MPRA
Munich Personal RePEc Archive

Application of Hierarchical Ascendant Clustering to the Distribution of Budgetary Ressources in the City-Province of Kinshasa

Mputu Losala Lomo, Denis-Robert

2022

Online at https://mpra.ub.uni-muenchen.de/113774/
MPRA Paper No. 113774, posted 09 Aug 2022 13:38 UTC

ENSEIGNEMENT SUPERIEUR ET UNIVERSITAIRE

UNIVERSITE PEDAGOGIQUE NATIONALE

B.P. 8815 KIN 1

Faculté des Sciences

Département de Mathématique et Informatique

MEMOIRE DE DEA

Domaine : Mathématiques

Orientation : Mathématiques Appliquées

Option : Analyse des données

APPLICATION DE LA CLASSIFICATION ASCENDANTE HIERARCHIQUE A LA REPARTITION DES RESSOURCES BUDGETAIRES DANS LA VILLE-PROVINCE DE KINSHASA

MPUTU LOSALA LOMO DENIS-ROBERT

Mémoire présenté et défendu le 11 Juillet 2022 en vue de l'obtention du Diplôme d'Etudes Approfondies en Sciences

Promoteur : MABELA MAKENGO MATENDO Rostin

Copromoteur : SUMATA MOTUKULA Claude

Jury

Président	:	TSHIBAKA KATUMONANGANI Emmanuel ; P., UPN
Secrétaire	:	MPAKASA KIBULU Dino ; P., UPN
Membres	:	MABELA MAKENGO MATENDO Rostin ; P.O., UNIKIN
	:	SUMATA MOTUKULA Claude ; P.O., UPN

Année Académique 2018-2019

Épigraphe

« L'élévation ne dépend pas que du prix payé par l'homme mais aussi et surtout de la grâce de Dieu manifestée en son temps »

Denis-Robert MPUTU

Dédicace

A ma très chère mère ITUNA BONDJOLONGO Thérèse qui n'a pas ménagé de sa force pour jeter dans ma vie des bases solides d'instruction qui portent ce jour leur fruit ;

A ma très chère épouse BOKOLO AMBA-MPUTU Bibiche pour son attachement et son soutien ;

A mes enfants : MPUTU LOSALA-LOMO Plajodi, MPUTU BOSOY-BAEKE Christivi, MPUTU BOKOPOLO Beldi, MPUTU ITUNA Miradi, MPUTU AMBA-MBOYO Sarah-Mervedi, MPUTU BOKOLO-MPIA Emmanuel-Merdi auxquels s'ajoute ma belle-sœur AMBA MPUTU Nenette pour avoir supporté tous les sacrifices consentis.

Remerciements

Nous sommes, certes, l'auteur de ce travail de recherche qui marque la fin de notre formation de DEA en Sciences, Orientation : Maths Appliquées, mais il y a bien de gens qui ont contribué pour qu'il soit apprêté. Nous nous donnons ici le devoir de leur montrer notre gratitude.

Nous voulons d'abord rendre grâce à l'Eternel Dieu Tout Puissant, Créateur de l'Univers, l'Inspirateur et le soutien incomparable de notre formation.

Nous tenons à remercier vivement le comité de gestion de l'UPN et toutes les autorités de la faculté des sciences plus particulièrement du département de Mathématique-Informatique de l'UPN pour leur sens de responsabilité ainsi que tout le personnel de l'UPN, de l'EIFI, du CIDEP-Université Ouverte, de la Faculté de Droit de l'UNIKIN pour tous les efforts fournis afin de nous assurer une formation de qualité.

Nous témoignons notre gratitude à l'endroit du Professeur Docteur Rostin MABELA MAKENGO MATENDO qui n'a pas hésité d'accepter la direction de notre mémoire en dépit de ses multiples occupations. Sa simplicité et son esprit d'ouverture nous ont permis de recevoir de lui d'importantes remarques et sages conseils qui nous ont beaucoup aidés lors de nos recherches et continueront certainement à nous aider dans l'avenir.

Notre gratitude va également à l'endroit du Professeur Docteur Claude SUMATA MOTUKULA de la Faculté de Sciences économiques de l'UPN et Professeur Docteur Alain MUSESA LANDA de l'UNIKIN pour avoir volontiers accepté de faire partie du comité d'encadrement et pour leurs remarques et sages conseils.

Nous adressons nos remerciements aux autorités de la ville de Kinshasa, plus particulièrement aux vingt-quatre Bourgmestres et leurs collaborateurs pour leur aval et courtoisie lors de la collecte des données dans leurs communes respectives.

Nos remerciements s'adressent aussi d'un côté aux Professeurs docteurs ENGOMBE WEDI et NGILAMBI-te-AKONAMBI pour leurs recommandations pour nos études post-universitaires et de l'autre, aux professeurs docteurs MOPONDI Alexandre, INDENGE José, BOLA Jean-Pierre pour leur encadrement en notre qualité de chercheur.

Nous remercions les autorités du CEF La Borne pour nous avoir laissé gratuitement l'Internet à notre pleine disposition. Il s'agit, plus particulièrement de celles du C.S. La

Borne, nous citons : Madame le Préfet KALUBI Véronique et la Directrice ULUNGI-DILA Adrienne sans oublier tous leurs collaborateurs : ALIMETI Florent, MWAMBA Jean Baptiste, NZUZI Liévin, MAWAKA Mélanie, KILUBU Gaby, DIYOKA Médard, LAMA Eugène ainsi que tous nos combattants de lutte : SHAKATANGA Vital, DIMWANY Lucienne, MUTUBA Freddy, ... pour leur sympathie et encouragement.

Notre gratitude va à l'endroit des autorités de l'INS pour avoir mis à notre disposition leur expertise ainsi qu'aux autorités du Secrétariat de la Division Urbaine de l'Intérieur et Sécurité, plus particulièrement, Monsieur NKOSI Stanley et son adjoint KATUMBA Roger pour avoir mis à notre disposition les données utiles à ce travail.

Nous tenons à remercier notre grand-frère Docteur MPUTU José, nos grandes-sœurs BOKETSU Catho, BOLOWA Denise, MPUTU Bébé, MPUTU Malou, MPUTU Marthe ; notre neveu Pasteur IPAKA Olivier ainsi que notre Belle-mère Evangéliste BOKOPOLO Louise, veuve du Professeur Docteur BOKOLO NGUBE N'SELE Arthur pour leur soutien et confiance en notre personne.

Nos remerciements vont à l'endroit de nos amis et camarades de la Licence Spéciale et du DEA ; de nos collègues chercheurs et enseignants de l'IREM, il s'agit plus particulièrement de : NSWEYA Lambert, LAY BUAMOKE, ALONGE Jean, ADIA Baudouin, KOMBI Charles, SAKODI Joseph-Pierre,... ; de tous les anciens de l'IFI et de tous nos collègues Enseignants pour leur soutien moral.

Nous tenons à remercier tous les dirigeants et membres de l'Eglise Assemblée des Enfants de Dieu (AED) que nous dirigeons et ceux de l'Eglise Siège de Dieu dirigée par le Pasteur NGOMAMBULA PERO Marcel ainsi qu'aux évangélistes BIKOPO Eddy, DONGO Enock, YEKINI Fréderick et Pasteur BINYONGO Abel pour leur soutien spirituel.

Enfin, nous adressons nos remerciements à tout celui qui, d'une manière ou d'une autre, a eu à contribuer dans notre formation de DEA ou à l'élaboration de ce mémoire.

Que Dieu vous bénisse tous.

Table des matières

Épigraphe **ii**

Dédicace **iii**

Remerciements **iv**

Table des matières **vi**

Liste des abréviations et sigles **ix**

Table des figures **xi**

Liste des tableaux **xii**

Résumé **xiv**

Abstract **xv**

Introduction générale **1**

Chapitre 1. Quelques méthodes factorielles **12**

1.1 Analyse en Composantes Principales 12

1.1.1 Matrice des données 13

1.1.2 Transformation des données 13

1.1.3 Matrice des variances-covariances et matrice des corrélations . . 14

1.1.4 Diagonalisation de la matrice des variances-covariances (ou de la matrice des corrélations) et détermination des facteurs 16

1.1.5 Présentation des résultats et interprétation 18

1.2 Analyse Factorielle des Correspondances 24

1.2.1 Tableau de contingence 25

1.2.2 Transformation des données de départ et présentation de tableaux de profils 26

1.2.3 Analyse en composantes principales du tableau des profils transformés centrés 30

1.3 Conclusion du premier chapitre 41

Chapitre 2. Classification Ascendante Hiérarchique **43**
2.1 Généralités sur la classification 43
2.1.1 Aperçu historique sur la classification 44
2.1.2 Données à utiliser dans une classification et objets à classifier . . 44
2.1.3 Subdivision de la Classification : classification supervisée et classification non supervisée (ou automatique) 45
2.2 Réalisation d'une CAH par des données quantitatives 46
2.2.1 Constitution du tableau des données et sa transformation éventuelle 47
2.2.2 Calcul des distances entre les individus deux à deux 48
2.2.3 Calcul de distance entre deux groupes (Choix d'un indice d'agrégation) 50
2.2.4 Représentation graphique : Le dendrogramme. 52
2.2.5 Découpe du dendrogramme 56
2.2.6 Interprétation d'une partition 61
2.3 Conclusion du deuxième chapitre 66

Chapitre 3. Procédés de la répartition des ressources **68**
3.1 Concepts de base du partage 69
3.1.1 Ressource 69
3.1.2 Partage 70
3.1.3 Principes de justice distributive 72
3.2 Répartition des Ressources Sans réduction des inégalités 74
3.2.1 Détermination des données à utiliser 74
3.2.2 Calcul des valeurs totales des individus et de la valeur globale de la population 75
3.2.3 Calcul des proportions et parts des individus 76
3.2.4 Représentation graphique des parts des individus 76
3.3 Répartition des Ressources avec réduction des inégalités 77
3.3.1 Détermination et présentation des résultats de la classification . . 77
3.3.2 Indice de niveau des inégalités et réduction des inégalités 85
3.3.3 Calcul des proportions des individus par rapport à l'ensemble de la population 89
3.3.4 Calcul de la part de la ressource commune revenant à chaque individu 90
3.3.5 Représentations graphiques des parts des individus 90

3.4 Conclusion du troisième chapitre 91

Chapitre 4. Application des procédés de la répartition des ressources 93

4.1 Quelques applications du Procédé PRRS 93

4.1.1 Cas de maintien des données de départ 93

4.1.2 Cas de transformation des données de départ 94

4.2 Application du Procédé PRRC 97

4.2.1 Présentation des objets, données et outils à utiliser 97

4.2.2 Réalisation de la classification Ascendante Hiérarchique après ACP 101

4.2.3 Réduction des inégalités entre les individus dans leurs classes et dans l'ensemble de la population 112

4.2.4 Proportions et parts des individus par rapport à la population . . 118

4.2.5 Représentation graphique des parts des communes 121

4.3 Conclusion du troisième chapitre 123

Conclusion générale et perspectives 124

Bibliographie 128

Annexes 135

Liste des abréviations et sigles

1. ACP : Analyse en Composantes Principales
2. AFC : Analyse Factorielle des Correspondances
3. AFCM(ou ACM) : Analyse Factorielle des Correspondances Multiples
4. Art. : Article
5. CAH : Classification Ascendante Hiérarchique
6. CDF : Franc congolais (unité monétaire de la RDC)
7. CDH : Classification Descendante Hiérarchique
8. CEF : Centre Évangélique Francophone
9. Cf.(ou Cfr) : Confère
10. Const. : Constitution
11. Contr. : Contribution d'un facteur à la variance totale expliquée
12. Coord : Coordonnée
13. CRIDUPN : Centre de Recherche InterDisciplinaire de l'Université Pédagogique Nationale
14. Ctr : Contribution d'un élément (individu, modalité ou variable) à la construction de l'axe
15. DEA : Diplôme d'Etude Approfondie
16. DTE/DTEs : Decentralized Territorial(s) Entity (ies)
17. ETD/ETDs : Entité(s) Territoriale(s) Décentralisée(s)
18. FC(ou CDF) : Franc congolais (Officiellement CDF)
19. HCPC : Hierarchical Clustering on Principal Components (En Français, Classification Hiérarchique sur Composantes Principales)
20. INS : Institut National de la Statistique
21. IREM : Institut de Recherche sur l'Enseignement des Mathématiques
22. L.O. : Loi Organique
23. P. (P.O.) : Professeur (Professeur Ordinaire)
24. PCSD : Procédé de Calcul Simultané des Distances

25. PDRC : Process of Distribution of Resource from the results of Classification)
26. PDRW : Process of Distribution of Resource Without reduction of inéqualities)
27. PRRC :Procédé de la Répartition des Ressources à partir des résultats de la Classification
28. PRRS : Procédé de la Répartition des Ressources Sans réduction des inégalités
29. PSCD : Process of Simultaneous Calculation of Distances
30. ODEP : Observatoire de la Dépense Publique
31. QLT : Qualité de la représentation d'un élément
32. R : Logiciel R
33. $\mathbb{R}$: Ensemble des nombres réels
34. RDC : République Démocratique du Congo
35. TDC : Tableau Disjonctif Complet
36. UNIKIN : Université de Kinshasa
37. UPN : Université Pédagogique Nationale

Table des figures

Figure 1.1. Cercle des corrélations (Exemple) 20
Figure 1.2. Représentation graphique des individus sur les axes factoriels (Exemple). 22

Figure 2.1. Subdivision des méthodes de la classification 46
Figure 2.2. Indices d'agrégation (Illustration). 52
Figure 2.3. Une Hiérarchie représentée par un Dendrogramme 53
Figure 2.4. Partition d'un ensemble 54
Figure 2.5. Dendrogramme (Exemple) 55
Figure 2.6. Représentation d'une distance ultramétrique 56
Figure 2.7. Représentation en mobile d'une hierarchie 56
Figure 2.8. Illustration de l'inertie intra-classe 57
Figure 2.9. Illustration de l'inertie inter-classe 58
Figure 2.10. Illustration de l'inertie totale 58
Figure 2.11. Illustration du théorème de Huygens 59
Figure 2.12. Illustration de la décomposition de l'inertie totale 60
Figure 2.13. Découpe du dendrogramme en deux classes (Exemple) 61
Figure 2.14. Illustration du parangon d'une classe 63
Figure 2.15. Illustration de l'extrême d'une classe 64

Figure 3.1. Deux matrices extraites du produit cartésien IxI 82

Figure 4.1. Dendrogramme des 24 communes de Kinshasa découpées en 6 classes 108
Figure 4.2. Graphique à bâtons des parts de 24 communes de la Ville-province de Kinshasa 122
Figure 4.3. Graphique en camembert des parts de 24 communes de la Ville-province de Kinshasa 123
Figure 4.4. Carte de la ville province de Kinshasa 135
Figure 4.5. Questionnaire d'enquête 136

Liste des tableaux

Tableau 1. Importance de considérer plusieurs variables au lieu d'une seule 4
Tableau 2. Même unité de mesure exprimée de différentes manières pour les variables homogènes 4
Tableau 3. Même unité de mesure exprimée de manière unique pour toutes les variables homogènes 5
Tableau 4. Quantité de viande achetée. Cas d'une seule unité de mesure . . 5
Tableau 5. Quantité de viande achetée. Cas de deux unités de mesure . . . 5
Tableau 6. Importance de la réduction des inégalités 6
Tableau 7. Nécessité de la mise en place d'un mécanisme de partage selon plusieurs variables hétérogènes 7

Tableau 1.1. Tableau de contingence (des données) (AFC) 25

Tableau 2.1. Tableau des données (Classification) 47

Tableau 3.1. Produit cartésien IxI 83

Tableau 4.1. Exemple de partage selon plusieurs variables homogènes 93
Tableau 4.2. Exemple de partage selon plusieurs variables homogènes (unité exprimée différemment) 94
Tableau 4.3. Exemple de partage selon plusieurs variables homogènes (une même notation de l'unité) 95
Tableau 4.4. Exemple de l'injustice remarquée lorsqu'on ne transforme pas les données 95
Tableau 4.5. Exemple des données d'un partage selon plusieurs variables hétérogènes 96
Tableau 4.6. Exemple de transformation de variables hétérogènes 96
Tableau 4.7. Exemple calcul des parts après transformation des variables hétérogènes 96
Tableau 4.8. Tableau T des données de 24 communes de la Ville-province de Kinshasa. Exercice 2015 102
Tableau 4.9. Tableau t des données réduites 104
Tableau 4.10. Tableau Tcr des données centrées réduites 105
Tableau 4.11. Tableau des valeurs propres et de contributions des axes factoriels 106

Tableau 4.12. Tableau des coordonnées des individus pour les deux premiers facteurs 107
Tableau 4.13. Tableau des numéros des classes respectives des individus . . . 109
Tableau 4.14. Tableau des parangons 110
Tableau 4.15. Tableau des extrêmes 111
Tableau 4.16. Tableau synthétique des parangons et des extrêmes des classes . 111
Tableau 4.17. Liste des variables caractérisant les classes 112
Tableau 4.18. Tableau des données réduites et des valeurs totales des individus 113
Tableau 4.19. Tableau des valeurs corrigées des individus par rapport à leurs classes 115
Tableau 4.20. Tableau des valeurs corrigées des individus par rapport à l'ensemble de la population 117
Tableau 4.21. Tableau des proportions des individus par rapport à l'ensemble de la population 119
Tableau 4.22. Tableau des parts des individus 121
Tableau 4.23. Quelques commandes de logiciel R 137
Tableau 4.24. Montants de rétrocession avec compléments. Exercice 2015 . . 138
Tableau 4.25. Montants estimés de capacité de production des communes . . . 140
Tableau 4.26. Table de fonction de répartition inverse de la loi normale 140
Tableau 4.27. Table de fonction de répartition de la loi normale centrée réduite (Probabilité $F(Z)$ de trouver une valeur inférieure à Z) 141
Tableau 4.28. Table de la loi Khi deux 142
Tableau 4.29. Tableau des variables caractérisant les classes 143

Résumé

Application de la Classification Ascendante Hiérarchique à la Répartition des Ressources Budgétaires dans la Ville-Province de Kinshasa.

Nous voulons plus particulièrement résoudre le problème de l'injustice constatée dans la répartition des ressources à caractère national allouées aux Entités Territoriales Décentralisées (ETDs), le cas des 24 communes de la ville-province de Kinshasa. En effet, cette répartition se fait dans la pratique en utilisant une seule variable (ou critère) qui est "Population" au lieu de trois (Capacité de production, superficie et population), proposées par le législateur, bien que toutes soient hétérogènes. Aussi, il n'existe aucun mécanisme de répartition des ressources communes utilisant plusieurs variables hétérogènes et prenant en compte les rapports entre les individus dans le sens à réduire les inégalités existant entre eux et conduisant au calcul de leurs parts respectives.

Élargissant le problème à d'autres cas qui se posent tant en Mathématiques qu'à la vie pratique, nous avons retenu quatre différentes causes d'injustice lors du partage notamment d'une somme d'argent. Il s'agit de : (1) *L'utilisation d'une seule variable* au lieu de plusieurs. (2) *L'utilisation directe des données de départ* issues *des variables homogènes pour lesquelles la même unité de mesure est exprimée différemment* pour chaque variable. (3) *L'utilisation directe des données de départ* issues *des variables hétérogènes*. (4) *L'absence de la réduction des inégalités* entre les individus.

Pour résoudre ces problèmes, nous avons proposé deux procédés qui utilisent plusieurs variables homogènes ou hétérogènes permettant le calcul des parts des individus : le premier est nommé "Procédé de la Répartition des Ressources Sans réduction des inégalités (PRRS)". Il s'occupe du cas où les individus ont contribué à la création de la ressource et ne permet pas la réduction des inégalités. Le deuxième, nommé "Procédé de la Répartition des Ressources à partir des résultats de la Classification (plus particulièrement, la Classification Ascendante Hiérarchique) (PRRC)" qui s'occupe du cas où les individus n'ont pas contribué à la création de la ressource et qui amène à réduire les inégalités entre eux. Ces procédés utilisent plusieurs formules intéressantes que nous avons proposées. Il s'agit notamment de l'« indice de niveau d'inégalités » pour mesurer le degré des inégalités entre les individus et de "la fonction des valeurs corrigées" pour la réduction des inégalités (écarts) entre les individus.

Dans la partie application, nous avons proposé quelques exemples sur le procédé PRRS. Quant au procédé PRRC, nous l'avons appliqué aux 24 communes de la Ville-province de Kinshasa pour répartir entre elles la somme (recettes à caractère national) leur allouée en 2015.

Au niveau de la classification, nous avons proposé deux formules de distances. La première calcule, elle seule, les distances entre deux individus isolés et entre deux groupes d'individus, contrairement au procédé classique qui utilise deux formules différentes. La deuxième, quant à elle, calcule simultanément les distances entre plusieurs individus et/ou groupes d'individus deux à deux en utilisant les matrices. A ce propos, nous avons proposé une démarche nommée « Procédé de Calcul Simultané de Distances (PCSD) » pour faciliter l'utilisation de cette formule.

Mots clés : Classification Ascendante Hiérarchique, Entités Territoriales Décentralisées, Partage, Réduction des inégalités, Répartition des ressources.

Abstract

Application of Hierarchical Ascendant Clustering to the Distribution of Budgetary Resources in the City-Province of Kinshasa.

More specifically, we want to resolve the problem of the injustice observed in the distribution of national resources allocated to Decentralized Territorial Entities (DTEs), the case of the 24 municipalities of the city-province of Kinshasa. Indeed, this distribution is done in practice by using a single variable (or criterion) which is “Population” instead of three (Production capacity, area and population), proposed by the legislator, although all are heterogeneous. Also, there is no mechanism for distributing common resources using several heterogeneous variables and taking into account the relationships between individuals in the sense of reducing the inequalities existing between them and leading to the calculation of their respective shares.

Extending the problem to other cases which arise both in Mathematics and in practical life, we have identified four different causes of injustice when sharing, in particular, a sum of money. These are : (1) Using a single variable instead of several. (2) The direct use of the starting data from homogeneous variables for which the same unit of measurement is expressed differently for each variable. (3) Direct use of initial data from heterogeneous variables. (4) The lack of reduction in inequalities between individuals.

To solve these problems, we have proposed two methods which use several homogeneous or heterogeneous variables allowing the calculation of the shares of individuals : the first is called "Process of Distribution of Ressource Without reduction of inequalities (PDRW)". It deals with the case where individuals have contributed to the creation of the resource and does not allow the reduction of inequalities. The second, called "Process of Distribution of Resource from the results of Clustering (more specifically, the Agglomerative Hierarchical Clustering) (PDRC)" which deals with the case where individuals have not contributed to the creation of resource and which leads to reducing inequalities between them. These methods use several interesting formulas that we have proposed. That concerns especially the "inequalities level index" to measure the degree of inequalities between individuals and "the function of corrected values" for the reduction of inequalities between individuals.

In the application part, we have proposed some examples on the PDRW process. As for the PDRC process, we applied it to the 24 municipalities of the city-province of Kinshasa to distribute among them the sum (national revenue) allocated to them in 2015.

In terms of clustering, we have proposed two distance formulas. The first calculates, by itself, the distances between two isolated individuals and between two groups of individuals, unlike the classic method which uses two different formulas. The second, for its part, simultaneously calculates the distances between several individuals and / or groups of individuals in pairs using the matrices. In this regard, we have proposed an approach called "Process of Simultaneous Calculation of Distances (PSCD)" to facilitate the use of this formula.

Keywords :Hierarchical Ascendant Clustering, Decentralized Territorial Entities, Sharing, Reduction of inequalities, Distribution of ressource.

Introduction générale

0.1. Définition des concepts de base

Dans toute société humaine, la conservation des ressources pose moins de problèmes que leur répartition entre ses membres. Ce qui fait que le besoin de la justice, de l'équité lors de la répartition des ressources n'est plus à démontrer. Une société organisée s'évertuerait à adopter des critères et des mécanismes permettant une répartition équitable de ses ressources.

Nous allons, avant toute chose, expliciter les concepts de base de ce travail. Il s'agit plus particulièrement de : ressource, partage, Classification Ascendante Hiérarchique, variables homogènes, variables hétérogènes, Réduction des inégalités.

Une ressource C est un ensemble fini d'objets (ou biens) physiques, quantités divisibles ou indivisibles finies. Elle joue un rôle central dans le problème de partage, d'où l'expression *partage de ressource* [50], ([73], pp.5,7).

Un partage d'une ressource C entre les éléments (individus) de I={1, ..., i, ..., n} est un n-uple $(C_1, \ldots, C_i, \ldots, C_n) \in \left(\mathscr{P}(C)\right)^n$, $\forall i$ et $\cup_{i=1}^{n} C_i = C$. La composante $C_i \in \mathscr{P}(C)$ est *la part* de l'individu i (une personne, un objet ou une entité) ([73],p. 16) ([73], p.16), ([77], p.63). Le concept "répartition" est utilisé en lieu et place de "partage" dans le cas où les individus eux-mêmes partagent entre eux la ressource.

La répartition d'une ressource commune ne peut pas se faire de manière arbitraire, il faut des *critères (ou variables)* qui sont des caractéristiques (intrinsèques) communes aux individus. Ils influencent le partage suivant leur nombre (un ou plusieurs) et leur type (homogène ou hétérogène). *Les variables hétérogènes* sont celles qui ont des unités de mesures différentes tandis que *les variables homogènes* en possèdent une même. Lorsqu'on considère plusieurs variables, les individus vérifient chacun autant de valeurs qu'il y a des variables. De ce fait, pour comparer deux individus, il convient d'utiliser pour chacun une seule valeur. C'est ce que nous avons appelée *valeur totale*. *La valeur totale* d'un individu est donc la somme de toutes ses valeurs vérifiées pour toutes les variables. *La somme de toutes les valeurs totales de tous les individus c'est-à-dire de toute la population constitue ce que nous avons appelé* valeur globale.

La ressource commune est répartie entre *les individus* qui en sont les bénéficiaires

suivant des critères bien définis. Dans le cas où les individus n'ont pas contribué à la création de cette ressource (cas des Entités Territoriales Décentralisées comme les communes), nous jugeons qu'il convient de tenir compte de leurs rapports de proximité et d'appartenance à une même population. Ce qui leur permettra de se solidariser et donc de procéder par la réduction des inégalités entre eux. En ce qui concerne la proximité des individus, la méthode (statistique) de classification aide à regrouper les individus dans des classes. Ceux qui sont les plus proches formeront une même classe. Elle est composée de différentes méthodes parmi lesquelles la Classification Ascendante Hiérarchique (CAH). Celle-ci est une méthode utilisée pour regrouper les individus en un certain nombre de classes ressorties à partir d'une hiérarchie de partitions [13].

Pour ce qui est des règles (ou principes) de partage, Forsé et Parodi [56] soutiennent qu'en se limitant à des sociétés modernes, trois critères de justice ressortent nettement des études déjà menées : -l'égalité absolue (elle garantit le même traitement pour tous), -l'équité (elle vise à récompenser proportionnellement des mérites individuels inégaux, -la satisfaction des besoins (de base) (elle vise à atribuer les parts suivant les besoins de chacun) [56], [62]. Nous nous intéressons au principe d'équité. Toutefois, nous jugeons que dans certains cas, notamment lorsque les individus n'ont pas contribué pour créer la ressource à partager, l'équité seule ne suffit pas. Il faut procéder par la réduction des inégalités entre les individus. Celle-ci consiste en un transfert des valeurs de certains individus (les riches) vers d'autres (les pauvres).

0.2. Problématique

La loi organique portant composition, organisation et fonctionnement des Entités Territoriales Décentralisées (ETDs) et leurs rapports avec l'Etat et les Provinces stipule que les ETDs ont droit à 40% de la part des recettes à caractère national allouées aux provinces et la répartition des ressources entre les ETDs est fonction des critères de capacité de production, de la superficie et de la population. La loi réserve la détermination du mécanisme de répartition à l'édit ([59], Art. 115, 116), ([60], Art. 226).

Il ressort de ces articles que le législateur a apporté des solutions au problème de la répartition (partage) des recettes à caractère national en proposant : (1) le mode d'acquisition de la ressource commune à partager : *la retenue à la source* au profit des provinces et *la rétrocession* au profit des ETDs ; (2) les proportions leur revenant chacun : les 60% reviennent au pouvoir central (respectivement la province) et 40% à la province (respectivement les ETDs). Plus précisément, dans ces 40% retenues par la

province, elle-même retire 60% et retrocède 40% restant à ses ETDs ; et (3) les trois critères de répartition pour les ETDs.

Toutefois, ils cachent quelques problèmes : (1) aucun critère clair permettant de différencier ou comparer le pouvoir central et une province donnée, et la province et ses ETDs n'est proposé. Les proportions de 60% et 40% ont été données de façon arbitraire. Nous laissons la solution à ce problème au législateur. (2) Pour les ETDs, *aucun mécanisme (procédé) de répartition de ces recettes utilisant les trois critères* (qui sont des variables hétérogènes) proposés par le législateur n'est donné. D'ailleurs, au lieu de trois critères, un seul , « Population », est utilisé dans la pratique en attribuant à chacun une part proportionnelle à sa population. Ce qui est injuste. (3) En outre, *aucun mécanisme prenant en compte les rapports entre les individus*, dans le sens réduire les inégalités entre eux, n'est proposé [61], [66]. Il y a donc un problème d'injustice dû à la non prise en compte du nombre et type de variables, et des rapports existant entre les individus bénéficiaires de la ressource.

De ce qui précède, prenant en compte aussi les problèmes qui se posent en Mathématiques sur le partage des ressources, nous retenons globalement quatre problèmes qui sont des motifs d'injustice dans le partage, notamment, d'une somme d'argent. Il s'agit de : (1) *L'utilisation d'une seule variable* au lieu de plusieurs (le cas où les individus vérifient ensemble plusieurs variables). Cela a pour conséquence : le déséquilibre entre les individus. (2) *L'utilisation directe des données de départ* issues *des variables homogènes pour lesquelles la même unité de mesure est exprimée différemment* pour chaque variable. Conséquence : l'influence de l'échelle d'une variable sur celles des autres. (3) *L'utilisation directe des données de départ* issues *des variables hétérogènes*. Conséquence : l'influence d'une unité de mesure sur les autres. (4) *L'absence de la réduction des inégalités* entre les individus, dans certains cas. Conséquence : l'absence de solidarité entre les individus.

Ces causes concernent d'une part les individus (comme ETDs) qui n'ont pas contribué à la création de la ressource à partager ((1), (2), (3) et (4)), et ceux (comme actionnaires d'une entreprise) ayant contribué à la création de la ressource ((2) et (3)), de l'autre. A partir de quelques exemples, nous allons ci-dessous justifier mathématiquement ces différentes causes d'injustice.

Exemple 0. 1 (Importance de considérer plusieurs variables au lieu d'une seule)**.**

Soit à partager entre deux élèves A et B une somme d'argent proportionnellement à leurs notes (points) obtenues dans trois branches (variables homogènes)

Tableau 1 – Importance de considérer plusieurs variables au lieu d'une seule

Individus	Math (50pts)	Français (50pts)	Culture générale (20pts)	Total
A	40	41	19	100
B	50	25	20	95

Source : Notre conception

Dans le cas où l'on considère une seule variable (Math ou Culture générale), B obtiendrait (50+20=70) plus que A (40+19=59) et pourtant en considérant l'ensemble des points, A vaut mieux que B. Il est donc injuste de considérer une seule variable (critère) lors du partage d'un bien entre plusieurs individus alors qu'ils vérifient ensemble un certain nombre de variables.

Exemple 0. 2 (Importance de la conversion, à une même expression, de l'unité exprimée différemment pour chaque variable homogène)**.**

Soit à partager une somme d'argent entre deux individus A et B suivant leurs poids au premier et au deuxième semestres d'une année donnée.

Tableau 2 – Même unité de mesure exprimée de différentes manières pour les variables homogènes

Individus	Poids au 1er semestre (Kg)	Poids au 2e semestre (g)	Total
A	40	50000	50040
B	50	40000	40050

Source : Notre conception

Les variables sont toutes homogènes ("Poids" aux 1er et 2e semestres) mais l'unité de mesure est exprimée différemment pour chaque variable (Kg=Kilogramme et g=gramme). Ces données ne peuvent pas être utilisées directement pour calculer les parts des individus car, dans ce cas, B bénéficierait d'une plus grande part que A vu que sa valeur totale est la plus grande. Et pourtant, si on ramenait les variables à une même expression (g), comme l'indique le tableau ci-dessous, A et B auraient une valeur totale égale qui est 90000. Comme pour dire qu'ils recevraient chacun une part égale à celle de l'autre.

Tableau 3 – Même unité de mesure exprimée de manière unique pour toutes les variables homogènes

Individus	Poids au 1er semestre (g)	Poids au 2e semestre (g)	Total
A	40000	50000	90000
B	50000	40000	90000

Source : Notre conception

Exemple 0. 3 (Importance de la transformation des variables hétérogènes)**.**

Un commerçant vend de la viande fraiche à 10000 CDF le kilo. Trois acheteurs A, B et C achètent chacun 6 kg au total pour deux commandes faites. Cette situation est explicitée dans le tableau suivant :

Tableau 4 – Quantité de viande achetée. Cas d'une seule unité de mesure

Acheteurs	1re Commande (en Kg)	2e Commande (en Kg)	Total
A	2	4	6
B	3	3	6
C	5	1	6

Source : Notre conception

Ce tableau montre que les variables sont homogènes et que tous les trois individus ont acheté la même quantité (6 kg) de viande. Dans le cas où le partage se fait sur base de ces variables, tous les trois individus auront la même valeur totale et donc la même part.

Supposons que la situation soit exprimée en Kilogramme et en CDF comme suit :

Tableau 5 – Quantité de viande achetée. Cas de deux unités de mesure

Acheteurs	1re Commande (en Kg)	2e Commande (en CDF)	Total
A	2	40000	40002
B	3	30000	30003
C	5	10000	10005

Source : Notre conception

On remarque que avec deux variables différentes (variables hétérogènes) les individus ont des valeurs totales différentes. Si ces variables sont utilisées directement, les trois individus auraient des parts complètement différentes. Ce qui est contradictoire. Il faut donc chercher à anéantir l'influence des unités de mesure sur le calcul des parts en transformant les variables (hétérogènes) de départ en variables réduites.

Dans ce cas, il convient de diviser chaque valeur du tableau 5 par l'écart-type (écart-type standard) (Cfr la formule 2.26) de la variable correspondante puis continuer.

Exemple 0. 4 (Importance de la réduction des inégalités)**.**

Considérons, le cas de deux groupes d'employés A (cadres) et B (agents d'entretien) qui doivent partager un don proportionnellement à leur salaire mensuel et leur effectif :

Tableau 6 – Importance de la réduction des inégalités

Individus	Salaire	Effectif	Total	%
A	2000000	20	2000020	90,9
B	200000	50	200050	9,1
Total	2200000	70	2200070	100

Source : Notre conception

Est-il juste d'accorder 90,9% aux cadres et 9,1% aux agents d'entretien, alors que l'obtention de cette somme d'argent n'est produit ni des salaires, ni des effectifs, ni des contributions des uns ou des autres ? D'où l'importance de la réduction des inégalités dans certains cas, notamment lorsque les individus n'ont pas contribué à la création de la ressource à partager. Ceci montre l'importance de la prise en compte des rapports existants entre les individus, en vérifiant s'ils sont contributeurs ou pas.

Nous soutenons que les individus ayant participé à la création de la ressource à partager (Exemple : Les actionnaires d'une entreprise) ont droit aux parts proportionnelles à leurs contributions respectives lors du partage. Cela est tout à fait juste. Par contre, pour des individus qui n'ont pas contribué à la création de la ressource (Exemple : Les communes qui reçoivent un don de la province), il est juste, que l'attribution des parts se fasse après réduction des inégalités entre eux car nul n'ayant contribué pour créer la ressource à partager.

Nous venons de montrer mathématiquement, à l'aide des exemples, que les causes d'injustice relevées ci-dessus sont vraies.

Exemple 0. 5 (Nécessité de la mise en place d'un mécanisme de partage selon plusieurs variables hétérogènes)**.**

Le problème de partage selon plusieurs variables homogènes ou hétérogènes (appelé aussi problème de partage proportionnel multi-critère), sans penser à la réduction des inégalités, est réel. En effet, en 2019, un internaute, prenons K, de 37 ans de niveau Master sollicite une solution à la question suivante : "1000 Euros à répartir entre A, B et C selon le capital investi dans l'entreprise et le temps passé au sein de celle-ci" [86] :

Tableau 7 – Nécessité de la mise en place d'un mécanisme de partage selon plusieurs variables hétérogènes

Individus	Capital (Euro)	Temps (mois)
A	300	6
B	200	12
C	100	8

Source : [86]

La réponse suivante lui est proposée par un autre internaute L : "dans le cas de plusieurs critères de répartition, (a) on ramène les critères de répartition à la même unité... ; (b) on calcule le produit des critères de répartition de chaque part ; (c) on effectue la répartition proportionnellement au produit des critères de répartition de chaque part". La même démarche est proposée par un autre internaute M [86] mais en parlant des critères d'ancienneté et nombre d'enfants.

En réaction contre cette démarche, la réflexion suivante fut postée par un quatrième internaute N : "Ce qui fait que quelqu'un ayant 20 ans d'ancienneté et sans enfant aura 0. Ce principe est bien absurde d'où l'impossibilité de résoudre de façon valable..."

Nous remarquons que l'internaute N a remis en question la théorie évoquée par L et M. En conséquence, la question de l'internaute K est restée sans suite favorable. Toutefois, nous proposons une solution à ce problème à l'exemple 4.4. du point 4.1.2.

Nous soutenons que pour qu'un partage soit juste il faut qu'il utilise un mécanisme adéquat qui prend en compte le nombre et le type de variables ainsi que les rapports existant entre les individus permettant ainsi de décider s'il faut ou pas réduire les inégalités entre eux.

De ce qui précède, une question mérite d'être posée, à savoir : Quelle(s) démarche(s) mettre en place pouvant permettre la répartition d'une (des) ressource (s) commune (s) entre plusieurs individus (notamment des ETDs) à partir de plusieurs variables notamment hétérogènes et prenant en compte les rapports entre les individus dans le sens à réduire ou pas les inégalités éventuelles existant entre eux puis à calculer leurs parts respectives ?

0.3. Hypothèse

Il est possible de mettre en place des démarches permettant la répartition d'une (des) ressource (s) commune (s) entre plusieurs individus utilisant plusieurs variables notamment hétérogènes et qui prendrait en compte les rapports entre les individus pouvant permettre ou pas la réduction des inégalités existant entre eux puis à calculer

leurs parts respectives.

0.4. Objectifs

0.4.1. Objectif général

L'objectif général de ce travail est de mettre en place des procédés de détermination des parts des individus d'une ou des ressources données, notamment à partir des résultats de la CAH, en utilisant plusieurs variables, prenant en compte les rapports entre les individus et permettant ou pas la réduction des inégalités, en vue de résoudre le problème d'injustice constatée lors de la répartition des ressources.

0.4.2. Objectifs spécifiques

Les objectifs spécifiques de ce travail sont les suivants : (1) présenter quelques méthodes factorielles qui forment avec la classification, les méthodes de l'Analyse des données et dont les résultats peuvent être utilisés dans la Classification ; (2) présenter la méthode de Classification Ascendante Hiérarchique (CAH) qui est une des méthodes de la Classification ; (3) déterminer les mécanismes de la répartition des ressources utilisant plusieurs variables notamment hétérogènes, prenant en compte les rapports entre les individus et permettant ou pas la réduction des inégalités entre eux ; (4) appliquer les mécanismes proposés aux données notamment des 24 communes de la Ville-province de Kinshasa.

0.5. Résultats des autres auteurs sur le problème de partage

En ce qui concerne l'injustice dans la répartition des ressources, Jean Salem Kapya [61] et Paulin Punga [66] ont remarqué l'injustice dans la répartition des recettes à caractère national suite à l'absence d'un mécanisme adéquat de correction des inégalités et l'utilisation d'une seule variable. Ils ont relevé des problèmes sans en proposer des solutions.

Quant à ce qui est des principes de justice, Forsé Michel et Parodi Maxime [56] font allusion aux trois principes de justice : -l'égalité absolue (ou principe d'égalité), -l'équité (ou principe de mérite), -la satisfaction des besoins (au moins ceux de base) (ou principe de besoin). Ces auteurs se sont intéressés aux principes de justice mais ne se sont pas préoccupés du nombre (un ou plusieurs) et de la nature (homogène ou hétérogène) des variables, ainsi qu'aux rapports entre les individus (leur rapprochement, la nature de leur société d'appartenance) qui peuvent modifier un principe ou un autre.

A propos de la Classification, après avoir réalisé une Classification Ascendante

Hiérarchique (CAH), il faut interpréter ses résultats. Différents auteurs notamment Chesneau [13], Husson et Josse [23] limitent cette interprétation au calcul des paramètres des classes issues de la CAH, à la détermination des individus les plus typiques (parangons et extrêmes). Toutefois, rien n'est proposé dans le sens à utiliser les résultats de la CAH pour déterminer les parts d'une ressource revenant aux individus, en passant par la réduction des inégalités entre eux.

0.6. Méthodologie utilisée

Dans la partie théorique, notre méthodologie consiste d'abord à présenter quelques méthodes de l'Analyse factorielle qui sont l'Analyse en Composantes Principales (ACP), l'Analyse Factorielle des Correspondances (AFC) et son extension l'Analyse Factorielle des Correspondances Multiples (AFCM) dont les résultats peuvent être utilisés pour classifier les individus ; ensuite, à présenter la CAH où il faudra tenir compte de type de variables : homogènes ou hétérogènes qui peut être réalisée à partir des résultats de l'ACP et enfin à présenter des démarches permettant la répartition des ressources, utilisant plusieurs variables et, au besoin, les résultats de la classification et la réduction des inégalités. Ces résultats seront interprétés.

Dans la partie pratique, notre méthodologie consiste au recueil des données plus particulièrement de 24 communes (ETDs) de la Ville/Province de Kinshasa, exercice 2015 relatives aux trois variables qui sont Superficie, production et population ; au traitement et à l'analyse de ces données où seront utilisées les démarches permettant la répartition des ressources notamment à partir des résultats de la classification et aboutissant aux parts des individus après réduction des inégalités et à leur interprétation.

