peut être décomposée point par point. Le quotient de l'inertie de la projection de l'élément i (de poids p_i) sur l'axe s [soit $p_i(OH_i^s)^2$] par l'inertie de la projection de l'ensemble du nuage sur l'axe s (soit λ_s) représente la contribution de l'élément i à l'inertie de l'axe s. Soit, en notant CTR_s (i) la contribution de l'élément i à l'axe de rang s :

$$CTR_s(i) = \frac{p_i(OH_i^s)^2}{\lambda_s} \qquad (49)$$

Cet indicateur se généralise à un sous-ensemble d'éléments. La contribution d'un ensemble de points à l'inertie d'un axe est la somme des contributions des points qui le composent. Ce rapport est précieux pour mettre en évidence le sous-ensemble d'éléments qui ont contribué principalement à la construction de l'axe et sur lequel s'appuiera en premier lieu l'interprétation.

Section 2 La Classification Ascendante Hiérarchique (CAH)

[4] Tout d'abord, la classification est l'action de constituer ou de construire des classes ou clusters (communément en statistique), un ensemble d'individus possédant des traits de caractères communs (groupes, catégories). Dans notre cas, les individus sont les mois et les saisons. Notons que nous avons en effet un grand échantillon, au total : 786 mois, 65 saison des pluies et 64 saison sèches puisque la dernière année se termine en Juin. L'examen visuelle du nuage de point d'un tel effectif sera pratiquement impossible sans regroupement des individus représentant les mêmes caractères. Donc, nous avons opter pour la CAH, pour procéder à l'automatisation de cette tâche. Effectivement, l'analyse sera traitée avec SPSS.

I. Principes généraux de la CAH

La CAH s'intéresse au même schéma de données que l'ACP : individus x variables quantitatives. Nous l'utiliserons donc en complémentaire à l'ACP.

1. Les objectifs

L'objectif principal est la production d'une structure (arborescence ou dendrogramme) sur la figure 37, permettant :

- La mise en évidence de liens hiérarchiques entre individus ou groupes d'individus.
- La détection d'un nombre de classes « naturel » au sein de la population.

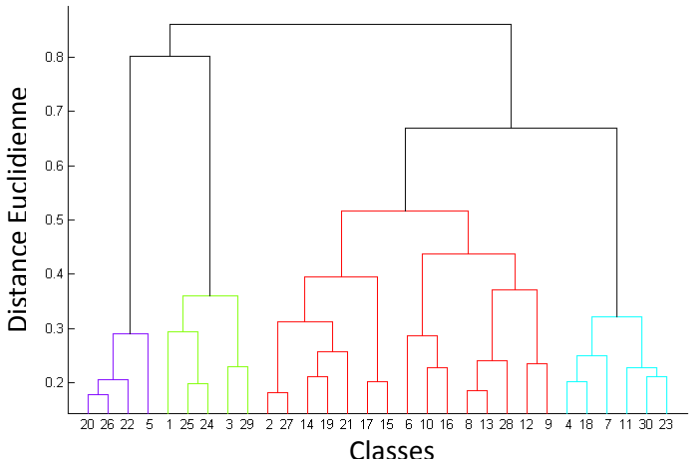

Figure 37 : Disposition d'un Arbre Hiérarchique.

II. Les critères

Il y a 3 critères pour opérer une CAH : Le critère principal est la mesure d'éloignement, le critère d'homogénéité, et le choix d'une méthode (ici, nous utiliserons la méthode de Ward).

1. La mesure d'éloignement

Principalement, on cherche la ressemblance entre les individus, ensuite sur les groupes d'individus définies par la distance euclidienne que nous avons déjà vu en ACP, à la formule 33. Donc, plus cette distance est petite plus les individus ou les groupes se ressemblent.

2. Le critère d'Homogénéité

L'homogénéité est basée sur l'inertie d'un individu par rapport à son centre de gravité dans la classe à laquelle il appartient, appelée inertie-inter ; et l'inertie de chaque classe par rapport à un autre centre de gravité, appelé inertie-intra. Donc, Une classe est homogène si son inertie est faible. Nous avons aussi vue les formule 41 et 42 calculant l'inertie en ACP.

Ces deux critères sont équivalents d'après le théorème de Huygens sur l'inertie totale, on a :

$$I_{totale} = I_{inter} + I_{intra} \qquad (50)$$

Cette formule peut être illustrer par la Figure 38 Ci-dessous.