0.7. Principaux résultats

En vue de résoudre les problèmes d'injustice relevés par Jean Salem Kapya et Paulin Punga suite à l'utilisation d'une seule variable et au manque de mécanisme adéquat de réduction des inégalités ainsi que ceux qui se posent en Mathématiques sur le partage, nous avons principalement proposé deux mécanismes utilisant plusieurs variables homogènes ou hétérogènes et prenant en compte les rapports entre les individus permettant ainsi de réduire ou pas les inégalités entre eux :

(-) l'un nommé Procédé de la Répartition des Ressources Sans réduction des inégalités (PRRS) adapté pour le cas où les individus (comme des actionnaires d'une entreprise) ont contribué à la création de la ressource commune à partager et n'autorise pas la réduction des inégalités. Quelques applications ont été proposées à cet effet.

(-) L'autre, quant à lui, est nommé Procédé de la Répartition des Ressources à partir des résultats de la Classification (PRRC) adapté pour le cas où les individus n'ont pas contribué à la création de la ressource. Il admet la réduction des inégalités et fait appel à la notion de classification permettant de retrouver les individus les plus proches qui devront se solidariser dans leurs classes respectives du fait de leur proximité et par la suite avec tous les autres du fait qu'ils appartiennent à une même population. Nous avons appliqué ce procédé aux données de 24 communes de la Ville-Province de Kinshasa.

Quant à ce qui est des principes de justice, nous avons mis l'accent sur plusieurs variables notamment hétérogènes ainsi que sur les rapports entre les individus en différenciant le cas des contributeurs (cas des actionnaires) où le principe d'équité doit s'appliquer, en attribuant aux individus des parts proportionnelles à leurs contributions respectives, de celui des non contributeurs (cas des entités étatiques) où l'équité de manière brute n'est pas adaptée. A ce propos, nous avons opté pour l'équité réduite (ou corrigée) multidimensionnelle selon lequel, le partage est fait suivant plusieurs variables et les parts sont attribuées proportionnellement aux valeurs des individus mais après réduction des inégalités entre eux.

En ce qui concerne la classification (plus particulièrement CAH), nous avons continué l'interprétation de ses résultats jusqu'à la détermination des parts respectives des individus. Aussi, au niveau de calcul des distances entre individus ou groupes d'individus, nous avons proposé deux formules : La première calcule, elle seule, les distances entre deux individus isolés et entre deux groupes d'individus, contrairement au procédé classique qui utilise deux formules différentes. La deuxième, quant à elle, calcule simultanément les distances entre plusieurs individus et/ou groupes d'individus deux à deux en utilisant les matrices. A ce propos, nous avons proposé le « Procédé de Calcul Simultané de Distances (PCSD) » pour faciliter l'utilisation de cette formule.

0.8. Division du travail

Hormis l'introduction générale et, la conclusion générale et les perspectives, nous allons d'abord parcourir quelques méthodes factorielles, plus particulièrement l'ACP et l'AFC. Ce qui permettra de réaliser combien elles ne sont pas adaptées pour classifier les individus, bien qu'elles puissent être utilisées avant de procéder par une classification (Chapitre 1).

Par la suite, nous allons aborder la méthode de Classification Ascendante Hiérarchique (CAH). Nous en présenterons les généralités, les étapes, la démarche conduisant

aux résultats qui seront interprétés (Chapitre 2).

Ensuite, nous allons présenter nos deux procédés de la répartition des ressources : l'un s'occupant du cas où les individus ont contribué pour créer la ressource à partager et l'autre de celui où les individus n'ont pas contribué pour créer la ressource à partager permettant ainsi la réduction des inégalités entre eux. Aussi nous ferons allusion aux deux théorèmes que nous avons proposés reprenant deux formules dont l'une permet le calcul de distances entre les individus deux à deux et entre deux groupes d'individus, et l'autre le calcul simultané de plusieurs distances entre plusieurs individus et/ou groupes d'individus. Au niveau de l'interprétation d'une partition, nous ajouterons un nouvel élément, celui de calcul des proportions et parts des individus (Chapitre 3).

Enfin, nous appliquerons toutes ces théories notamment aux données des 24 communes de la ville-province de Kinshasa en vue de la répartition entre celles-ci des ressources budgétaires (recettes à caractère national) leurs alloués (Chapitre 4).

Chapitre 1

Quelques méthodes factorielles

Introduction

[1], [10], [15], [17], ([24], p.2), [39], [40], ([76], p.34), [84].

Une étude statistique portant sur plusieurs variables amène à procéder par les méthodes d'Analyse des données qui comprend notamment la méthode de l'Analyse Factorielle et celle de la Classification.

Considérons la méthode de l'Analyse Factorielle. Elle comprend principalement les méthodes suivantes : 1) L'Analyse en Composantes Principales (ACP) (1.1.) qui utilise les données quantitatives ; 2) l'Analyse Factorielle de Correspondance (AFC) (1.2.) qui porte sur les données de deux variables qualitatives et l'Analyse en Composantes Multiples (ACM) appelée aussi Analyse Factorielle des Correspondances Multiples (AFCM) qui est une généralisation d'une AFC. Elle utilise plusieurs variables qualitatives. Ces trois méthodes seront développées dans la suite et suivies d'une conclusion (1.3.)

1.1. Analyse en Composantes Principales

Définition 1. 1 (Analyse en Composantes Principales).

L'Analyse en Composantes Principales (ACP) est une des méthodes de la Statistique multidimensionnelle, classée parmi les méthodes factorielles et qui a pour objectif de réduire le nombre m de variables en nombre q de composantes principales afin d'aboutir à une dimension réduite $(q < m)$ *de l'espace multidimensionnel qui préserve les distances entre les individus. Ce qui permet de représenter graphiquement les individus et les variables sur un espace à q dimensions afin d'étudier la liaison entre les variables quantitatives([2], pp.5-6), ([24], p.3), ([36], p.4), [70].*

L'ACP utilise les variables quantitatives et la distance euclidienne entre les individus. Elle se réalise en suivant les étapes ci-après : *(1) Construire la matrice de données* (1.1.1) ; *(2) Transformer les données de départ* en données centrées ou centrées réduites (1.1.2) ; *(3) Déterminer : soit la matrice des variances-covariances soit la matrice*

des corrélations (1.1.3) ; *(4) Diagonaliser, suivant le modèle choisi*, l'une de ces deux matrices (1.1.4) ; *(5) Présenter les résultats* (1.1.5).

1.1.1. Matrice des données

Soit m variables $X_1, \ldots, X_j, \ldots, X_m$ observées sur un ensemble Γ de n individus $1, \ldots, i, \ldots, n$. Etant donnée X_{ij}, la valeur vérifiée par un individu i pour la variable X_j, alors *la matrice des données* se présente comme suit :

$$M = \begin{pmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & X_{ij} & \vdots \\ X_{nm} & \cdots & X_{nm} \end{pmatrix} = (X_{ij})_{1 \leq i \leq n, 1 \leq j \leq m} \tag{1.1}$$

1.1.2. Transformation des données

La transformation des données consiste à centrer et/ou à réduire ces dernières. Elles sont centrées pour amener l'origine des axes au centre de gravité du nuage des individus (Cas des variables homogènes ou hétérogènes) tandis qu'elles sont réduites afin d'anéantir l'influence des unités de mesure (Cas des variables hétérogènes). Les variables hétérogènes sont celles qui sont exprimées sur des unités de mesures différentes. Dans le cas contraire, elles sont dites homogènes ([3], pp. 12, 14, 16), ([30], pp.57, 75-76), ([55], p.43).

A la lumière de ce qui précède, on a :

1) La matrice centrée

$$\widehat{M} = (\widehat{X}_{ij})_{1 \leq i \leq n, 1 \leq j \leq m} \tag{1.2}$$

$$\text{où} \quad \widehat{X}_{ij} = X_{ij} - \overline{X}_j \tag{1.3}$$

$$\text{et} \quad \overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij} \tag{1.4}$$

sont respectivement la variable centrée et la moyenne arithmétique (simple) de la variable X_j.

2) La matrice réduite :

$$T = (t_{ij})_{1 \leq i \leq n, 1 \leq j \leq m} \tag{1.5}$$

$$\text{où} \quad t_{ij} = \frac{X_{ij}}{\sigma_j} \tag{1.6}$$

est la variable réduite.

3) La matrice centrée-réduite :

$$\widetilde{M} = (\widetilde{X}_{ij})_{1 \leq i \leq n, 1 \leq j \leq m} \tag{1.7}$$

$$\text{où} \quad \widetilde{X}_{ij} = \frac{X_{ij} - \overline{X}_j}{\sigma_j} \tag{1.8}$$

$$\text{et} \quad \sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2} \tag{1.9}$$

sont respectivement la variable centrée-réduite et l'écart-type de la variable X_j.

1.1.3. Matrice des variances-covariances et matrice des corrélations

1.1.3.1. Matrice des variances-covariances

([3], p.19), ([16], p.148), ([36], p.12), ([72], p.6)

Définition 1. 2 (Matrice des variances-covariances).

Une matrice des variances-covariances des variables centrées $\widehat{X}_1, \ldots, \widehat{X}_j, \ldots, \widehat{X}_m$ est une matrice symétrique d'ordre m où les termes de la diagonale principale sont les variances tandis que les autres sont les différentes covariances.

La matrice des variances-covariances est calculée en utilisant la formule :

$$\widehat{V} = \frac{1}{n} \widehat{M}^t \widehat{M} = \widehat{M}^t \acute{D} \widehat{M} \tag{1.10}$$

Où $\widehat{M}$ et $\widehat{M}^t$ sont respectivement la matrice centrée et sa transposée, $\acute{D} = diag(\frac{1}{n}, \ldots, \frac{1}{n})$ est la matrice diagonale d'ordre égal à celui de $\widehat{M}$ et de terme de la diagonale principale $\frac{1}{n}$.

La variance de $\widehat{X}_j$ et la covariance de $\widehat{X}_k$ et $\widehat{X}_j$ sont respectivement données par :

$$Var\widehat{X}_j = \sigma^2_{\widehat{X}_j} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{X}_{ij} - \overline{\widehat{X}_j})^2 \tag{1.11}$$

$$Cov(\widehat{X}_j, \widehat{X}_k) = \sigma^2_{\widehat{X}_j \widehat{X}_k} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{X}_{ij} - \overline{\widehat{X}_j})(\widehat{X}_{ik} - \overline{\widehat{X}_k}) \tag{1.12}$$

Plus la variance est grande plus les données $\widehat{X}_j$ sont dispersées.

1.1.3.2. Matrice des corrélations

Définition 1. 3 (Matrice des corrélations).

Une matrice des corrélations des variables centrées-réduites $\widetilde{X}_1, \ldots, \widetilde{X}_j, \ldots, \widetilde{X}_m$ est une matrice symétrique d'ordre m où les termes de la diagonale principale sont l'unité tandis que les autres sont les différents coefficients des corrélations entre ces variables deux à deux ([72], p.7).

Elle est donnée par la formule :

$$\widetilde{R} = \frac{1}{n}\widetilde{M}^t\widetilde{M} = \widetilde{M}^t\acute{D}\widetilde{M} \tag{1.13}$$

Où $\widetilde{M}$ et $\widetilde{M}^t$ sont respectivement la matrice centrée réduite et sa transposée.

Le coefficient de corrélation de deux variables centrées-réduites $\widetilde{X}_k$ et $\widetilde{X}_j$ est :

$$R_{\widetilde{X}_j\widetilde{X}_k} = \frac{Cov(\widetilde{X}_j, \widetilde{X}_k)}{\sqrt{Var\widetilde{X}_j Var\widetilde{X}_k}} \tag{1.14}$$

Le coefficient de corrélation de deux variables $\widetilde{X}_j$ et $\widetilde{X}_k$ égale le cosinus de l'angle formé par ces variables : $R_{\widetilde{X}_j\widetilde{X}_k} = cos(\widetilde{X}_j, \widetilde{X}_k)$

Théorème 1. 1 (Propriétés du coefficient de corrélation).

Le coefficient de corrélation vérifie les propriétés suivantes :

1) $R_{\widetilde{X}_j\widetilde{X}_k} \in [-1,\ 1]$

2) $\widetilde{X}_j = \widetilde{X}_k \Rightarrow R_{\widetilde{X}_j\widetilde{X}_k} = R_{\widetilde{X}_j\widetilde{X}_j} = R_{\widetilde{X}_k\widetilde{X}_k} = 1$

3) $R_{\widetilde{X}_j\widetilde{X}_k} = R_{\widetilde{X}_k\widetilde{X}_j}$ *(Le coefficient de corrélation est symétrique)*

Démonstration. 1) Montrons que $R_{\widetilde{X}_j\widetilde{X}_k} \in [-1,\ 1]$

On sait que $R_{\widetilde{X}_j\widetilde{X}_k} = cos(\widetilde{X}_j, \widetilde{X}_k)$.

Or $cos(\widetilde{X}_j, \widetilde{X}_k) \in [-1,\ 1]$. Donc $R_{\widetilde{X}_j\widetilde{X}_k} \in [-1,\ 1]$

2) Montrons que $\widetilde{X}_j = \widetilde{X}_k \Rightarrow R_{\widetilde{X}_j\widetilde{X}_k} = R_{\widetilde{X}_j\widetilde{X}_j} = R_{\widetilde{X}_k\widetilde{X}_k} = 1$

On sait que $R_{\widetilde{X}_j\widetilde{X}_k} = \frac{Cov(\widetilde{X}_j,\widetilde{X}_k)}{\sqrt{Var\widetilde{X}_j Var\widetilde{X}_k}}$

En prenant $\widetilde{X}_j = \widetilde{X}_k$,

on a : $R_{\widetilde{X}_j\widetilde{X}_k} = R_{\widetilde{X}_j\widetilde{X}_j} = \frac{Cov(\widetilde{X}_j,\widetilde{X}_j)}{\sqrt{Var\widetilde{X}_j Var\widetilde{X}_j}} = \frac{Var\widetilde{X}_j}{\sqrt{(Var\widetilde{X}_j)^2}} = \frac{Var\widetilde{X}_j}{Var\widetilde{X}_j} = 1$

3) Prouvons que $R_{\widetilde{X}_j\widetilde{X}_k} = R_{\widetilde{X}_k\widetilde{X}_j}$

En effet, $R_{\widetilde{X}_j\widetilde{X}_k} = \frac{Cov(\widetilde{X}_j,\widetilde{X}_k)}{\sqrt{Var\widetilde{X}_j Var\widetilde{X}_k}} = \frac{Cov(\widetilde{X}_k,\widetilde{X}_j)}{\sqrt{Var\widetilde{X}_k Var\widetilde{X}_j}} = R_{\widetilde{X}_k\widetilde{X}_j}$ □

1.1.4. Diagonalisation de la matrice des variances-covariances (ou de la matrice des corrélations) et détermination des facteurs

Les matrices $\widehat{V}$ des variances- covariances et $\widetilde{R}$ des corrélations étant symétriques, elles sont diagonalisables. Dans la suite, nous allons nous servir de la matrice des variances-covariances. On procédera de manière analogue pour le cas de la matrice des corrélations ([3], pp. 5-9).

1.1.4.1. Détermination des valeurs propres

Pour déterminer les m valeurs propres $\lambda_1, \ldots, \lambda_k, \ldots, \lambda_m$ de la matrice $\widehat{V}$, on résout *l'équation caractéristique* de $\widehat{V}$:

$$Det(\widehat{V} - \lambda \acute{I}) = 0 \tag{1.15}$$

où λ est l'inconnue représentant les différentes valeurs propres et $\acute{I}$ la matrice unité.

Les termes $\sigma^2_{\widehat{X}_1}, \ldots, \sigma^2_{\widehat{X}_j}, \ldots, \sigma^2_{\widehat{X}_m}$ de la matrice des variances-covariances sont les variances empiriques tandis que les m valeurs propres $\lambda_1, \ldots, \lambda_j, \ldots, \lambda_m$ sont les variances expliquées.

1.1.4.2. Détermination des vecteurs propres et construction de la matrice de passage

Les m vecteurs propres $v_{k(1\leq k\leq m)}$ de $\widehat{V}$ associés aux m valeurs propres $\lambda_{k(1\leq k\leq m)}$ sont déterminés en résolvant le système d'équations en $v = (v_1,\ldots, v_k,\ldots, v_m)$ suivant :

$$(\widehat{V} - \lambda \acute{I})v = 0 \tag{1.16}$$

Dans ces équations, on remplace λ par λ_1 puis après résolution, on détermine le 1er vecteur propre $v_1 = (v_{1j})_{1\leq j\leq m}$. On fait de même pour le reste de valeurs propres ([36], pp. 15-16).

Par la suite, on construit la matrice de passage Q, inversible en plaçant les unes à côté des autres, les colonnes vecteurs propres formant une base de chacun des sous-espaces v_k :

$$Q = (v_{kj})_{1\leq k\leq m, 1\leq j\leq m} \tag{1.17}$$

Les matrices Q et $\widehat{V}$ sont semblables, et vérifient l'égalité :

$$D = Q^{-1}\widehat{V}Q \tag{1.18}$$

D est la matrice diagonale dont les valeurs sont les différentes valeurs propres de la matrice $\widehat{V}$.

1.1.4.3. Détermination des composantes principales et leurs contributions à la variance totale

1.1.4.3.1. Détermination des composantes principales

Aux valeurs propres λ_1, ..., λ_j, ..., λ_m sont associées m nouvelles variables notées F_1, F_2, ..., F_m appelées Composantes principales qui sont optimales. Chaque valeur propre représente la variance du facteur correspondant. La 1ère composante principale F_1 doit récupérer le maximum d'inertie du tableau des données. Quant à la 2e composante F_2, elle traitera de l'inertie (résiduelle) non expliquée par F_1. Elle doit avoir une corrélation linéaire nulle (orthogonalité) avec la première. On détermine les autres composantes de la même manière ([36], pp. 8-9).

Les composantes principales sont des combinaisons linéaires des variables originelles qui sont centrées (ou centrées-réduites), tout dépend du modèle choisi. Ceci donne l'idée d'un changement de base ou changement de repère. On a :

$$(F_k)_{1\leq k\leq m} = (v_{kj})_{1\leq k\leq m, 1\leq j\leq m}.(\widehat{X}_j)_{1\leq k\leq m} \tag{1.19}$$

La k^e composante principale associée à la valeur propre λ_k est :

$$F_k = \widehat{X}_1 v_{k1} + \ldots + \widehat{X}_j v_{kj} + \ldots + \widehat{X}_m v_{km} = \sum_{j=1}^{m} \widehat{X}_j v_{kj} \tag{1.20}$$

1.1.4.3.2. Contribution des k premiers facteurs à la variance totale

La contribution des k ($k \leq m$) premiers facteurs dans l'inertie totale, se calcule par la formule suivante :

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_k + ... + \lambda_m} \tag{1.21}$$

En multipliant cette expression par 100, on trouve le résultat en pourcentage. Si on prend tous les « m » facteurs possibles, cela permettra de récupérer toute l'information disponible contenue dans le tableau des données de départ auquel cas la contribution égale 1.

En particulier, la contribution du k^e facteur dans l'inertie totale est donnée par :

$$\frac{\lambda_k}{\lambda_1 + \ldots + \lambda_k + ... + \lambda_m} = \frac{\lambda_k}{\sum_{j=1}^{m} \lambda_k} = \frac{\lambda_k}{tr(D)} \tag{1.22}$$

L'ACP est réalisable si deux ou au maximum trois composantes principales restituent à eux seuls la quasi-totalité de l'inertie du tableau de données c'est-à-dire qu'elles expliquent près de 70% de la variance totale. Ces facteurs seront retenus tandis que les autres pourront être négligés lors de la représentation graphique.

1.1.5. Présentation des résultats et interprétation

1.1.5.1. Analyse du nuage des variables

1.1.5.1.1. Corrélation de deux variables

La liaison entre deux variables quantitatives $\widehat{X}_j$ et $\widehat{X}_{j'}$ se mesure au moyen de leur coefficient de corrélation $R_{jj'} \in [-1, 1]$. Plus il est proche des extrémités de l'intervalle, plus les variables sont corrélées (ou liées). Elles sont corrélées positivement si $R_{jj'} > 0$ et négativement si $R_{jj'} < 0$. Si $R_{jj'} = 1$ (respectivement -1), il existe une relation linéaire positive (respectivement négative) entre les variables : $\widehat{X}_j$ =$a\widehat{X}_{j'} + b$ où a et b sont des réels. Si $R_{jj'} = 0$, la corrélation est nulle ([16], pp. 150-151), ([36], pp. 26-27).

Théorème 1. 2 (Corrélation des facteurs avec les variables de départ).

Les corrélations respectives de la composante principale F_k avec la variable originelle (ou ancienne) centrée $\widehat{X}_j$ et avec la variable centrée-réduite $\widetilde{X}_j$) sont respectivement données par les formules suivantes :

$$R(\widehat{X}_j, F_k) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma^2_{\widehat{X}_j}}}.v_{kj} \tag{1.23}$$

$$R(\widetilde{X}_j, F_k) = \sqrt{\lambda_k}.v_{kj} \tag{1.24}$$

Démonstration. $\star$ **La covariance** de $\widehat{X}_j$ et F_k est

$$Cov(\widehat{X}_j, F_k) = \sigma^2_{\widehat{X}_j F_k} = \frac{(\widehat{X}_j - \overline{\widehat{X}_j})^t(\widehat{X}_j - \overline{F}_k)}{n} = \frac{(\widehat{X}_j - \overline{\widehat{X}_j})^t(\widehat{X}_j - \overline{\widehat{X}_j})}{n}$$

$$= \sigma^2_{\widehat{X}_j}.Q = I.\sigma^2_{\widehat{X}_j}.Q = Q.Q^{-1}.\sigma^2_{\widehat{X}_j}.Q = Q.(Q^{-1}.\sigma^2_{\widehat{X}_j}.Q) = Q.D = v_{kj}\lambda_k$$

⋆ **Le coefficient de corrélation de $\widehat{X}_j$ et F_k** est

- Pour le modèle centré :

$$R_{\widehat{X}_j F_k}=\frac{Cov(\widehat{X}_j,F_k)}{\sigma_{\widehat{X}_j}.\sigma_{F_k}}=\frac{\lambda_k.v_{kj}}{\sqrt{\sigma^2_{\widehat{X}_j}}.\sqrt{\sigma^2_{F_k}}}=\frac{\lambda_k.v_{kj}}{\sqrt{\sigma^2_{\widehat{X}_j}}.\sqrt{\lambda_k}}=\frac{\sqrt{\lambda_k}.v_{kj}}{\sqrt{\sigma^2_{\widehat{X}_j}}} \qquad (1)$$

- Pour modèle centré-réduit :

$\sigma^2_{\widehat{X}_j}=1$,
alors (1) devient :
$R_{\widetilde{X}_j F_k}=\sqrt{\lambda_k}.v_{kj}$ (Formule (1.24)).
Il en ressort

$$\sum_{j=1}^{m} R_{\widetilde{X}_j F_k}=\lambda_k \qquad (1.25)$$

□

1.1.5.1.2. Coordonnées des variables

La variable centrée $\widehat{X}_j$ est représentée dans le plan factoriel (F_1, F_2) par le point suivant de coordonnées $F_1(\widehat{X}_j), F_2(\widehat{X}_j)$.

$$\widehat{X}_j=(F_1(\widehat{X}_j),F_2(\widehat{X}_j))=\left(\frac{\sqrt{\lambda_1}}{\sqrt{\sigma^2_{\widehat{X}_j}}}.v_{1j},\frac{\sqrt{\lambda_2}}{\sqrt{\sigma^2_{\widehat{X}_j}}}.v_{2j}\right)=(R_{\widehat{X}_j F_1},R_{\widehat{X}_j,F_2}) \qquad (1.26)$$

La variable $\widetilde{X}_j$ centrée-réduite est représentée dans le plan factoriel par le point :

$$\widetilde{X_j}=(F_1(\widetilde{X}_j),F_2(\widetilde{X}_j))=\left(\sqrt{\lambda_1}.v_{1j},\sqrt{\lambda_2}.v_{2j}\right)=(R_{\widetilde{X}_j F_1},R_{\widetilde{X}_j,F_2}) \qquad (1.27)$$

1.1.5.1.3. Représentation graphique des variables

On représente graphiquement la corrélation des anciennes variables avec les deux (ou trois) facteurs en utilisant les points-variables explicités ci-dessus. Ce qui produit un graphique appelée *cercle des corrélations*, de centre égal à l'origine des axes et de rayon égal à l'unité.

Exemple 1. 1 (Représentation graphique d'une variable dans un cercle des corrélations).

La variable $\widehat{X}_j$ est représentée de la manière suivante :

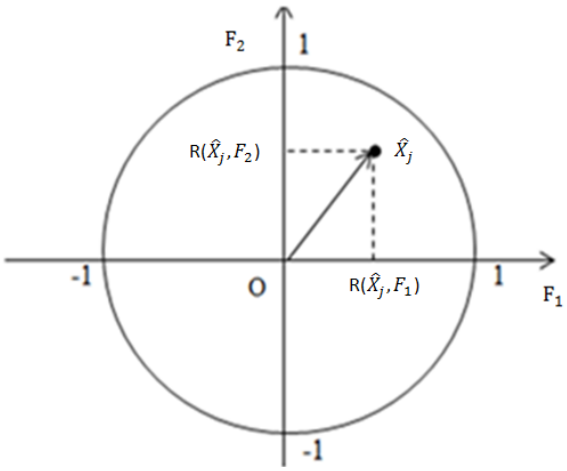

Figure 1.1 – Cercle des corrélations (Exemple)
Source :([16], p.151)

1.1.5.1.4. Interprétation des axes en fonction des anciennes variables

1) Interprétation des facteurs en fonction des variables originelles

A ce niveau, sur base du cercle des corrélations, si la valeur de la variable sur l'axe est positive et que la variable est proche de l'axe, l'axe est corrélé positivement avec cette variable. Si par contre, elle est négative et la variable proche de l'axe, l'axe est corrélé négativement avec la variable.

2) Interprétation des positions des variables originelles entre elles

La corrélation de deux variables centrées $\widehat{X}_j$ et $\widehat{X}_k$ (respectivement centrées-réduites) est le cosinus de l'angle formé par leurs vecteurs :

$$R_{\widehat{X}_j\widehat{X}_k} = Cos(O\widehat{X}_j, O\widehat{X}_k) \tag{1.28}$$

Si l'angle $(O\widehat{X}_j, O\widehat{X}_{j'}) = arccos(R_{\widehat{X}_j\widehat{X}_{j'}})$ (arc cosinus est la réciproque de la fonction cosinus) égale 0^0, la corrélation est de 1, les variables sont très corrélés positivement. S'il est égal à 180^0, la corrélation est de -1, les variables sont très corrélés négativement. Si, enfin, il est égal à 90^0, la corrélation est nulle auquel cas les variables ne sont pas corrélées.

1.1.5.2. Analyse du nuage des individus

1.1.5.2.1. Coordonnées d'un individu sur les axes

Les coordonnées des individus sur les axes sont données par la matrice $\widehat{L}$, produit de la matrice centrée $\widehat{M}$ et de la matrice de passage Q ([3], pp. 29-35), ([16], pp. 156-157) :

$$\widehat{L} = \widehat{M}.Q \tag{1.29}$$

La coordonnée d'un individu i sur un axeF_k est :

$$F_k(i) = v_{ik}\widehat{X}_{i1} + \ldots + v_{ik}\widehat{X}_{ij} + \ldots + v_{ik}\widehat{X}_{im} = \sum_{j=1}^{m} \widehat{X}_{ij} v_{ik} \tag{1.30}$$

L'individu i est représenté sur le plan factoriel par le point suivant :

- Modèle centré

$$i = (F_1(X_i), F_2(X_i)) = \Big(\sum_{j=1}^{m} \widehat{X}_{ij} v_{i1}, \sum_{j=1}^{m} \widehat{X}_{ij} v_{i2}\Big) \tag{1.31}$$

- Modèle centré-réduit

$$i = \Big(\sum_{j=1}^{m} \widetilde{X}_{ij} v_{i1}, \sum_{j=1}^{m} \widetilde{X}_{ij} v_{i2}\Big) \tag{1.32}$$

1.1.5.2.2. Proximité entre deux individus

La proximité ou la ressemblance entre deux individus se mesure à l'aide de la distance euclidienne ([3], pp. 29-35), ([16], pp. 156-157).

1) Distance entre deux individus

La distance euclidienne (au carré) entre deux individus i et i' est donnée par :

$$d^2(i, i') = \sum_{k=1}^{2} \big(F_k(i) - F_k(i')\big)^2 \tag{1.33}$$

Plus la distance entre deux individus est petite, plus ils se ressemblent, plus ils sont proches

2) Distance des individus au centre de gravité G_Γ

$$d^2(i, G_\Gamma) = \sum_{j=1}^{m} (\widehat{X}_{ij} - G_\Gamma)^2 = \sum_{j=1}^{m} (\|\widehat{X}_{ij} - G_\Gamma\|)^2 \tag{1.34}$$

3) Inertie du nuage des individus

Définition 1. 4 (Inertie du nuage des points).

L'inertie $I_{\widehat{X}}$ du nuage $\Gamma_{\widehat{X}}$ des points-individus $\widehat{x}_1, \ldots, \widehat{x}_i, \ldots, \widehat{x}_n$ relatifs aux individus $1, \ldots, i, \ldots, n$ est la mesure de leur dispersion dans l'espace à m dimensions autour du barycentre du nuage.

Elle traduit la quantité d'information disponible dans le tableau des données. Plus elle est élevée, plus le nuage est dispersé autour de son barycentre. Lorsque les variables sont centrées-réduites, l'inertie $I_{\widehat{X}} = m$ (Nombre de variables).

$$I_{\widehat{X}} = \frac{1}{n}\sum_{i=1}^{n} d^2(i, G_\Gamma) = tr(D) = \sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{m} \sigma^2_{\widehat{X}_j} \tag{1.35}$$

1.1.5.2.3. Représentation graphique des individus sur les nouveaux axes

La représentation graphique des individus se fait en utilisant les coordonnées des points telles qu'indiquées précédemment suivant le modèle choisi.

Exemple 1. 2 (Graphique de nuage de dix individus).

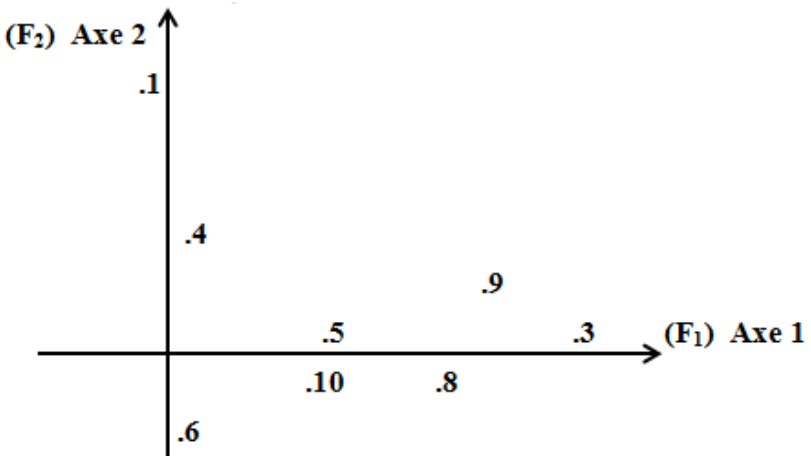

Figure 1.2 – Représentation graphique des individus sur les axes factoriels (Exemple).
Source : ([16], p.150)

Ces deux axes portent le maximum d'inertie du nuage des points. Celle-ci est portée essentiellement par l'axe 1 tandis que l'axe 2 qui lui est perpendiculaire n'en porte qu'une faible partie non prise en compte par le premier.

1.1.5.2.4. Interprétation des résultats des individus

Après la représentation graphique des individus sur le plan factoriel, on vérifie si les individus sont proches du barycentre auquel cas ils sont bien représentés. Sinon, ils sont mal représentés. Les individus similaires sont regroupés ensemble.

1.1.5.3. Aides à l'interprétation

En parlant des aides à l'interprétation, l'allusion est faite à la qualité de représentation d'un élément (individu ou variable) sur un axe ou un plan et à la contribution d'un élément à l'inertie d'un axe ([36], p. 30).

1.1.5.3.1. Qualité de la représentation d'un élément sur un axe

1) Cosinus au carré (Qualité de la représentation d'un élément sur un axe)

La qualité de la représentation (ou de la projection) d'un élément X_i sur un axe F_k, notée QLT, $0 \leq QLT \leq 1$, est une quantité qui égale le carré de cosinus de l'angle formé par la droite Gi (passant par le centre de gravité G et l'élément X_i) et l'axe F_k. Elle permet de vérifier si l'élément est plus proche de l'axe. Certains auteurs parlent de la contribution relative d'un élément sur un axe et la notent CR ou encore CTR ([87], p. 17).

$$QLT(X_i, F_k) = cos^2(Gi, F_k) = \frac{F_1^2(X_i)}{d^2(G, X_i)} = \frac{F_k^2(X_i)}{\sum_{k=1}^{m} F_k^2(X_i)} \tag{1.36}$$

Si $Cos^2(Gi, F_k) = 1$, le point se trouve sur l'axe, l'élément est parfaitement représenté par sa projection sur l'axe. Si $Cos^2(Gi, F_k)$ est proche de 1, l'élément est bien représenté par sa projection sur l'axe. Si $Cos^2(Gi, F_k) = 0$, le point se trouve dans une direction orthogonale à l'axe. Il est mal représenté par sa projection sur axe.

Si deux éléments sont bien représentés en projection sur un axe (ou un plan) et ont des projections proches, alors on pourra dire qu'ils sont aussi proches dans l'espace.

QLT peut être exprimée en $\%$ et $\sum_{i=1}^{n} Cos^2(Gi, F_k) = 1$.

2. Qualité de représentation d'un élément sur un plan

La qualité de représentation d'un élément X_i sur un plan (F_1, F_2) est mesurée par :

$$QLT(X_i, (F_1, F_2)) = cos^2(Gi, F_1) + cos^2(Gi, F_2) \tag{1.37}$$

1.1.5.3.2. Contributions d'un élément à la construction des axes

1) Inertie d'un élément sur un axe

L'inertie d'un élément X_i sur l'axe F_k est la quantité

$$\lambda_k = \frac{1}{n}\sum_{i=1}^{n} F_k^2(X_i) \Longleftrightarrow \sum_{i=1}^{n} F_k^2(X_i) = n\lambda_k \tag{1.38}$$

2) Contribution d'un élément à un axe

Par définition, la contribution d'un élément X_i à l'axe F_k est donnée par

$$Ctr(X_i, F_k) = \frac{F_k^2(X_i)}{\sum_{i=1}^{n} F_k^2(X_i)} = \frac{F_k^2(X_i)}{n\lambda_k} \tag{1.39}$$

En particulier, la contribution de X_i à la composante principale F_1 est :

$$Ctr(X_i, F_1) = \frac{F_1^2(X_i)}{n\lambda_1} \tag{1.40}$$

Remarque : $0 \leq Ctr \leq 1$ et $\sum_{i=1}^{n} Ctr(X_i, F_k) = 1, \forall k$

3) Contribution d'un élément à un plan

$$Ctr(X_i, (F_k, F_{k'})) = \frac{F_k^2(X_i)}{n\lambda_k} + \frac{F_{k'}^2(X_i)}{n\lambda_{k'}} = \frac{1}{n}.\frac{F_k^2(X_i) + F_{k'}^2(X_i)}{\lambda_k + \lambda_{k'}} \tag{1.41}$$

1.1.5.4. Points supplémentaires ou inactifs

Il peut arriver qu'un ou plusieurs points (individus et/ou variables) se situent en dehors du graphique ou éloignés de tous les autres c'est-à-dire qu'ils possèdent des caractéristiques spécifiques par rapport à tous les autres lorsqu'on considère le tableau de départ (centré ou centré-réduit). Leurs positions dans le plan factoriel seront donc isolées et empêcheront une étude précise des proximités des autres points projetés [81].

De ce fait, il convient de rendre ce ou ces points inactifs (les mettre en supplémentaire). Ce qui revient à réaliser une ACP du tableau de départ après avoir éliminé la ou les colonnes (lignes) qui représentent ces points. Toutefois, ces points pourront être représentés sur un plan factoriel en utilisant leurs nouvelles coordonnées qu'il faut calculer.

1.2. Analyse Factorielle des Correspondances

([2], pp. 15-), [5], [6],([16], pp. 153-154), ([19], pp. 63-82), [78].

Définition 1. 5 (**Analyse Factorielle des Correspondances).**

Une Analyse Factorielle des Correspondances (AFC) est une des méthodes de la statistique multidimensionnelles, comptée parmi les méthodes factorielles qui a pour but d'étudier la liaison (ou correspondance) entre deux variables qualitatives.

L'AFC est une ACP réalisée sur les profils associés à un tableau de contingence. Elle poursuit le même objectif que l'ACP. Un tableau utilisé dans l'AFC ne peut pas comporter des cases vides et seules les valeurs positives sont permises.

Pour la réaliser, on procède comme suit : (1) Construire le tableau de contingence. (1.2.1.). (2) Transformer les données de départ : construire le tableau des fréquences, les tableaux des profils (lignes et colonnes) (1.2.2.). (3) Réaliser l'ACP du tableau des profils : construire le tableau des profils transformés et centrés, déterminer la matrice des variances-covariances ou des corrélations à partir de ce tableau, diagonaliser soit la matrice des variances-covariances soit la matrice des corrélations, présenter les résultats (1.2.3.).

1.2.1. Tableau de contingence

Définition 1. 6 (**Tableau de contingence).**

Soit à étudier pour N individus donnés, les caractères ou variables qualitatives X et Y de modalités respectives $X_1, ..., X_i \ldots, X_n$ et $Y_1, ..., Y_j \ldots, Y_m$. Un tableau de contingence (tableau des données) $K_{(n,m)} = \{n_{ij}\}$ est un tableau rectangulaire de dimension nxm qui croise les variables X et Y, et qui contient à l'intersection de la ligne i et de la colonne j le nombre n_{ij} d'individus ayant réalisé à la fois les modalités X_i et Y_j.

Le tableau de contingence se présente comme suit :

Tableau 1.1 – Tableau de contingence (des données) (AFC)

Modalités	$Y_1 \ldots Y_j \ldots Y_m$	**Total**
X_1		
$\ldots$	$\vdots$	
X_i	n_{ij}	$n_{i.}$
$\ldots$	$\vdots$	
X_n		
Total	$n_{.j}$	N

Source :([16], p.154)

Notons ce tableau $K_{(n,m)} = \{n_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq m}$.

Dans ce tableau,

$$n_{i.} = \sum_{j=1}^{m} n_{ij} \tag{1.42}$$

est l'effectif marginal de la ligne i (i est fixé) c'est-à-dire le nombre d'individus ayant vérifié la modalité X_i ;

$$n_{.j} = \sum_{i=1}^{n} n_{ij} \tag{1.43}$$

l'effectif marginal de la colonne j (j est fixé) c'est-à-dire le nombre d'individus ayant vérifié la modalité Y_j et

$$N = \sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij} = \sum_{j=1}^{m} n_{.j} = \sum_{i=1}^{n} n_{i.} \tag{1.44}$$

l'effectif global des répondants aux questions (ou effectif total de la population observée).

1.2.2. Transformation des données de départ et présentation de tableaux de profils

1.2.2.1.Tableau des fréquences

Il est question de transformer le tableau de contingence $K_{(n,m)}$ en tableau des fréquences en divisant chacune de ses cases par l'effectif total N.

Notons le tableau des fréquences par $F_{(n,m)} = \{f_{ij}\}_{1\leq i\leq n, 1\leq j\leq m}$.

Dans ce tableau,

- $f_{ij} = \dfrac{n_{ij}}{N}$ (1.45)

est la fréquence relative de n_{ij} ;

- $f_{i.} = \displaystyle\sum_{j=1}^{m} f_{ij} = \frac{n_{i.}}{N}$ (1.46)

la fréquence marginale (ou l'effectif total) de la ligne i ;

- $f_{.j} = \displaystyle\sum_{i=1}^{n} f_{ij} = \frac{n_{.j}}{N}$ (1.47)

la fréquence marginale de la modalité j et

- $f = \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = \sum_{j=1}^{m} f_{.j} = \sum_{i=1}^{n} f_{i.} = 1$ (1.48)

la fréquence totale.

Ce tableau s'interprète en disant par exemple $100.f_{i.}\%$ d'individus vérifient les modalités X_i et $100.f_{.j}\%$ d'individus vérifient les modalités Y_j.

1.2.2.2. Tableaux des profils

Il est question de transformer le tableau des fréquences en tableaux des profils (ligne ou colonne). Contrairement au tableau de contingence, les tableaux des profils- lignes et colonnes n'ont plus le même sens si on remplace les unes par les autres ([29], p. 49).

1.2.2.2.1. Tableau des profils-lignes

Le tableau des profils-lignes est obtenu en divisant chaque ligne du tableau des fréquences $F_{(n,m)}$ par la fréquence marginale de cette ligne.

Notons le par $X_{(n,m)} = \{X_{ij}\}_{1\leq i\leq n, 1\leq j\leq m}$.

Dans ce tableau,

- $X_{ij} = \dfrac{f_{ij}}{f_{i.}}$ (1.49)

est la fréquence conditionnelle de j sachant i ;

- $\displaystyle\sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}} = 1;$ (1.50)

- $\displaystyle\sum_{j=1}^{m}\sum_{i=1}^{n} \frac{f_{ij}}{f_{i.}} = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}} = n$ (1.51)

le nombre de lignes.

Avec le tableau de profils-lignes, on se situe dans l'espace $\mathbb{R}^m$ des variables (modalités colonnes) dans lequel on peut représenter le nuage Γ_X de n points "individus" (modalités lignes) $X_1, \ldots, X_i, \ldots, X_n$ de m coordonnées avec

$$X_i = \left(\frac{f_{i1}}{f_{i.}}, \ldots, \frac{f_{ij}}{f_{i.}}, \ldots, \frac{f_{im}}{f_{i.}}\right) \tag{1.52}$$

La moyenne pondérée de la variable Y_j étant $\overline{Y}_j = f_{.j}$, le centre de gravité du nuage Γ_X de ces points est $G_Y = (f_{.1}, \ldots, f_{.j}, \ldots, f_{.m})$.

Démonstration. Montrons que $\overline{Y}_j = f_{.j}$

En effet, $\overline{Y}_j = \dfrac{1}{N}\displaystyle\sum_{i=1}^{n} n_{i.}\frac{f_{ij}}{f_{i.}} = \sum_{i=1}^{n}\frac{n_{i.}}{N}.\frac{f_{ij}}{f_{i.}} = \sum_{i=1}^{n} f_{i.}\frac{f_{ij}}{f_{i.}} = \sum_{i=1}^{n} f_{ij} = f_{.j}$

□

1.2.2.2.2. Tableau des profils-colonnes

Le tableau des profils-colonnes $Y_{(m,n)} = \{Y_{ij}\}$ est obtenu en divisant chaque colonne du tableau des fréquences par la fréquence marginale de la colonne.

Dans ce tableau,

- $Y_{ij} = \dfrac{f_{ij}}{f_{.j}}$ (1.53)

est la fréquence conditionnelle de i sachant j ;

- $\displaystyle\sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} = 1;$ (1.54)

- $\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{f_{ij}}{f_{.j}} = \sum_{j=1}^{m}\sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} = m$ (1.55)

le nombre de colonnes (le nombre de modalités de la variable Y),

- $\displaystyle\sum_{i=1}^{n} f_{i.} = \sum_{i=1}^{n} \frac{1}{n} = 1.$ (1.56)

Ce tableau comprend m points (profils-colonnes) $Y_1, \ldots, Y_j, \ldots, Y_m$ de n coordonnées avec

$$Y_j = \Big(\frac{f_{i1}}{f_{.j}}, \ldots, \frac{f_{ij}}{f_{.j}}, \ldots, \frac{f_{im}}{f_{.j}}\Big), \tag{1.57}$$

La moyenne pondérée de la modalité X_i étant

$$\overline{X}_i = f_{i.} = \frac{1}{n}, \tag{1.58}$$

le centre de gravité de nuage Γ_Y de ces points est donné par
$G_X = (f_{1.}, \ldots, f_{i.}, \ldots, f_{n.})$.

1.2.2.3. Tableaux des profils transformés et centrés

1.2.2.3.1. Tableaux des profils transformés

Les tableaux de profils transformés viennent résoudre le problème d'échelle qui se pose en AFC, contrairement à l'ACP, dans le cas où les deux variables ne sont pas de même dimension. Il faut, à cet effet pondérer les profils-lignes (respectivement les profils-colonnes) par les composantes du centre de gravité du nuage des n points de l'espace à m dimensions (respectivement par celles des m points de l'espace à n dimensions) [64].

1. Tableau des Profils-lignes transformés

Il est constitué en divisant, sur la colonne, chaque case du tableau des profils-lignes par la racine carrée de $f_{.j}$ correspondante (Cfr 1.2.2.2.1) . Le terme de la i^e ligne et j^e colonne du tableau des profils-lignes transformés est :

$$\grave{X}_{ij} = \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} \tag{1.59}$$

Avec $f_{.j}$ = $\overline{Y}_j$ la moyenne pondérée de Y_j du tableau des profils-lignes.

Les coordonnées de $\grave{X}_i$ se présente comme suit :

$$\grave{X}_i = \Big(\frac{f_{i1}}{f_{i.}\sqrt{f_{.1}}}, \ldots, \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}, \ldots, \frac{f_{im}}{f_{i.}\sqrt{f_{.m}}}\Big), \text{i étant fixé} \tag{1.60}$$

Le centre de gravité (profil moyen) $G_{\grave{X}}$ du nuage $\Gamma_{\grave{X}}$ de ces points est :

$$G_{\grave{X}} = \Big(\sqrt{f_{.1}}, \ldots, \sqrt{f_{.j}}, \ldots, \sqrt{f_{.m}}\Big) = \Big(\sqrt{\overline{Y}_1}, \ldots, \sqrt{\overline{Y}_j}, \ldots, \sqrt{\overline{Y}_m}\Big) \tag{1.61}$$

2. Tableau des Profils-colonnes transformés

Chaque case du tableau des profils-colonnes est divisée, sur la ligne, par la racine carrée de $f_{i.}$ correspondante. Le terme de la i^e ligne et j^e colonne de ce tableau est

$$\grave{Y}_{ij} = \frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} \tag{1.62}$$

Avec $f_{i.}$ = $\overline{X}_i$ la moyenne pondérée de X_i du tableau des profils-colonnes.

Ce tableau comprend m points (profils-colonnes transformés) $\grave{Y}_1, \ldots, \grave{Y}_j$, ..., $\grave{Y}_m$ avec

$$\grave{Y}_j = \Big(\frac{f_{1j}}{f_{.j}\sqrt{f_{1.}}}, \ldots, \frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}}, \ldots, \frac{f_{nj}}{f_{.j}\sqrt{f_{n.}}}\Big), \text{j étant fixé} \tag{1.63}$$

$$\text{et } G_{\grave{Y}} = \Big(\sqrt{f_{1.}}, \ldots, \sqrt{f_{i.}}, \ldots, \sqrt{f_{n.}}\Big) = (\sqrt{\overline{X}_1}, \ldots, \sqrt{\overline{X}_i}, \ldots, \sqrt{\overline{X}_n}) \tag{1.64}$$

le centre de gravité (profil moyen) $G_{\grave{Y}}$ du nuage $\Gamma_{\grave{Y}}$ de ces points.

1.2.2.3.2. Tableau des profils-lignes transformés et centrés

Il est question de centrer les profils transformés de sorte que les axes passent par le centre de gravité $G_{\grave{X}}$. Chaque profil-ligne transformé $\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}$ est diminué du barycentre $G_X = \sqrt{f_{.j}}$ du tableau des profils-lignes.

Notons le tableau des profils-lignes transformés centrés par

$\breve{M}_{(n,m)} = \left\{\breve{X}_{ij}\right\}_{1\leq i\leq n, 1\leq j\leq m}$.

Ainsi, le terme de la i^e ligne et j^e colonne de ce tableau est

$$\breve{X}_{ij} = \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}} \tag{1.65}$$

Dans ce tableau $\breve{X}_i = \left(\frac{f_{i1}}{f_{i.}\sqrt{f_{.1}}} - \sqrt{f_{.1}}, \ldots, \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}}, \ldots, \frac{f_{im}}{f_{i.}\sqrt{f_{.m}}} - \sqrt{f_{.m}}\right)$ est le point-ligne d'un individu i et $G_{\breve{Y}} = (0, \ldots, 0, \ldots, 0)$ le barycentre du nuage des points.

La moyenne pondérée de la colonne $\breve{Y}_j$ notée $\bar{\breve{Y}}_j$ égale 0. On procède de manière analogue pour les profils-colonnes transformés et centrés.

En effet, $\bar{\breve{Y}}_j = \sum_{i=1}^n f_{i.}\breve{X}_{ij} = \sum_{i=1}^n f_{i.}\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}}\right) = \sum_{i=1}^n \left(\frac{f_{ij} - f_{i.}f_{.j}}{\sqrt{f_{.j}}}\right)$
$= \frac{1}{\sqrt{f_{.j}}}\left(\sum_{i=1}^n f_{ij} - f_{.j}\sum_{i=1}^n f_{i.}\right) = \left(\frac{f_{.j} - f_{.j}}{\sqrt{f_{.j}}}\right) = 0$

1.2.3. Analyse en composantes principales du tableau des profils transformés centrés

1.2.3.1. Matrice des variances-covariances

La matrice des variances-covariances $\breve{V}$ est constituée des variances et des covariances du tableau $\breve{M}$ des profils-lignes transformés et centrés ([29], p.49).

On a :

$\breve{V} = \left(\breve{V}_{jj'}\right)_{1\leq j\leq m, 1\leq j'\leq m}$

où

$$\breve{V}_{jj'} = \sigma^2_{\breve{Y}_j\breve{Y}_{j'}} = \sum_{i=1}^n \frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}} - \sqrt{f_{.j}}\sqrt{f_{.j'}} \tag{1.66}$$

est la covariance de $\breve{Y}_j$ et $\breve{Y}_{j'}$

En effet, la covariance de $\breve{Y}_j$ et $\breve{Y}_{j'}$ égale

$\breve{V}_{jj'} = \sigma^2_{\breve{Y}_j\breve{Y}_{j'}} = \sum_{i=1}^n f_{i.}\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}}\right)\left(\frac{f_{ij'}}{f_{i.}\sqrt{f_{.j'}}} - \sqrt{f_{.j'}}\right)$

$$=\sum_{i=1}^{n} f_{i.}\left(\frac{f_{ij}f_{ij'}}{f_{i.}^2\sqrt{f_{.j}}\sqrt{f_{.j'}}} - \frac{f_{ij}\sqrt{f_{.j'}}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{ij'}\sqrt{f_{.j}}}{f_{i.}\sqrt{f_{.j'}}} + \sqrt{f_{.j}}\sqrt{f_{.j'}}\right)$$

$$=\sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}} - \sum_{i=1}^{n}\frac{f_{ij}\sqrt{f_{.j'}}}{\sqrt{f_{.j}}} - \sum_{i=1}^{n}\frac{f_{ij'}\sqrt{f_{.j}}}{\sqrt{f_{.j'}}} + \sum_{i=1}^{n} f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}$$

$$=\sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}} - \frac{f_{.j}\sqrt{f_{.j'}}}{\sqrt{f_{.j}}} - \frac{f_{.j'}\sqrt{f_{.j}}}{\sqrt{f_{.j'}}} + \sqrt{f_{.j}}\sqrt{f_{.j'}}$$

$$=\sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}} - \sqrt{f_{.j}}\sqrt{f_{.j'}} - \sqrt{f_{.j'}}\sqrt{f_{.j}} + \sqrt{f_{.j}}\sqrt{f_{.j'}}$$

$$=\sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}} - \sqrt{f_{.j}}\sqrt{f_{.j'}}$$

Si dans la formule de la covariance, on considère $\breve{Y}_j = \breve{Y}_{j'}$, on trouve alors la formule de la variance de $\breve{Y}_j$:

$$\sigma^2_{\breve{Y}_j} = \breve{V}_{jj} = \sum_{i=1}^{n}\frac{f_{ij}^2}{f_{i.}f_{.j}} - f_{.j} \tag{1.67}$$

1.2.3.2. Diagonalisation de la matrice des variances-covariances ou de la matrice d'inertie et détermination des axes

1) Matrice des variances-covariances et matrice d'inertie

C'est la matrice des variances-covariances $\breve{V}$ d'ordre (m, m) issue du tableau des profils-lignes transformés et centrés qui sera diagonalisée. Elle permettra de déterminer la matrice de changement de base à l'aide des vecteurs propres normés calculés à partir des valeurs propres de cette matrice. Toutefois, en AFC, on n'utilise pas toujours cette matrice mais une matrice plus simple appelée la matrice d'inertie et notée S qui est utilisée dans beaucoup de logiciels informatiques. La matrice d'inertie est déterminée à partir de la matrice des variances-covariances.

Les vecteurs propres de ces deux matrices sont identiques. Ce qui permet de diagonaliser indifféremment l'une ou l'autre de ces deux matrices. Toutefois, il existe une différence entre elles : la première valeur propre de $\breve{V}$ est 0 et celle de S est 1. Le premier vecteur propre associé à cette première valeur propre définit un axe principal (axe trivial) pour lequel les projections des individus (modalités lignes) et des variables (modalités colonnes) possèdent une variance ou dispersion nulle c'est-à-dire toutes les projections ont les mêmes coordonnées. En conséquence, cet axe factoriel est exclu de l'analyse.

2) Détermination de la matrice d'inertie

Soit la matrice des variances-covariances $\breve{V} = (\breve{V}_{jj'})_{1\leq j\leq m, 1\leq j'\leq m}$. Le premier vecteur propre noté v_o issue de cette matrice est associée à sa première valeur propre $\lambda_0 = 0$.

On a :

$$v_0 = \begin{pmatrix} v_{01} \\ ... \\ v_{0j} \\ ... \\ v_{0m} \end{pmatrix} = \begin{pmatrix} \sqrt{f_{.1}} \\ ... \\ \sqrt{f_{.j}} \\ ... \\ \sqrt{f_{.m}} \end{pmatrix}$$

Considérons un vecteur propre quelconque v_k associé à la valeur propre λ_k de la matrice $\breve{V}$. Les vecteurs propres étant deux à deux orthogonaux, alors le produit scalaire

$$v_k v_0 = 0 \iff \begin{pmatrix} v_{k1} ... v_{kj} ... v_{km} \end{pmatrix} \begin{pmatrix} v_{01} \\ ... \\ v_{0j} \\ ... \\ v_{0m} \end{pmatrix} \iff \sum_j^m v_{kj} v_{0j} = 0$$

$$\iff \sum_j^m v_{kj} \sqrt{f_{.j}} = 0 \qquad (1)$$

Or v_k et λ_k vérifie la relation

$\breve{V} v_k = \lambda_k v_k$

ou encore

$$\left(\breve{V}_{jj'}\right) \begin{pmatrix} v_{k1} \\ ... \\ v_{kj} \\ ... \\ v_{km} \end{pmatrix} = \lambda_k \begin{pmatrix} v_{k1} \\ ... \\ v_{kj} \\ ... \\ v_{km} \end{pmatrix}$$

Pour le jè terme :

$$\breve{V}_{j1} v_{k1} + ... + \breve{V}_{jj'} v_{kj'} + ... + \breve{V}_{jm} v_{km} = \lambda_k v_{kj} \iff \sum_{j'=1}^m \breve{V}_{jj'} v_{kj'} = \lambda_k v_{kj} \qquad (2)$$

Or $\breve{V}_{jj'} = \sum_{i=1}^n \frac{f_{ij} f_{ij'}}{f_{i.} \sqrt{f_{.j}} \sqrt{f_{.j'}}} - \sqrt{f_{.j}} \sqrt{f_{.j'}}$

(2) devient

$$\sum_{j'=1}^m \left(\sum_{i=1}^n \frac{f_{ij} f_{ij'}}{f_{i.} \sqrt{f_{.j}} \sqrt{f_{.j'}}} - \sqrt{f_{.j}} \sqrt{f_{.j'}} \right) v_{kj'} = \lambda_k v_{kj}$$

$$\iff \sum_{j'=1}^m \sum_{i=1}^n \frac{f_{ij} f_{ij'}}{f_{i.} \sqrt{f_{.j}} \sqrt{f_{.j'}}} v_{kj'} - \sum_{j'=1}^m \sqrt{f_{.j}} \sqrt{f_{.j'}} v_{kj'} = \lambda_k v_{kj}$$

$$\iff \sum_{j'}^m \left(\sum_{i=1}^n \frac{f_{ij} f_{ij'}}{f_{i.} \sqrt{f_{.j}} \sqrt{f_{.j'}}} v_{kj'} \right) - \sqrt{f_{.j}} \sum_{j'=1}^m \sqrt{f_{.j'}} v_{kj'} = \lambda_k v_{kj}$$

Or d'après (1), $\sum_{j'=1}^{m}\sqrt{f_{.j'}}v_{kj'} = 0$

D'où $\lambda_k v_{kj} = \sum_{j'}^{m}\big(\sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}}v_{kj'}\big)$ (3)

On pose,

$$\sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}} = S_{jj'} \tag{1.68}$$

$\Longleftrightarrow S_{jj'} = \sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{f_{i.}\sqrt{f_{.j}}\sqrt{f_{.j'}}} = \sum_{i=1}^{n}\frac{f_{ij}f_{ij'}}{\sqrt{f_{i.}}\sqrt{f_{.j}}\sqrt{f_{i.}}\sqrt{f_{.j'}}}$

La relation (3) s'écrit $\lambda_k v_{kj} = \sum_{j'}^{m} S_{jj'}v_{kj'} = \sum_{j'}^{m}\breve{V}_{jj'}v_{kj'}$

Appelons S, la matrice des termes $S_{jj'}$, on a alors

$\lambda_k v_k = \breve{V}v_k = Sv_k$

La matrice S porte le nom de **matrice d'inertie**.

3) Part de variance totale expliquée par la kè composante principale

La part de la variance totale expliquée par la kè composante principale est donnée par la formule

$$\frac{\lambda_j}{I_X - 1}.100 = \frac{\lambda_j}{tr(\breve{V}) - 1}.100 = \frac{\lambda_j}{tr(\breve{S}) - 1}.100 = \frac{\lambda_j}{\sum_{k=1}^{m}\lambda - 1}.100 \tag{1.69}$$

où -1 au dénominateur se justifie par le fait que la première valeur propre de $\breve{V}$ est nulle), I_X est l'inertie du nuage Γ_X, $tr(\breve{V})$ et $tr(S)$ sont les traces respectives de $\breve{V}$ et de S.

Comme en ACP, l'AFC est réalisable si avec 1, 2 ou au maximum 3 axes principaux, on explique près de 70% de la variance totale.

1.2.3.3. Présentation des résultats

1.2.3.3.1. Analyse du nuage des variables (qualitatives)

([16], pp. 156-157), ([19], p. 74-79), ([29], p. 49), ([45], pp. 25-26).

Lorsqu'on mesure la liaison entre les variables quantitatives, on parle de corrélation. Par contre, dans le cas des variables qualitatives, on utilise l'expression « association ».

1. Association entre deux variables qualitatives

La relation entre les variables qualitatives est mesurée à l'aide de différents coefficients d'association.

1) Coefficient du Khi-deux

L'indice (ou coefficient) du Khi-deux noté χ^2 est une mesure classique de la liaison entre deux variables qualitatives. Il est une mesure de "l'écart à la situation d'indépendance". Il mesure la distance globale entre les effectifs observés n_{ij} (Cfr Tableau de contingence) et les effectifs théoriques e_{ij} attendus lorsque les variables X et Y sont indépendantes.

Ainsi,

$$e_{ij} = \frac{n_{i.}n_{.j}}{N}, \tag{1.70}$$

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{N}\right)^2}{\frac{n_{i.}n_{.j}}{N}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(nf_{ij} - Nf_{i.}f_{.j}\right)^2}{Nf_{i.}f_{.j}}$$

$$= N\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}} - \sqrt{f_{i.}}\sqrt{f_{.j}}\right)^2 \tag{1.71}$$

Le coefficient Khi-deux est positif ou nul : La valeur 0 correspond à l'indépendance des variables et plus la valeur du coefficient est grande plus l'écart à l'indépendance observé (sur l'échantillon) est grand.

2) Coefficient phi-deux

On utilise aussi le coefficient phi-deux qui élimine l'effet de la taille N :

$$\phi^2 = \frac{\chi^2}{N} \tag{1.72}$$

ou encore phi :

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{1.73}$$

Il varie entre 0 et 1. Cette liaison est plus intense si les modalités de l'une s'associent plus à celles de l'autre.

2. Test (du Khi-deux) d'indépendance pour deux variables qualitatives

Etant données deux variables qualitatives X et Y, le test du Khi-deux suit les étapes suivantes :

• *Poser l'hypothèse :* L'hypothèse testée est H_0 (Hypothèse nulle) : X et Y ne sont pas liées (elles sont indépendantes) contre H_1 (hypothèse alternative) : X et Y sont liées.

• *Calculer la statistique du Khi-deux :* χ^2 (valeur observée) suit une loi (une distribution) du Khi2 avec un nombre de degré de liberté égal à $(n-1)(m-1)$ où n est le nombre de modalités de X et m celui de Y.

• *Déterminer la valeur critique :* Dans la table, on lit la valeur critique $\chi^2_{\alpha,k}$ qui est la valeur ayant une probabilité α (5%, 10%) d'être dépassée pour une distribution du khi2 avec comme degré de liberté $k = (n-1)(m-1)$.

• *Décider sur l'acceptation ou le rejet de l'hypothèse nulle :* Si $\chi^2 > \chi^2_{\alpha,k}$, on rejette H_0 par conséquent les deux variables sont liées (elles ne sont pas indépendantes). Mais, si $\chi^2 \leq \chi^2_{\alpha,k}$, on accepte H_0 et on conclut que les deux variables ne sont pas liées (elles sont indépendantes).

• *Confirmer le résultat avec la valeur p :* La valeur p (p-valeur ou p-value) notée pval correspond à la quantité

$$pval = P(\chi^2 \geq \chi^2_{\alpha,k}) \tag{1.74}$$

C'est la probabilité sous H_0 d'observer une valeur de χ^2 aussi grande que $\chi^2_{\alpha,k}$ (valeur observée sur l'échantillon). Si $pval \leq \alpha$, on rejette H_0 et on conclut que les variables sont liées. Mais si $pval > \alpha$, on accepte l'hypothèse nulle et on conclut que les variables ne sont pas liées (elles sont indépendantes).

1.2.3.3.2. Analyse du nuage des individus

1. Proximités entre deux individus

En ACP, l'information est donnée par la distance euclidienne, appliquée aux données quantitatives, entre les points des nuages de deux espaces à n et m dimensions ($\mathbb{R}^n$ et $\mathbb{R}^m$). Par contre, en AFC, c'est la métrique de Khi-deux qui est utilisée aux données qualitatives non transformées. Ce qui équivaut à la distance euclidienne entre les profils transformés et centrés (l'AFC étant une ACP).

La distance du Khi-deux sera donc utilisée pour mesurer la proximité entre les individus :

$$d^2(\breve{X}_i, \breve{X}_{i'}) = \sum_{j=1}^{m} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.}\sqrt{f_{.j}}} \right)^2 \tag{1.75}$$

Comme cela est souligné ci-dessus, cette distance euclidienne équivaut à *la métrique du χ^2 (Khi-deux)* sur les données non transformées.

En effet, on sait qu'étant données trois distributions de probabilités x, y et z (de valeurs respectives x_i, y_i et z_i) sur un ensemble discret des valeurs, la distance du chi-deux entre x et y centrée sur Z est :

$$d^2_{\chi^2}(x, y) = \chi^2_y(x, y) = \sum_i \frac{(x_i - y_i)^2}{z_i} \tag{1.76}$$

Ainsi, la distance entre $X_i = \frac{n_{ij}}{n_{i.}}$ et $X_{i'} = \frac{n_{i'j}}{n_{i'.}}$ centrée sur $n_{.j}$ est

$$d^2_{\chi^2}(X_i, X_{i'}) = \sum_{j=1}^{m} \frac{\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2}{n_{.j}} = \sum_{j=1}^{m} \frac{\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2}{f_{.j}} = \sum_{j=1}^{m} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.}\sqrt{f_{.j}}} \right)^2 \tag{1.77}$$

2. Distance des individus au centre de gravité

$$d^2(\breve{X}_i, G_{\breve{X}}) = \sum_{j=1}^{m} (\breve{X}_{ij} - G_{\breve{X}})^2 = \sum_{j=1}^{m} (\breve{X}_{ij} - \sqrt{f_{.j}})^2 \tag{1.78}$$

3. Inertie du nuage des individus

L'inertie $I_{\breve{X}}$ du nuage $N_{\breve{X}}$ des points $\breve{X}_i$ est calculée par la formule suivante :

$$\begin{aligned} I_{\breve{X}} &= \textstyle\sum_{i=1}^{n} f_{i.} d^2(\breve{X}_i, G_{\breve{X}}) = \sum_{i=1}^{n} \sum_{j=1}^{m} f_{i.} (\breve{X}_{ij} - \sqrt{f_{.j}})^2 \\ &= \sum_{j=1}^{m} S_{ij} = tr(S) = \sum_{j=1}^{m} \breve{V}_{ij} = tr(\breve{V}) = \sum_{j=1}^{m} \lambda_j \end{aligned} \tag{1.79}$$

1.2.3.3.3. Coordonnées des individus et des variables sur les axes

Soit v_{kj} et u_{ki} les vecteurs propres correspondant respectivement aux valeurs propres λ_k de $\breve{B}^t \breve{B}$ et γ_k de $\breve{B}\breve{B}^t$.

– Le vecteur de coordonnées des projections des points lignes $\breve{X}_i$ sur l'axe F_k :

$$F_k(\breve{X}_i) = \sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} v_{kj} \tag{1.80}$$

– Le vecteur de coordonnées des projections des points colonnes $\breve{Y}_j$ sur l'axe G_k de rang k :

$$G_k(\breve{Y}_j) = \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} u_{ki} \tag{1.81}$$

Dans le cas où deux facteurs sont retenus, les coordonnées $F_1(\breve{X}_i)$ et $F_2(\breve{X}_i)$ d'un individu $\breve{X}_i$ sur le plan factoriel sont données par :

$$\Big(F_1(\breve{X}_i), F_2(\breve{X}_i)\Big) = \left(\sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} v_{1j}, \sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} v_{2j}\right) \tag{1.82}$$

1.2.3.3.4. Relations entre les vecteurs propres de deux nuages

Soit à diagonaliser la matrice d'inertie $S = B^t B$ (issue du tableau des profils transformés et centrés) avec B la matrice de terme général B_{ij} = $\sum_{j=1}^{m} b_{ij}$ =$\sum_{j=1}^{m} \frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}}$

et B^t sa transposée avec B^t_{ij} = $\sum_{i=1}^{n} b_{ij}$ =$\sum_{i=1}^{n} \frac{f_{ij}}{\sqrt{f_{.j}}\sqrt{f_{i.}}}$

v_j étant le vecteur propre de S associé à la valeur propre λ_j, on a :

$B\ v_j$ = $\lambda_j\ v_j$ (1) et $v_j^t v_j = 1 \Leftrightarrow B^t B\ v_j$ = $\lambda_j\ v_j$ (2)

En multipliant les deux membres de (2) par B, on obtient :

$BB^t B\ v_j$ = $\lambda_j\ B\ v_j \Leftrightarrow BB^t(B\ v_j)$ = $\lambda_j\ (B\ v_j)$

En posant $(B\ v_j)$ = u'_i (3), on a : $BB^t u'_i$ = $\lambda_j\ u'_i \Leftrightarrow BB^t u'_i$ = $\lambda_j\ u'_i$

$\Leftrightarrow (BB^t)u'_i$ = $\lambda_j\ u'_i \Leftrightarrow S^* u'_i$ = $\lambda_j\ u'_i$ (4) (où $S^* = BB^t$)

En comparant (1) et (4), on comprend que

$S^* u'_i$ = $\lambda_j\ u'_i$

est l'écriture des vecteurs propres non normés u'_i = $(B\ v_j)$ associés aux valeurs propres λ_j de la matrice S^*.

En effet, les vecteurs propres u'_i sont non normés c'est-à-dire $\|u'_i\|^2 \neq 1$ (ou $u'^t_i u'_i \neq 1$).

Cela se justifie plus simplement :

$\|u'_i\|^2$ = $u'^t_i u'_i$ = $(Bv_j)^t\ (Bv_j)$= $v_j^t B^t\ Bv_j$ = $v_j^t S v_j$= $\lambda_j v_j^t v_j$ = $\lambda_j \neq 1$

Puis que le vecteur propre u'_i est non normé, il faut le normer. Ainsi, on lui impose

$(\alpha u'^t_i)(\alpha u'_i) = 1$ avec $(\alpha \in \mathbb{R})$

$\Leftrightarrow \alpha^2 {u'}_i^t u'_i = 1 \Leftrightarrow \alpha^2(\lambda_j) = 1 \Leftrightarrow \alpha = \pm\frac{1}{\sqrt{\lambda_j}}$.

A la lumière de (3), on retrouve ainsi le vecteur propre normé noté u_i qui s'écrit :

$$u_i = \frac{1}{\sqrt{\lambda_j}} B v_j \tag{1.83}$$

Ce qui signifie que lorsqu'on connait les vecteurs propres normés $v_1, ..., v_j, ..., v_m$ associés aux valeurs propres $\lambda_1, ..., \lambda_j, ..., \lambda_m$ de $\mathbb{R}^m$, il est possible de calculer les vecteurs propres normés $u_1, ..., u_i, ..., u_n$ associés aux valeurs propres $\lambda_1, ..., \lambda_j, ..., \lambda_m$ de la matrice $S = BB^t$

On peut prouver réciproquement que lorsqu'on connait les vecteurs propres normés $u_1, ..., u_i, ..., u_n$ asssociés aux valeurs propres $\lambda_1, ..., \lambda_j, ..., \lambda_m$ de $S = BB^t$, il est possible de déterminer les valeurs propres $\lambda_1, ..., \lambda_j, ..., \lambda_m$ et les vecteurs propres normés $v_1, ..., v_j, ..., v_m$ de la matrice S^* = $B^t B$.

En définitive, on a les deux relations suivantes entre les vecteurs propres de deux nuages des points :

$$u_{ki} = \frac{1}{\sqrt{\lambda_k}} B v_{kj} \tag{1.84}$$

$$v_{kj} = \frac{1}{\sqrt{\lambda_k}} B^t u_{ki} \tag{1.85}$$

Ce qui permet d'écrire

$u_{ki} = \frac{1}{\sqrt{\lambda_k}} B v_{kj} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{m} \frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}} v_{kj}$

$$\Leftrightarrow \sqrt{\lambda_k} u_{ki} = \sum_{j=1}^{m} \frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}} v_{kj} \tag{1.86}$$

et

$v_{kj} = \frac{1}{\sqrt{\lambda_k}} B^t u_{ki} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} \frac{f_{ij}}{\sqrt{f_{.j}}\sqrt{f_{i.}}} u_{kj}$

$$\Leftrightarrow \sqrt{\lambda_k} v_{kj} = \sum_{i=1}^{n} \frac{f_{ij}}{\sqrt{f_{.j}}\sqrt{f_{i.}}} u_{ki} \tag{1.87}$$

1.2.3.3.5. Formules de transition

La métrique du Khi-deux met en correspondance les modalités des lignes et colonnes, et joue le rôle symétrique entre elles. Ce qui fait que l'ACP des profils-lignes (Première analyse) est équivalente à l'ACP des profils-colonnes (Deuxième analyse) : Dualité des

deux ACP.

Les valeurs propres de ces deux analyses sont identiques. Les composantes principales d'une analyse sont celles de l'autre. Ainsi, dans la pratique, on n'effectue qu'une seule analyse d'entre les deux ACP, celle faite avec moins de modalités. Les résultats de l'autre se déduisent au moyen des relations entre les facteurs sur les deux nuages. Ces relations sont nommées formules de transition. ([72], p. 30)

Théorème 1. 1 (Formules de transition).

Soient F_k et G_k deux facteurs de même rang k respectivement de nuages $\Gamma_{\breve{X}}$ des individus et $\Gamma_{\breve{Y}}$ des variables. $F_k(\breve{X}_i)$ et $G_k(\breve{Y}_j)$ étant respectivement la coordonnée de l'individu $\breve{X}_i$ sur l'axe F_k et la coordonnée de la variable $\breve{Y}_j$ sur l'axe G_k. Alors les formules de transition relatives au tableau de profils transformés et centrés s'écrivent de la manière suivante :

$$F_k(\breve{X}_i) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}} G_k(\breve{Y}_j) \tag{1.88}$$

$$G_k(\breve{Y}_j) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} F_k(\breve{X}_i) \tag{1.89}$$

Démonstration. Considérons les quatre relations suivantes :

- Les coordonnées des individus et des variables sur les axes

$F_k(\breve{X}_i) = \sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} v_{kj}$ (1)

$G_k(\breve{Y}_j) = \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} u_{ki}$ (2)

(-) Les relations entre les vecteurs propres de deux nuages des points

$\sqrt{\lambda_k} u_{ki} = \sum_{j=1}^{m} \frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}} v_{kj}$ (3)

$\sqrt{\lambda_k} v_{kj} = \sum_{i=1}^{n} \frac{f_{ij}}{\sqrt{f_{.j}}\sqrt{f_{i.}}} u_{ki}$ (4)

A partir de (1) et (3), écrivons u_{ki} en fonction de $F_k(\breve{X}_i)$:

$$F_k(\breve{X}_i) = \sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} v_{kj} = \sum_{j=1}^{m} \frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{i.}}\sqrt{f_{.j}}} v_{kj}$$

$$= \frac{1}{\sqrt{f_{i.}}} \left(\sum_{j=1}^{m} \frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}} v_{kj} \right) = \frac{\sqrt{\lambda_k}}{\sqrt{f_{i.}}} u_{ki}$$

$\Leftrightarrow u_{ki} = \frac{\sqrt{f_{i.}}}{\sqrt{\lambda_k}} F_k(\breve{X}_i)$ (5)

De manière analogue, à partir de (2) et (4), écrivons v_{ki} en fonction de $G_k(\breve{Y}_j)$:

$G_k(\breve{Y}_j) = \frac{1}{\sqrt{f_{.j}}}\left(\sum_{i=1}^{n} \frac{f_{ij}}{\sqrt{f_{.j}}\sqrt{f_{i.}}} u_{ki}\right) = \frac{\sqrt{\lambda_k}}{\sqrt{f_{.j}}} v_{kj}$

$v_{kj} = \frac{\sqrt{f_{.j}}}{\sqrt{\lambda_k}} G_k(\breve{Y}_j) \qquad (6)$

De (1) et (6), déterminons $F_k(\breve{X}_i)$,

$F_k(\breve{X}_i) = \sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} . \frac{\sqrt{f_{.j}}}{\sqrt{\lambda_k}} G_k(\breve{Y}_j)$

$\Leftrightarrow F_k(\breve{X}_i) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}} G_k(\breve{Y}_j) \qquad (7)$

De (2) et (5), déterminons $G_k(\breve{Y}_j)$,

$G_k(\breve{Y}_j) = \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} . \frac{\sqrt{f_{i.}}}{\sqrt{\lambda_k}} F_k(\breve{X}_i)$

$\Leftrightarrow G_k(\breve{Y}_j) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} F_k(\breve{X}_i) \qquad (8)$

D'où (7) et (8), les formules de transition permettant le passage des coordonnées projetées des nuages $\Gamma_{\breve{X}}$ et $\Gamma_{\breve{Y}}$. □

1.2.3.3.6. Représentation simultanée (ou barycentrique) des éléments

Nous avons considéré le tableau des profils-lignes transformés et centrés qui vient résoudre le problème d'échelle entre les colonnes et les lignes, et de ramener l'origine des axes au centre de gravité du nuage des points. Ce qui permet représenter sur un même graphique, les lignes et les colonnes en utilisant les coordonnées des individus sur les axes. Les coordonnées des variables peuvent s'exprimer en fonction de celles des individus et vice versa grâce aux formules de transition.

La représentation simultanée de deux nuages de points (individus et variables) s'obtient en superposant les projections de ces deux nuages sur le plan engendré par des axes de même rang. Pour y arriver, on utilise la propriété barycentrique : la projection $F_k(\breve{X}_i)$ de la ligne X_i sur l'axe F_k est le barycentre des projections $G_k(\breve{Y}_j)$ des colonnes Y_j sur l'axe G_k. La formulation symétrique est aussi possible. D'où l'expression double propriété barycentrique.

1.2.3.4. Aides à l'interprétation et éléments supplémentaires

Les indices d'aides à l'interprétation définis en ACP s'appliquent aussi en AFC du fait qu'ils sont valables pour n'importe quel nuage. Toutefois, puis qu'en AFC les éléments n'ont pas tous le même poids, contrairement à l'ACP où ils ont généralement des poids égaux, le poids de chacun intervient dans sa contribution à l'inertie d'un axe.

Quant à ce qui est des éléments supplémentaires, la référence est faite à la technique utilisée dans l'ACP. Les lignes (respectivement les colonnes) supplémentaires sont reliées aux colonnes (respectivement aux lignes) au moyen des relations barycentriques.

1.2.3.5. Extension de l'AFC

L'AFC est réalisée dans le cas des tableaux de contingence issus de deux variables qualitatives mais lorsque le nombre de variables dépasse deux, on fait appel à l'Analyse Factorielle des Correspondance Multiple (FCM ou ACM). Celle-ci est donc une extension de l'AFC. Pour la réaliser, il convient de procéder de l'une ou l'autre manière suivante : Réaliser une ACP sur le tableau disjonctif complet (TDC) ou Réaliser une ACP sur le tableau de Burt. Les documents ci-après proposent des éléments intéressants sur l'AFCM : ([2], p. 27-30), ([19], pp. 85-105), [38].

Pour constituer le TDC, on procède comme suit : en partant de tableau des données de départ, on trace un tableau où on considère toutes les modalités comme des caractères. Si un individu vérifie une modalité, on note 1 dans la case formée par cet individu et cette modalité. Dans le cas contraire, on marque 0. Le tableau de Burt quant à lui est construit à partir du TDC en multipliant la transposée de ce dernier par lui-même.

L'ACP de l'un ou l'autre tableau sera réalisée comme en AFC (Cf. 1.2.3.) tout en prenant en compte les particularités du TDC qui influent sur le tableau des fréquences et les tableaux des profils. Dans la pratique, à titre d'exemple, il est question de remplacer dans les formules utilisées en AFC :
$f_{i.}$ par $\frac{1}{n}$, $f_{.j}$ par $\frac{n_{i.}}{n.m'}$, $\frac{f_{ij}}{f_{i.}}$ par $\frac{n_{ij}}{m'}$, etc. où n est l'effectif total des individus et m' le nombre de modalités ([16], pp. 156-157),([19], p. 74-79), ([29], p. 49), ([45], pp. 25-26).

1.3. Conclusion du premier chapitre

Ce chapitre a porté plus particulièrement sur l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle de Correspondances (AFC) qui a pour extension l'Analyse Factorielle des Correspondances Multiples (AFCM). Ce sont des méthodes de la Statistique multidimensionnelle, classée parmi les méthodes factorielles qui consistent, dans l'ensemble, à rechercher des facteurs en nombre réduit (2 ou 3) pouvant résumer le mieux possible les données considérées. Elles aboutissent à des représentations graphiques du type nuage des points (ou diagramme de dispersion). Elles analysent les proximités entre les individus et les corrélations ou associations entre des variables et mo-

dalités suivant le cas en présence. Les données (individus ou variables) sont représentées par rapport aux facteurs considérés comme des axes.

L'ACP utilise les variables quantitatives et la distance euclidienne. Quant à l'AFC est une ACP appliquée au tableau de profils transformés et centrés partant du tableau de contingence. Elle utilise deux variables qualitatives et la distance du Chi-deux correspondant à la distance euclidienne. L'AFC s'étend à l'AFCM qui intervient dans le cas de plus de deux variables qualitatives. Elle est une AFC appliquée au tableau disjonctif complet (TDC) ou au tableau de Burt. Le tableau de Burt est construit à partir du TDC. Celui-ci comprend dans chacune de ses cases le chiffre 0 ou 1. A partir du TDC, on construit le tableau de profils transformés et centrés sur lequel sera réalisée l'ACP. On tiendra tout de même compte des particularités du TDC.

Les cinq étapes suivantes permettent de réaliser une ACP : (1) Transformer les données de départ en données centrées ou en données centrées réduites. (2) Déterminer : soit la matrice des variances-covariances, dans le cas où la matrice est centrée sur le centre de gravité soit la matrice de corrélation dans le cas où on utilise la matrice centrée-réduite. (3) Diagonaliser, suivant le modèle choisi, soit la matrice de variances-covariances soit la matrice de corrélation. (4) Extraire les facteurs (ou composantes principales) que l'on cherche en petit nombre (2 ou 3). (5) Présenter les résultats (sur les variables et sur les individus).

L'ACP est réalisable si deux ou trois composantes principales, expliquent près de 70% de la variance totale.

Toutefois, les méthodes factorielles ne permettent pas de regrouper les individus en classes en vue d'une éventuelle répartition d'une ressource commune entre eux. Il faut donc choisir une autre méthode pour le faire.

Chapitre 2

Classification Ascendante Hiérarchique

Introduction

[1], [10], [15], [17], ([24], p.2), [39], [40], ([76], p.34), [84].

Une étude statistique portant sur plusieurs variables amène à procéder par les méthodes d'Analyse des données qui comprennent notamment la méthode de l'Analyse Factorielle et celle de la Classification.

La méthode de l'Analyse Factorielle comprend principalement les méthodes suivantes : L'Analyse en Composantes Principales (ACP) ; l'Analyse Factorielle de Correspondance (AFC) et l'Analyse en Composantes Multiples (ACM) appelée aussi Analyse Factorielle des Correspondances Multiples (AFCM). Quant à ce qui est de la classification, elle composée entre autres de la Classification Ascendante Hiérarchique (CAH) et de la Classification Descendante Hiérarchique (CDH).

Les méthodes de l'Analyse Factorielle permettent de déterminer les facteurs (axes factoriels) en nombre inférieur à celui des variables de départ. Ces facteurs, comme nouvelles variables, pourront être utilisés pour déterminer, à l'aide de la méthode de classification les différentes classes auxquelles appartiennent les individus.

2.1. Généralités sur la classification

Définition 2. 1 (Classification).

La classification est une méthode de l'Analyse des données qui consiste à retrouver des classes qui sont telles que les individus d'une même classe soient les plus semblables possibles (homogénéité intra-classe) tandis que ceux des classes différentes les plus dissemblables (hétérogénéité inter-classe). Une classe est un ensemble d'individus qualifiés de plus semblables entre eux à l'issue de la réalisation d'une classification.

2.1.1. Aperçu historique sur la classification

La Classification a évoluée dans le temps, à travers différents domaines et avec le concours des personnages dont la plus part n'étaient pas au départ des Statisticiens. En effet, la classification est issue de l'Anthropologie par, d'abord Czekanowski Jan en 1911 puis Driver Harnold E. et Kroeber Alfred Louis en 1932 en vue des analyses par grappes (analyse d'agrégats). Ces dernières sont utilisées en psychologies par Zubin Joseph en 1938, Tryon Robert Choate en 1939 et à partir de 1943 par Cattell ([71], pp. 16-19).

Les problèmes de calcul ont empêché le développement initial de ces idées. Mais au début des années 1960 des techniques de regroupement sont introduites dans d'autres disciplines, notamment biologie avec Sokal Robert R. et Sneath Peter H. A. où elles sont désignées par le terme taxonomie ou taxinomie. Actuellement, la classification apparaît comme une notion de Statistique utilisée aussi dans d'autres domaines de la vie : domaine médical où elle est désignée par nosologie, domaine commercial, en marketing, en linguistique, en géographie, ... [43], [67], ([71], pp. 16-19).

En ce qui nous concerne, nous appliquons la classification dans le domaine politico-administratif en vue de la répartition des ressources budgétaires allouées aux 24 communes de la Ville-province de Kinshasa considérées comme ETDs.

2.1.2. Données à utiliser dans une classification et objets à classifier

Dans la classification, on peut utiliser les données qualitatives [44] tout comme les données quantitatives. Il y a aussi des types particuliers de données comme les données floues [25]. La classification est faite soit sur les observations (individus) soit sur les variables [7] qui constituent des objets à classifier. Elle est aussi faite sur les deux lorsqu'on suppose que les deux contribuent simultanément à la mise en évidence de structures explicites de classes. On parle dans ce cas de la classification conjointe sur eux deux (cas relativement rare).