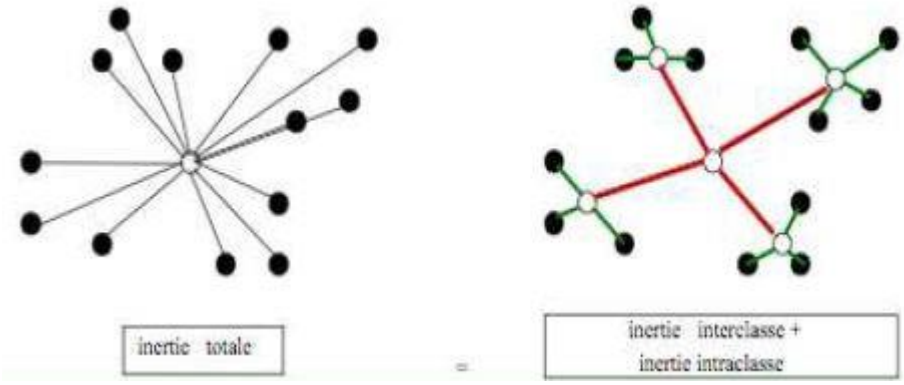

Figure 38 : Schéma expliquant le fonctionnement des inerties dans une CAH.

3. Le choix de la méthode

La méthode choisie est celle de Ward qui considère chaque individu comme étant une classe, ce qui rendra les partitions optimales.

III. L'algorithme de Ward

Cette méthode permet d'avoir un résultat optimal sur la classification si le nombre d'individu n'est pas trop grands. Il se base sur la qualité de la partition et la coupure de l'arbre (le choix du nombre de classes)

1. Qualité d'une Partition

On définit qu'une partition est optimale si les individus d'une même clase sont proches et les individus de 2 classes différentes sont éloignés. Mathématiquement, on traduit que la variabilité intra-classe est petite et celle inter-classe est grande. Ce qui revient à minimiser l'inertie inter et maximiser l'inertie intra. Et la qualité de la partition est défini par :

$$0 \leq \frac{Inertie\ inter}{Inertie\ totale} \leq 1 \qquad (51)$$

Plus la valeur est proche de 1, puis on a une partition optimale. Donc, L'idée de Ward consiste à choisir à chaque pas le regroupement de classes tel que l'augmentation de l'inertie intra soit minimum.

2. Coupure de l'arbre Hiérarchique

Le fait de couper l'arbre hiérarchique revient à choisir le nombre de classe qui rend la partition la meilleure. Dans le cas de la méthode de Ward, la hauteur d'une branche est proportionnelle à la distance entre les deux classes regroupées. Ce qui veut dire que si on monte dans l'arbre (agréger), on diminue de nombre de classe. Alors on perd de l'inertie implique, on perd de l'information. Sur ce, la méthode dit de couper l'arbre avant une forte perte d'inertie qu'on détectera sur le diagramme des indices d'agrégations.

IV. Mise en œuvre conjointe de l'ACP et de la CAH

Ici, la CAH est mise en œuvre à partir des facteurs obtenus en ACP et non des données brutes. L'utilisation des contributions des chaque individu aux facteurs permet d'éliminer les résidus d'informations qui pourrait faussés la représentation qui influera sur l'interprétation.

Synthèse de la partie

Nombreux sont les combinaisons de méthode qu'on pourrait utilisées pour traiter le sujet. Le choix des méthodes réside surtout sur la disponibilité des données à exploiter. Pour notre cas, on a suivi la procédure méthodologique suivante :

–Au sixième chapitre, la présentation des données et matériels utilisés. On insiste surtout à la nature des données et les logiciels avec lesquels on les traitera dans ce chapitre.

–Au septième chapitre, les méthodes de caractérisation de la zone et de la période étudiées.

–Le huitième chapitre, on procède au processus de traitement de données pluviométriques en suivant les normes de l'OMM. La mise en évidence des traitements automatique des données sur les progiciels est à la page.

–Le neuvième, dixième et onzième chapitres : la statistique descriptive et inférentielle, les normales climatologiques, et l'utilisation des indices SPI contribuent à la description de l'anomalie de précipitation de l'année 2016-2017.

–Dans le douzième, treizième et quatorzième chapitres, la contribution à l'étude des relations avec les variabilité climatiques ENSO et IOD.

Notons aussi qu'on va faire 2 ACP normée l'un à partir des indices mensuels, l'autre avec les indices saisonniers pour mieux observer les liaisons, et d'avoir une idée précise sur les impacts individuels et combinés des phénomènes ENSO et IOD. Les procédures des ACP et de la CAH sous SPSS seront présentées dans l'Annexe XI.