Quel que soit le type de données (variables qualitatives ou quantitatives) à utiliser ou le type d'objets (individus ou variables) à classifier, il y a toujours besoin de se choisir une méthode de classification à utiliser qui peut être précédée d'une ACP dans le cas où l'on voudrait réduire le nombre de variables [31].

2.1.3. Subdivision de la Classification : classification supervisée et classification non supervisée (ou automatique)

La classification se subdivise en classification supervisée et classification non supervisée. Dans la classification supervisée, les classes sont déjà définies et on place de nouveaux individus un après l'autre dans l'une ou l'autre de ces classes déjà étiquetées. Quant à ce qui est de la Classification non supervisée ou automatique, les classes ne sont pas fixées au préalable mais elles sont déterminées progressivement.

La classification non supervisée (clustering) comprend la classification non hiérarchique (ou par partitionnement ou classification plat) qui construit les classes sans lien hiérarchique entre les groupes [32] et la classification hiérarchique. [20], [40].

La classification non hiérarchique comprend la méthode de centres mobiles (k-means) et la méthode des nuées dynamiques. Dans la méthode de k-means, on recalcule le barycentre à chaque fois qu'un nouvel individu est introduit dans le groupe et un centre de groupe est représenté par un point. Tandis que dans la méthode de nuées dynamiques, qui généralise la première méthode, le centre de chaque classe n'est plus défini par un point (barycentre) mais un noyau de points (d'individus).

La méthode de centres mobiles (k-means) est mieux adaptée pour de très grands nombres d'individus et sa réalisation comprend les étapes ci-après consistant à : (1) choisir une métrique pour le calcul des distances, définir un nombre k de classes et choisir au hasard k individus comme centres initiaux des classes. (2) calculer les distances entre chaque individu restant et chaque centre initial puis affecter chaque individu au centre qui lui est le plus proche afin de trouver k groupes correspondant au k centres choisis. (3) déterminer les barycentres de ces k groupes qui constituent les k nouveaux centres puis calculer les distances entre ces derniers et tous les individus afin d'affecter chaque individu au centre (barycentre) qui lui est le plus proche et former k groupes des individus les plus proches. (4) arrêter l'algorithme si les centres ne changent plus ou encore si la variance intra-classe ne diminue plus ou la variance interclasse n'augmente, sinon revenir à l'étape (3) pour continuer.

Quant à la Classification hiérarchique, elle regroupe les individus par la construction d'une hiérarchie. Elle comprend la Classification Ascendante Hiérarchique (CAH) (ou Classification par agglomération) et la Classification Descendante Hiérarchique (CDH) (ou par division). La CAH permet de regrouper les individus en un certain nombre de

classes ressorties à partir d'une hiérarchie de partitions. La CDH regroupe les individus en classes en scindant, à chaque étape, un groupe choisi en deux autres [13].

La représentation en arbre hiérarchique et la procédure de sa découpe sont les mêmes, tant par l'approche "Ascendante" que par l'approche "Descendante". Toutefois, la revue de la littérature est très moins nourrie quant à ce qui est de la CDH.

Dans le cas où l'étude porte sur un très grand nombre d'individus, les méthodes non hiérarchique et les méthodes hiérarchiques sont utilisées ensemble d'où l'expression : Méthodes mixtes. En effet, dans le cas où le nombre d'individus est trop élevé, on fait une partition par des méthodes de partitionnement afin de retrouver un nombre réduit de classes d'individus en vue de la construction d'une CAH tandis que dans le cas de nombre de variables est trop élevé, on réduit le nombre avec une ACP puis on procède par la CAH (CAH indirecte).

Nous synthétisons la subdivision de la classification comme suit :

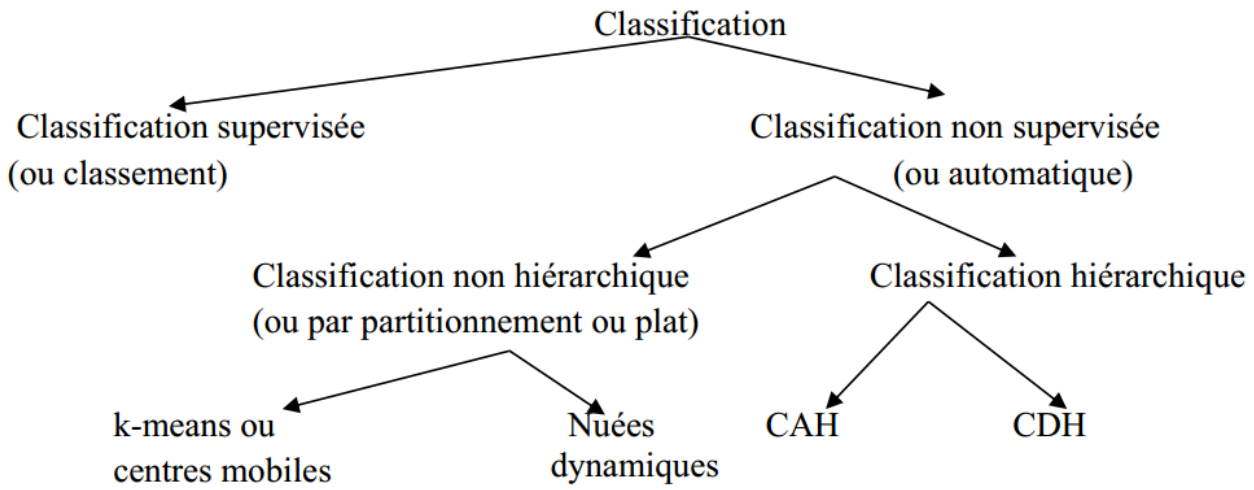

Figure 2.1 – Subdivision des méthodes de la classification
Source :Les notions précédentes

Nous nous intéressons aux données quantitatives pour une CAH réalisée sur les individus qui sont, dans notre cas, les 24 communes de la ville-province de Kinshasa en leur qualité d'ETDs. La réalisation de la CAH passe par des étapes bien connues.

2.2. Réalisation d'une CAH par des données quantitatives

La CAH se réalise par les trois grandes étapes suivantes consistant à : (1) calculer (après avoir constitué le tableau des données) les distances entre individus deux à deux et regrouper les deux individus les plus proches (1er regroupement). (2) calculer la distance entre le groupe nouvellement constitué et les n-2 individus isolés restants en utilisant un critère (un indice) d'agrégation choisi. Ceci permet de former le 2è regroupement en

agrégeant les deux objets (individus ou groupe d'individus) les plus proches. On répétera cette procédure jusqu'à ce que tous les individus se retrouvent regroupés dans un même groupe. Cette étape fait de la CAH une méthode itérative c'est-à-dire qui répète une action. (3) représenter graphiquement la hiérarchie de partitions obtenue à travers un dendrogramme, à partir des regroupements successifs déterminés à l'étape (2) [13], [20].

Donc les résultats d'une CAH sont une hiérarchie de partitions et non une partition. Toutefois, pour obtenir une partition permettant de ressortir les différentes classes au nombre voulu, il faut découper l'arbre à un certain niveau choisi. On peut aller plus loin en interprétant la partition obtenue après découpe du dendrogramme. Il convient, dans la suite, de détailler les étapes de la CAH citées ci-dessus.

2.2.1. Constitution du tableau des données et sa transformation éventuelle

Soit à classifier n individus $X_1, \ldots, X_i, \ldots, X_n$ (notés aussi $1, \ldots, i, \ldots, n$) sur base des m variables $Y_1, \ldots, Y_j, \ldots, Y_m$. La valeur de l'individu X_i pour la variable Y_j étant X_{ij}, alors le tableau des données se présente comme suit :

Tableau 2.1 – Tableau des données (Classification)

	$Y_1 \ldots$	$Y_j \ldots$	Y_m
X_1			
$\ldots$		$\vdots$	
X_i		X_{ij}	
$\ldots$		$\vdots$	
X_n			

Source : Référence au tableau 1.1.

La CAH ne passe pas nécessairement par la transformation des données. En effet, lorsque les variables sont homogènes, le tableau de données peut être directement utilisé pour le calcul des distances entre les individus deux à deux. Par contre, si elles sont hétérogènes, il convient de les transformer en données réduites pour éliminer l'influence des unités de mesure et/ou en données centrées pour ramener l'orige des axes au centre de gravité du nuage des points. Cette transformation permet une meilleure visibilité des points lors de la représentation graphique.

Nous soutenons que, la transformation des variables hétérogènes en données réduites résout le problème d'injustice dans la répartition de ressource, du fait qu'il anéantit l'influence des unités de mesure.

Une répartition faite sur base des variables ayant des unités de mesure différentes (Exemples : km^2 et CDF) sans pouvoir les réduire au préalable ne peut être juste. Est-il juste d'attribuer une part égale aux individus A et B qui vérifient pour deux variables retenues (superficie et production), les valeurs respectives 3 km^2 et 102 CDF, et 100 km^2 et 5 CDF, sous prétexte qu'ils ont la même valeur totale : 105 ? La réponse est sans doute non.

2.2.2. Calcul des distances entre les individus deux à deux

Cette étape de la CAH consiste à calculer les distances entre individus deux à deux conduisant au regroupement de deux individus ayant la plus petite distance (1er regroupement). Il convient donc de faire le choix d'un indice de distance qui permettra de conclure sur la ressemblance entre les individus ([3], pp. 29-30), [35].

Définition 2. 2 (Distance).

*Une distance sur un ensemble I des points est une application d :I*x*I*$\longrightarrow \mathbb{R}^+$ *vérifiant* $\forall X_i, X_j$ *et* $X_k \in I$ *les trois propriétés suivantes :*

- $d(X_i, X_j) = 0 \Leftrightarrow X_i = X_j$ *(La séparation)*
- $d(X_i, X_j) = d(X_j, X_i)$ *(La symétrie)*
- $d(X_i, X_k) \leq d(X_i, X_j) + d(X_j, X_k)$ *(L'inégalité triangulaire).*

2.2.2.1. Différentes expressions de distances

Il existe différentes expressions de distances [13], il s'agit de :

1) Distance de Manhattan (du city-block)

Cette distance est donnée par la fonction :

$$d(X_i, X_k) = \sum_{j=1}^{m} |X_{ij} - X_{kj}| \tag{2.1}$$

C'est **la 1- distance**.

2) Distance euclidienne

Elle est donnée par :

$$d(X_i, X_k) = \sqrt{\sum_{j=1}^{m} (X_{ij} - X_{kj})^2} \tag{2.2}$$

C'est **la 2-distance**.

3) Distance de Minkowski (distance à la puissance)

Elle est donnée par l'expression :

$$d(X_i, X_k) = \sqrt[p]{\sum_{j=1}^{m} |X_{ij} - X_{kj}|^p} \tag{2.3}$$

C'est la p-distance. Si p = 2, cette distance équivaut à la distance euclidienne [13]

4) Distance de Tchebychev

Elle est donnée par :

$$d(X_i, X_k) = lim_{p\to\infty} \sqrt[p]{\sum_{j=1}^{m} |X_{ij} - X_{kj}|^p} = sup_{1\leq j\leq m} |X_{ij} - X_{kj}| \tag{2.4}$$

C'est l'∞- distance.

5) Distance euclidienne au carré

Elle est donnée par :

$$d(X_i, X_k) = \sum_{j=1}^{m} (X_{ij} - X_{kj})^2 \tag{2.5}$$

6) Distance de Canberra

La distance de Canberra ([69], p. 15) est donnée par l'expression :

$$d(X_i, X_k) = \sum_{j=1}^{m} \frac{|X_{ij} - X_{kj}|}{|X_{ij}| + |X_{kj}|} \tag{2.6}$$

ou encore

$$d(X_i, X_k) = \sum_{j=1}^{m} \frac{|X_{ij} - X_{kj}|}{|X_{ij} + X_{kj}|} \qquad [13] \tag{2.7}$$

2.2.2.2. Tableau de distances

Après choix d'une mesure de distance et calcul des distances entre les individus deux à deux, pour permettre de retrouver plus facilement la plus petite de toutes ces distances et par conséquent, regrouper les individus concernés, on les place dans un tableau (de distances).

2.2.2.3. Nombre de distances à calculer

Pour ne pas omettre de calculer la distance entre un certain nombre d'individus, il importe de se rendre compte de nombre de distances à calculer entre les individus. Il est déterminé par la formule suivante :

Proposition 2. 1 (Nombre de distances à calculer)**.**

Le nombre N de distances à calculer entre n individus donnés est déterminé par la formule :

$$N = \frac{n(n-1)}{2} \tag{2.8}$$

Démonstration. Le nombre de distances entre n individus deux à deux correspond à la combinaison de n éléments pris deux à deux du fait que ces distances diffèrent entre elles par leur nature et non par leur ordre : d(A, B) = d(B, A), par exemple.

Ainsi,

$$N = C_n^2 = \frac{n!}{2!(n-2)!} = \frac{n(n-1)(n-2)!}{2!(n-2)!} = \frac{n(n-1)}{2}$$

□

2.2.3. *Calcul de distance entre deux groupes (Choix d'un indice d'agrégation)*

Le calcul de distance entre deux groupes d'individus se fait en utilisant un indice d'agrégation qu'il faut choisir parmi tant d'autres. Un de ces groupes peut être constitué d'un seul élément (Cas d'un individu isolé c'est-à-dire qui n'est pas encore regroupé).

Deux groupes les plus proches sont fusionnés en un nouveau groupe pour l'itération en cours. Cette opération continuera jusqu'à retrouver un seul groupe contenant tous les individus. Les individus isolés conservent leurs distances avec d'autres individus ou groupes déjà calculées à l'étape précédente ([3], pp. 38-39), ([69], p. 15).

Les différents indices d'agrégation (ou méthodes d'agrégation) sont les suivants :

1) Distance minimale ou écart simple ou Saut minimum

C'est la méthode du plus proche voisin. Pour cet indice, la distance entre deux groupes I_1 et I_2, est la plus petite distance parmi toutes les distances calculées entre les

individus respectifs X_i et X_k de ces deux groupes pris deux à deux. Elle est déterminée à partir de la formule :

$$d(I_1, I_2) = min_{(X_i,X_k)\in I_1 X I_2} d(X_i, X_k) \tag{2.9}$$

2) Distance maximale ou écart complet ou Diamètre

C'est la méthode du voisin le plus éloigné. Ici, la distance entre deux groupes I_1 et I_2, est la plus grande distance parmi toutes les distances calculées entre les individus respectifs X_i et X_k de ces deux groupes pris deux à deux. Elle est donnée par la formule :

$$d(I_1, I_2) = max_{(X_i,X_k)\in I_1 X I_2} d(X_i, X_k) \tag{2.10}$$

3) Distance moyenne ou écart moyen

C'est la méthode de distance moyenne. Pour cet indice, la distance entre deux groupes I_1 et I_2, est la distance moyenne de toutes les distances calculées entre les individus respectifs de ces deux groupes pris deux à deux. Elle est déterminée à partir de la formule :

$$d(I_1, I_2) = \frac{1}{n_1.n_2}\sum_{i=1}^{n_1}\sum_{k=1}^{n_2} d(X_i, X_k) \tag{2.11}$$

Où n_1 (respectivement n_2) est le nombre d'individus dans I_1 (respectivement dans I_2).

4)Distance médiane

Elle consiste d'abord à calculer toutes les distances entre les individus respectifs de deux groupes étudiés et ensuite à déterminer leur valeur médiane c'est-à-dire la valeur qui partage la distribution (série ordonnée) de ces distances en deux parties égales.

5) Distance des barycentres

Pour cet indice d'agrégation, la distance entre deux groupes I_1 et I_2 est celle qui est définie entre leurs barycentres respectifs G_1 et G_2 :

$$d(I_1, I_2) = d(G_1, G_2) \tag{2.12}$$

6) Méthode de Ward

La distance (ou l'écart) de Ward entre ces deux groupes est donnée par la formule suivante :

$$d(I_1, I_2) = \frac{n_1.n_2}{n_1 + n_2} d^2(G_1, G_2) \tag{2.13}$$

où G_1 (respectivement G_2) est le barycentre du groupe I_1 (respectivement I_2) d'effectif n_1 (respectivement n_2).

La méthode de Ward prend en compte à la fois la dispersion à l'intérieur d'un groupe (inertie intraclasse) et la dispersion entre les groupes (inertie interclasse) en ce sens qu'elle diminue l'inertie intraclasse et augmente l'inertie interclasse. Elle est utilisée par défaut dans la plupart des logiciels statistiques avec la distance euclidienne. Toutefois, elle se justifie mieux lorsque la distance entre les individus est la distance euclidienne au carré.

Les différents indices d'agrégation présentés ci-dessus s'illustrent comme suit :

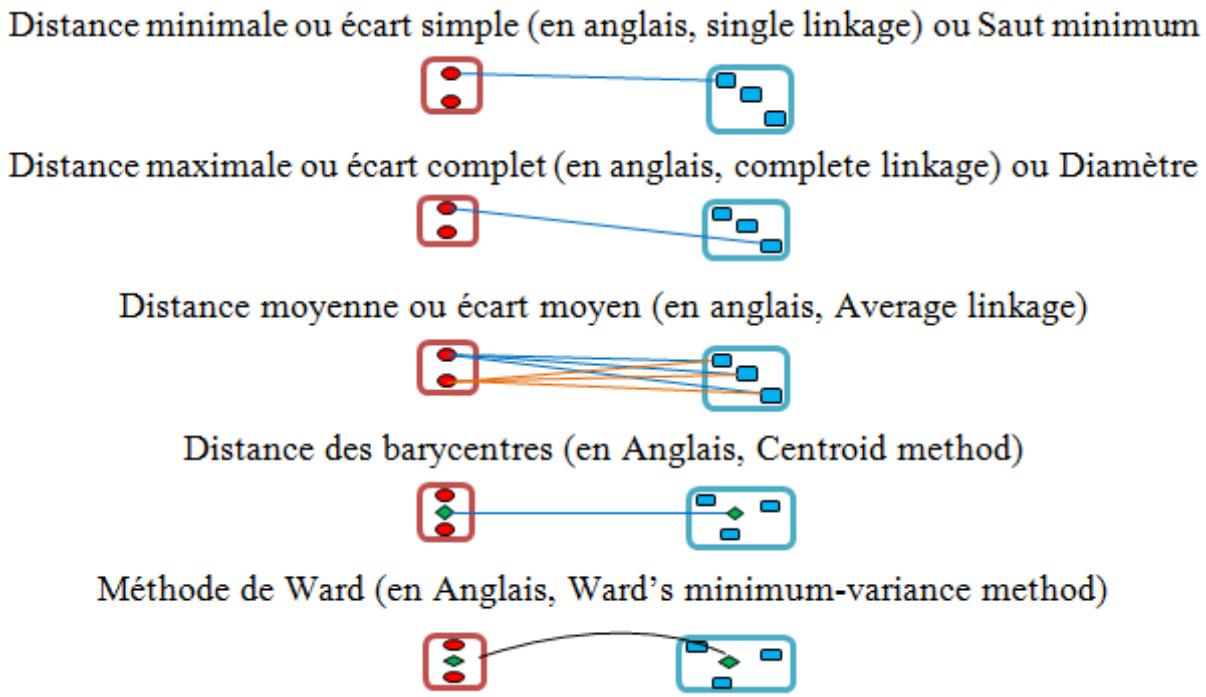

Figure 2.2 – Indices d'agrégation (Illustration).
Source : [13]

2.2.4. *Représentation graphique : Le dendrogramme.*

Après l'agrégation vient la représentation graphique d'une Classification Ascendante Hiérarchique appelée dendrogramme (ou arbre de classification ou arbre hiérarchique). Cet arbre est généralement binaire.

Il est positionné comme un arbre tourné vers le bas. En allant de bas en haut, l'on retrouve en bas les individus isolés qui constituent les éléments terminaux de l'arbre, puis viennent, des groupes d'individus, issus des regroupements, appelés nœuds de l'arbre. Chacun des nœuds constitue une partition ([4], p. 280), ([11], p.34), ([12], pp. 104, 106), ([14], p. 22), ([23], p.8), ([26], p. 389) et ([74], p. 5).

2.2.4.1. Hiérarchie

Définition 2. 3 (Hiérarchie).

Soit $\Gamma = \{I_1, \ldots, I_k, \ldots, I_q\}$ un ensemble de q groupes et $I = \{1, \ldots, i, \ldots, n\}$ un ensemble de n individus. Une hiérarchie (de clusters) est un ensemble Γ de groupes (clusters) tels que :

• Tout individu appartient à au moins un groupe (ou Tout groupe est non nul) : $i \in I_k(I_k \neq \emptyset), \forall i \in I, \exists I_k \in \Gamma$

• Pour deux groupes donnés de la hiérarchie, soit l'un est inclus dans l'autre soit les deux sont disjoints (On n'admet pas de chevauchement) : $I_k \subset I_{k'}$ $(ou I_{k'} \subset I_k)$ sinon $I_k \cap I_{k'}$=$\emptyset$, $\forall I_k$ et $I_{k'} \in \Gamma$

• Tout groupe est la réunion des groupes qui lui sont inclus : $I_{k'} \subset I_k \Rightarrow \cup I_{k'} = I_k$, $\forall I_{k'} \in \Gamma$.

Exemple 2. 1 (Une hiérarchie).

Une hiérarchie *représentée par* **un dendrogramme** *se présente comme suit :*

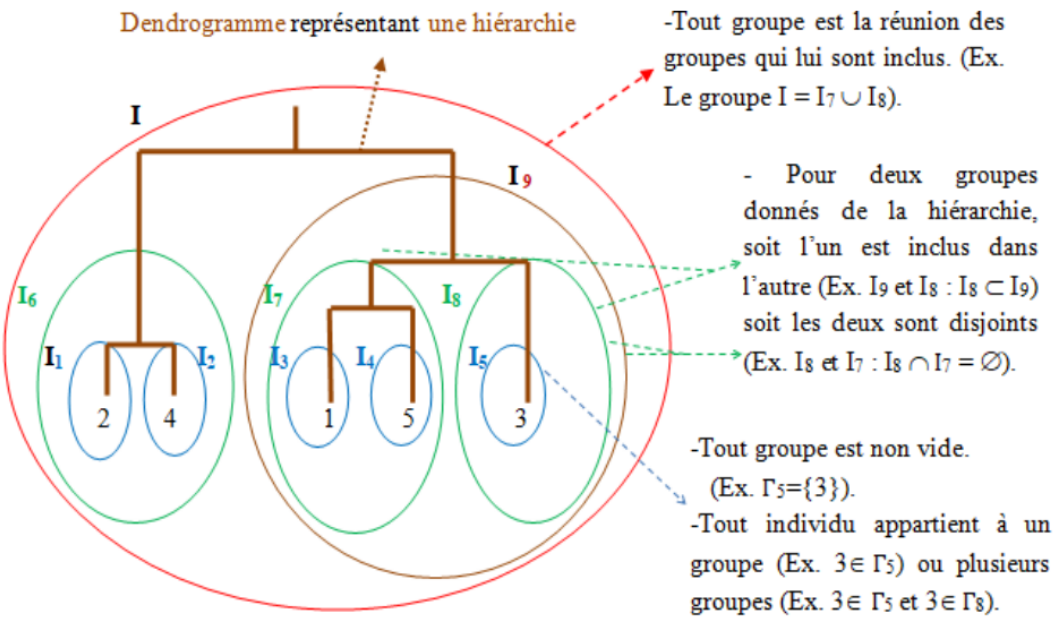

Figure 2.3 – Une Hiérarchie représentée par un Dendrogramme
Source : Notre conception à partir du dendrogramme et hiérarchie

Faisons remarquer que le résultat d'une CAH représenté par un dendrogramme n'est pas une partition de l'ensemble des individus. C'est donc une hiérarchie. Pour obtenir une partition qui sera constituée des classes, il faut couper l'arbre à un certain niveau.

2.2.4.2. Partition d'un ensemble

Définition 2. 4 (Partition d'un ensemble).

Soit $I = \{1, \ldots, i, \ldots, n\}$ un ensemble d'individus et $\mathcal{P}(I) = \{I_1, \ldots, I_k, \ldots, I_q\}$ l'ensemble des q parties de I. Une partition d'un ensemble I est une partie K de $\mathcal{P}(I)$ vérifiant les trois conditions suivantes :

• *Chaque partie est non vide (c'est-à-dire que chacune comprend au moins un individu) :*
$I_k \neq \emptyset, \quad \forall I_k \in K$

• *Les parties sont deux à deux disjointes c'est-à-dire qu'elles ne comprennent pas d'individu commun :* $I_k \cap I_{k'}$=$\emptyset$, $\quad \forall I_k, I_{k'} \in K$

• *Leur réunion forme l'ensemble I :* $I_1 \cup \ldots \cup I_k \cup \ldots \cup I_q, \quad \forall I_1, \ldots, I_k, \ldots, I_q \in K$

Exemple 2. 2 (Partition).

Soit l'ensemble des individus $I = \{1, 2, 3, 4, 5\}$. Les parties I_1=$\{1\}$, I_2=$\{2, 3\}$, I_3=$\{4, 5\}$ de I forment une partition de I que nous représentons comme suit :

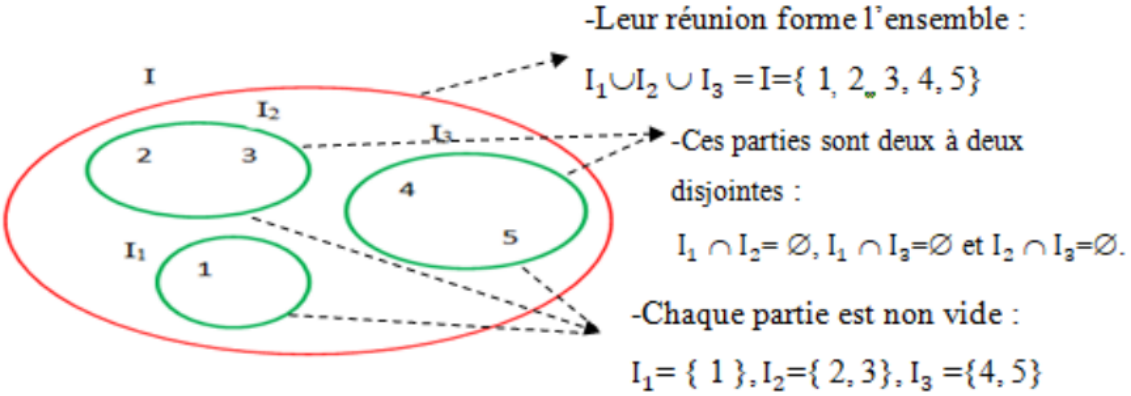

Figure 2.4 – Partition d'un ensemble
Source : Notre conception à partir de la définition d'une partition

2.2.4.3. Dendrogramme

Définition 2. 5 (Dendrogramme).

Un dendrogramme est le résultat de la classification. C'est une hiérarchie de partitions qui indique l'ordre dans lequel les regroupements successifs ont été effectués ([24], p.90)

Sa construction part du niveau le plus bas au niveau le plus élevé. Au plus bas niveau sont repris les individus simples tandis qu'au niveau le plus élevé les individus sont rassemblés dans un seul groupe.

Il est constitué des branches (qui se trouvent à la base de l'arbre et aboutissent aux individus) et des nœuds (qui raccordent les branches entre elles ou encore les nœuds entre eux). Lorsque deux objets (individus ou groupes) groupés constituent un nœud, l'un sera appelé Ainé (celui à partir de la branche duquel commence le nœud) et l'autre Benjamin.

La hauteur à laquelle se trouvent reliés deux objets exprime la plus petite distance ayant permis le regroupement des objets concernés. Elle est appelée l'indice de niveau d'agrégation ou encore l'indice de niveau de nœud.

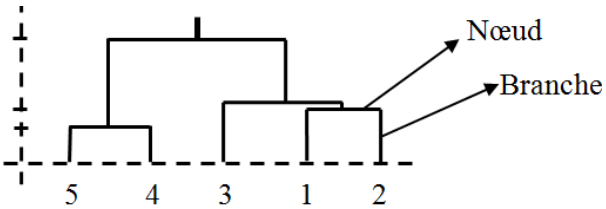

Figure 2.5 – Dendrogramme (Exemple)
Source : Logiciel R à partir de nos données

Rappelons que lorsqu'il est question d'agréger les individus, ceux ayant la plus petite distance sont regroupés. Ceci permet d'associer à chaque groupe I_i un nombre $V(I_i)$ qui est le niveau de noeud. Sur le dendrogramme, cette distance (ce nombre) est l'indice de niveau d'agrégation.

Cet indice est croissant pour la relation d'inclusion entre les groupes c'est-à-dire si un groupe I_i est inclus dans un groupe I_k alors l'indice $V(I_i)$ est inférieur à l'indice $V(I_k)$. On dit alors que l'ensemble des groupes forme une hiérarchie indicée. L'indice $V(I_i)$ fournit une nouvelle distance δ entre individus appelée distance ultramétrique et définie comme suit :

Définition 2. 6 (Distance ultramétrique).

Soit I un ensemble des points. Une distance ultramétrique [27] sur I est une application $\delta : I\text{x}I \rightarrow \mathbb{R}^+$ *vérifiant,* $\forall\ X_i, X_j, X_k \in$ *I, les trois propriétés suivantes :*

- $\delta(X_i, X_j) = 0 \Leftrightarrow X_i = X_j$ *(Séparation)*
- $\delta(X_i, X_j) = \delta(X_j, X_i)$ *(Symétrie)*
- $\delta(X_i, X_k) \leq max\{\delta(X_i, X_j), \delta(X_j, X_k)\}$ *ou encore* $\delta(X_i, X_j) = \delta(X_j, X_k) > \delta(X_i, X_k)$ *(Inégalité ultratriangulaire).*

La distance $\delta(X_i, X_j)$ entre les individus X_i et X_j est l'indice correspond au plus petit groupe contenant à la fois X_i et X_j.

Cette inégalité ultratriangulaire [82] signifie simplement qu'étant donné un triangle, la longueur d'un côté est inférieure sinon égale à la plus grande des longueurs des deux autres côtés. A cet effet, lorsqu'une distance vérifie l'inégalité ultratriangulaire, tous les triangles sont isocèles comme le montre la figure suivante :

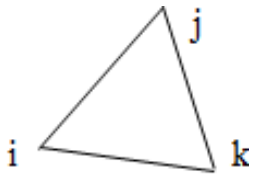

Figure 2.6 – Représentation d'une distance ultramétrique
Source : [26]

Plus précisément, dans le cas de trois individus 1, 2 et 3 ; si les individus 1 et 3 étaient regroupés en premier lieu et que l'on calcule la distance entre ce groupe et l'individu 2, alors les distances entre 1 et 2, et entre 2 et 3 seront égales et les deux, plus grande que celle calculée entre 1 et 3. En faisant allusion aux côtés d'un triangle isocèle, la mesure des côtés égaux [1,2] et [2,3] est plus grande que celle de la base [1,3].

En effet, $\delta(1,3) \leq max\{\delta(1,2); \delta(2,3)\}$ ou encore $\delta(1,2) = \delta(2,3) > \delta(1,3)$.

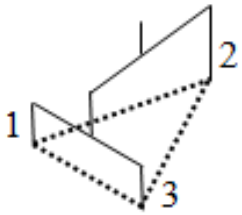

Figure 2.7 – Représentation en mobile d'une hierarchie
Source : [26]

Le dendrogramme aide aussi à déterminer les distances ultramétriques entre les individus. Il suffit de regarder le niveau où les individus sont regroupés. Un arbre de Classification hiérarchique correspond donc à une matrice de distances ultramétriques.

2.2.5. *Découpe du dendrogramme*

1. Mesure de dispersion des individus : Inertie

Définition 2. 7 (Inertie).

L'inertie est une mesure de dispersion du nuage des points.

Définition 2. 8 (Nuage des points).

Un nuage des points est l'ensemble des points représentant une série statistique et qui sont graphiquement représentés.

On parle de l'inertie intra-classe, l'inertie interclasse et l'inertie totale. Ces éléments sont présentés avec plus de détail dans ([24], pp. 86-87).

Soit à classifier les n individus $1, .., i, \ldots, n$ d'une population. Considérons le nuage I de n points (issus de n individus), réparti en m groupes $I_1, \ldots, I_k, \ldots, I_m$ d'effectifs respectifs $n_1, \ldots, n_k, \ldots, n_m$. Notons par G le barycentre du nuage I ; $G_1, \ldots, G_k, \ldots, G_m$ les barycentres respectifs de chacun de ces groupes.

2.2.5.1. Inertie intra-classe

Définition 2. 9 (Inertie intra-classe).

L'inertie intra-classe ou intragroupe est une mesure de dispersion des n_k individus autour du centre de gravité G_k du groupe I_k auquel ils appartiennent.

$$I_{intra} = \frac{1}{N}\sum_{i=1}^{n_k}\sum_{i\in I_k} d^2(i, G_k) = \frac{1}{N}\sum_{i=1}^{n_k}\sum_{i\in I_k} \|i - G_k\|^2 \tag{2.14}$$

Considérant $X_i \in \mathbb{R}^m$ (ligne i de la matrice des données) comme le vecteur décrivant l'individu i, G_k le centre de gravité du groupe I_k, une inertie intra-classe s'illustre comme suit.

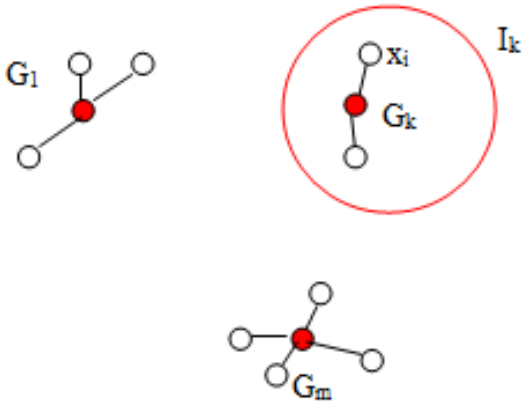

Figure 2.8 – Illustration de l'inertie intra-classe
Source : [24]

2.2.5.2. Inertie inter-classe

Définition 2. 10 (Inertie inter-classe).

L'inertie interclasse ou intergroupe (pour q groupes) est une mesure de dispersion des barycentres G_k du k^e groupe autour du barycentre G de la population.

$$I_{inter} = \sum_{k=1}^{q}\frac{n_k}{N} d^2(G_k, G) = \sum_{k=1}^{q}\frac{n_k}{N}\|G_k - G\|^2 \tag{2.15}$$

Où $\frac{n_k}{N}$ est le poids du k^e groupe.

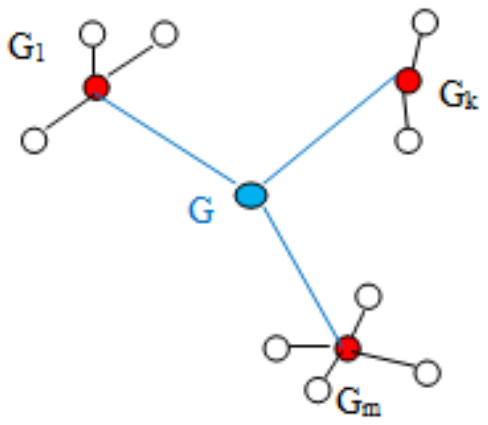

Figure 2.9 – Illustration de l'inertie inter-classe
Source : [24]

2.2.5.3. Inertie totale

Définition 2. 11 (Inertie totale).

L'inertie totale du nuage est la mesure de dispersion des N individus autour du centre de gravité G de l'ensemble I des individus.

$$I_{tot} = \frac{1}{N}\sum_{i=1}^{N} d^2(i, G) = \frac{1}{N}\sum_{i=1}^{N} \|i - G\|^2 \tag{2.16}$$

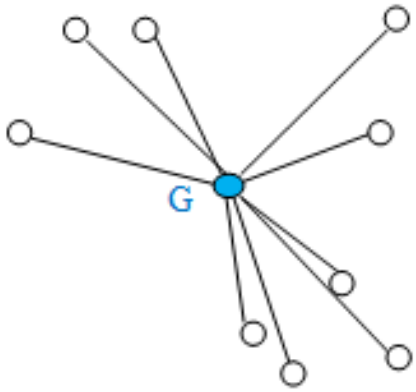

Figure 2.10 – Illustration de l'inertie totale
Source : [24]

On peut remarquer que

$$I_{tot}(I) = \sum_{k=1}^{q} \sigma_k^2 \tag{2.17}$$

Avec σ_k^2 la variance de X_k .

En particulier, si la série est constituée de deux variables X_1 et de X_2, on a :

$$I_{tot}(I) = \sigma_1^2 + \sigma_2^2 \tag{2.18}$$

où σ_1^2 et σ_2^2 sont respectivement les variances de X_1 et de X_2.

Théorème 2. 1 (Théorème de Huygens).

Soit G le barycentre d'un nuage des points $X_1, \ldots, X_i, \ldots, X_n$ et X un point quelconque de $\mathbb{R}^m$. L'inertie du nuage au point X est donnée par :

$$I_X = I_G + d^2(G, X) \tag{2.19}$$

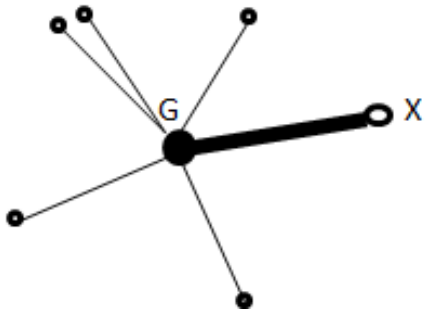

Figure 2.11 – Illustration du théorème de Huygens
Source : Les théories précédentes.

Démonstration. Montrons que $I_X = I_G + d^2(G, X) = \sum_{i=1}^{N} p_i \|X_i - G\|^2 + \|G - X\|^2$.

En effet, $I_X = \sum_{i=1}^{N} p_i d^2(X, X_i) = \sum_{i=1}^{N} p_i \|X_i - X\|^2 = \sum_{i=1}^{N} p_i \|X_i + G - G - X\|^2$

$= \sum_{i=1}^{N} p_i((X_i - G)^2 + 2(G - X)(X_i - G) + (G - X)^2)$

$= \sum_{i=1}^{N} p_i \|X_i - G\|^2 + \sum_{i=1}^{N} p_i 2\langle G - X | X_i - G\rangle + \sum_{i=1}^{N} p_i \|(G - X)\|^2$

$= \sum_{i=1}^{N} p_i \|X_i - G\|^2 + 2(G - X)\sum_{i=1}^{N} p_i (X_i - G) + \sum_{i=1}^{N} p_i \|(G - X)\|^2$

$= \sum_{i=1}^{N} p_i \|X_i - G\|^2 + 2(G - X)(\sum_{i=1}^{N} p_i X_i - G\sum_{i=1}^{N} p_i) + \sum_{i=1}^{N} p_i \|(G - X)\|^2$

$= \sum_{i=1}^{N} p_i \|X_i - G\|^2 + 2(G - X)(G - G) + \sum_{i=1}^{N} p_i \|(G - X)\|^2$

$= \sum_{i=1}^{N} p_i \|X_i - G\|^2 + \|(G - X)\|^2 = I_G + d^2(G, X)$ □

L'inertie du nuage au point X est donc minimale quand X = G.

Proposition 2. 2 (Décomposition d'inertie totale selon la Relation de Huygens).

L'inertie totale se décompose par la somme des inerties intra et interclasses

$$I_{tot} = I_{intra} + I_{inter} \tag{2.20}$$

Au fur et à mesure que les regroupements sont effectués, l'inertie intra-classe augmente et l'inertie interclasse diminue. Leur somme (I_{tot}) est constante.

La décomposition de l'inertie totale est illustrée comme suit :

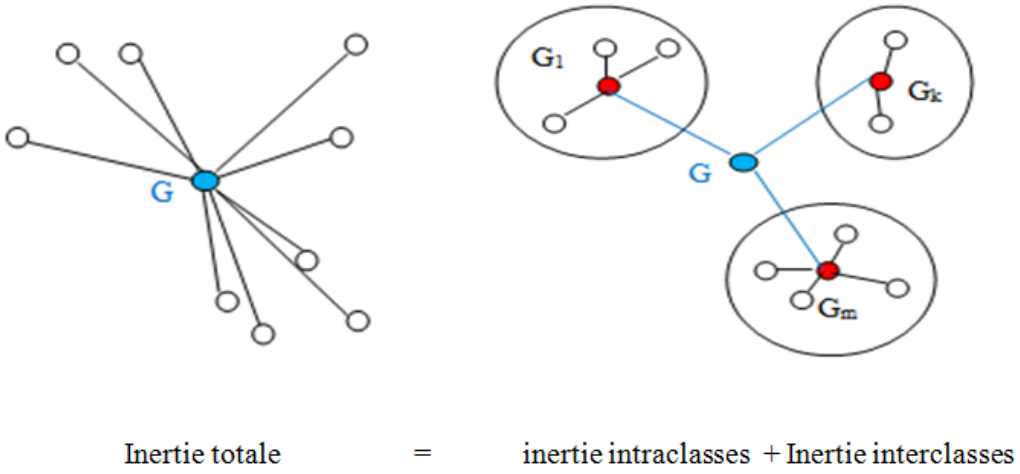

Figure 2.12 – Illustration de la décomposition de l'inertie totale
Source : [24]

Démonstration. ([24], p. 86)

$$I_{tot} = \frac{1}{N}\sum_{i=1}^{N} d^2(i,G) = \frac{1}{N}\sum_{i=1}^{N} \|i-G\|^2$$

$$= \frac{1}{N}\sum_{k=1}^{q}\sum_{i\in I_k} \|i-G\|^2 = \frac{1}{N}\sum_{k=1}^{q}\sum_{i\in I_k} \|i-G_k+G_k-G\|^2$$

$$= \frac{1}{N}\sum_{k=1}^{q}\sum_{i\in I_k} \|i-G_k\|^2 + \frac{1}{N}\sum_{k=1}^{q}\sum_{i\in I_k} \|G_k-G\|^2 \text{ (Théorème de Huygens)}$$

$$= \frac{1}{N}\sum_{k=1}^{q}\sum_{i\in I_k} d^2(i,G_k) + \frac{1}{N}\sum_{k=1}^{q}\sum_{i\in I_k} d^2(G_k,G)$$

$$I_{tot} = I_{intra} + I_{inter}$$ □

2.2.5.4. Obtention de la meilleure partition

Pour un même dendrogramme, on peut déterminer plusieurs partitions alors qu'on a besoin que d'une seule, la meilleure. Il se pose le problème de la qualité d'une partition. A cet effet, l'inertie joue un rôle important. Elle est une mesure de dispersion du nuage des points. Une partition est meilleure, si les individus d'une même classe sont les plus proches (l'inertie intra-classe est petite : la minimisation de l'inertie intra-classe) et ceux de deux classes différentes sont les plus éloignés (l'inertie inter-classe est grande : La maximisation de l'inertie inter-classe) ([4], p. 288), [23], pp. 10-11).

Pour obtenir une meilleure partition, on se sert de l'indice de niveau d'agrégation. On découpe l'arbre (le dendrogramme) au niveau où cet indice fait un saut important lorsqu'on passe d'une partition à une autre moyennant une droite horizontale. C'est à ce niveau qu'il faut couper le dendrogramme pour obtenir la meilleure partition.

Ceci conduit à la détermination des classes qui forment une partition. Le nombre de classes dépend aussi de l'appréciation du chercheur [13], ([54], p. 14).

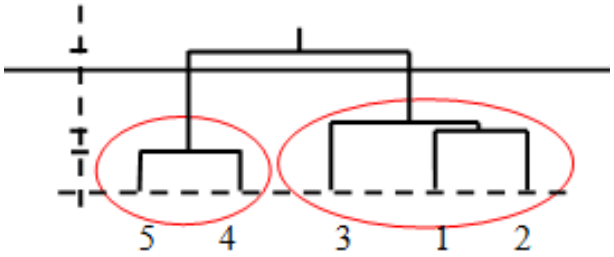

Figure 2.13 – Découpe du dendrogramme en deux classes (Exemple)
Source : Les théories précédentes

Dans cette illustration, le dendrogramme est découpé en deux classes : les individus 5 et 4 forment une classe, la 1ère, tandis que 3, 1, 2 la deuxième classe.

La réalisation de la Classification Ascendante Hiérarchique conduit au dendrogramme qui est découpé en vue de l'obtention de la partition des classes des individus. Cette partition devra être interprétée.

2.2.6. Interprétation d'une partition

Après avoir déterminé la meilleure partition après découpe du dendrogramme, on procède par son interprétation. Celle-ci revient à caractériser (ou décrire) les classes par leurs individus ou par leurs variables.

Il sera donc question de calculer les paramètres de position et de dispersion pour chaque classe : la moyenne, la variance, l'écart-type de chaque variable mais aussi de déterminer les individus les plus typiques (Parangons et extrêmes), les variables les plus importantes et de représenter graphiquement les classes des individus.

Nous trouvons que les résultats de la CAH peuvent être utilisés au niveau de l'interprétation pour calculer, à partir des valeurs totales des individus, les proportions et parts de ces derniers dans la répartition d'une ressource. C'est donc un nouvel élément dans l'interprétation.

2.2.6.1. Paramètres de position et de dispersion

Ce sont les paramètres de position et de dispersion : Il s'agit de la moyenne, la variance et l'écart-type. Un paramètre est une caractéristique de la population tandis qu'une statistique est celle d'un échantillon. Les paramètres et les statistiques sont généralement notés différemment ([3], pp. 12, 14) :

1) Paramètres d'une population

Soit N la taille d'une population (si elle est finie), on a :

• **Moyenne**

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2.21}$$

• **Variance**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2 \tag{2.22}$$

• **Ecart-type**

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2} \tag{2.23}$$

C'est l'écart-type de Pearson.

2) Statistiques d'un échantillon

Soit n la taille d'un échantillon, on a :

• **Moyenne**

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.24}$$

• **Variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{2.25}$$

• **Ecart-type**

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2} \tag{2.26}$$

C'est l'écart-type standard.

2.2.6.2. Description des classes par les individus

La description d'une classe par les individus se fait par la détermination des individus les plus typiques de la classe, il s'agit des parangons et des extrêmes ([24], p. 91).

1) Parangon

Définition 2. 12 (Parangon).

Le parangon d'une classe est l'individu le plus proche du barycentre (centre de gravité) de cette classe.

L'utilité du parangon d'une classe s'explique par le fait qu'il convient mieux pour remplacer cette classe étant donné qu'il est un individu réel contrairement au barycentre qui est un individu fictif.

Pour déterminer le parangon d'une classe, il faut calculer le barycentre de cette classe et la distance entre celui-ci et chaque individu. L'individu qui aura la plus petite distance sera retenu comme le parangon de la classe.

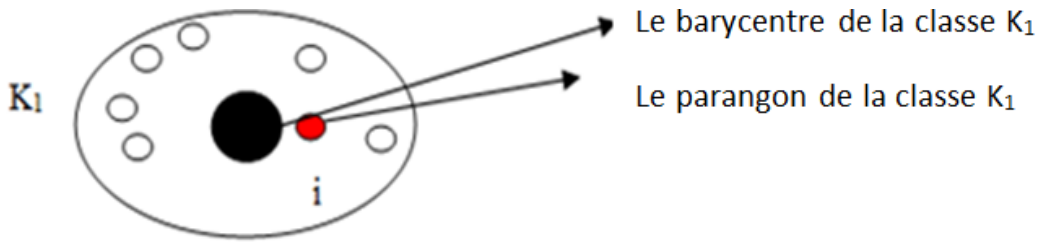

Figure 2.14 – Illustration du parangon d'une classe
Source : Notre conception à partir de la définition

2) Extrême

Définition 2. 13 (Extrême).

L'extrême d'une classe est l'individu le plus éloigné des barycentres des autres classes.

Pour le déterminer, il faut, d'abord calculer les distances entre les différents barycentres des classes pour retrouver, pour chacune, celle qui lui est la plus proche. Ensuite calculer les distances entre le barycentre de la classe la plus proche et les individus de la classe dont on cherche l'extrême. L'individu ayant la plus grande distance de ce barycentre est l'extrême.

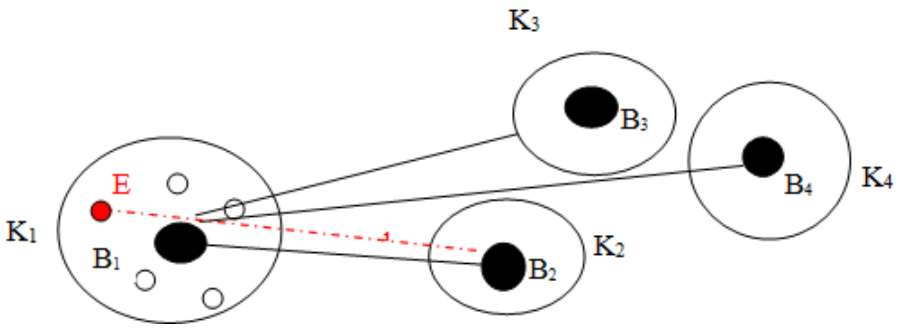

Figure 2.15 – Illustration de l'extrême d'une classe
Source : Notre conception à partir de la définition

Le point E est l'extrême de la classe K_1, les gros points (B_1, B_2, B_3, B_4) au milieu des classes (K_1, K_2, K_3, K_4) sont les barycentres respectifs des classes. La classe K_2 est la plus proche de la classe K_1, E est donc l'individu de K_1 ayant la plus grande distance d_1 avec K_2.

Les extrêmes sont des individus auxquels il faut porter une attention particulière du fait qu'ils soient éloignés des autres classes.

2.2.6.3. Description des classes par les variables

Pour caractériser une classe par les variables, on procède par un test de conformité. C'est un test dit « à un échantillon » qui consiste à vérifier si l'échantillon (dans notre cas, la classe) est extrait de la population d'où il est ressorti c'est-à-dire vérifier si la classe peut être considérée comme un échantillon représentatif de la population (Acceptation de l'hypothèse nulle) pour une variable donnée, auquel cas, celle-ci n'est pas plus importante pour cette classe. Si, par contre, la classe ne constitue pas un échantillon représentatif de la population pour une variable (Rejet de l'hypothèse nulle) alors celle-ci est plus importante pour cette classe et par conséquent la caractérise. Les documents suivants fournissent des éléments intéressants sur l'inférence statistique : [13], [18], [33] et [34].

Il est question de supposer que, pour la variable Y_j, la classe K est un échantillon de moyenne (observée) $\overline{Y}_{j,k}$, appartenant à une population (de référence) de moyenne (théorique) $\overline{Y}_j$ et variance σ_j^2. Il faut par la suite, comparer (égaler) ces deux moyennes (Hypothèse nulle). C'est le cas d'un test bilatéral. S'il était unilatéral, on utiliserait la supériorité ou l'infériorité. Dans le cas où l'hypothèse nulle est rejetée, on conclut que la variable caractérise la classe.

Deux cas se présentent, celui où la variance de la population est connue (test Z) et celui où elle n'est pas connue (test t). Nous sommes dans le cas où la variance σ_j^2 de la

population (de référence) est connue. En ce qui nous concerne, le test sera bilatéral pour le cas où la variance de la population est connue.

De ce fait, considérant pour une variable donnée Y_j, la moyenne $\overline{Y}_{j,k}$ de la classe K (échantillon) ; la moyenne $\overline{Y}_j$, et la variance σ_j^2 de la population (de référence) et n l'effectif de la population.

On procède par suite comme suit :

1) Poser l'hypothèse : L'hypothèse testée est $H_0 : \overline{Y}_{j,k} = \overline{Y}_j$ (Hypothèse nulle) contre $H_1 : \overline{Y}_{j,k} \neq \overline{Y}_j$ (Hypothèse alternative).

2) Calculer la valeur test de la variable aléatoire Z : On pose

$$Z_{obs} = \frac{|\overline{Y}_{j,k} - \overline{Y}_j|}{\sqrt{\frac{\sigma_j^2}{n}}} \tag{2.27}$$

La moyenne empirique $\overline{Y}_{j,k}$ suit la loi normale $N(\overline{Y}_j, \frac{\sigma_j}{\sqrt{n}})$ et Z suit donc une loi normale $N(0,1)$, de moyenne 0 et de variance 1.

3) Calculer la valeur critique : Z_α lue sur la table de la loi normale centrée réduite pour un risque d'erreur $\alpha = 0,05$; $0,01$ ou $0,001$ (c'est un seuil de significativité). (Cf. Annexe 5 et 6).

4) Décider sur l'acceptation ou le rejet de l'hypothèse nulle : Si $|Z_{obs}| > Z_\alpha$ (ou plus simplement si $Z_{obs} > 1,96$ au seuil de 0,05 (Cf. Annexe 5)), l'hypothèse nulle est rejetée et par conséquent la variable caractérise la classe. Mais si $Z_{obs} \leq Z_\alpha$, l'hypothèse nulle est acceptée et par conséquent la variable ne caractérise pas la classe.

La région de rejet de H_0 est construite en cherchant Z_α tel que

$P(|Z_{obs}| > Z_\alpha) = \alpha \Leftrightarrow P(Z_{obs} > Z_\alpha \text{ ou } Z_{obs} < -Z_\alpha) = \alpha$

$\Leftrightarrow P(Z_{obs} > Z_\alpha) + P(Z_{obs} < -Z_\alpha) = \alpha \Leftrightarrow P(Z_{obs} > -Z_\alpha) = \frac{\alpha}{2}$

$\Leftrightarrow F(-Z_\alpha) = \frac{\alpha}{2} \Leftrightarrow 1 - F(Z_\alpha) = \frac{\alpha}{2}$

$\Leftrightarrow F(Z_\alpha) = 1 - \frac{\alpha}{2} \Leftrightarrow F^{-1}(Z_\alpha) = 1 - \frac{\alpha}{2} = 0,9750$ pour $\alpha = 0,05$

D'après la table de la fonction de répartition inverse de la loi normale, on déduit que $Z_\alpha = 1,96$. A cet effet, il faut regarder dans la table de la fonction de répartition de la loi normale (Annexe 5), la case comprenant la valeur 0,9750 puis considérer la valeur de z sur la première colonne (1,9) et la décimale (0,06) sur la première ligne formant cette case et conclure que $Z_\alpha = 1,9 + 0,06 = 1,96$. Puis que pour $\alpha = 0,05$; $Z_{0,05} = 1,96$;

on peut conclure pour $\alpha = 0,01$ (deux fois 0,05) et $Z_{0,01} = 3,92$ (deux fois 1,96) et enfin, pour $\alpha = 0,001$ (0,01 fois 10^{-1}) et $Z_{0,001} = 0,392$ (3,92 fois 10^{-1}) [13], ([23], pp. 32-33).

5) Confirmer le résultat avec la valeur p : Le résultat peut être confirmé avec la valeur p (p-valeur ou p-value) notée pval. C'est la probabilité, sous H_0, d'observer une statistique aussi extrême (grande) que la valeur observée sur l'échantillon. On compare pval et $\alpha(= 0,05)$, afin d'accepter ou de rejeter H_0. Si $pval \leq \alpha$, l'hypothèse nulle est rejetée et la variable caractérise la classe. Mais si $pval > \alpha$; on accepte l'hypothèse nulle et la variable ne caractérise pas la classe.

La p valeur correspond à la quantité

$$P(|Z| \geq Z_{obs}) = 2P(Z \geq Z_{obs}) = 2(1 - F(Z_{obs})) \tag{2.28}$$

où F est une fonction de répartition.

La table de fonction de répartition de la loi normale centrée-réduite donne les valeurs de $F(Z_{obs})$. (Cfr Annexe 5).

Les variables les plus importantes, peuvent servir comme des variables explicatives au moment de la construction d'un modèle économique.

2.3. Conclusion du deuxième chapitre

La classification est une méthode qui forme avec l'Analyse Factorielle les deux familles de méthodes d'Analyse de données. Celle-ci s'occupe des études portant sur plusieurs variables. Nous avons opté pour la classification réalisée à partir des résultats de l'ACP.

La Classification est issue de l'*Anthropologie* et considérée actuellement comme une notion de Statistique utilisée dans presque tout le domaine de la vie : dans le domaine médical où elle est désignée par *nosologie*, dans le domaine commercial, en marketing,... Elle utilise les données quantitatives ou qualitatives et est appliquée sur les individus ou sur les variables. Nous avons opté pour une classification sur les individus appliquée aux données quantitatives.

Nous avons mis l'accent sur la Classification Ascendante Hiérarchique (CAH) qui est comptée parmi les méthodes de la classification non supervisée (ou automatique) hiérarchique. Ces dernières utilisent une hiérarchie de partition permettant de déterminer les classes dont le nombre n'est pas connu au départ. Ces classes aident pour la répartition

des ressources communes dans le cas où les individus n'ont pas contribué à la création de ces ressources dans le sens à réduire les inégalités entre eux, dans leurs classes respectives.

La CAH se réalise en suivant les étapes ci-après : 1) Constituer le tableau des données (éventuellement transformer les données de départ en données réduites si les variables sont hétérogènes) ; 2) choisir un indice de distance, calculer les distances entre les individus deux à deux et constituer le 1er regroupement ; 3) choisir un indice d'agrégation, calculer les distances entre le groupe nouvellement constitué et le reste de groupes et/ou individus isolés ; continuer cette opération jusqu'à ce que tous les individus forment un seul groupe ; 4) Représenter graphiquement le dendrogramme et le découper.

Après toutes ces étapes vient l'interprétation des résultats qui s'occupe de calcul des paramètres de position et de dispersion, de la description des classes. Cette dernière permet de déterminer les individus les plus typiques : Parangon et extrême, et les variables les plus importantes.

Nous précisons que la classification intervient dans la répartition des ressources par la présentation des individus dans leurs classes respectives permettant ainsi à ceux qui appartiennent à une même classe de se solidariser, du fait de leur rapprochement, dans le sens à réduire les inégalités entre eux.

Chapitre 3

Procédés de la répartition des ressources

Introduction

Nous voulons résoudre les problèmes d'injustice due à l'utilisation directe des données de départ des variables homogènes pour lesquelles la même unité de mesure est exprimée différemment pour chaque variable, l'utilisation directe des données de départ issues des variables hétérogènes sans pour autant les transformer au préalable et l'absence d'un mécanisme de partage permettant notamment la réduction des inégalités entre les individus.

Nous soutenons que les classes issues de la classification pourront être utilisées dans la répartition des ressources. En effet, dans le cas où les individus n'ont pas contribué pour créer la ressource à partager, les individus d'une même classe pourront se solidariser dans le sens à réduire les inégalités entre eux par un transfert des valeurs des individus plus riches aux plus pauvres suite à leur rapprochement. Les résultats de cette première réduction des inégalités (au niveau des classes) pourront ensuite être utilisés pour une deuxième réduction des inégalités entre les individus au niveau de la population du fait de leur appartenance à une même population. Les valeurs des individus issues de cette dernière réduction seront utilisées pour le partage des ressources données.

Nous mettons en place deux procédés de la répartition des ressources qui utilisent plusieurs variables. Le premier procédé, que nous avons nommé Procédé de la Répartition des Ressources Sans réduction des inégalités (PRRS), est adapté au cas où les individus (comme actionnaires) ont contribué pour créer la ressource à partage et n'admet pas la réduction des inégalités. Pour ce cas, le principe d'équité (brute) qui est l'un de principes du partage convient mieux. Le deuxième quant à lui concerne le cas où les individus n'ont pas contribué à la création de la ressource à partager. Il utilise les résultats de la classification, autorise la réduction des inégalités entre les individus. Nous l'avons nommé Procédé de la Répartition des Ressources à partir des résultats de la classification (PRRC). Pour ce cas, le principe d'équité est inadapté. Ceci nous amène à proposer le principe « d'équité réduite (ou corrigée) multidimensionnelle » du

fait qu'on utilise le principe d'équité après réduction des inégalités à partir de plusieurs variables.

Dans la suite, nous parlerons d'abord des concepts de base du partage (3.1) puis de la Répartition des Ressources Sans réduction des inégalités (3.2.) et enfin de la Répartition des Ressources avec réduction des inégalités (3.3.).

3.1. Concepts de base du partage

La notion de partage (répartition) regroupe un certain nombre de terminologies qu'il convient d'expliciter. Nous ferons autant pour d'autres éléments importants y afférents.

3.1.1. Ressource

Définition 3. 1 (Ressource).

Une ressource C est un ensemble fini d'objets (ou biens) physiques, quantités divisibles ou indivisibles finies.

Elle joue un rôle central dans le problème de partage, d'où l'expression partage de ressource. Plusieurs individus peuvent avoir en commun une ressource, on parle en ce moment-là de la ressource commune [50], ([73], pp. 5, 7). Dans ressource se trouvent cacher différents termes comme gain et coût ([41], p.1), [51].

Les bénéficiaires ou demandeurs de la ressource sont des « personnes physiques ou morales, machines, ... impliqués dans le problème de partage de cette ressource » ([73],pp. 15-16).

Les contraintes sont des facteurs physiques ou légaux extérieurs aux individus qui limitent l'ensemble des allocations possibles de la ressource. Elles peuvent être formulées sur les objets de la ressource tout comme sur les quantités demandées ou les valeurs des variables. Comme exemples de contraintes sur les objets : la contrainte de préemption pour laquelle un objet déterminé de la ressource ne peut être attribué à plus d'un individu et la contrainte d'exclusion où deux objets donnés ne peuvent être attribués simultanément. Comme exemple de contraintes sur la quantité demandée : la somme des quantités d'objets demandées doit être égale à 100, en est un exemple ([73], pp. 5, 7).

Suivant la nature des objets qui constituent la ressource, il existe deux grands types de ressources : la ressource continue et la ressource discrète. Une ressource est continue si elle peut être divisée indéfiniment (Exemple : Somme d'argent, volume d'un liquide)

tandis qu'une ressource est discrète si elle ne peut être divisée qu'un nombre fini de fois jusqu'à arriver à un ensemble des éléments simples qu'on ne peut pas diviser (Exemple : ensemble des voiture) ([73], p. 16).

Une ressource peut être divisible (Exemple : Monnaie) ou indivisible (Exemples : Voiture) selon que les objets qui la composent sont divisibles ou indivisibles. D'où les expressions ressource divisible et ressource indivisible ([73], p. 16).

Selon la valeur que les individus accordent sur un objet, on distingue la ressource homogène dans le cas où chacun des individus a la même valeur des objets de la ressource ou les objets sont de même nature (Exemples : monnaie, temps) et la ressource hétérogène si les valeurs des objets sont différentes pour les agents (Exemple : ressource composée des voitures) [52], ([73], pp. 16-17).

En ce qui concerne plus particulièrement les ETDs de la République Démocratique du Congo (RDC), il y a une ressource commune que le législateur a prévue pour être partagée entre elles dans leur province d'origine à partir de trois critères : Superficie, Capacité de production et Population. Cette ressource commune est constituée de 40% de recettes à caractère national allouées aux provinces. Ces dernières constituent donc des ressources budgétaires .

3.1.2. Partage

Définition 3. 2 (Partage).

Soit C une ressource, $I = \{1, \ldots, i, \ldots, n\}$ l'ensemble des individus bénéficiaires de C et $\mathscr{P}(C)$ l'ensemble des parties de C. Un partage de C entre les éléments de I est un n-uple $(C_1, \ldots, C_i, \ldots, C_n) \in \big(\mathscr{P}(C)\big)^n$, $\forall i$ et $\cup_{i=1}^{n} C_i = C$. La composante C_i $\in \mathscr{P}(C)$ est la part de l'individu i (une personne, un objet ou une entité) ([73],p. 16), ([77], p. 63).

Précisons que dans cette définition, nous n'avons pas supposé que les parts sont disjointes deux à deux [73], [77]. Mais dans la suite, nous considérerons les parts deux à deux disjointes, auquel cas, C_i appartiendra à une partition de C.

Nous distinguons la part individuelle de la part commune (ou collective). En effet, la part individuelle est celle qui revient à chaque individu tandis que la part commune est celle qui revient à tous les individus concernés.

Dans [47], trois concepts ressortent de celui de partage : Partage comme distribution, partage comme répartition et enfin partage considéré comme don. L'allusion sera faite

au terme distribution lorsqu'un individu qui n'est pas bénéficiaire, appelé « arbitre », partage une ressource entre plusieurs individus bénéficiaires, la procédure de partage est dite centralisée. Lorsque, par contre, plusieurs individus bénéficiaires se partagent entre eux une ressource (un héritage, une somme d'argent), c'est le concept de répartition qui est utilisé, la procédure de partage est décentralisée c'est-à-dire les individus eux-mêmes se partagent la ressource. Et enfin, si un individu partage une ressource qu'elle possède avec un autre individu, il s'agit d'un don [28], [47].

La responsabilité des individus dans le partage se retrouve exprimée à travers leurs caractéristiques intrinsèques choisies par eux-mêmes par le moyen d'un contrat ou par le législateur (en tant que leur représentant) à travers une loi. La définition des caractéristiques (ou critères ou encore variables) est très importante car ce sont elles qui permettent de différencier les individus notamment en ce qui concerne leurs parts comme cela est appuyé par Aristote à travers son interrogation après une affirmation : « … pour des personnes égales la chose doit être égale. Mais égales en quoi ? » [46] ; cité par [56].

Un partage peut être juste ou injuste. Le partage d'une ressource sera dit juste lorsqu'on alloue à chaque individu une part et que cela est accepté par tous. « Une distribution est équitable lorsqu'aucune personne n'en envie une autre ou lorsqu'aucune paire d'individus ne se trouve dans cette situation » [57]. Nous considérons qu'un partage est équitable lorsqu'on alloue à chacun une part correspondant à la valeur (proportion) qu'il vérifie.

Un partage peut être égal ou inégal. Un partage est égal si, à son issue tous les bénéficiaires obtiennent chacun une part égale à celle de l'autre. Il est inégal dans le cas contraire. Un partage égal pour des individus vérifiant les mêmes valeurs est en principe juste. Un partage inégal, quant à lui, est soit juste (partage inégal juste qui est acceptable) soit injuste (partage inégal injuste qui n'est pas souhaitable) tout dépend des motivations. En ce qui concerne plus particulièrement le partage inégal et juste, cela est soutenu dans la théorie de la justice distributive qui admet que toute inégalité n'est pas injustice et appuyé par les enquêtés de [Forsé et Parodi (2007)].

Nous soutenons que pour qu'un partage soit juste il faut qu'il utilise un mécanisme adéquat qui emploie plusieurs variables (critères) tout en tenant compte des particularités de ces dernières et qui prend en compte les rapports entre les individus, s'ils sont des contributeurs ou non à la création de la ressource commune à partager permettant ainsi

de décider sur la réduction ou non des inégalités entre eux. Dans le cas où ils ne sont pas des contributeurs, il faudra en plus tenir compte de leur proximité et de leur appartenance à une même population.

3.1.3. Principes de justice distributive

Les théories de la justice distributive sont du domaine de la philosophie morale [68]. Platon dans « Les Lois » (ses derniers écrits, 346 avant J.C) désigne la justice par deux égalités : l'une, « la meilleure », donnait à chacun en proportion de sa nature. Et puis l'autre justice, « la moins souhaitable », donnant la même part à tous.

Aristote, quant à lui, continue l'idée de son maître platon dans son livre intitulé « l'Ethique à Nicomaque (v -340) » (335 avant J.C), consacré à la vertu de justice, « vertu de relations aux autres ». Il distingue deux sens du mot justice, d'un côté, un sens large : « la vertu toute entière » qui consiste au respect des règles morales et sociales, et de l'autre, un sens restreint : « la vertu particulière » impliquant une certaine égalité. En ce dernier sens, il distingue la justice distributive de la justice corrective (ou rectificative). Ce sont, selon sa théorie, les deux façons de concevoir les rapports entre les personnes et le rôle de la société.

La justice distributive (justice comme proportionnalité) consiste à donner à chacun selon son mérite. Il s'agit d'une égalité proportionnelle (géométrique). Par là Aristote entend l'attribution des charges, des honneurs et des biens dans la cité en proportion des mérites et des apports personnels de chaque citoyen. La justice consiste à traiter inégalement des personnes inégales. Quant à ce qui est de la justice corrective (ou commutative) (justice comme égalité), elle consiste à attribuer « à chacun la même part ». Elle exige une égalité simple (arithmétique) [8].

Jeremy Bentham élabore l'utilitarisme (raffiné par John Stuart Mill) avec le principe *d'"utilité"* fondé sur la maximisation du bonheur ou bien-être collectif c'est-à-dire du plus grand nombre. Ce qui entraine entre autres la satisfaction des besoins de base. Pour cette doctrine, la société juste est celle qui maximise la somme des utilités de ses membres [53], [90].

Quant à John Rawls, il parle de deux principes de justice qui doivent être garantis par les institutions. Le premier est le principe de liberté selon lequel chaque citoyen doit avoir accès aux mêmes libertés, et la liberté de chacun devra être compatible avec celles des autres membres de la société ; le deuxième est le principe de différence pour lequel

certaines différentes peuvent être tolérées dans une société juste sous deux conditions :1) les fonctions qui procurent des avantages doivent être accessibles à tous les membres de la même manière (égalité de chances). A ce niveau, Rawls introduit le concept de "biens premiers" (liberté et droits fondamentaux) et soutient qu'ils constituent les soubassements d'une société juste et qu'il convient de les mettre à la disposition de tout un chacun. 2) les inégalités sont justifiées dans le cas où elles viennent améliorer la situation des plus désavantagés. A cet effet, Rawls s'oppose à la philosophie utilitariste et postule que la société doit s'occuper de maximiser l'utilité des plus désavantagés. [8], [89].

Armatya Sen de son côté considère que l'accès aux "biens premiers" (égalité des "biens premiers") soutenu par Rawls n'est pas suffisant pour garantir la justice dans une société mais il faut préconiser l'égalité des "capacités" ("capabilités") des citoyens à profiter de ces biens (la santé, la réflexion, une longue espérance de vie, etc.). Il pose donc un principe de justice : l'égalité des "capabilités" et non l'égalité des utilités comme dans l'utilitarisme ou l'égalité des « biens premiers » comme chez Rawls. Selon lui, il faut non seulement prendre en compte ce que possèdent les individus, mais aussi leur capacité à utiliser leurs biens pour choisir leur propre mode de vie.

Deux principaux concepts sont utilisés dans sa théorie : « modes de fonctionnement » (functionings) et de « capabilités » (capabilities). Ces dernières sont les différentes combinaisons possibles des premiers (pour un individu). Ceux-là quant à eux, constituent ce qu'un individu peut réaliser considérant les biens qu'il possède (se nourrir suffisamment, se déplacer sans entraves, savoir lire et écrire), ce qui décrit son état. Une capabilité est donc un vecteur de modes de fonctionnement exprimant la liberté, pour un individu, de choisir entre différentes conditions de vie. [65].

En ce qui concerne les critères pour un partage juste, Forsé et Parodi [56] soutiennent qu'en se limitant à des sociétés modernes, trois principes de justice ressortent nettement de ces études : (1) L'égalité absolue qui garantit le même traitement pour tous. (2) L'équité qui garantit à chacun un traitement proportionnel à ses mérites. Elle introduit une égalité relative. (3) La satisfaction des besoins pour laquelle chacun est traité suivant ses besoins (au moins ceux de base) [56], [62]. Dans sa théorie de justice sociale, David Miller parle plus simplement du principe d'égalité, principe de mérite (répartition selon le mérite) et principe de besoin. Ces principes équivalent aux trois critères ci-dessus énumérés [49]. Maroy [63] quant à lui aborde l'aspect équité en utilisant le terme

méritocratie.

Nous nous intéressons au principe d'équité ou encore de mérite. La règle de partage est la proportionnalité. C'est une inégalité juste. Toutefois, il convient avant tout de considérer plusieurs variables pour garantir un équilibre entre les individus. Dans le cas des variables hétérogènes, il faut réduire leurs écarts suscités par les unités de mesure qui sont différentes.

Dans certaines situations de la vie et plus particulièrement dans un cadre social ou étatique ou plus simplement si les individus n'ont pas contribué pour créer la ressource à partager, il faut aussi tenir compte des rapports entre les individus au niveau de leurs classes c'est-à-dire parmi ceux qui leur sont les plus proches ainsi que dans la population. Il s'agit des rapports de proximité et d'appartenance à une même population. Ce qui permettra la solidarité entre les individus dans le sens à réduire les inégalités entre eux pour garantir la paix sociale. Ceci nous amène à considérer le principe de justice que nous avons nommé l'équité-réduite (ou corrigée) multidimensionnelle dont la règle de partage sera la proportionnalité réduite (ou corrigée) multidimensionnelle.

3.2. Répartition des Ressources Sans réduction des inégalités

Nous mettons en place une démarche dénommée Procédé de la Répartition des Ressources Sans réduction des inégalités (PRRS). C'est une démarche qui utilise une ou plusieurs variables et qui conduit à l'attribution des parts de la ressource sans passer par l'étape de la réduction des inégalités. Il convient pour le cas où les individus (par exemple les actionnaires d'une entreprise) ont contribué à la création de la ressource à partager.

Le PRRS comporte trois grandes étapes : (1) Détermination des données à utiliser (avec, au besoin, transformation des variables) ; (2) calcul des valeurs totales des individus ; (3) calcul des parts des individus à partir de leurs proportions ; (4) représentation graphique des individus. Nous proposerons, à cet effet, quelques applications (Cfr 4.1).

3.2.1. Détermination des données à utiliser

3.2.1.1. Cas des variables homogènes

Les variables homogènes ont toutes la même unité de mesure qui peut s'exprimer de manière unique ou différemment pour chaque variable.

Dans le cas où l'unité de mesure est exprimée de manière unique, les données de

départ sont maintenues et peuvent être directement utilisées à l'étape suivante (Exemple : la variable "poids" considérée plusieurs fois, cela constitue des variables homogènes et en plus toutes ont la même unité de mesure exprimée de manière unique (kg)).

Par contre, si elle est exprimée différemment pour chaque variable, il convient de convertir ces différentes expressions à une même expression en vue d'anéantir l'influence de l'échelle d'une variable sur celles des autres, après quoi vient à l'étape suivante. C'est le cas par exemple des variables homogènes "poids au 1er trimestre", "poids au $2^{è}$ trimestre" et "poids au $3^{è}$ trimestre" pour lesquelles la même unité de mesure est exprimée différemment pour chaque variable :(hectogramme (hg), kilogramme (kg), décagramme (dag)). Toutes ces expressions peuvent être exprimées en "gramme (g)".

Ceci vient résoudre deux problèmes d'injustice due à l'utilisation : (-) d'une variable au lieu de plusieurs et (-) des variables homogènes ayant une même unité de mesure exprimée différemment pour chaque variable, sans au préalable les convertir à une même expression.

3.2.1.2. Cas des variables hétérogènes

Les variables hétérogènes n'ont pas toutes la même unité de mesure. Nous ne pouvons donc pas les utiliser directement. Il convient de les transformer en données réduites en vue de réinitialiser l'influence de l'unité de mesure reflétée dans l'écart-type. Cette transformation se fait en divisant chaque valeur par l'écart-type de la variable d'où elle est issue.

Cette étape résout le problème d'injustice due à l'utilisation : (-) d'une variable au lieu de plusieurs et (-) des variables hétérogènes sans pouvoir les transformer au préalable.

3.2.2. Calcul des valeurs totales des individus et de la valeur globale de la population

Il est question de déterminer une seule valeur pour chaque individu en lieu et place de plusieurs, valeur que nous avons nommée valeur totale. Pour les classes et l'ensemble de la population, nous parlons de la valeur globale.

Définition 3. 3 (Valeur totale d'un individu).

La valeur totale W_i d'un individu quelconque i est la somme de toutes ses valeurs vérifiées pour chaque variable.

$$W_i = \sum_{j=1}^{m} Y_{ij} \tag{3.1}$$

où Y_{ij} est la valeur de l'individu i pour la variable Y_j

Définition 3. 4 (Valeur globale d'une classe).

Soit $n_1, \ldots, n_k, \ldots, n_q$ les nombres d'individus respectivement des q classes K_1, ..., K_k, ..., K_q issues d'une population. La valeur globale de la classe K_k est la somme de toutes les valeurs totales de ses individus.

$$V_k = \sum_{i=1}^{n_k} W_{ik} \tag{3.2}$$

où W_{ik} est la valeur totale de l'individu i de la k^e classe.

Définition 3. 5 (Valeur globale d'une population).

La valeur globale V d'une population est la somme de toutes les valeurs des individus. C'est aussi la somme de toutes les valeurs globales des classes issues d'elle.

$$V = \sum_{i=1}^{n} W_i = \sum_{k=1}^{q} V_k \tag{3.3}$$

où W_i est la valeur totale de l'individu i et V_k la valeur globale de la k^e classe.

3.2.3. Calcul des proportions et parts des individus

La part C_i d'un individu i est proportionnelle à sa valeur totale W_i :

$$C_i = \frac{W_i}{V}.C \tag{3.4}$$

où C est la ressource à partager.

$$P_i = \frac{W_i}{V} \tag{3.5}$$

est la proportion de i par rapport à la valeur globale. Il suffit de multiplier P_i par 100 pour obtenir le résultat en pourcentage.

3.2.4. Représentation graphique des parts des individus

On peut aller plus loin dans le sens à représenter graphiquement les parts des individus sur un diagramme à bâtons ou en camembert.

3.3. Répartition des Ressources avec réduction des inégalités

Le Procédé de la Répartition des Ressources à partir des résultats de la Classification(PRRC) que nous avons proposé fait appel avant tout à la notion de classification qui permet de retrouver les individus répartis dans leurs classes respectives. Ceux d'une même classe vont se solidariser à travers la réduction des inégalités entre eux suite à leurs rapports de proximité, et eux tous formant la population devront aussi se solidariser suite au fait qu'ils appartiennent à une même population. Le calcul des part vient en dernier lieu. Ceci vient ajouter un plus dans l'interprétation d'une partition issue d'une CAH.

Dans la vie pratique, il vient résoudre le problème d'injustice dû à la non utilisation de plusieurs variables (ou critères) notamment hétérogènes et la non prise en compte des rapports entre les individus lors d'un partage, plus particulièrement entre les individus n'ayant pas contribué à la création de la ressource à partager. C'est le cas entités étatiques telles que les ETDs qui doivent se répartir les recettes à caractère national leur allouées.

Le procédé PRRC se réalise à l'aide de cinq grandes étapes suivantes : (1) Détermination et présentation des résultats de la Classification (les classes) (3.3.1.). (2) Réduction des inégalités entre les individus dans leurs classes et dans l'ensemble de la population. Cela se fait à partir de leurs valeurs totales en vue de la détermination de leurs valeurs corrigées (3.3.2.). (3) Calcul des proportions corrigées des individus (3.3.3.). (4) Calcul des parts respectives des individus (3.3.4.). (5) Représentation graphique des parts des individus (3.3.5.). A ce propos, une application est proposée (Cfr 4.2).

3.3.1. Détermination et présentation des résultats de la classification

Considérons que les valeurs de n individus $1, \ldots, i, \ldots, n$ sont vérifiées à partir de m variables $X_1, \ldots, X_j, \ldots, X_m$. Notons par X_{ij}, la valeur de la variable X_j vérifiée par l'individu i et W_{ij} la valeur totale de i.

Supposons que l'opération de centrage des variables hétérogènes ait été effectuée. Nous soutenons que celle-ci vient résoudre le problème d'injustice due à l'utilisation directe des variables hétérogènes sans pouvoir les transformer au départ.

Par la suite, il est question de regrouper les individus dans des classes. Ainsi, après avoir suivi toutes les étapes de la classification (CAH) jusqu'au niveau de la découpe du dendrogramme, on retrouve une partition (la meilleure) constituée des classes (Cfr 2.2.). Chaque classe étant composée d'un ou plusieurs individus qui se ressemblent le plus.

Au niveau de la réalisation de la CAH, la démarche classique présente toujours deux formules différentes pour calculer, du moins manuellement, la distance entre deux individus et l'indice d'agrégation entre deux groupes d'individus. En plus, elles ne permettent pas un calcul simultané de plusieurs distances et/ou indices d'agrégation. Ce qui occasionne une perte de temps. Nous résolvons ce problème en proposant deux formules et la manière dont il faut les utiliser.

3.3.1.1. Expression commune de calcul de la distance euclidienne et de l'indice d'agrégation centroïde

Nous proposons le premier théorème ci-dessous que nous allons démontrer. Il donne l'expression commune de la distance euclidienne et de l'indice d'agrégation centroïde (barycentre) de deux groupes en utilisant des points en coordonnées cartésiennes.

Théorème 3. 2 (**Distance et indice d'agrégation utilisant les points en coordonnées cartésiennes).**

Etant donnés deux groupes d'individus I_1 et I_2 pouvant contenir chacun un seul élément. Considérons p et q respectivement le nombre d'individus de I_1 et de I_2, m la dimension de l'espace dans lequel se calcule la distance (nombre de variables) et, X_{ij} et Y_{ij} respectivement les différentes coordonnées des points (individus) de I_1 et de I_2. Alors l'expression de la distance entre ces deux groupes est :

$$d(I_1, I_2) = \frac{1}{pq}\sqrt{\sum_{j=1}^{m}(p\sum_{i=1}^{q}Y_{ij} - q\sum_{i=1}^{p}X_{ij})^2} \tag{3.6}$$

Démonstration. Il est question de prouver que

$d(I_1, I_2) = \frac{1}{pq}\sqrt{\sum_{j=1}^{m}(p\sum_{i=1}^{q}Y_{ij} - q\sum_{i=1}^{p}X_{ij})^2}$.

En effet, soit I_1 un ensemble de p individus (points) :

$A_1 = (X_{11}, \ldots, X_{1m}), \ldots, A_i = (X_{i1}, \ldots, X_{im}), \ldots, A_p = (X_{p1}, \ldots, X_{pm})$

de barycentre $G_X = (G_{X_1}, \ldots, G_{X_m}) = (\frac{1}{p}\sum_{i=1}^{p}X_{i1}, \ldots, \frac{1}{p}\sum_{i=1}^{p}X_{im})$

et I_2 un ensemble de q individus (points) :

$B_1 = (Y_{11}, \ldots, Y_{1m}), \ldots, B_i = (Y_{i1}, \ldots, Y_{im}), \ldots, B_q = (Y_{q1}, \ldots, Y_{qm})$

de barycentre $G_Y = (G_{Y_1}, \ldots, G_{Y_m}) = (\frac{1}{q}\sum_{i=1}^{q}Y_{i1}, \ldots, \frac{1}{q}\sum_{i=1}^{q}Y_{im})$

La distance entre les centres de gravité de ces deux groupes d'individus constitue leur distance selon l'indice d'agrégation centroïde.

Ainsi, $d(I_1, I_2) = d(G_X, G_Y)$

$= \sqrt{(\frac{1}{q}\sum_{i=1}^{q} Y_{i1} - \frac{1}{p}\sum_{i=1}^{p} X_{i1})^2 + \ldots + (\frac{1}{q}\sum_{i=1}^{q} Y_{im} - \frac{1}{p}\sum_{i=1}^{p} X_{im})^2}$

$= \sqrt{\sum_{j=1}^{m}(\frac{1}{q}\sum_{i=1}^{q} Y_{ij} - \frac{1}{p}\sum_{i=1}^{p} X_{ij})^2} = \frac{1}{pq}\sqrt{\sum_{j=1}^{m}(p\sum_{i=1}^{q} Y_{ij} - q\sum_{i=1}^{p} X_{ij})^2}.$

□

Corollaire 3. 1.

L'expression $d(I_1, I_2) = \frac{1}{pq}\sqrt{\sum_{j=1}^{m}(p\sum_{i=1}^{q} Y_{ij} - q\sum_{i=1}^{p} X_{ij})^2}$ *est à la fois une mesure de distance et un indice d'agrégation.*

En effet, comme mesure de distance, elle équivaut à la distance euclidienne (cas particulier : $p = q = 1 \in \mathbb{N}$) entre deux individus $X \in I_1$ et $Y \in I_2$:

$$d(I_1, I_2) = d(X, Y) = \sqrt{\sum_{j=1}^{m}(Y_j - X_j)^2} \tag{3.7}$$

et comme indice d'agrégation, elle correspond à la distance des barycentres G_1 et G_2 respectivement des groupes I_1 et I_2 (cas général : p $(p \neq 1)$, nombre d'éléments dans I_1 et q $(q \neq 1)$ nombre d'éléments dans I_2)

Exemple 3. 1 (Distance entre deux groupes).

Soit les groupes d'individus I_1 =$\{A_1(1, 5); A_2(5, 8); A_3(1, 2)\}$ et I_2 =$\{B_1(6, 3); B_2(2, -5)\}$. Calculer la distance entre ces deux groupes.

Solution

Suivant les données ci-dessus, on a :

$m = 2$, $p = 3$, $q = 2$; $X_{11} = 1$, $X_{12} = 5$, $X_{21} = 5$, $X_{22} = 8$, $X_{31} = 1$, $X_{32} = 2$, $Y_{11} = 6$, $Y_{12} = 3$, $Y_{21} = 2$, $Y_{22} = -5$.

La distance entre les groupes I_1 et I_2 est égale à :

$d(I_1, I_2) = \frac{1}{pq}\sqrt{\sum_{j=1}^{m}(p\sum_{i=1}^{q} Y_{ij} - q\sum_{i=1}^{p} X_{ij})^2}$

=$\frac{1}{6}\sqrt{\sum_{j=1}^{2}(3\sum_{i=1}^{2} Y_{ij} - 2\sum_{i=1}^{3} X_{ij})^2}$

= $\frac{1}{6}\sqrt{(3\sum_{i=1}^{2} Y_{i1} - 2\sum_{i=1}^{3} X_{i1})^2 + (3\sum_{i=1}^{2} Y_{i2} - 2\sum_{i=1}^{3} X_{i2})^2}$

= $\frac{1}{6}\sqrt{(3(6+2) - 2(1+5+1))^2 + (3(3-5) - 2(5+8+2))^2}$ = $\frac{1}{3}\sqrt{349}$

=$6,227180564$

Vérification

$G_A = (G_{A_1}, G_{A_2}) = (\frac{1}{3}\sum_{i=1}^{3} X_{i2}, \frac{1}{3}\sum_{i=1}^{3} X_{i2}) = (\frac{1+5+1}{3}, \frac{5+8+2}{3}) = (\frac{7}{3}, 5)$

$G_B = (G_{B_1}, G_{B_2}) = (\frac{1}{2}\sum_{i=1}^{2} Y_{i2}, \frac{1}{2}\sum_{i=1}^{2} Y_{i2}) = (\frac{6+2}{2}, \frac{3-5}{2}) = (4, -1)$

$d(G_A, G_B) = \sqrt{(4-\frac{7}{3})^2 + (-1-5)^2} = \sqrt{\frac{25+324}{9}} = \frac{1}{3}\sqrt{349} = 6,227180564$

3.3.1.2. Expression de calcul simultané de plusieurs distances entre les individus et/ou groupes d'individus deux à deux.

Nous proposons ici un théorème, que nous allons démontrer, qui présente une formule de calcul de plusieurs distances simultanément en utilisant les matrices.

Théorème 3. 3 (Calcul simultané de plusieurs distances entre les individus et/ou groupes d'individus en utilisant les matrices).

Etant donné un ensemble I d'individus et I_1 et I_2 deux sous-ensembles de I constitués à partir de son produit cartésien. Les distances entre les individus de I_1 et I_2 pris deux à deux sont données par la formule :

$$d(I_1, I_2) = \sqrt{diag(\Omega.\Omega^t)} \tag{3.8}$$

où $diag(\Omega.\Omega^t)$ est la diagonale de la matrice symétrique $\Omega.\Omega^t$ avec $\Omega = I_1 - I_2$ et Ω^t sa transposée.

Démonstration. Considérons deux groupes I_1 et I_2 formés chacun de p individus.

$A_1 = (X_{11}, \ldots, X_{1m}), \ldots, A_p = (X_{p1}, \ldots, X_{pm})$ les individus de I_1 et $B_1 = (Y_{11}, \ldots, Y_{1m}), \ldots, B_p = (Y_{p1}, \ldots, Y_{pm})$ ceux de I_2.

Soit la matrice

$$\Omega = I_1 - I_2 = \begin{pmatrix} X_{11} - Y_{11} & \cdots & X_{1m} - Y_{1m} \\ & \cdots & \\ X_{i1} - Y_{i1} & \cdots & X_{im} - Y_{im} \\ & \cdots & \\ X_{p1} - Y_{p1} & \cdots & X_{pm} - Y_{pm} \end{pmatrix}$$

et Ω^t sa transposée.

On calcule le produit $\Omega.\Omega^t$ qui est une matrice symétrique et la distance

$d(I_1, I_2) = \sqrt{diag(\Omega.\Omega^t)}$ où $diag(\Omega.\Omega^t)$ est la diagonale de $\Omega.\Omega^t$.

□

Corollaire 3. 2.

L'expression

$$d(I_1, I_2) = \sqrt{diag(\Omega.\Omega^t)} \tag{3.9}$$

est à la fois une mesure de distance et une expression de calcul de plusieurs distances simultanément.

En effet, comme mesure de distance, cette expression équivaut à la distance euclidienne (cas particulier : I_1 et I_2 sont des singletons). Ainsi, la distance entre deux individus $X \in I_1$ et $Y \in I_2$:

$$d(I_1, I_2) = d(X, Y) = \sqrt{diag(\Omega.\Omega^t)} = \sqrt{(\Omega.\Omega^t)} \tag{3.10}$$

où $\Omega = X - Y$ et $(\Omega.\Omega^t)$ est une matrice uniligne unicolonne.

Dans le cas général, elle est une expression de calcul simultané de distances entre les individus (et/ou groupes d'individus) deux à deux. Elle passe par la construction des matrices I_1 et I_2 considérées comme des sous-ensembles de $I = \{1, \ldots, i, \ldots, n\}$ issus du produit cartésien IxI.

Pour permettre l'utilisation de la formule ci-dessus, nous proposons une démarche nommée : Procédé de Calcul Simultané de plusieurs Distances (PCSD). Il suit les étapes ci-après :

1) Constituer les deux matrices à m colonnes issues de l'ensemble des individus.

Soit m variables vérifiées par les individus 1, . . . , i, . . . , n de I. Les deux matrices à m colonnes I_1 et I_2 sont respectivement formées des premières et deuxièmes composantes de I x I, en considérant soit la partie inférieure soit la partie supérieure de son tableau.

D'une manière générale, ces deux matrices se présenteront comme suit :

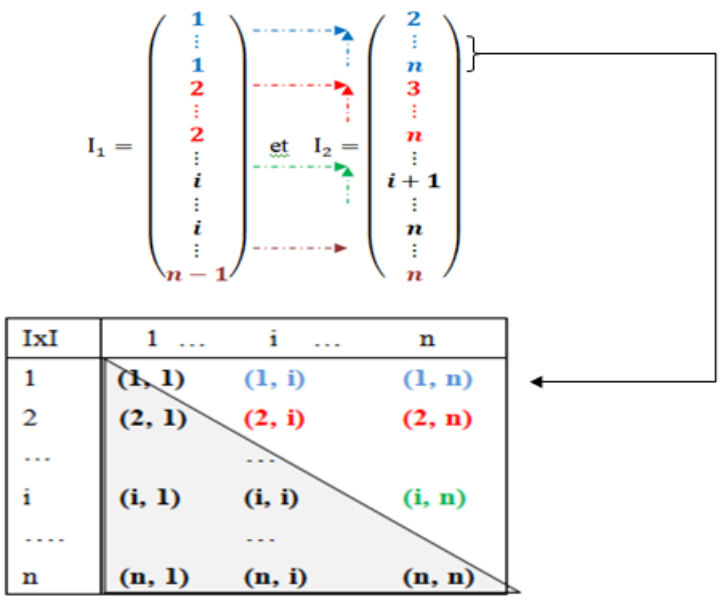

Figure 3.1 – Deux matrices extraites du produit cartésien IxI
Source : Notre conception à partir des tableaux existants

Considérant la partie supérieure de la diagonale principale de ce tableau, la première ligne comprend les n-1 couples $(1, 2), ..., (1, i), ..., (1, n)$ dont les premières composantes forment les n-1 premiers termes de la matrice I_1 et les 2è ceux de I_2. La 2è ligne reprend les n-2 couples $(2, 3), ..., (2, i), ..., (2, n)$ dont les composantes forment les n-2 termes suivants de I_1 et de I_2 respectivement. On continuera ainsi jusqu'à la $(n-1)^{\text{è}}$ ligne correspondant à un $(n-(n-1)=1)$ couple $(n-1, n)$.

2) Calculer plusieurs distances simultanément entre les individus deux à deux (et même entre les groupes deux à deux)

• Calculer la matrice $\Omega = I_1 - I_2$.

• Calculer la transposée de Ω^t.

• Calculer de la matrice symétrique $(\Omega.\Omega^t)$ et extraire les éléments de sa diagonale principale : $diag(\Omega.\Omega^t)$.

• Calculer les différentes distances entre les individus deux à deux avec la formule (3.8). On tiendra compte des positions des individus dans les matrices I_1 et I_2 pour déterminer la distance entre tel individu de I_1 et tel autre de I_2.

La première valeur de $d(I_1, I_2)$ correspondra à la distance des individus de I_1 et I_2 se trouvant à la première ligne.

Exemple 3. 2.

Soit I l'ensemble de quatre individus a, b, c, d représentés respectivement par les

points (1, 5), (4, 2), (3, 3) et (9, 0) dont on veut simultanément calculer les distances deux à deux.

En suivant le procédé PCSD, on a :

1) Constitution de deux matrices I_1 et I_2 à 2 colonnes issues de l'ensemble des individus. Alors, le tableau de I*x*I se présente comme suit :

Tableau 3.1 – Produit cartésien IxI

IxI	a	b	c	d
a	(a, a)	(a, b)	(a, c)	(a, d)
b	(b, a)	(b, b)	(b, c)	(b, d)
c	(c, a)	(c, b)	(c, c)	(c, d)
d	(d, a)	(d, b)	(d, c)	(d, d)

Source : Nos enquêtes

De ce tableau, nous extrayons les matrices I_1 et I_2 formées respectivement des premières et deuxièmes composantes des couples retenus (de la partie du tableau se trouvant au-dessus de la diagonale principale).

Ces matrices se présentent comme suit :

$$I_1 = \begin{pmatrix} a \\ a \\ a \\ b \\ b \\ c \end{pmatrix}, \quad I_2 = \begin{pmatrix} b \\ c \\ d \\ c \\ d \\ d \end{pmatrix}$$

Il suffit pour la suite, de remplacer chacun des individus a, b, c et d par les valeurs qu'il vérifie suivant les variables retenues.

De ce fait, on a :

$$I_1 = \begin{pmatrix} 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 4 & 2 \\ 4 & 2 \\ 3 & 3 \end{pmatrix}, \quad I_2 = \begin{pmatrix} 4 & 2 \\ 3 & 3 \\ 9 & 0 \\ 3 & 3 \\ 9 & 0 \\ 9 & 0 \end{pmatrix}$$

2) Calculons plusieurs distances simultanément entre les individus deux à deux

• Calculer la matrice $\Omega = I_1 - I_2$.

$$\Omega = \begin{pmatrix} 1-4 & 5-2 \\ 1-3 & 5-3 \\ 1-9 & 5-0 \\ 4-3 & 2-3 \\ 4-9 & 2-0 \\ 3-9 & 3-0 \end{pmatrix} = \begin{pmatrix} -3 & 3 \\ -2 & 2 \\ -8 & 5 \\ 1 & -1 \\ -5 & 2 \\ -6 & 3 \end{pmatrix}$$

• Calculer la transposée de Ω^t.

$$\Omega^t = \begin{pmatrix} -3 & -2 & -8 & 1 & -5 & -6 \\ 3 & 2 & 5 & -1 & 2 & 3 \end{pmatrix}$$

• Calculer de la matrice symétrique $(\Omega.\Omega^t)$ et extraire les éléments de sa diagonale principale : $diag(\Omega.\Omega^t)$.

$$diag(\Omega.\Omega^t) = \begin{pmatrix} 18 & 8 & 89 & 2 & 29 & 45 \end{pmatrix}$$

• Calculer les différentes distances entre les individus deux à deux avec la formule proposée ci-dessus.

$$\begin{aligned} d(I_1, I_2) = \sqrt{(diag\Omega.\Omega^t)} &= \begin{pmatrix} \sqrt{18} & \sqrt{8} & \sqrt{89} & \sqrt{2} & \sqrt{29} & \sqrt{45} \end{pmatrix} \\ &= \begin{pmatrix} 4,2 & 2,8 & 9,4 & 1,4 & 5,4 & 6,7 \end{pmatrix} \end{aligned}$$

Ceci étant, nous avons les distances suivantes :

$d(a,b) = 4,2$; $d(a,c) = 2,8$; $d(a,d) = 9,4$; $d(b,c) = 1,4$; $d(b,d) = 5,4$; $d(c,d) = 6,7$

3.3.1.3. Nouvel élément dans l'interprétation des résultats de la Classification

Hormis la représentation graphique des classes, nous jugeons qu'il ne convient pas que l'interprétation d'une partition se limite au calcul des paramètres des classes, à la détermination de leurs individus les plus typiques et aux variables les caractérisant. Elle doit aller plus loin jusqu'à la détermination des proportions et parts des individus après réduction des inégalités existant entre eux. Ce qui est utile pour la répartition d'une (ou des) ressource (s) commune (s) entre un nombre déterminé d'individus n'ayant pas contribué pour créer la ressource à partager. D'où l'importance du Procédé PRRC.

3.3.2. Indice de niveau des inégalités et réduction des inégalités

3.3.2.1. Indice de niveau des inégalités

Les valeurs totales des individus et les valeurs globales des classes et de la population ayant été calculées (Cfr 3.2.2.), pour réduire les inégalités entre les individus, nous proposons un indice que nous avons nommé « Indice de niveau des inégalités».

Définition 3. 6 (Indice de niveau des inégalités).

L'Indice de niveau des inégalités est un indice qui permet de mesurer et de réduire les inégalités entre les individus. C'est donc l'indice de mesure de niveau des inégalités permettant la réduction des inégalités.

D'une manière générale, nous notons l'indice de niveau des inégalités J_M. Il est calculé à l'aide de la formule :

$$J_M = \frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i} \tag{3.11}$$

Où W_i est la valeur totale de l'individu i et W_1 la plus petite valeur totale de la distribution avec $W_1 \leq \ldots \leq W_i \leq \ldots \leq W_n$ et $\sum_{i=1}^{n} W_i$ la somme de toutes les valeurs totales des individus qui équivaut à la valeur globale de la classe ou encore de la population.

La mesure des inégalités peut se faire même après le partage, en utilisant les parts des individus. Dans ce cas, il suffit d'utiliser la formule suivante :

$$J_M = \frac{\sum_{i=1}^{n}(C_i - C_1)}{C} \tag{3.12}$$

où C_1 est la plus petite part et $C_{i(1 \leq i \leq n)}$ la part correspondant à un individu i parmi les n individus concernés.

3.3.2.2. Fonction de réduction des inégalités entre les individus

La fonction de réduction des inégalités que nous avons proposée permet de calculer les valeurs corrigées des individus.

Elle est fonction des valeurs totales des individus. Elle s'écrit :

$$Z_i = W_1 + W_i.J_M \tag{3.13}$$

Les notations Z_i et J_M seront utilisées dans le cas où la réduction des inégalités se fait au niveau des classes. Si c'est au niveau de l'ensemble de la population, nous

utiliserons les notations Z'_i et $J_{M'}$.

Il convient de souligner que les valeurs de Z_i se répartissent en deux parties. *Celles des individus ayant transféré des valeurs aux autres ("les riches") et celles des individus qui les ont reçues ("les pauvres").* Les « riches » verront leurs valeurs diminuées tandis que les pauvres les verront augmentées. En faisant la différence $Z_i - W_i = E_i$, on trouve la valeur perdue $E_i(-)$ ou gagnée $E_i(+)$ par l'individu i, selon qu'elle est négative ou positive.

On vérifiera que la somme de toutes les valeurs gagnées et perdues égale zéro :

$$\sum_{i=1}^{n}(Z_i - W_i) = \sum_{i=1}^{n} E_i = 0 \tag{3.14}$$

Dans le cas où l'indice de niveau des inégalités est encore grand, il y a lieu de tenter une deuxième réduction voire une troisième ou une quatrième tout dépend du niveau des inégalités cherché. Toutefois, la première réduction des inégalités est exigée lorsque celles-ci ne sont pas nulles c'est-à-dire lorsque les individus n'ont pas la même valeur totale.

Les formules ci-dessus peuvent aussi être utilisées dans le cas de la répartition des sièges entre différentes circonscriptions en leur attribuant des quotas ainsi que dans le cas où les individus doivent apporter leurs contributions.

3.3.2.3. Justification de la réduction des inégalités

Nous justifions la réduction des inégalités entre les individus par le fait que ces derniers n'ont pas contribué pour créer la ressource à partager. Ils doivent donc se solidariser en prenant en compte leur proximité et leur appartenance à une même société, surtout qu'ils ne sont pas nécessairement égaux entre eux. Il convient donc que la réduction des inégalités soit faite d'abord dans les classes et ensuite dans l'ensemble de la population. On se situe dans un cadre social, étatique.

Cette réduction des inégalités vient répondre à la préoccupation concernant le manque de mécanisme de réduction des inégalités dans le partage des ressources, soulevée par Marcel Kapya [61] et Paulin Punga [66] (Cfr Problématique). Ainsi, les individus subiront une perte ou bénéficieront d'un gain parmi leurs semblables et dans l'ensemble de la population par la diminution ou l'augmentation de leurs valeurs.

Proposition 3. 1 (Indice de niveau des inégalités et valeurs de départ).

Si $W_1, \ldots, W_i, \ldots, W_n$ les valeurs totales respectives de n individus. Alors l'indice de niveau des inégalités J_M vérifie les propriétés suivantes :

1) $J_M = 0$ si $W_i = W_1, \forall i \in [1, n]$.

C'est l'égalité complète : Tous les individus ont la même valeur.

2)$J_M = 1$ si $\exists i \in [1, n]$, $W_i = 0$.

C'est l'inégalité totale : Un individu au moins a la valeur 0.

3) $J_M = 1 - \frac{W_1.n}{\sum_{i=1}^{n} W_i}$.

4) $J_M \in [0, 1]$.

Ses valeurs sont comprises entre 0 et 1, les deux inclus. Plus elles approchent 1 plus l'inégalité est grande.

5) $J_M = \frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i} = \frac{\overline{(W_i - W_1)}_{1 \leq i \leq n}}{(\overline{W_i})_{1 \leq i \leq n}}$

Démonstration. 1) Montrons que $J_M = 0$ si $W_i = W_1$, $\forall i \in [1, n]$.

On sait que $J_M = \frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i}$.

Si $W_i = W_1$, on a : $J_M = \frac{\sum_{i=1}^{n}(0 - W_1)}{\sum_{i=1}^{n} W_i} = 0$.

2) Montrons que $J_M = 1$ si $\exists i \in [1, n]$, $W_i = 0$.

Supposons qu'il existe $i' \in [1, n]$ tel que $W_{i'} = 0$.

Considérons $W_1 = 0$, on a :

J_M=$\frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i} = \frac{\sum_{i=1}^{n}(W_i - 0)}{\sum_{i=1}^{n} W_i} = 1$.

Il suffit donc de considérer $W_{i'} = W_1 = 0$.

D'où l'existence de $i' = 1$.

3) Montrons que $J_M = 1 - \frac{W_1.n}{\sum_{i=1}^{n} W_i}$.

En effet, $J_M = \frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i}$

$= \frac{\sum_{i=1}^{n} W_i - W_1.n}{\sum_{i=1}^{n} W_i} = 1 - \frac{W_1.n}{\sum_{i=1}^{n} W_i}$.

Donc $J_M = 1 - \frac{W_1.n}{\sum_{i=1}^{n} W_i}$.

4) Montrons que $J_M \in [0, 1]$.

Ceci revient à montrer que $J_M = 0$, $J_M = 1$ et $0 < J_M < 1$.

En effet, $\exists\, W_1 = W_i\ \forall\ i\ tel\ que\ J_M = 0$; $\exists$ aussi $W_1 = 0$ tel que $J_M = 1$.

En outre, on sait que $\sum_{i=1}^{n} W_i < \sum_{i=1}^{n} W_i + W_1.n$

$\iff 0 < \sum_{i=1}^{n} W_i < \sum_{i=1}^{n} W_i + W_1.n$

$\iff 0 < \frac{\sum_{i=1}^{n} W_i - W_1.n}{\sum_{i=1}^{n} W_i} < 1$

$\iff 0 < \frac{\sum_{i=1}^{n} (W_i - W_1)}{\sum_{i=1}^{n} W_i} < 1 \iff 0 < J_M < 1$

Donc $J_M \in [0, 1]$.

5) Montrons enfin que $J_M = \frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i} = \frac{\overline{(W_i - W_1)}_{1 \leq i \leq n}}{(\overline{w_i})_{1 \leq i \leq n}}$.

On sait, par définition de la moyenne arithmétique, que

$\overline{(W_i - W_1)} = \frac{\sum_{i=1}^{n}(W_i - W_1)}{n}$ (1) et

$\overline{W_i} = \frac{\sum_{i=1}^{n} W_i}{n}$ (2)

En divisant (1) par (2) membre à membre, on a : $\frac{\overline{(W_i - W_1)}}{\overline{W_i}} = \frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i} = J_M$. $\square$

Proposition 3. 2 (Valeurs corrigées et valeurs de départ).

Si $W_1, \ldots, W_i, \ldots, W_n$ sont les n valeurs totales respectivement des n individus $1, \ldots, i, \ldots, n$ dont les valeurs corrigées respectives sont $Z_1, \ldots, Z_i, \ldots, Z_n$. Alors,

1) $\sum_{i=1}^{n}(Z_i - W_i) = 0$

2) $\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} W_i$

3) $\overline{Z}_i = \overline{W_i}$

4) $\lim_{J_M \to 0} Z_i = Z' = \overline{W_i}$

(Z' *est la valeur constante des individus à la toute dernière réduction des inégalités)*

Démonstration. 1) Montrons que $\sum_{i=1}^{n}(Z_i - W_i) = 0$.

En effet, $\sum_{i=1}^{n}(Z_i - W_i) = \sum_{i=1}^{n}(W_1 + W_i J_M - W_i)$

$= \sum_{i=1}^{n}(W_1 + W_i(J_M - 1)) = W_1.n + \sum_{i=1}^{n}(W_i(J_M - 1))$

$= W_1.n + (J_M - 1)\sum_{i=1}^{n} W_i = W_1.n + (\frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i} - 1))\sum_{i=1}^{n} W_i$

$= W_1.n + \sum_{i=1}^{n}(W_i - W_1) - \sum_{i=1}^{n} W_i = W_1.n + \sum_{i=1}^{n} W_i - W_1.n - \sum_{i=1}^{n} W_i = 0$

Donc $\sum_{i=1}^{n}(Z_i - W_i) = 0$.

2) Montrons que $\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} W_i$

En effet, $\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n}(W_1 + W_i J_M)$

$= W_1.n + J_M \sum_{i=1}^{n} W_i$

$= W_1.n + (\frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i})\sum_{i=1}^{n} W_i)$

$= W_1.n + \sum_{i=1}^{n}(W_i - W_1)$

$= W_1.n + \sum_{i=1}^{n} W_i - W_1.n = \sum_{i=1}^{n} W_i$

Donc $\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} W_i$

3) Montrons que $\overline{Z_i} = \overline{W_i}$

On sait que $\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} W_i \qquad (1)$

Divisons (1) membre à membre par n (nombre d'individus), on a :

$\frac{\sum_{i=1}^{n} Z_i}{n} = \frac{\sum_{i=1}^{n} W_i}{n}$

$\iff \overline{Z_i} = \overline{W_i}$

Donc $\overline{Z_i} = \overline{W_i}$

4) Montrons que $\lim_{J_M \to 0} Z_i = Z' = \overline{W_i}$
(Z' est la valeur constante des individus à la toute dernière réduction des écarts)

On sait que si $J_M \to 0$, $Z_i \to Z'$ (Z'_i étant la toute dernière fonction des valeurs corrigées). On sait aussi que $J_M \in [0, 1]$.

Considérant $J_M = 0$, $W_i = W_1$ pour tout i, et se trouvant dans le cas de Z'_i, on a : Z'_1=$Z'_i = Z'_n$

Ceci implique $\lim_{J_M \to 0} Z_i$=Z' $(Z' = Z'_1$=$Z'_i = Z'_n)$

Or Z'=Z'_1=$\frac{nZ'_1}{n} = \frac{\sum_{i=1}^{n} Z'_1}{n} = \overline{Z'_1}$ et

$\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} W_i \Rightarrow \overline{Z_i} = \overline{W_i}$

Donc $\lim_{J_M \to 0} Z_i$=Z'=$\overline{W_i}$

□

3.3.3. Calcul des proportions des individus par rapport à l'ensemble de la population

Il s'agit des proportions corrigées des n individus par rapport à l'ensemble de la population connaissant leurs valeurs corrigées $Z_{i'} = W_1 + W_i.J_M$. Elles se calculent en utilisant ces dernières valeurs à partir de la formule

$$P_i = \frac{Z_{i'}}{V} \tag{3.15}$$

où V est la valeur globale.

Ce sont ces proportions corrigées qui seront utilisées dans la suite pour calculer les parts des individus.

3.3.4. Calcul de la part de la ressource commune revenant à chaque individu

La part C_i de la ressource C (par exemple : recettes à caractère national allouées aux ETDs) de l'individu i sera

$$C_i = P_i.C \tag{3.16}$$

Ou encore directement

$$C_i = \frac{Z_{i'}}{V}.C \tag{3.17}$$

où P_i est la proportion corrigée de l'individu i calculée après réduction des inégalités, $Z_{i'}$ la valeur corrigée par rapport à l'ensemble de la population, V la valeur globale de la distribution (somme de toutes les valeurs corrigées des individus au niveau de la population) et C la ressource commune à partager.

On vérifiera que

$$C = \sum_{i=1}^{n} C_i \tag{3.18}$$

Soulignons que le partage d'une ressource entre plusieurs individus doit se faire sur base d'un principe de justice précis. En ce qui concerne notre procédé PRRC qui conduit au calcul des parts des individus, nous jugeons que suite au fait que les proportions sont calculées sur base du principe d'"équité" à partir de plusieurs variables et après réduction des inégalités, l'emploi de l'expression "principe d'équité" n'est pas adapté. Nous proposons le principe d'équité réduite multidimensionnelle.

3.3.5. Représentations graphiques des parts des individus

Après calcul des différentes proportions corrigées et parts des individus, il sied de représenter graphiquement ces dernières à travers un diagramme à bâtons (en barres) ou en camembert (ou en secteurs). ([24], p. 92)

3.4. Conclusion du troisième chapitre

La répartition des ressources est un partage de celles-ci par les individus eux-mêmes. Le problème de partage fait intervenir entre autres une ressource (ou des ressources) commune (s) limitée (s), des critères (variables) sur base desquels se fera le partage, une règle (un principe) de partage. A l'issue du partage, chaque bénéficiaire s'attend à une portion de la ressource qui est sa part.

En ce qui concerne les principes de justice, Forsé et Parodi [56] en ont retenu globalement trois : l'égalité, l'équité et la satisfaction de besoin mais ils ne se sont pas préoccupés du nombre (un ou plusieurs) et des types (homogène ou hétérogène) des variables, ainsi que des rapports entre les individus (leur rapprochement, la nature de leur société d'appartenance).

Quant à nous, nous avons mis l'accent sur plusieurs variables notamment hétérogènes ainsi que sur les rapports entre les individus (contributeurs (actionnaires) ou non contributeurs (entités étatiques)). Ce qui permet la transformation des variables hétérogènes et, au besoin, la réduction des inégalités entre les individus les plus proches puis entre eux tous formant une même population en étude. Dans le cas où les individus sont des contributeurs, le principe d'équité convient et permet d'attribuer aux individus des parts proportionnelles à leurs contributions respectives alors que dans le cas des non contributeurs, l'équité seule est inadaptée. Il faut proposer un autre principe de justice. Nous avons ainsi proposé le principe d'équité réduite (ou corrigée multidimensionnelle) selon lequel, le partage est fait suivant plusieurs variables et les parts sont attribuées proportionnellement aux valeurs des individus mais après réduction des inégalités entre eux.

De ce fait, pour résoudre les problèmes d'injustice soulevés à l'introduction, nous avons proposé deux mécanismes qui utilisent plusieurs variables homogènes ou hétérogènes et prennent en compte les rapports entre les individus permettant de réduire ou pas les inégalités. Il s'agit de :

(-) Procédé de la Répartition des Ressources Sans réduction des inégalités (PRRS) adapté pour le cas où les individus (comme des actionnaires d'une entreprise) ont contribué à la création de la ressource commune à partager et n'autorise pas la réduction des inégalités. Quelques applications ont été proposées à cet effet.

Il suit les quatre étapes suivantes : (1) Détermination des données à utiliser ; (2) calcul

des valeurs totales des individus ; (3) calcul des parts des individus ; (4) Représentation graphique des parts des individus.

(-) Procédé de la Répartition des Ressources à partir des résultats de la Classification (PRRC) adapté pour le cas où les individus n'ont pas contribué à la création de la ressource. Il admet la réduction des inégalités et fait appel à la notion de classification permettant de retrouver les individus les plus proches qui devront se solidariser dans leurs classes respectives et par la suite avec tous les autres du fait qu'ils appartiennent à une même population.

Ce procédé se réalise en suivant les cinq grandes étapes ci-après : (1) Détermination et présentation des résultats de la Classification (les classes). (2) Réduction des inégalités entre les individus dans leurs classes et dans l'ensemble de la population. (3) Calcul des proportions corrigées des individus. (4) Calcul des parts respectives des individus. (5) Représentation graphique des parts des individus.

En ce qui concerne le calcul des distances et d'indices d'agrégation, nous avons proposé deux formules différentes (3.6.) et (3.8.). L'une calcule à la fois la distance entre deux individus simples (distance euclidienne) et entre deux groupes d'individus (distance entre deux barycentres). La deuxième, quant elle, calcule simultanément plusieurs distances à travers la démarche que nous avons proposée : Procédé de Calcul Simultané des Distances (PCSD).

Quant à ce qui est de l'interprétation des résultats de la CAH, nous soutenons qu'elle doit aller plus loin jusqu'à la détermination des proportions et parts des individus. Pour le cas où les individus n'ont pas contribué à la création de la ressource commune à partager, il faut procéder par la réduction des inégalités entre les individus les plus proches puis entre eux tous formant population.

Nos deux procédés proposés dans ce chapitre peuvent être appliqués à des cas concrets afin de se rendre compte de leur pertinence.

Chapitre 4

Application des procédés de la répartition des ressources

Nous allons présenter quelques exemples d'application de nos procédés de répartition des ressources. Plus particulièrement nous appliquerons le procédé PRRC aux 24 communes de la ville-province de Kinshasa en vue de la répartition des ressources budgétaires de la ville-province de Kinshasa qui sont ici les recettes à caractère national (4.2). Nous proposons aussi quelques applications relatives au procédé PRRS (4.1)

4.1. Quelques applications du Procédé PRRS

4.1.1. Cas de maintien des données de départ

C'est le cas où les données sont homogènes et l'unité de mesure est exprimée de manière unique.

Exemple 4. 1 (Partage suivant plusieurs variables homogènes)**.**

Soit à partager 50000 CDF entre deux individus (Ind.) A et B suivant leurs poids pour trois trimestres (tr.)

Tableau 4.1 – Exemple de partage selon plusieurs variables homogènes

Ind.	**Poids au 1er tr. (kg)**	**Poids au 2e tr. (kg)**	**Poids au 3e tr. (kg)**	**Total**
A	40	50	40	130
B	50	50	70	170
Total	90	100	110	300

Source : Notre conception

Ces données seront utilisées directement :

• La part de A : $C_1 = \frac{W_1}{V}.C = \frac{130}{300}.50000 = 21666,7$

• La part de B : $C_2 = \frac{W_2}{V}.C = \frac{170}{300}.50000 = 28333,3$

On vérifie que $\sum_{i=1}^{2} C_i = C_1 + C_2 = 21666,7 + 28333,3 = 50000$

4.1.2. Cas de transformation des données de départ

(1) Les données sont homogènes mais la même unité de mesure est exprimée différemment

Exemple 4. 3 (Partage suivant plusieurs variables homogènes (Unité exprimée différemment)).

Soit à partager 50000 CDF entre deux individus A et B suivant leurs poids pour trois trimestres (tr.)

Tableau 4.2 – Exemple de partage selon plusieurs variables homogènes (unité exprimée différemment)

Ind.	**Poids au 1er tr. (kg)**	**Poids au 2e tr. (g)**	**Poids au 3e tr. (hg)**
A	40	50000	400
B	50	50000	700
Total	90	100000	1100

Source : Notre conception

où kg=kilogramme, g=gramme et hg=hectogramme. Ces données ne peuvent pas être utilisées directement.

(-) Démarche incorrecte (à ne pas suivre)

Si les données de départ sont utilisées sans les convertir à une même notation, on aboutirait à un partage incorrect et donc injuste :

• La part de A :

$C_1 = \frac{W_1}{V}.C = \frac{50440}{101190}.50000 = 24923,4$

• La part de B :

$C_2 = \frac{W_2}{V}.C = \frac{50750}{101190}.50000 = 25076,6$

La somme $C_1 + C_2 = 24923,4 + 25076,6 = 50000$ mais les parts sont incorrectes (à comparer avec la correction ci-dessous)

(-) Démarche correcte (à suivre)

Il faut les convertir à une même notation (par exemple : g) pour aboutir à un partage correct et donc juste :

Tableau 4.3 – Exemple de partage selon plusieurs variables homogènes (une même notation de l'unité)

Ind.	**Poids au 1er tr. (g)**	**Poids au 2e tr. (g)**	**Poids au 3e tr. (g)**	**Total**
A	40000	50000	40000	130000
B	50000	50000	70000	170000
Total	90000	100000	110000	300000

Source : Notre conception

• La part de A :

$C_1 = \frac{W_1}{V}.C = \frac{130000}{300000}.50000 = 21666,7$

• La part de B :

$C_2 = \frac{W_2}{V}.C = \frac{170000}{300000}.50000 = 28333,3$

On vérifie que $\sum_{i=1}^{2} C_i = C_1 + C_2 = 21666,7 + 28333,3 = 50000$

On peut donc bien remarquer l'injustice en comparant les résultats de ces deux démarches :

Tableau 4.4 – Exemple de l'injustice remarquée lorsqu'on ne transforme pas les données

Ind.	**Part (Démarche incorrecte)**	**Part (Démarche correcte)**	**Différence**
A	24923,4	21666,7	3256,7
B	25076,6	28333,3	-3256,7
Total	50000	50000	0

Source : Notre conception

On attribue abusivement la somme de 3256,7 CDF à l'individu A et pourtant cela devait revenir à B.

(2) Les données sont hétérogènes (les unités de mesure sont complètement différentes)

Exemple 4. 4 (Partage selon plusieurs variables hétérogènes).

"1000 Euros à répartir entre A, B et C selon le capital investi dans l'entreprise et le temps passé au sein de celle-ci" [86] :

Tableau 4.5 – Exemple des données d'un partage selon plusieurs variables hétérogènes

Ind.	Capital (Euro)	Temps (mois)
A	300	6
B	200	12
C	100	8

Source : [86]

On remarque que les unités mesure sont différentes "Euro" et "mois". Les variables "Capital" et "Temps" sont hétérogènes. Il faut donc les transformer en données réduites en divisant chaque valeur par l'écart-type (Cfr la formule 2.26) de la variable correspondante.

L'écart-type de la variable "Capital" est 100 et celui de "Temps" est 3,055050463. Après division de chaque valeur par l'écart-type de la variable correspondante, les données transformées se présentent comme suit :

Tableau 4.6 – Exemple de transformation de variables hétérogènes

Ind	Capital(Euro)	Temps (mois)	Valeur totale
A	3	1,963961012	4,963961012
B	2	3,927922024	5,927922024
C	1	2,618614683	3,618614683

Source : Notre conception à partir du tableau 4.5

L'écart-type de chacune de ces variables transformées est 1. Ces données peuvent donc maintenant être utilisées pour calculer les parts des individus.

Tableau 4.7 – Exemple calcul des parts après transformation des variables hétérogènes

	Valeur totale	Proportion	Part (en Euro)
A	4,963961012	0,342094469	342,1
B	5,927922024	0,408526444	408,5
C	3,618614683	0,249379088	249,4
Total	14,51049772	1	1000

Source : Notre conception à partir du tableau 4.6

Les parts des individus A, B et C sont calculées en multipliant chaque proportion correspondante par la somme (1000 Euros) à partager. Les proportions quant elles sont

déterminées en divisant chaque valeur totale du tableau ci-dessus par le total des valeurs totales.

4.2. Application du Procédé PRRC

Nous comptons, dans la suite, appliquer le Procédé de la Répartition des Ressources à partir des résultats de la Classification (PRRC) aux 24 communes de la Ville-province de Kinshasa considérées comme ETDs pour répartir entre elles les recettes à caractère national leur alloués. Ce procédé utilisera les résultats de la CAH qui sont une hiérarchie de partitions représentée à travers un dendrogramme. Nous utiliserons les données issues de l'ACP (4.2.1.). Par la suite, nous allons réduire les inégalités entre les communes à l'aide de l'indice de niveau des inégalités que nous avons proposé (4.2.2.). Les parts des individus seront représentées graphiquement (4.2.5.) après être calculées (4.2.4.) à partir des proportions corrigées des individus (4.2.3.).

4.2.1. Présentation des objets, données et outils à utiliser

Nous comptons présenter d'une part la population, les individus et les variables, et de l'autre, l'outil informatique principal (logiciel R) que nous avons utilisé pour appliquer les données collectées.

4.2.1.1. Présentation de la population, des individus et des variables à utiliser

La ville-province de Kinshasa est la population sur laquelle porte notre étude. « La ville de Kinshasa est la capitale de la République Démocratique du Congo (RDC) et le siège des institutions nationales. Elle forme avec les 25 provinces (Bas-Uele, Equateur, Haut-Lomami, Haut-Katanga, Haut-Uele, Ituri, Kasai, Kasai oriental, Kongo central, Kwango, Kwilu, Lomami, Lualaba, Kasaï central, Mai-Ndombe, Maniema, Mongala, Nord-Kivu, Nord-Ubangi, Sankuru, Sud-Kivu, Sud-Ubangi, Tanganyika, Tshopo, Tshuapa) la République Démocratique du Congo. Elle a le statut de province » ([58], Art.2). Sa superficie de 9964,9 Km^2 et sa population est évaluée à 8 002 962 (Huit million deux mille neuf cent soixante-deux) en 2015.

Elle a deux organes dirigeants ([58], Art.195) : l'assemblée provinciale ([58], Art.197) et le gouvernement provincial ([58], Art.198). Elle est subdivisée en 24 communes.

En ce qui concerne les individus, ce sont les 24 communes de la ville-province de Kinshasa qui sont les individus à étudier dans notre travail. La commune est tout chef-lieu de territoire ; toute subdivision de la ville ou toute agglomération ayant une population

d'au moins 20.000 habitants à laquelle un décret du Premier ministre aura conféré le statut de commune. La commune est subdivisée en quartiers et/ou en groupements incorporés ([59], Art.46). Le quartier est donc un simple démembrement de la commune constitué des rues ou/et des avenues.

Soulignons en passant que le territoire, le quartier, le groupement et le village sont des entités territoriales déconcentrées dépourvues de la personnalité juridique. Tandis que la ville, la commune, le secteur et la chefferie sont des entités territoriales décentralisées (ETDs) dotées de la personnalité juridique. Elles jouissent de la libre administration et de l'autonomie de gestion de leurs ressources humaines, économiques, financières et techniques ([59], Art.5).

Les 24 communes de la ville de Kinshasa sont réparties dans les 4 district [83] comme suit : 1) District de la Funa (7 communes) : Bandalungwa, Bumbu, Kalamu, Kasa-Vubu, Makala, Ngiri-Ngiri et Selembao. 2) District de la Lukunga (6 communes) : Barumbu, Gombe, Kinshasa, Kintambo, Lingwala et Ngaliema. 3) District du Mont-Amba (6 communes) : Kisenso, Lemba, Limete, Matete, Ngaba et Mont-Ngafula. 4) District de la Tshangu (5 communes) : Kimbanseke, Maluku, Masina, Ndjili et Nsele.

En ce qui concerne les variables retenues pour l'étude, on a : la superficie, la capacité de production et la population. Ce sont les trois critères prévus par le législateur pour la répartition entre les ETDs des 40% de la part des recettes à caractère national leur allouées par la province ([58], Art.175), ([59], Art.115, 116). Rappelons, dans la pratique, une seule variable est utilisée : population.

Les données ont été collectées de façon exhaustive à travers une enquête au moyen d'un recensement. Nous avons parcouru toutes les 24 communes de la ville de Kinshasa en vue d'y collecter des données relatives à notre travail. Nous l'avons fait soit auprès du secrétariat de la commune en utilisant le rapport annuel de la commune soit auprès des services de la population et de l'Etat civil en utilisant leurs rapports annuels. Nous avons collecté, pour chaque commune, la superficie, la capacité de production (Cfr Annexe 4) et la population.

En outre, nous avons fait un déplacement jusqu'à la division urbaine de l'intérieur et sécurité pour y recueillir aussi des données en vue de les comparer avec celles déjà collectées dans les communes. Il importe de faire observer qu'en ce qui concerne la période ciblée pour l'enquête, nous avions au départ retenu l'année 2014 (de Janvier à décembre). Toutefois, contrairement à nos attentes, pour les communes de Maluku,

Ngaba et Gombe nous n'avons pas pu collecter les données souhaitées pour les raisons suivantes : Pour les deux premières, les données fournies étaient celles de l'exercice 2013 et pour la dernière c'était celles de l'exercice 2012.

Nous avons aussi fait un déplacement jusqu'à l'Institut National de la Statistique(INS) pour nous rendre compte des méthodes que leurs experts utilisent pour le traitement des données. L'INS utilise habituellement les méthodes factorielles, selon un de ses experts. La classification n'est pas utilisée.

Compte tenu des raisons évoquées ci-dessus à propos des données de l'exercice 2014 et vu que nous sommes en possession aussi des données de l'exercice 2015 pour les variables superficie et population, nous jugeons mieux utiliser finalement les données de l'exercice 2015 fournies par la Division Urbaine de l'Intérieur et Sécurité. Quant à ce qui est de la variable capacité de production, suite aux difficultés rencontrées lors de la collecte des données, nous l'avons simplement estimée sur base du rapport de l'ODP pour l'exercice 2015 à partir du montant total des recttes de la ville-province de Kinshasa et des populations des communes (Cfr Annexe 4).

4.2.1.2. Présentation des logiciels R

[9], [21], [22].

Au niveau de l'application de nos théories sur le procédé PRRC, nous avons fondamentalement utilisé les logiciels R et Excel pour des raisons de gain de temps. Dans la suite, nous allons présenter ces deux logiciels.

R, développé par Ross Ihaka et Robert Gentleman, est un logiciel d'application et un langage de programmation qui est "orienté objet" c'est-à-dire dans la pratique l'on crée des objets (une lettre, un mot, une phrase,une fonction) en leur attribuant une ou plusieurs valeurs numériques puis on les utilise. Il est utilisé plus particulièrement pour le traitement et l'analyse de données statistiques.

Le set up de R est téléchargé gratuitement au site internet de CRAN (Comprehensive R Archive Network) puis installé dans l'ordinateur. Il faut donc disposer de la connexion internet pour le téléchargement et/ou installation de certains packages. Un package étant un ensemble de fichiers et de répertoires nécessaires pour un produit logiciel.

Une fois que le logiciel R est installé, après lancement, la console apparaît. Dans la console, se trouvent reprises les informations relatives à R notamment sa version : « R version 3.1.3 (2015-03-09) – "Smooth Sidewalk" Copyright (C) 2015 The R Foundation

for Statistical Computing Platform : i386-w64-mingw32/i386 (32-bit)...» ainsi que quelques commandes, notamment : demo ( ) pour des démonstrations, par exemple demo (graphics) affiche différents graphiques, help ( ) pour l'aide en ligne, q( ) pour quitter R et le symbole > invitant à saisir une commande.

Lorsqu'on a saisi une commande, si la ligne précédente est incomplète et qu'on valide, le symbole + apparaît en début de ligne. Le symbole # quant à lui est utilisé pour expliquer ou commenter une opération. La ligne qui le porte n'est pas prise en compte par R pour des calculs. En ce qui concerne le symbole <-, il est utilisé en lieu et place de = (égal).

Une ligne tapée est exécutée en appuyant sur la touche ENTER du clavier. Il en est de même lorsqu'on veut passer d'une ligne à une autre.

Dans R le séparateur décimal est le point « . », ainsi les données décimales sont écrites avec un point « . » au lieu d'une virgule « , ». Cette dernière est par contre utilisée pour séparer les termes d'un vecteur ou d'une matrice.

La commande c( ) sert à entrer des données dans un objet. Par exemple, c(1 :8). Il y a possibilité de naviguer dans l'historique des commandes en utilisant les touches Haut et Bas du clavier. Les commandes tapées précédemment dans la console peuvent, dans ce cas, être exécutées de nouveau ou modifiées.

R utilise différentes commandes (ou fonctions). Les éléments sur les commandes de logiciel R sont fournis dans [9], [21] et [22], [33], [75]. Nous en avons utilisés quelques unes (cfr Annexe 3).

R utilise les données provenant d'Excel et vice versa. En effet, si les données sont directement saisies dans la console de R, elle s'exécute directement lorsqu'on valide. Par contre, si elles sont saisies sous Excel, avant de l'utiliser sous R, il faut l'y importer.

Pour importer les données de l'Excel vers R, on procède comme suit : D'abord sous Excel, (1) Saisir les données dans un tableau. Eviter les espaces (sinon les remplacer par des traits d'union), les signes de ponctuations et les caractères accentués afin de permettre l'enregistrement au format capable d'être lu par le logiciel R. Au lieu de la virgule dans Excel, placer un point. (2) Supprimer toutes les feuilles (Sheet) ne contenant pas les données. Ne garder qu'une seule, celle les contenant.(3) Enregistrer le fichier, sous le nom par exemple « TABLEAUBRUT2015 », au format, par exemple, « .txt » avec comme séparateur : tabulation. Confirmer deux fois afin de permettre l'enregistrement

dans le répertoire de travail de R . Ce répertoire est connu en tapant dans la console de R : $> getwd()$. On obtient ainsi un fichier $.txt$, à une seule feuille sur Excel, lisible dans R. (4) Ensuite sous R, saisir la fonction read.table("nomdufichiercréé précédement.txt"). Dans notre cas : read.table("TABLEAUBRUT2015.txt") puis valider. Le tableau créé sous Excel apparaitra dans la console de R.

Pour exporter les données d'un tableau T de R vers Excel, on procède comme suit : (1) Ecrire *write.table(T, 'clipboard', sep='')* sous R puis (2) aller dans la fenêtre Excel sur le fichier et la feuille où l'on souhaite voir s'afficher ce tableau et cliquer dans une cellule à partir de laquelle sera affiché le tableau enfin (3) cliquer sur le bouton *Coller* ou au clavier appuyer sur Ctrl+V.

4.2.2. Réalisation de la classification Ascendante Hiérarchique après ACP

Il est question de classifier les 24 communes de la ville-province de Kinshasa par la méthode de CAH à partir des résultats de l'ACP. Les résultats de cette CAH indirecte seront présentés et interprétés.

4.2.2.1. Tableau des données

Nous avons saisi les données de l'exercice 2015 de 24 communes de la Ville-Province de Kinshasa sous Excel dans un fichier nommé *TABLEAUBRUT2015*, enregistré au format *.txt* et l'avons importé vers R (Cf. Annexe 3, (N° 1)) en vue du centrage et de la réduction des données.

Précisons que pour des raisons d'économie d'espace, les unités de mesure des variables ne seront pas reprises dans les tableaux.

INDIVIDUS	SUPERFICIE	PRODUCTION	POPULATION
Bandalungwa	6,82	27813458131	273218
Barumbu	4,60	18051309794	111758
Bumbu	5,30	38037720000	365716
Gombe	29,33	18051309800	57308
Kalamu	6,64	42716619500	195385
Kasa-Vubu	5,05	20734456481	72940
Kimbanseke	237,78	115326208025	1036732
Kinshasa	2,87	20803773513	169793
Kintambo	2,70	18051309800	84875

Kisenso	16,60	50774724194	359675
Lemba	23,70	49128444931	356853
Limete	67,60	43409789800	294810
Lingwala	2,88	18051309794	123619
Makala	5,60	29199798725	223502
Maluku	7948,80	21081041631	656672
Masina	69,73	68883798181	631364
Matete	4,88	37431195994	223685
Mont-Ngafula	358,92	22467382225	340378
Ndjili	11,40	45749239550	395890
Ngaba	4,00	24338942025	208283
Ngaliema	224,30	88812444194	667608
Ngiri-Ngiri	3,40	22120797081	105664
Nsele	898,70	18051309794	555440
Selembao	23,30	45575946975	491794
Total	**9964,90**	**904662330138**	**8002962**
Ecart-type	**1582,473839**	**23892282493**	**234681,5641**

Tableau 4.8 – Tableau T des données de 24 communes de la Ville-province de Kinshasa. Exercice 2015
Source : Division Urbaine de l'Intérieur et Sécurité-Ville de Kinshasa et ODEP [85]

Les unités de mesure des trois variables considérées : Superficie, Production et Population sont respectivement Km^2, CDF et Habitants (Nombre d'habitants). En ce qui concerne la variable Population, nous aurions dû parler, par exemple d'habitants/Km^2 comme unité exprimant la densité. Celle-ci mesure la population, le nombre d'habitants occupant une surface donnée. La densité se calcule en divisant la population totale par la superficie de la région considérée. Toutefois, dans notre cas le nombre d'habitants est donné de manière brute et servira à estimer la capacité de production de chaque commune.

Puis que les trois variables retenues ont des unités de mesures différentes, elles sont par conséquent hétérogènes. Elles ne peuvent donc pas être utilisées directement. Il faut au préalable les transformer en données réduites ou encore centrées-réduites.

Il ressort de ce tableau que la Ville-province de Kinshasa a une superficie de 9964,9 Km^2 ; une capacité de production des communes estimée à 904662330138 CDF (Cf.

Annexe 4) et une population évaluée à 8002962 (huit million deux mille neuf cent soixante-deux) habitants en 2015.

Plus particulièrement, du point de vue de la superficie, la commune de Maluku est la plus grande (7948,8 Km^2) et celle de Kintambo, la plus petite (2,7 Km^2) de la Ville de Kinshasa. En ce qui concerne la population, la commune de Kimbaseke est la plus peuplée (1036732 habitants) suivie de la commune de Ngaliema (667608 habitants) et de Maluku (656672 habitants) tandis que la commune de la Gombe (57308 habitants) la moins peuplée des communes de la Ville de Kinshasa. Quant à ce qui est de la production, la commune de Kimbanseke aurait plus produit (115326208025 CDF) tandis que les communes de Barumbu, Lingwala et Nsele auraient produit (18 051 309 794) moins que toutes les autres communes. Rappelons que les productions des communes sont une estimation à partir du rapport de l'ODEP- Exercice 2015 (Cf. Annexe 4).

4.2.2.2. Transformation des données

La transformation des données est autorisée dans le cas des variables hétérogènes, qui n'ont pas la même unité de mesure. Mais dans le cas où les variables sont homogènes c'est-à-dire qu'elles ont la même unité de mesure, les données sont directement utilisées. Il convient tout de même de préciser que si les variables homogènes ont la même unité de mesure exprimée différemment pour chaque variable, il faudra convertir cette unité de mesure à une même expression pour toutes les variables.

1. Tableau des données réduites

Les variables que nous avons considérées sont hétérogènes. Le tableau t des données réduites est calculé en divisant chaque valeur du tableau T (tableau 4.8) ci-dessus par l'écart type de la variable correspondante. Le logiciel R calcule le tableau réduit comme cela est précisé à l'Annexe 3, (N° 2). Excel produit le tableau t des données réduites suivant :

INDIVIDUS	SUPERFICIE	PRODUCTION	POPULATION	VALEUR TOTALE
Bandalungwa	0,00430971	1,16411892	1,16420734	2,33263597
Barumbu	0,00290684	0,75552890	0,47621125	1,23464698
Bumbu	0,00334919	1,59205049	1,55834993	3,15374961
Gombe	0,01853427	0,75552890	0,24419473	1,01825789
Kalamu	0,00419596	1,78788358	0,83255368	2,62463322

Kasa-Vubu	0,00319121	0,86783071	0,31080413	1,18182605
Kimbanseke	0,15025841	4,82692301	4,41761160	9,39479302
Kinshasa	0,00181362	0,87073194	0,72350379	1,59604934
Kintambo	0,00170619	0,75552890	0,36166028	1,11889536
Kisenso	0,01048990	2,12515168	1,53260867	3,66825025
Lemba	0,01497655	2,05624745	1,52058387	3,59180787
Limete	0,04271793	1,81689589	1,25621287	3,11582668
Lingwala	0,00181994	0,75552890	0,52675207	1,28410090
Makala	0,00353876	1,22214354	0,95236284	2,17804514
Maluku	5,02302143	0,88233686	2,79814055	8,70349884
Masina	0,04406392	2,88309826	2,69030080	5,61746299
Matete	0,00308378	1,56666472	0,95314262	2,52289111
Mont-Ngafula	0,22680944	0,94036148	1,45038236	2,61755328
Ndjili	0,00720391	1,91481243	1,68692416	3,60894050
Ngaba	0,00252769	1,01869472	0,88751326	1,90873567
Ngaliema	0,14174010	3,71720216	2,84473986	6,70368213
Ngiri-Ngiri	0,00214853	0,92585533	0,45024414	1,37824801
Nsele	0,56790828	0,75552890	2,36678157	3,69021875
Selembao	0,01472378	1,90755935	2,09558003	4,01786317
Total	**6,30**	**37,86**	**34,10**	**78,26**
Ecart-type	**1**	**1**	**1**	

Tableau 4.9 – Tableau t des données réduites
Source : Notre conception à partir du tableau 4.8.

C'est le tableau des données réduites que nous utiliserons lors de calcul des parts des individus. Ce qui montrera la possibilité et l'importance de l'utilisation de plusieurs variables notamment hétérogènes au lieu d'une seule.

2. Tableau des données centrées-réduites

Ceci permet à ce que le centre de gravité soit l'origine des axes (données centrées) et que l'influence des unités de mesures soit anéantie (données réduites). Le tableau (TABLEAUBRUT2015) des données étant importé d'Excel vers R. Il est nommé T (Cfr Annexe 3 (N° 1)) et utilisé par R pour créer le tableau Tcr des données centrées-réduites (Cfr Annexe 3 (N° 3)) ci-dessous :

INDIVIDUS	SUPERFICIE	PRODUCTION	POPULATION
Bandalungwa	-0.25263333	-0.40484895	-0.25127847
Barumbu	-0.25400666	-0.80483611	-0.92478881
Bumbu	-0.25357363	0.01407253	0.13456546
Gombe	-0.23870827	-0.80483611	-1.15192023
Kalamu	-0.25274469	0.20578235	-0.57594916
Kasa-Vubu	-0.25372829	-0.69489881	-1.08671328
Kimbanseke	-0.10975758	3.18083484	2.93362540
Kinshasa	-0.25507687	-0.69205867	-0.68270302
Kintambo	-0.25518204	-0.80483611	-1.03692791
Kisenso	-0.24658326	0.53594927	0.10936618
Lemba	-0.24219108	0.46849583	0.09759455
Limete	-0.21503380	0.23418381	-0.16121012
Lingwala	-0.25507068	-0.80483611	-0.87531212
Makala	-0.25338805	-0.34804603	-0.45866259
Maluku	4.66040946	-0.68069808	1.34825229
Masina	-0.21371615	1.27793731	1.24268311
Matete	-0.25383345	-0.01077875	-0.45789923
Mont-Ngafula	-0.03481833	-0.62389517	0.02887114
Ndjili	-0.24980007	0.33003872	0.26043255
Ngaba	-0.25437783	-0.54721124	-0.52214676
Ngaliema	-0.11809654	2.09447916	1.39387046
Ngiri-Ngiri	-0.25474900	-0.63809590	-0.95020918
Nsele	0.29909868	-0.80483611	0.92597557
Selembao	-0.24243853	0.32293836	0.66048418

Tableau 4.10 – Tableau Tcr des données centrées réduites
Source : Notre conception à partir du tableau 4.8.

4.2.2.3. Analyse en Composantes Principales à partir du tableau des données centrées-réduites

Nous nous limiterons à présenter d'abord les valeurs propres et les pourcentages de variance en vue de nous rendre compte de la possibilité ou l'impossibilité de réalisation de l'ACP. Ensuite, nous nous servirons des résultats de l'ACP sur les individus, plus précisément les coordonnées des individus afin d'en extraire les deux ou trois premières composantes principales qui restituent la quasi-totalité de l'inertie du nuage des points.

Nous laisserons de côté le cercle de corrélations, le graphique des individus ainsi que les résultats sur les variables et d'autres détails sur les résultats des individus.

4.2.2.3.1. Valeurs propres et contributions des axes factoriels à l'inertie totale

Les différentes valeurs propres de la matrice des corrélations et les contributions des axes factoriels sont produites par le logiciel R à partir des résultats de l'ACP (Cfr Annexe 3 (N° 5))

	Eigen value	**Percentage of variance**	**Cumulative percentage of variance**
Comp 1	1.81070696	60.356899	60.35690
Comp 2	1.09768689	36.589563	96.94646
Comp 3	0.09160615	3.053538	100.00000

Tableau 4.11 – Tableau des valeurs propres et de contributions des axes factoriels
Source : Notre conception à partir du tableau Tcr (Tableau 4.8.)

Les deux premiers axes restituent $96,95\%$ de l'inertie totale ou de l'information contenue dans le tableau des données ($60,36\%$ contribution de la première composante et $36,59\%$ celle de la deuxième). Ainsi, ces deux axes conservés seront utilisés lors de la représentation graphique et lors de la classification.

4.2.2.3.2. Résultats de l'ACP sur les individus

Il s'agit de : coordonnées des individus, les corrélations individus - axes factoriels, la qualité de projection des individus, la contribution des individus aux axes factoriels. Le logiciel R renvoie tous ces éléments (Cf. Annexe 3 (N° 7)). Toutefois, nous ne nous sommes intéressé qu'aux coordonnées des individus et en avons extrait les deux premières composantes principales dans le tableau ci-dessous nommé res.acp2 (Cfr Annexe 3 (N° 8)).

Individus	**Dim.1**	**Dim.2**
Bandalungwa	-0.509589699	-0.10664698
Barumbu	-1.278338234	-0.02771781
Bumbu	0.059428388	-0.22777099
Gombe	-1.442862389	-0.03914023
Kalamu	-0.334129395	-0.38259931
Kasa-Vubu	-1.322997446	-0.08885199
Kimbanseke	4.303435151	-1.01535786

Kinshasa	-1.023384346	-0.04542019
Kintambo	-1.361266408	-0.04153313
Kisenso	0.396815161	-0.42837335
Lemba	0.343155237	-0.39918870
Limete	-0.001630463	-0.31137075
Lingwala	-1.242057783	-0.02310479
Makala	-0.624079459	-0.15310316
Maluku	1.439171358	4.78467752
Masina	1.743175451	-0.55954880
Matete	-0.394433847	-0.28545994
Mont-Ngafula	-0.409417495	0.21488855
Ndjili	0.367681962	-0.33365312
Ngaba	-0.806420451	-0.08326324
Ngaliema	2.428120385	-0.77248363
Ngiri-Ngiri	-1.183931427	-0.09656719
Nsele	0.194238632	0.70020927
Selembao	0.659317115	-0.27862019

Tableau 4.12 – Tableau des coordonnées des individus pour les deux premiers facteurs

Source : Notre conception à partir des résultats de l'ACP

4.2.2.4. Classification Ascendante Hiérarchique à partir des résultats de l'ACP

Comme nous l'avons signalé précédemment, les deux premières composantes restituent la quasi-totalité de l'inertie totale $(96,95\%)$. De ce fait, elles seront utilisées pour réaliser la CAH. Nous allons construire et découper le dendrogramme avec la commande *HCPC* [88] du package *FactoMineR* du logiciel R. Cette commande propose une coupure du dendrogramme en trois classes (Cf. Annexe 3 (N° 9)). Toutefois, nous optons pour un découpage en six classes pour éviter qu'il y ait des classes dont l'effectif approche ou dépasse la moitié de l'effectif de la population.

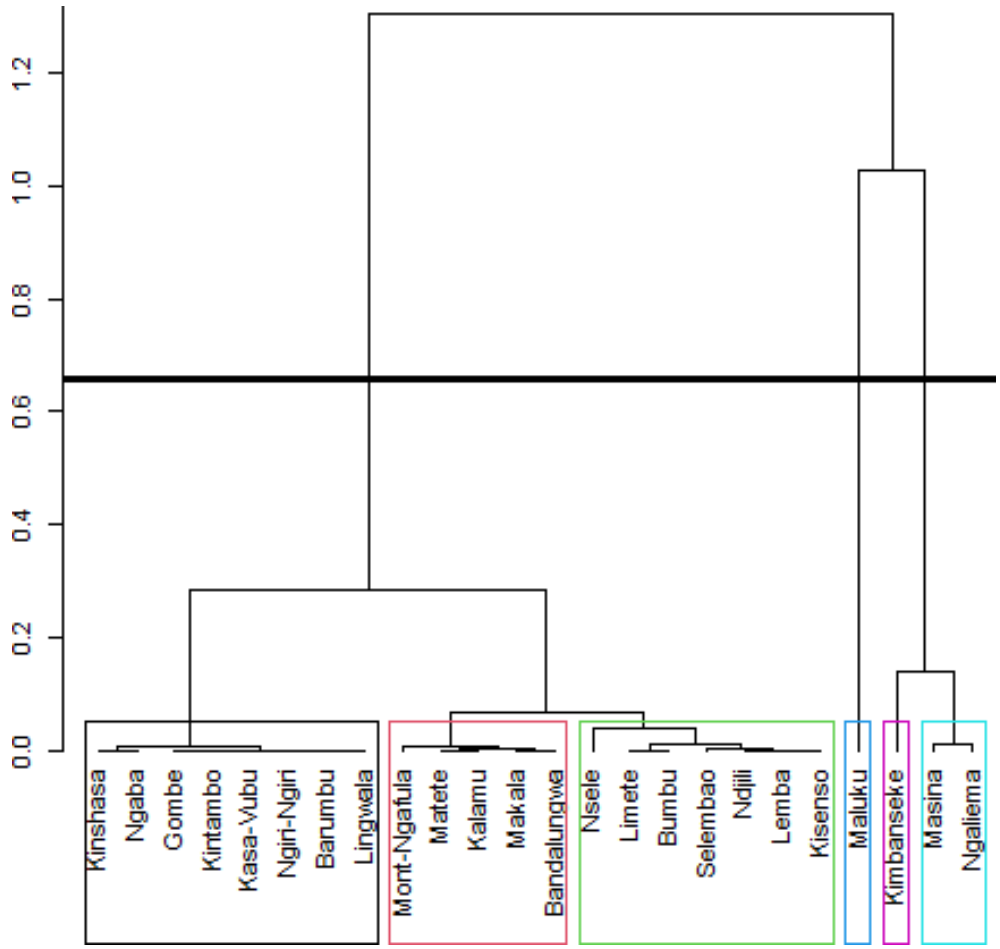

Figure 4.1 – Dendrogramme des 24 communes de Kinshasa découpées en 6 classes
Source : Notre conception à partir des résultats de l'ACP

Le logiciel R adjoint automatiquement chaque individu au numéro de sa classe d'appartenance (Cfr Annexe 3 (N° 10)).

Individus	Dim.1	Dim.2	clust
Gombe	-1.442862389	-0.03914023	1
Kintambo	-1.361266408	-0.04153313	1
Kasa-Vubu	-1.322997446	-0.08885199	1
Barumbu	-1.278338234	-0.02771781	1
Lingwala	-1.242057783	-0.02310479	1
Ngiri-Ngiri	-1.183931427	-0.09656719	1
Kinshasa	-1.023384346	-0.04542019	1
Ngaba	-0.806420451	-0.08326324	2
Makala	-0.624079459	-0.15310316	2
Bandalungwa	-0.509589699	-0.10664698	2
Mont-Ngafula	-0.409417495	0.21488855	2
Matete	-0.394433847	-0.28545994	2

Kalamu	-0.334129395	-0.38259931	2
Limete	-0.001630463	-0.31137075	3
Bumbu	0.059428388	-0.22777099	3
Nsele	0.194238632	0.70020927	3
Lemba	0.343155237	-0.39918870	3
Ndjili	0.367681962	-0.33365312	3
Kisenso	0.396815161	-0.42837335	3
Selembao	0.659317115	-0.27862019	3
Maluku	1.439171358	4.78467752	4
Masina	1.743175451	-0.55954880	5
Ngaliema	2.428120385	-0.77248363	5
Kimbanseke	4.303435151	-1.01535786	6

Tableau 4.13 – Tableau des numéros des classes respectives des individus

Source : Notre conception à partir des résultats de la classification

Ce tableau montre que les individus Barumbu, Gombe, Kasa-vubu, Kinshasa, Kintambo, Lingwala et Ngiri-Ngiri appartiennent à la 1ère classe bien que le logiciel ait omis Ngaba qu'il a mentionné dans la $2^{è}$ classe. En réalité, ce qui importe ce n'est pas le numéro d'une classe mais les individus regroupés dans cette classe.

Les individus se répartissent dans les 6 classes de la manière suivante : 1ère Classe : 1) Barumbu, 2) Gombe, 3) Kasa-Vubu, 4) Kinshasa, 5) Kintambo, 6) Lingwala, 7) Ngaba, 8) Ngiri-Ngiri ; $2^{è}$ Classe : 1) Bandalungwa, 2) Kalamu, 3) Makala, 4) Matete, 5) Mont-Ngafula ; $3^{è}$ Classe : 1) Bumbu, 2) Kisenso, 3) Lemba, 4) Limete, 5) Ndjili, 6) Nsele, 7) Selembao ; $4^{è}$ Classe : 1) Maluku ; $5^{è}$ Classe : 1) Masina, 2) Ngaliema ;$6^{è}$ Classe : 1) Kimbanseke.

4.2.2.5. Interprétation de la partition obtenue

1. Description des classes par les individus

La description d'une classe par les individus se fait par la détermination des individus les plus typiques de la classe. On détermine pour chaque classe le parangon et l'extrême. Le parangon d'une classe étant l'individu le plus proche du centre de gravité de cette classe et l'extrême, l'individu le plus éloigné des centres des autres classes.

1) Détermination des parangons

Pour chaque classe, l'individu qui aura la plus petite distance du barycentre sera retenu comme parangon. R renvoie automatiquement les parangons des classes (Cfr Annexe 3 (N° 12)).

Cluster : 1				
Barumbu 0.02750742	Lingwala 0.03669511	Kasa-Vubu 0.06886248	Ngiri-Ngiri 0.09260589	Kinstambo 0.09683135
Cluster : 2				
Bandalungwa 0.02627417	Makala 0.11292670	Matete 0.19338336	Ngaba 0.29754397	Kalamu 0.30732699
Cluster : 3				
Njili 0.1705096	Lemba 0.2233169	Bumbu 0.2333979	Kisenso 0.2685370	Limete 0.3173260
Cluster : 4				
Maluku 0				
Cluster : 5				
Masina 0.3586401	Ngaliema 0.3596401			
Cluster : 6				
Kimbanseke 0				

Tableau 4.14 – Tableau des parangons

Source : Notre conception à partir des résultats de la classification

Le parangon de chaque classe est le premier individu de la ligne correspondante de ce tableau. Ainsi, les parangons des classes sont les suivants : 1ère Classe : Barumbu, $2^{è}$ Classe : Bandalungwa, $3^{è}$ Classe : Ndjili, $4^{è}$ Classe : Maluku, $5^{è}$ Classe : Masina et $6^{è}$ classe : Kimbanseke

Le parangon peut être utilisé en lieu et place du barycentre pour remplacer sa classe car c'est un individu réel alors que le barycentre est fictif.

2) Détermination des extrêmes

Cluster : 1				
Gombe 0.9345454	Kintambo 0.853195	Kasa-Vubu 0.8111716	Barumbu 0.7724930	Lingwala 0.7372372
Cluster : 2				
Mont-Ngafula 0.8031513	Bandalungwa 0.7573784	Matete 0.6905547	Kalamu 0.6538706	Makala 0.3488601
Cluster : 3				
Selembao 1.1813756	Nsele 1.0926740	Kisenso 0.9566657	Ndjili 0.9033297	Lemba 0.8966825
Cluster : 4				
Maluku 5.098908				
Cluster : 5				
Ngaliema 1.890977	Masina 1.502769			
Cluster : 6				
Kimbanseke 2.245132				

Tableau 4.15 – Tableau des extrêmes
Source : Notre conception à partir des résultats de la classification

L'extrême de chaque classe est le premier individu de chaque ligne correspondante du tableau ci-dessus. Ainsi, les extrêmes des classes sont les suivants : 1ère Classe : Gombe, $2^{è}$ Classe : Mont-Ngafula, $3^{è}$ Classe : Selembao, $4^{è}$ Classe : Maluku, $5^{è}$ Classe : Ngaliema et $6^{è}$ Classe : Kimbanseke.

Tableau 4.16 – Tableau synthétique des parangons et des extrêmes des classes

Classes	**Parangons**	**Extrêmes**
K_1	Barumbu	Gombe
K_2	Bandalungwa	Mont-Ngafula
K_3	Ndjili	Selembao
K_4	Maluku	Maluku
K_5	Masina	Ngaliema
K_6	Kimbanseke	Kimbanseke

Source :Notre conception en référence aux tableaux précédents

2. Caractérisation des classes par les variables

Nous allons chercher à retrouver la variable ou les variables qui caractérisent chaque classe. Une telle variable est la plus importante de la classe. C'est la variable dont la valeur test est supérieure à la valeur critique (lue sur la table). Cela se confirme avec une valeur p qui doit être inférieure au seuil de significativité α (Cf. Annexe 5). Toutefois, R renvoie automatiquement les variables qui caractérisent les classes (Annexe 3 (N° 13)). Il faut utiliser les résultats de la commande ***HCPC*** du package ***FactoMineR*** (Cf. Annexe 8).

Les variables caractérisant les classes se présentent comme suit :

Tableau 4.17 – Liste des variables caractérisant les classes

Classes	**Variables les plus importantes**
K_1	Production, Population
K_2	Aucune
K_3	Aucune
K_4	Superficie
K_5	Production
K_6	Production, Population

Source : Notre conception en référence aux tableaux précédents

4.2.3. Réduction des inégalités entre les individus dans leurs classes et dans l'ensemble de la population

Nous allons utiliser le procédé PRRC (Procédé de la Répartition des Ressources à partir des résultats de la Classification) que nous avons proposé pour répartir les recettes à caractère national entre les 24 communes de la ville/province de Kinshasa.

4.2.3.1. Valeurs totales des individus

Les résultats de la classification de 24 communes de la ville-province de Kinshasa sont déterminés au point (4.2.2.).

Nous allons utiliser le tableau 4.9. (Cfr 4.2.2.2.) des données réduites pour lesquelles l'influence des unités de mesure est anéantie. Nous y avons calculé les valeurs totales des individus.

N^o	**Individu**	**Superficie**	**Production**	**Population**	**Valeur totale**
Classe 1 (8 Individus)					
1	Barumbu	0,00290684	0,7555289	0,47621125	1,23464698
2	Gombe	0,01853427	0,7555289	0,24419473	1,01825789
3	Kasa-Vubu	0,00319121	0,86783071	0,31080413	1,18182605
4	Kinshasa	0,00181362	0,87073194	0,72350379	1,59604934
5	Kintambo	0,00170619	0,7555289	0,36166028	1,11889536
6	Lingwala	0,00181994	0,7555289	0,52675207	1,2841009
7	Ngaba	0,00252769	1,01869472	0,88751326	1,90873567
8	Ngiri-Ngiri	0,00214853	0,92585533	0,45024414	1,37824801
Classe 2 (5 Individus)					
9	Bandalungwa	0,00430971	1,16411892	1,16420734	2,33263597
10	Kalamu	0,00419596	1,78788358	0,83255368	2,62463322
11	Makala	0,00353876	1,22214354	0,95236284	2,17804514
12	Matete	0,00308378	1,56666472	0,95314262	2,52289111
13	Mont-Ngafula	0,22680944	0,94036148	1,45038236	2,61755328
Classe 3 (7 Individus)					
14	Bumbu	0,00334919	1,59205049	1,55834993	3,15374961
15	Kisenso	0,0104899	2,12515168	1,53260867	3,66825025
16	Lemba	0,01497655	2,05624745	1,52058387	3,59180787
17	Limete	0,04271793	1,81689589	1,25621287	3,11582668
18	Ndjili	0,00720391	1,91481243	1,68692416	3,6089405
19	Nsele	0,56790828	0,7555289	2,36678157	3,69021875
20	Selembao	0,01472378	1,90755935	2,09558003	4,01786317
Classe 4 (1 Individu)					
21	Maluku	5,02302143	0,88233686	2,79814055	8,70349884
Classe 5 (2 Individus)					
22	Masina	0,04406392	2,88309826	2,6903008	5,61746299
23	Ngaliema	0,1417401	3,71720216	2,84473986	6,70368213
Classe 6 (1 Individu)					
24	Kimbanseke	0,15025841	4,82692301	4,4176116	9,39479302
Total		**6,30**	**37,86**	**34,10**	**78,26**

Tableau 4.18 – Tableau des données réduites et des valeurs totales des individus

Source : Notre conception à partir du tableau 4.9

La valeur totale, par exemple, de la commune Barumbu (W_{11} : premier individu de la première classe) est
$W_{11} = \sum_{j=1}^{m} X_{1j} = 0,00290684 + 0,7555289 + 0,47621125 = 1,23464698$
et la valeur globale (valeur toale de l'ensemble de la population en étude) est
$V = \sum_{i=1}^{n} W_i = 78,26.$

4.2.3.2. Réduction des inégalités entre les individus dans leurs classes et dans la population

Cette réduction des inégalités concernent les classes ayant au moins deux individus. Pour celles ayant un seul individu, celui-ci gardera sa valeur totale. C'est le cas de la $4^{\grave{e}}$ classe et $6^{\grave{e}}$ classe.

1. Réduction des inégalités dans les classes

Nous résumons les calculs sur la réduction des inégalités au niveau des classes dans le tableau suivant :

N^o	**Individu**	**Valeur totale** : W_i	**Différence à la plus petite valeur totale :** $W_i - W_1$	**Valeur corrigée :** Z_i= $W_1 + W_i$xJ_M	**Ecart entre valeur corrigée et valeur totale :** $Z_i - W_i$
Classe 1 (8 Individus)					
1	Gombe	1,01825789	0	1,26280264	0,244545
2	Kintambo	1,11889536	0,10063747	1,28697173	0,168076
3	Kasa-Vubu	1,18182605	0,16356816	1,30208516	0,120259
4	Barumbu	1,23464698	0,21638909	1,31477063	0,080124
5	Lingwala	1,2841009	0,26584301	1,32664749	0,042547
Sous-total		5,83772718	0,74643773	6,49327766	0,655550
6	Ngiri-Ngiri	1,37824801	0,35999012	1,34925785	-0,028990
7	Kinshasa	1,59604934	0,57779145	1,40156500	-0,194484
8	Ngaba	1,90873567	0,89047778	1,47665973	-0,432076
Sous-total		4,88303302	1,82825935	4,227482587	-0,655550
Total		10,7207602	2,57469708	10,72076025	0
J_M =0,24015994					
Classe 2 (5 Individus)					
1	Makala	2,17804514	0	2,4238754	0,245830
2	Bandalungwa	2,33263597	0,1545908	2,4413237	0,108688

Sous-total		4,51068111	0,15459083	4,86519913	0,354518
3	Matete	2,52289111	0,3448460	2,4627973	-0,060094
4	Mont-Ngafula	2,61755328	0,4395081	2,4734816	-0,144072
5	Kalamu	2,62463322	0,4465881	2,4742807	-0,150353
Sous-total		7,76507761	1,23094219	7,41055954	-0,354518
Total		16,7864398	1,5401239	17,1409578	0
J_M = 0,1128674					
Classe 3 (7 Individus)					
1	Limete	3,11582668	0	3,49653163	0,380705
2	Bumbu	3,15374961	0,03792293	3,50116521	0,347415
Sous-total		6,26957629	0,03792293	6,9976968	0,728120
3	Lemba	3,59180787	0,47598119	3,55468903	-0,037119
4	Ndjili	3,6089405	0,49311382	3,55678237	-0,052158
5	Kisenso	3,66825025	0,55242357	3,56402909	-0,104221
6	Nsele	3,69021875	0,57439207	3,56671329	-0,123505
7	Selembao	4,01786317	0,90203649	3,60674628	-0,411117
Sous-total		18,5770805	2,99794714	17,84896005	-0,728120
Total		24,8466568	3,03587007	24,84665689	0
J_M = 0,12218425					
Classe 4 (1 Individu)					
1	Maluku	8,70349884	0	8,70349884	0
J_M = 0					
Classe 5 (2 Individus)					
1	Masina	5,61746299	0	6,11269257	0,4952296
2	Ngaliema	6,70368213	1,08621914	6,20845250	-0,4952296
Total		12,3211451	1,08621914	12,32114507	0
J_M =0,08815894					
Classe 6 (1 Individu)					
1	Kimbanseke	9,39479302	0	9,39479302	0
J_M = 0					
Total général		**78,26**		**78,26**	**0**

Tableau 4.19 – Tableau des valeurs corrigées des individus par rapport à leurs classes

Source : Notre conception à partir du tableau précédent

Nous venons de réduire les inégalités entre les individus dans leurs classes respectives. Considérons la 1ère classe (Cfr Le tableau ci-dessus). Sachant que la plus petite valeur totale est celle de Gombe : W_1=$1,01825789$, alors :

(1) L'indice de niveau des inégalités égale :

$J_M = \frac{\sum_{i=1}^{n}(W_i - W_1)}{\sum_{i=1}^{n} W_i} = \frac{2{,}57469708}{10{,}7207602} = 0,240159936$
Donc l'inégalité dans la 1ère classe est de 24%.

(2) La valeur corrigée, par exemple, de Kintambo est :

$Z_2 = W_1 + W_2.J_M = 1,01825789 + 1,11889536\text{x}0,240159936 = 1,28697173$

(3) L'écart entre la valeur corrigée et la valeur totale est :

• L'individu Kintambo (Pauvre) a bénéficié de :
$Z_2 - W_2$ = $1,28697173 - 1,11889536 = 0,168076$ au détriment des autres (les riches).

• Par contre l'individu Ngaba (Riche) a perdu 0,432076 (d'où -0,432076) au profit des autres (les pauvres).

Cet écart a permis de séparer la 1ère classe en deux parties : les pauvres (Gombe, Kintambo, Kasa-Vubu, Barumbu et Lingwala) d'un côté et les riches (Ngiri-Ngiri, Kinshasa et Ngaba) de l'autre.

2. Réduction des inégalités dans l'ensemble de la population

Les valeurs corrigées Z_i des individus déterminées au niveau des classes seront utilisées pour calculer les valeurs corrigées Z'_i au niveau de l'ensemble de la population. Ce qui permettra la réduction des inégalités à ce niveau.

De ce fait, les calculs se trouvent résumés dans le tableau ci-dessous :

N^o	**Individu**	**Valeur corrigée au niveau des classes :** Z_i	**Différence à la plus petite valeur corrigée :** $Z_i - Z_1$	**Valeur corrigée au niveau de la population :** Z'_i=$Z_1 + Z_i$xJ'_M	**Ecart entre valeur corrigée et valeur totale :** $Z'_i - Z_i$
1	Gombe	1,26280264	0	2,03658388	0,7737812
2	Kintambo	1,28697173	0,02416909	2,05139347	0,7644217
3	Kasa-vubu	1,30208516	0,03928252	2,06065422	0,7585691
4	Barumbu	1,31477063	0,05196799	2,06842723	0,7536566

5	Lingwala	1,32664749	0,06384485	2,07570476	0,7490573
6	Ngiri-Ngiri	1,34925785	0,08645521	2,08955924	0,7403014
7	Kinshasa	1,40156500	0,13876236	2,12161040	0,7200454
8	Ngaba	1,47665973	0,21385709	2,16762463	0,6909649
9	Makala	2,42387540	1,16107276	2,74803023	0,3241548
10	Bandalungwa	2,44132370	1,17852106	2,75872166	0,3173980
11	Matete	2,46279730	1,19999466	2,77187959	0,3090823
12	Mont-Ngafula	2,47348160	1,21067896	2,77842639	0,3049448
13	Kalamu	2,47428070	1,21147806	2,77891604	0,3046353
Sous-total		**22,99651893**	**6,58008461**	**30,50753175**	**7,5110128**
14	Limete	3,49653163	2,23372899	3,40529942	-0,0912322
15	Bumbu	3,50116521	2,23836257	3,40813865	-0,0930266
16	Lemba	3,55468903	2,29188639	3,44093532	-0,1137537
17	Ndjili	3,55678237	2,29397973	3,44221801	-0,1145644
18	Kisenso	3,56402909	2,30122645	3,44665844	-0,1173707
19	Nsele	3,56671329	2,30391065	3,44830318	-0,1184101
20	Selembao	3,60674628	2,34394364	3,47283336	-0,1339129
21	Masina	6,11269257	4,84988993	5,00834982	-1,1043428
22	Ngaliema	6,20845250	4,94564986	5,06702663	-1,1414259
23	Maluku	8,70349884	7,44069620	6,59586416	-2,1076347
24	Kimbanseke	9,39479302	8,13199038	7,01945408	-2,3753389
Sous-total		**55,2660938**	**41,3752648**	**47,7550811**	**-7,5110128**
Total		**78,2626128**	**47,9553494**	**78,2626128**	**0**
J_M**=0,61274915**					

Tableau 4.20 – Tableau des valeurs corrigées des individus par rapport à l'ensemble de la population

Source :Notre conception à partir du tableau précédent

Nous venons de réduire les inégalités entre les individus dans l'ensemble de la population. Sachant que la plus petite valeur corrigée (issue de la première opération de réduction des inégalités) est celle de Gombe : Z_1=$1,26280264$, alors :

(1) L'indice de niveau des inégalités pour l'ensemble de la population égale :

$J'_M = \frac{\sum_{i=1}^{n}(Z_i - Z_1)}{\sum_{i=1}^{n} Z_i} = \frac{47{,}9553494}{78{,}2626128} = 0,61274915$

Donc l'inégalité dans la population est de 61,3%.

(2) La valeur corrigée, par exemple, de Kintambo :

$Z'_2 = Z_1 + Z_2.J'_M = 1,26280264 + 1,28697173\text{x}0,61274915 = 2,05139347$

(3) L'écart entre la valeur corrigée et la valeur totale :

• L'individu Kintambo (Pauvre) a bénéficié de :
$Z'_2 - Z_2 = 2,05139347 - 1,28697173 = 0,7644217$ au détriment des autres (les riches).

• Par contre l'individu Limete (Riche) a perdu 0,0912322 (d'où -0,0912322) au profit des autres (les pauvres).

Cet écart a permis de diviser la population en deux parties : les pauvres qui sont au nombre de 13 (Gombe, Kintambo, Kasa-Vubu, Barumbu, Lingwala, Ngiri-Ngiri, Kinshasa, Ngaba, Makala, Bandalungwa, Matete, Kalamu et Mont-Ngafula) d'un côté et de l'autre les riches au nombre de 11 (Limete, Bumbu, Lemba, Ndjili, Kisenso, Selembao, Nsele, Masina, Ngaliema Kimbanseke et Maluku).

4.2.4. Proportions et parts des individus par rapport à la population

4.2.4.1. Proportions et parts des individus par rapport à la population

Les valeurs corrigées des individus au niveau de l'ensemble de la population vient résoudre le problème des inégalités entre les individus dans leur ensemble, sachant que le problème a d'abord été résolu au niveau des classes. Ces valeurs corrigées seront utilisées pour calculer les proportions (corrigées) des individus par rapport à l'ensemble de la population.

De ce fait, les proportions des individus par rapport à l'ensemble de la population se trouvent calculées dans le tableau ci-dessous :

N^o	**Individu**	**Valeur corrigée au niveau de la population :** Z'_i	**Proportion par rapport à la population :**$P_i = \frac{Z'_i}{V}$
1	Gombe	2,03658388	0,0260
2	Kintambo	2,05139347	0,0262
3	Kasa-vubu	2,06065422	0,0263
4	Barumbu	2,06842723	0,0264
5	Lingwala	2,07570476	0,0265
6	Ngiri-Ngiri	2,08955924	0,0267
7	Kinshasa	2,1216104	0,0271
8	Ngaba	2,16762463	0,0277
9	Makala	2,74803023	0,0351
10	Bandalungwa	2,75872166	0,0352
11	Matete	2,77187959	0,0354
12	Mont-Ngafula	2,77842639	0,0355
13	Kalamu	2,77891604	0,0355
Sous-total		30,5075318	0,3898
14	Limete	3,40529942	0,0435
15	Bumbu	3,40813865	0,0435
16	Lemba	3,44093532	0,0440
17	Ndjili	3,44221801	0,0440
18	Kisenso	3,44665844	0,0440
19	Nsele	3,44830318	0,0441
20	Selembao	3,47283336	0,0444
21	Masina	5,00834982	0,0640
22	Ngaliema	5,06702663	0,0647
23	Maluku	6,59586416	0,0843
24	Kimbanseke	7,01945408	0,0897
Sous-total		47,7550811	0,6102
Total		**78,2626128**	**1**

Tableau 4.21 – Tableau des proportions des individus par rapport à l'ensemble de la population
Source : Notre conception à partir du tableau précédent

Sachant que la proportion P_i d'un individu i est le rapport de sa valeur corrigée Z'_i par rapport à l'ensemble de la population et de la valeur globale V, alors, à titre d'exemple, la proportion P_2, de l'individu Kintambo est calculée comme suit : $P_2 = \frac{Z'_2}{V} = \frac{2{,}05139347}{78{,}2626128} = 0,0262$, soit 2,62%

4.2.4.2. Parts des individus à partir de leurs proportions corrigées

A partir des proportions corrigées P_i trouvées ci-dessus, nous déterminons les parts C_i de la ressource commune $C = 144745972822CDF$ revenant aux communes (C constitue les recettes à caractère national allouées aux ETDs qui sont dans notre cas les communes de la ville-province de Kinshasa). La part C_i revenant à l'individu i est le produit de sa proportion corrigée au niveau de la population et de la ressource commune.

Les parts des individus (communes) sont présentées dans le tableau ci-après :

N^o	**Individu :i**	**Valeur corrigée :** Z'_i	**Proportion par rapport à la population :**P_i	**Part :** C_i
1	Gombe	2,03658388	0,026	3763395293
2	Kintambo	2,05139347	0,0262	3792344488
3	Kasa-vubu	2,06065422	0,0263	3806819085
4	Barumbu	2,06842723	0,0264	3821293683
5	Lingwala	2,07570476	0,0265	3835768280
6	Ngiri-Ngiri	2,08955924	0,0267	3864717474
7	Kinshasa	2,1216104	0,0271	3922615863
8	Ngaba	2,16762463	0,0277	4009463447
9	Makala	2,74803023	0,0351	5080583646
10	Bandalungwa	2,75872166	0,0352	5095058243
11	Matete	2,77187959	0,0354	5124007438
12	Mont-Ngafula	2,77842639	0,0355	5138482035
13	Kalamu	2,77891604	0,0355	5138482035
14	Limete	3,40529942	0,0435	6296449818
15	Bumbu	3,40813865	0,0435	6296449818
16	Lemba	3,44093532	0,044	6368822804
17	Ndjili	3,44221801	0,044	6368822804
18	Kisenso	3,44665844	0,044	6368822804

19	Nsele	3,44830318	0,0441	6383297401
20	Selembao	3,47283336	0,0444	6426721193
21	Masina	5,00834982	0,064	9263742261
22	Ngaliema	5,06702663	0,0647	9365064442
23	Maluku	6,59586416	0,0843	12202085509
24	Kimbanseke	7,01945408	0,0897	12983713762
Sous-total		47,7550811	0,6102	88323992616
Total		**78,2626128**	**1**	**144745972822**

Tableau 4.22 – Tableau des parts des individus
Source : Notre conception à partir du tableau 4.21

Soit à calculer la part de l'individu Kintambo, on a :
$C_2 = 0,0262\text{x}144745972822CDF = 3792344488CDF$

En comparant ce tableau des parts à celui publié par l'ODEP (Cfr Annexe 4), nous remarquons, par exemple, que la commune de la Gombe aurait une rétrocession de 3763395293 CDF au lieu de 2888209568 CDF, montant lui alloué réellement en 2015. Il y a donc un manque à gagner de 8751857 CDF. De même pour la commune de Maluku qui aurait 12202085509 CDF au lieu de 3133129739 CDF soit un manque à gagner de 9068955770 CDF. Par contre, pour la commune de Selembao, elle aurait 6426721193 CDF au lieu de 7292151516 CDF soit un surplus de 865430323 CDF. Toutefois, ces écarts peuvent être corrigés dans les jours à venir en utilisant notre procédé qui tient compte de plusieurs variables notamment hétérogènes et des rapports entre les individus.

Les proportions corrigées calculées ci-dessus peuvent aussi être utilisées pour déterminer les quotas des communes dans la répartition des sièges électoraux [48]. Dans ce cas, on pourra tenir compte de trois critères (au lieu d'un seul : nombre d'électeurs ou encore population) et réduire les inégalités. Elles peuvent aussi être utilisées dans le cas où les communes doivent apporter leurs contributions à la ville-province de Kinshasa.

4.2.5. Représentation graphique des parts des communes

Nous allons utiliser le tableau des parts des 24 communes de la Ville-province de Kinshasa de la ressource commune (les recettes à caractère national allouées aux ETDs exercice 2015) (Annexe 3 (N° 14)) pour les représenter graphiquement.

4.2.5.1. Graphique à bâtons des parts des communes

Le logiciel R renvoie automatiquement le graphique à bâtons des parts (corrigées) des communes (Cfr Annexe 3 (N° 15))

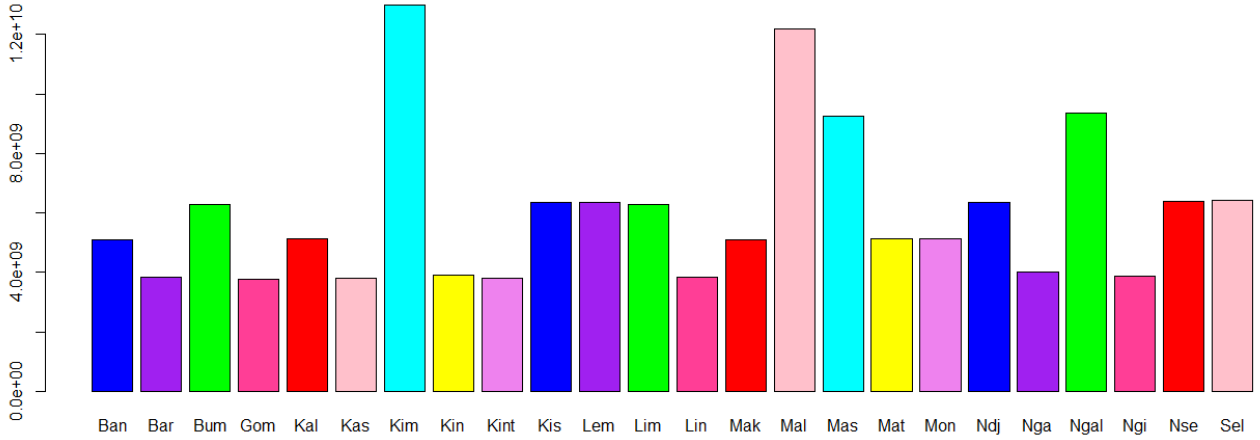

Figure 4.2 – Graphique à bâtons des parts de 24 communes de la Ville-province de Kinshasa
Source : Notre conception à partir du tableau 4.22

Légende :1)"Ban" : Bandalungwa, 2) "Bar" : Barumbu, 3) "Bum", : Bumbu, 4) "Gom" : Gombe, 5) "Kal" : Kalamu, 6) "Kas" : Kasa-vubu, 7) "Kim" : Kimbanseke, 8) "Kin" : Kinshasa, 9) "Kint" : Kintambo, 10) "Kis" : Kisenso, 11) "Lem" : Lemba, 12) "Lim" : Limete, 13) "Lin" : Lingwala, 14) "Mak" : Makala, 15) "Mal" : Maluku, 16) "Mas" : Masina, 17) "Mat" : Matete, 18) "Mon" : Mont-Ngafula, 19) "Ndji" : Ndjili, 20) "Nga" : Ngaba, 21) "Ngal" : Ngaliema, 22) "Ngi" : Ngiri-Ngiri, 23) "Nse" : Nsele, 24) "Sel" : Selembao.

A la lumière du graphique ci-dessus, la part (corrigée) de Kimbanseke est la plus grande suivie de celle de Maluku puis de Ngaliema puis de Masina, . . . Par contre, celle de Gombe est la plus petite précédée de celle de Kintambo, Kasa-vubu. . .

4.2.5.2. Graphique en Camembert des parts des individus

Nous utilisons la même matrice qui a servi à la construction du graphique à bâtons. Le logiciel R renvoie automatiquement le graphique en Camembert des parts (corrigées) des communes (Annexe 3 (N° 16))

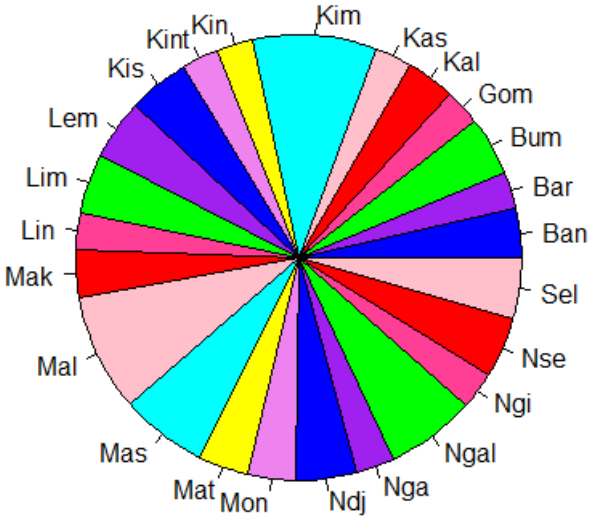

Figure 4.3 – Graphique en camembert des parts de 24 communes de la Ville-province de Kinshasa
Source : Notre conception à partir du tableau 4.22

4.3. Conclusion du troisième chapitre

Nous avons appliqué nos deux procédés PRRS et PRRC qui utilisent plusieurs variables notamment hétérogènes pour résoudre le problème d'injustice due à l'utilisation directe des données issues de plusieurs variables homogènes pour lesquelles la même unité de mesure est exprimée différemment pour chaque variable, l'utilisation directe des données issues des variables hétérogènes, l'absence d'un mécanisme de répartition des ressources utilisant plusieurs variables notamment hétérogènes, l'absence d'un mécanisme de réduction des inégalités entre les individus.

Le premier procédé est appliqué au cas où l'on utilise directement les données de départ pour calculer les parts des individus et celui où il faut au préalable transformer des données de départ en données réduites du fait que les variables sont hétérogènes. Quant au deuxième, nous l'avons appliqué aux données de 24 communes de la Ville-Province de Kinshasa partant de trois variables hétérogènes en vue de la répartition des ressources budgétaires qui sont ici les recettes à caractère national leur alloués. Il s'en suit que la part de Kimbanseke est la plus grande suivie de celle de Maluku puis de Ngaliema puis de Masina. Par contre, celle de Gombe est la plus petite précédée de celle de Kintambo, Kasa-vubu, ...

Conclusion générale et perspectives

Nous avons voulu résoudre le problème de l'injustice constatée dans la répartition des ressources à caractère national allouées aux ETDs, plus précisément les 24 communes de la ville-province de Kinshasa. En effet, cette répartition se fait dans la pratique en utilisant une seule variable (critère) au lieu de trois (Capacité de production, superficie et population), proposées par le législateur bien qu'elles n'aient pas la même unité de mesure, et ne tiennent pas compte de mécanisme de réduction des inégalités entre les individus. Ceci nous a permis d'envisager l'utilisation de plusieurs variables hétérogènes et la prise en compte des rapports entre les individus. Nous avons mené la réflexion plus loin en considérant aussi d'autres problèmes qui se posent en Mathématiques sur le partage.

Pour résoudre ces problèmes, nous avons d'abord jeté un regard sur les principes de justice et ensuite nous avons proposé des mécanismes de partage qui prennent en compte ces principes. Forsé et Parodi [56] ont retenu globalement trois principes : l'égalité, l'équité et la satisfaction de besoin mais ils ne se sont pas préoccupés de mettre en place des mécanismes prenant en compte le nombre (un ou plusieurs) et les types (homogène ou hétérogène) des variables, ainsi que des rapports entre les individus (leur rapprochement, la nature de leur société d'appartenance).

Quant à nous, considérant le principe d'équité, nous avons mis l'accent sur plusieurs variables notamment hétérogènes ainsi que sur les rapports entre les individus. A cet effet, nous avons différencié le cas des contributeurs (comme actionnaires) de celui des non contributeurs (comme entités étatiques qui reçoivent un don). Dans le premier cas, le principe d'équité doit s'appliquer de manière brute, en attribuant aux individus des parts proportionnelles à leurs contributions respectives tandis que dans le second, puis que les individus doivent se solidariser, l'équité brute n'est pas adaptée, il faut proposer un autre principe de justice. Nous avons opté pour l'équité réduite (ou corrigée) multidimensionnelle selon lequel, le partage est fait suivant plusieurs variables et les parts sont attribuées proportionnellement aux valeurs des individus mais après réduction des inégalités entre eux (d'abord dans leur classe respective suite à leur proximité et ensuite entre eux tous du fait de leur appartenance à une même population).

De ce fait, nous avons proposé deux mécanismes qui utilisent plusieurs variables

homogènes ou hétérogènes et prennent en compte les rapports entre les individus permettant de réduire ou pas les inégalités :

1) l'un nommé Procédé de la Répartition des Ressources Sans réduction des inégalités (PRRS) adapté pour le cas où les individus (comme des actionnaires d'une entreprise) ont contribué à la création de la ressource commune à partager Il n'autorise pas la réduction des inégalités. Quelques applications ont été proposées à cet effet.

Le procédé PRRS suit les quatre étapes suivantes :(1) Détermination des données à utiliser ; (2) calcul des valeurs totales des individus ; (3) calcul des proportions et parts des individus ; (4) Représentation graphique des parts des individus.

2) L'autre, quant à lui, est nommé Procédé de la Répartition des Ressources à partir des résultats de la Classification (PRRC), adapté pour le cas où les individus n'ont pas contribué à la création de la ressource. Il admet la réduction des inégalités et fait appel à la notion de classification permettant de retrouver les individus les plus proches qui devront se solidariser dans leurs classes respectives et par la suite avec tous les autres du fait qu'ils appartiennent à une même population.

Le procédé PRRC se réalise en suivant les cinq grandes étapes ci-après : (1) Détermination et présentation des résultats de la Classification (les classes). 2) Réduction des inégalités entre les individus dans leurs classes et dans l'ensemble de la population. (3) Calcul des proportions corrigées des individus par rapport à l'ensemble de la population ; (4) Calcul des parts respectives des individus ; (5) Représentation graphique des parts des individus. Il utilise des formules intéressantes que nous avons proposées, plus particulièrement, en ce qui concerne la réduction des inégalités entre les individus.

Nous avons appliqué ce procédé aux données de 24 communes de la Ville-Province de Kinshasa que nous avons réparties en 6 classes partant de trois variables hétérogènes (Superficie, Production et Population). Il nous a permis de déterminer les parts (corrigées), en CDF, des recettes à caractère national (144745972822 (en CDF)) leur allouées en 2015 dans le cadre de la rétrocession. Ces parts qui ont été finalement représentées graphiquement sont les suivantes : Bandalungwa : 5095058243, Barumbu : 3821293683, Bumbu : 6296449818, Gombe : 3763395293, Kalamu : 5138482035, Kasa-vubu : 3806819085, Kimbanseke : 12983713762, Kinshasa : 3922615863, Kintambo : 3792344488, Kisenso : 6368822804, Lemba : 6368822804, Limete :6296449818, Lingwala : 3835768280, Makala : 5080583646, Maluku :12202085509, Masina : 9263742261, Matete : 5124007438, Mont-Ngafula : 5138482035, Ndjili : 6368822804,

Ngaba : 4009463447, Ngaliema : 9365064442, Ngiri-Ngiri : 3864717474, Nsele : 6383297401, Selembao : 6426721193.

Nous avions un moment donné besoin de regrouper les individus dans des classes afin de réduire les inégalités entre eux. A cet effet, les méthodes factorielles : ACP, AFC (et AFCM) ont été présentées. Toutefois, elles ne conviennent pas pour déterminer les classes, bien que leurs résultats peuvent être utilisés dans la classification. de ce fait, l'appel devrait être fait à la classification et plus particulièrement la Classification Ascendante Hiérarchique (CAH) pour retrouver les individus qui se ressemblent les plus afin de réduire les inégalités entre eux. La CAH est une des méthodes de l'Analyse des données et plus précisément de la Classification non supervisée qui comporte de façon classique quatre étapes : (1) Constituer le tableau des données ; (2) choisir un indice de distance permettant de calculer les distances entre les individus deux à deux, présenter le tableau de distance et regrouper les deux individus ayant la plus petite distance ; (3) choisir un indice d'agrégation, calculer les distances entre le (les) groupe (s) et le (les) reste de groupes et/ou individus isolés, et agréger les objets ayant la plus petite valeur de l'indice d'agrégation ; continuer cette opération jusqu'à ce que tous les individus formeront un seul groupe ; (4) Représenter graphiquement le dendrogramme. Le dendrogramme est découpé au niveau voulu pour obtenir les différentes classes des individus qui forment une partition.

Plus particulièrement, au niveau de calcul des distances entre individus ou groupes d'individus, nous avons proposé deux formules : La première qui calcule, elle seule, les distances entre deux individus isolés et entre deux groupes d'individus, contrairement au procédé classique qui utilise deux formules différentes. La deuxième, quant à elle, calcule simultanément les distances entre plusieurs individus et/ou groupes d'individus deux à deux en utilisant les matrices.

A ce propos, nous avons proposé une démarche nommée « Procédé de Calcul Simultané de Distances (PCSD) » pour faciliter l'utilisation de cette formule. Cette démarche a quelques avantages : (i) D'abord, le gain de temps par le calcul simultanément de plusieurs distances entre les individus isolés et/ou groupes ; (ii) ensuite, l'utilisation d'une même formule pour la distance entre deux individus et/ou entre deux groupes d'individus, question de représenter les groupes par leurs barycentres. Par contre, dans le procédé classique il y a perte de temps suite au calcul une après l'autre des distances entre les individus et à l'utilisation de deux formules complètement différentes pour la

distance entre deux individus d'une part et entre deux groupes, d'autre part.

Après classification, on procède par l'interprétation de la partition (et donc des classes) obtenue à l'issue du découpage du dendrogramme par le calcul des paramètres des classes (moyenne, variance, écart-type), la détermination des individus les plus typiques des classes : les parangons et les extrêmes, et la détermination des variables les plus importantes des classes. Elle se termine par la représentation graphique des classes et des individus. Toutefois, nous avons estimé que l'interprétation de la partition doit aller plus loin jusqu'au calcul des proportions et des parts revenant aux individus après réduction des inégalités comme nous l'avons fait dans ce mémoire.

Ce travail laisse plusieurs ouvertures aux chercheurs : (1) Poursuivre la même étude mais en considérant carrément d'autres variables (notamment qualitatives ou au nombre encore plus grand) ou les données floues ; (2) Etendre cette étude aux 26 provinces (25 provinces + Ville de Kinshasa) de la République Démocratique du Congo ; (3) Mettre en place un programme informatique qui utilise les deux procédés PRRS et PRRC que nous avons proposés ; (4) Approfondir ces deux procédés.

En ce qui nous concerne, nous projetons approfondir ces deux procédés (PRRS et PRRC) avec les données floues, considérant en ce moment que chaque individu a la possibilité d'appartenir à plusieurs classes à la fois.

Bibliographie

Ouvrages

[1] AURAY J. P. , DURU G.et ZIGHED A. , *Analyse des données multidimensionnelles* , 4 tomes, édition Alexandre, Lacassagne, (1990)

[2] BACCINI A, *Statistique Descriptive Multidimensionnelle* , Institut de Math de Toulouse, Toulouse, (2010)

[3] BEATRICE de Tilière , *Analyse Statistiques Multivariée*, (http ://proba.jussien.fr/detiline/cours/polycop-Bio.pdf (Consulté le 20/02/2016)), (2009)

[4] BENZECRI F, *la classification ascendante hiérarchique d'après un exemple de données économiques* , Les cahiers de l'analyse des données, tome 3, n°3, Dunod, (1985)

[5] BENZECRI J.P. et Cie , *l'Analyse des données : 1 La taxinomie* , 3è édition, Collection Bordas, Dunod, Paris, (1980)

[6] BENZECRI J.P. et Cie , *l'Analyse des données : 2 L'Analyse des correspondances* , 3è édition, Collection Bordas, Dunod, Paris , (1980)

[7] BERTRAND F. et MAUMY-BERTRAND M. , *Notions fondamentales en statistique* , Université de Strasbourg, (2013)

[8] BILLAUDOT B., *Justice distributive et justice commutative dans la société moderne*, (https ://halshs.archives-ouvertes.fr/halshs-00644799 (Submitted on 25 Nov 2011))

[9] BOUCHIER A. , *Programmer avec R* , Montpellier, (mai 2011)

[10] BOUROCHE J. M. et BERTIER P. , *Analyse des données Multidimensionnelles* , 2è édition, PUF , (1977)

[11] BROSSIER G. , *Représentation ordonnée des classifications hiérarchiques* , Tome 5, n°2, Université de Haute-bretagne, (1980)

[12] CARPENTIER F.-G , *Introduction aux analyses multidimensionnelles* , (2007)

[13] CHESNEAU C., *Eléments de classification* , université de Caen, (2018)

[14] CHESSEL D., THIOULOUSE J., DUFOUR AB , *Introduction à la classification hiérarchique* , (2004)

[15] DAGNELIE P., *Théorie et méthodes statistiques : Analyse statistique à plusieurs dimensions* , Les presses Agronomiques de Gembloux, (1986)

[16] DELGADO P. , *Mathématique appliquées, cours et application* ,2è édition, eska, Paris , (2001)

[17] DIDAY E. et cie , *Eléments d'analyse de données* , Edition Dunod, Paris, (1982)

[18] DUSART P. , *Cours de statistique inférentielle* , SI-MASS , (2018)

[19] ESCOFIER B. et PAGES J. , *Analyses factorielles simples et multiples* ,4è édition, Dunod, Paris , (2008)

[20] GAMBETTE P., *Classification supervisée et non supervisée, cours, Master 1*, université Marne-la-Vallée, (2014)

[21] GOULET Vincent , *Introduction à la programmation en R* , 4è édition, école d'actuariat, Université Laval, Québec, (2014)

[22] GOULET V., *Introduction à la programmation en R* , 5è édition, école d'actuariat, Université Laval, Québec, (2016)

[23] HUSSON F. et JOSSE Juill, *Analyse de données avec R, Complémentarité des méthodes d'Analyse Factorielle et de Classification* , Agrocampus Rennes, Marseille, (2010)

[24] IOOSS B et VERRIER V , *Introduction à l'analyse des correspondances et à la classification, EDF R&D, Cours* , Toulouse, (2011)

[25] KAUFFMAN A., *Introduction à la théorie des sous-ensembles floues : éléments théoriques de base* , Masson, Paris, 424 P., (1977)

[26] LEBART L., MORINEAU A., FENELON J.-P., *Traitement des données statistiques : méthodes et programmes* , 2è éd, Dunod, Paris, (1982)

[27] LEBART L., MORINEAU A. et PIRON M , *Statistique exploratoire multidimensionnelle* ,3è édition, Dunod, Paris , (2000)

[28] LEMAITRE M , *Partages et allocations équitables, cours* , (https ://sites.google.com/site/michellemaitre31/enseignement), (2016)

[29] MARTIN A., *l'analyse de données, cours* , (Septembre 2004)

[30] MASIERI W., *Notions essentielles de statistique et calcul des probabilités* , Sirey, Paris, (1969)

[31] MORICE E , *Dictionnaire de Statistique* , Collection Dunod, Paris, (1968)

[32] NAKACHE J-P, COFAIS Josiane , *Approche pragmatique de la classification : Arbres hiérarchiques, Partitionnements* , Editions Technip, Paris , (2005))

[33] PHILIPPE A., *Méthodes de statistique inférentielle* , Université de Nantes , (2016)

[34] PONGER L., *Les tests statistiques élémentaires avec R* , INSERM, (2012)

[35] PONTIER J., DUFOUR A-B et NORMAND M, *Statistique et Mathématiques Appliquées*, éditions ellipses, Bruxelles, (1990)

[36] RAKOTOMALALA R, *Analyse en Composantes Principales (ACP)*, Université Lumière, Lyon 2 (http ://tutoriels-data-mining.blogspot.fr/, Consulté le 20/05/2017)

[37] RAKOTOMALALA R., *Analyse Factorielle des Correspondances (AFC)*, Université Lumière, Lyon 2 (http ://tutoriels-data-mining.blogspot.fr/, Consulté le 25/07/2020)

[38] ROUX E. et cie , *Méthodes d'Analyses Factorielles : ACP et AFCM* , LTSI, INSERM U 642, Avril , (2004)

[39] SABITI J., *Analyse des données Multidimensionnelles, cours* , UPN,Kinshasa , (2010)

[40] SAPORTA G. , *Analyse discriminante, classification supervisée* , scoring, Editions technip, Paris, (2009)

[41] TEJEDO C et TRUCHON M, *Serial costsharing in multidimensional contexts* , CREFA, Québec, (2002)

[42] TRIGAN J. , *Exercices progressifs corrigés pour une initiation au calcul matriciel* , Gauthier, Paris, (1969)

[43] TRYON R. Choate , *Cluster analysis* , Ann Arbor : Edwards Bros, (1939)

Articles

[44] ABDALLAH H. and SAPORTA G , Classification d'un ensemble de variables qualitatives, *Revue de Statistique Appliquée, 46 (4), 5-26*, (1998)

[45] AMBAPOUR S, Application de l'Analyse des données au traitement d'enquêtes. Mesure de satisfaction de clientèle pour les grands services publics : le cas de la Société Nationale d'Electricité , Brazzaville, *BAMSI REPRINT, (05/2003)*

[46] ARISTOTE, La politique, III, *12*

[47] AUTANT E., Le partage : nouveau paradigme ? , *Dans revue de MAUSS,* 2010/1 (n°35), (2010)

[48] BARTHELEMY F. et MARTIN M., Critères pour une meilleure répartition des sièges au sein des structures intercommunales, une application au cas du val-d'oise, *Presse des Sciences Po « Revue économique »,* n°2, vol.58, pages 399 à 425, (2007)

[49] BONIN P. Y., Le retour de la méritocratie : la théorie de la justice sociale de David Miller, études critiques, Dialogue XLI, 741-64 (David Miller, principles of social justice, *Cambridge, MA, Harvard university Press,* 1999, XI, 337 p), (2002)

[50] BOUVERET S., FARGIER H., LANG J. et LEMAITRE M., Un modèle général et des résultats de complexité pour le partage de biens indivisibles, Dans Andreas Herzig, Yves Lespérance et Abdel-Illah Mouaddib, éditeurs : Actes des troisièmes journées francophones Modèles Formels de l'Interaction, *Cépaduès Éditions*, (2005)

[51] BOYER M., MOREAUX M. et TRUCHON M., Partage des coûts et tarification des infrastructures, *CIRANO*, Quebec, (2006)

[52] BOYER M., MARCHETTI N., Principes de choix d'une méthode économique d'allocation : partage des coûts et tarification à Gaz de France, Rapport de projet, *collection CIRANO*, Montréal , (2007)

[53] CAILL*É* A., l'aspiration anti-utilitariste de la sociologie classique, *Presses universitaires de Paris Nanterre* , p. 91-106 (Internet : https ://books.openedition.org/pupo/6948 ?lang=fr (Consulté le 31/12/2019)

[54] CHAVALIER F. et LE BALLAC J., la Classification, *Université Rennes*, 2012-2013

[55] ESBENSEN K. et GELADI P. , principal component analysis, chemometrics and intelligent *laboratory systems* , 37-52, (1987)

[56] FORSE M. et PARODI M., Perception des inégalités économiques et sentiment de justice sociale, *Revue de l'OFCE*, 2007/3 (n°102), pages 483-540, , (2007)

[57] GUIBET LAFAYE C. P. Savidan, dictionnaire des inégalités et de la justice sociale, *PUF, hal-01566318* , (2017)

[58] *JOURNAL OFFICIEL de la RDC* , Constitution de la RDC du 18 février 2006, 52è année, numéro spécial, Kinshasa, (2006)

[59] *JOURNAL OFFICIEL de la RDC* , Loi organique n°08/016 du 07/10/2008 portant composition, organisation et fonctionnement des Entités Territoriales Décentralisées et leurs rapports avec l'Etat et les Provinces, Kinshasa, (2008)

[60] *JOURNAL OFFICIEL de la RDC*, Loi n° 11/011 du 13 juillet 2011 relative aux finances publiques *, Kinshasa, (20011)*

[61] KAPYA Kabesa J. S. I. M. , A propos de la répartition des recettes à caractère national entre le pouvoir central et les provinces de la RDC : Modalités et contraintes , Université de Lubumbashi,*Lubumbashi, RDC, (https ://www.hamann-legal.de/upload/4Jean_Salem_Franz.pdf. (Consulté le 05/12/2019))*

[62] LESEUR A., La recherche de l'équité dans la répartition de ressources publiques entre entreprises, Centre National de la recherche scientifique *, cahier n°2005-2006, Paris, (2005-2006)*

[63] MAROY C. , La méritocratie seule en cause (Discussion de l'ouvrage de Marie-Duru-Bellet, Le mérite contre la justice , Paris, Presses de sciences Po*, (2009) ((http ://sociologies.revues.org/index 3778.html (Consulté le 30/11/2019))*

[64] MICROSOFT ENCARTA 2009 [DVD]*, "inertie ",* Microsoft Corporation*, (2008)*

[65] MOMMET E., la théorie des « capabilités » d'Amartya Sen face au problème du relativisme, Tracés. Revue de Sciences humaines *, 12/2007, p. 103-120 (Internet : https ://doi.org/10.4000/traces.211, consulté le 30/12/2019)*

[66] PUNGA Kumakinga P., Problématique de la conformité à la constitution de la loi organique sur les entités territoriale décentralisées en République Démocratique du Congo. Regard sur la commune de Mont-Ngafula dans la ville de Kinshasa *(https ://www.hamann-legal.de/upload/6paulin_franz.pdf (Consulté ce 10/12/2019))*

[67] REINERT A., Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte, Collection les chaires de l'analyse des données , tome 8, *Dunod*, Vol. VIII, 1983, n°2, P.187-198, (1983)

[68] SEN A., Un nouveau modèle économique ; Développement, justice, liberté, *éditions Odile Jacob*, Paris, (2000) (https ://deshautsetdebats.blog/2018/10/09/les-grandes-thories-de-la-justice-sociale-1-2/ (Consulté le 27/12/2019))

[69] SOUSA L. et GAMA J., The application of hierarchical clustering algorithms for recognition using Biometrics of the hand, IJAERS, vol-1, Issue 7, (dec 2014)

[70] ZOU H., HASTIE T. and TIBISHRANI R., Sparse principal Component Analysis, (april 26, 2004)

Mémoires et thèses

[71] ABDELLAOUI NEZHA N., *La classification hiérarchique ascendante*, Mémoire, Algerie, (2014)

[72] BENDJABALLAH I., *Analyses factorielles des correspondances*, Master en Math, option statistique, Université Mohamed Khider, Biskra, Algerie, (juin 2019)

[73] BOUVERET S., *Allocation et partage équitables de ressources indivisibles : modélisation, complexité et algorithmique* , Thèse de doctorat, Université de Toulouse, (Novembre 2007)

[74] CHELCEA S. T., *Agglomerative 2-3 Hierarchical Classification : Theoritical and application study* , Université de Nice Sophia Antipolis, (2007)

[75] CHIQUET J., *Introduction au logiciel R et à la pratique des statistiques en vue de l'analyse des données issues de la Biologie, école doctorale « du génome aux organismes »,* , Université d'evry, (30 janvier-3 février 2012)

[76] KASIAMA NGI-ONKOR J., *Data mining des données médicales par la méthode d'analyse discriminante et application,* Dissertation DEA, Université Pédagogique Nationale (UPN), Kinshasa,,(2012-2013)

[77] NGOIE R.-B., *Choix social et partage équitable : une Analyse mathématique postérieure aux élections législatives et présidentielles en République Démocratique du Congo de 2006 et 2011*, Dissertation DEA, UPN, Kinshasa,(Décembre 2012)

Webographie

[78] Internet : *www.foad.refer.org/MG/pdf/M05-3.pdf*, (Consulté le 01/02/2013 à 10h00')

[79] Internet : *http ://stadon.voila.net/Tables.htm*, (Consulté le 30/05/2014 à 20h35')

[80] Internet : *http ://www.jybaudot.fr/Stats/inertie.html*, (Consulté le 02/07/2014 à 19h30')

[81] Internet : *www.foad.org/IMG/pdf/M.5-3.pdf* , (Consulté le 15/08/2016 à 15h00')

[82] Internet :*https ://fr.wikipedia.org/wiki/Distance_ ultram% C3%A9trique*, (Consulté le 20/02/2017 à 20h35')

[83] Internet : *https ://fr.m.wikipedia.org/wiki/Liste_des_communes_de_Kinshasa*, (Consulté le 14/12/2018 à 20h20')

[84] Internet : *https ://fr.wikipedia.org/wiki/Analyse des donn C3 A9es*, (Consulté le 20/09/2019 à 19h30')

[85] Internet :*http ://www.odeprdc.org/index.php/17-recettes-publiques/13-la-retrocession-un-appas-pour-l-hotel-de-ville-de-kinshasa*, (Consulté le 30/12/2019 à 18h25')

[86] Internet :*https ://www.ilemaths.net/sujet-repatition-proportionnelle multi-critere-827199.html*, (Consulté le 31/12/2019 à 21h30')

[87] Internet : *http ://www.foad-mooc.auf.org/IMG/pdf/M03-5.pdf*, (Consulté le 20/07/2019 à 22h00')

[88] Internet : *http ://factominer.free.fr/factomethods/classification-hierarchique-sur-composantes-principales.html*, (Consulté le 20/08/2019 à 20h40')

[89] Internet : *hptts ://www.economie.gouv.fr*, (Consulté le 30/12/2019 à 22h00')

[90] Internet : *hptts ://www.fr.m.wikipedia.org/wiki/Utilitarisme*, (Consulté le 31/12/2019 à 20h30')

Annexes

Annexe 1. Carte de la ville de Kinshasa

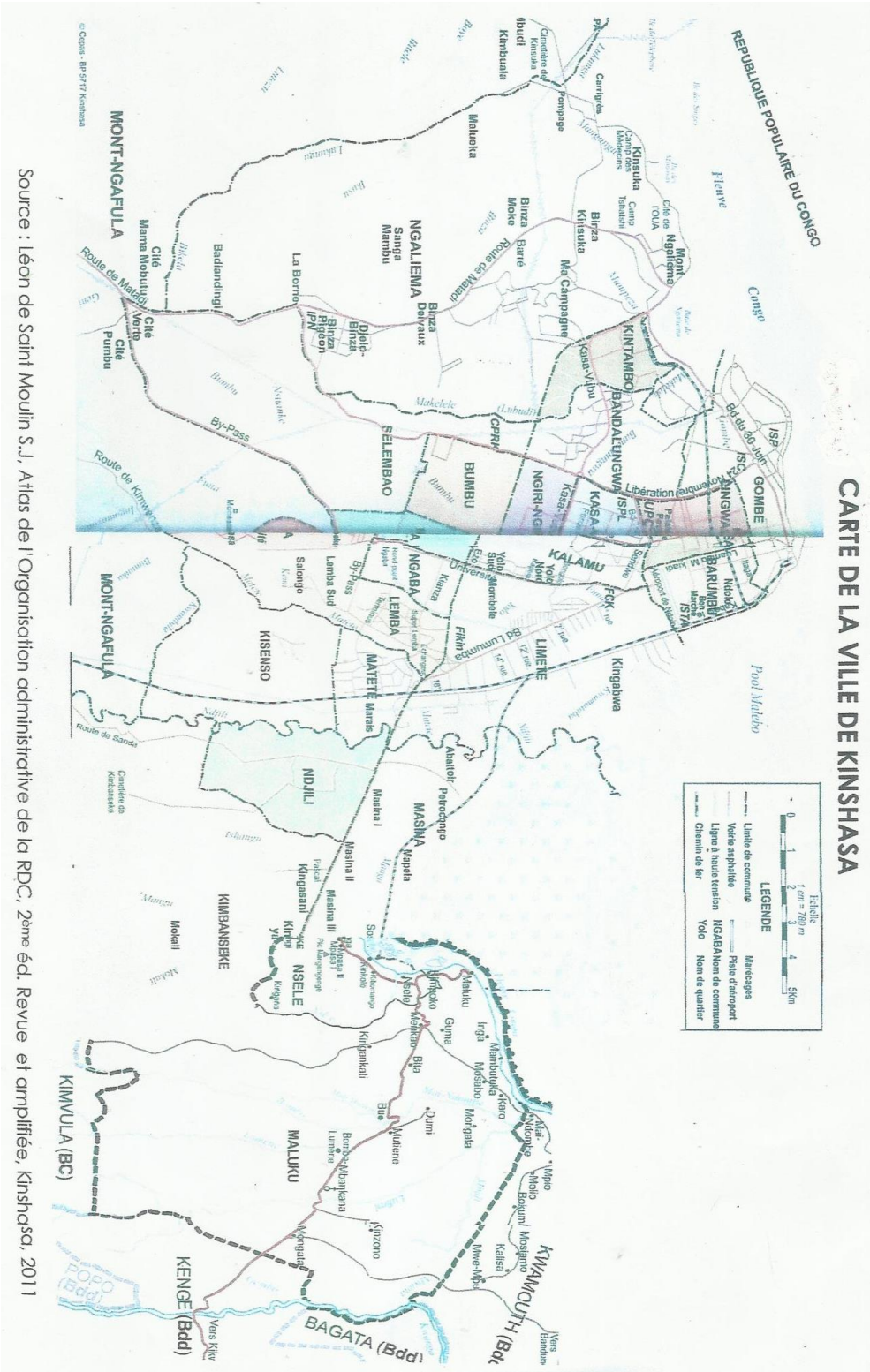

Figure 4.4 – Carte de la ville province de Kinshasa

Annexe 2. Questionnaire d'enquête

QUESTIONNAIRE D'ENQUETE N°01

Dans le cadre des recherches relatives à la réalisation de notre mémoire de DEA dans le domaine de Mathématique appliquée, plus précisément la Statistique (notion de classification), nous venons vous soumettre le présent questionnaire d'enquête et vous prions de bien vouloir y répondre.

Commune :

Service (concerné par l'enquête) **:**

Epoque ciblée pour l'enquête : (de Janvier à Décembre) 2014/2015

1) Combien d'habitants comporte-t-elle votre commune (tout dernier recensement) ? (Veuillez les répartir par tranche d'âges).....
2) Quelle est la superficie de votre commune?.....
3) Qu'avez-vous réalisé cette année comme recette ?.....
4) Avez-vous un commentaire à faire sur votre commune ?.....

.....

Fait à Kinshasa, le / /

Figure 4.5 – Questionnaire d'enquête

Annexe 3. Commandes de logiciel R

Voici ci-dessous quelques commandes de R que nous avons utilisées dans ce mémoire :

N°	**Commandes et commentaires**
1	> library(gdata) # Charge la librairie gdata. Elle permet l'importation > T=read.table("TABLEAUBRUT2015.txt") # Importe le tableau « TABLEAUBRUT2015 » sous le nom T. > T
2	> t=scale(T, center = FALSE, scale = TRUE) # Transforme les données de départ en données réduites > t
3	> Tcr=scale(T, center = TRUE, scale = TRUE) # Transforme les données de départ en données centrées-réduites > Tcr
4	> install.packages("FactoMineR") # télécharge le package FactoMineR > library(FactoMineR) # charge la librairie FactoMineR déjà installé > res.acp=PCA(Tcr) # réalise l'ACP du tableau Tcr > res.acp # renvoie les résultats de l'ACP
5	> res.acp$eig # affiche les valeurs propres et contributions des axes factoriels
6	> res.acp$var # affiche les coordonnées ($coord), les cosinus carré ($cos2), les contributions ($contrib) et les distances au centre de gravité des individus ($dist)
7	> res.acp$ind # affiche les coordonnées ($coord), les cosinus carré ($cos2), les contributions ($contrib) et les distances au centre de gravité des individus ($dist)

8	> res.acpind = res.acpindcoord #renvoie les coordonnées des individus pour toutes les composantes principales à part des résultats de l'ACP (res.acp) > res.acpind > res.acp2=res.acpind [, 1 : 2] # extrait les coordonnées des individus pour les deux premières composantes
9	> res.hcpc= HCPC(res.acp2) # propose une coupure du dendrogramme en trois classes c'est-à-dire au niveau du plus grand saut d'inertie
10	> res.acpdeux = PCA(Tcr, ncp = 2, graph = F) # réalise l'ACP du tableau Tcr (centré réduit) et extrait les premières composante sans afficher les graphiques de l'ACP > res.hcpc = HCPC(res.acpdeux) # produit les résultats de la classification à partir des résultats de l'ACP pour les 2 premiers axes
11	> res.hcpc$call # adjoint automatiquement chaque individu au numéro de sa classe d'appartenance se servant des résultats de HCPC
12	> res.hcpc$desc.ind # produit les éléments de la caractérisation des classes par les individus (parangons et les extrêmes) des classes à partir des résultats de la commande HCPC > res.hcpc$desc.ind$para # produit les parangons des classes > res.hcpc$desc.ind$dist # renvoie les extrêmes des classes
13	> res.hcpc$desc.var # renvoie les variables caractérisant les classes
14	> Part=matrix(c(5095058243, 3821293683, 6296449818, 3763395293, 5138482035, 3806819085, 12983713762, 3922615863, 3792344488, 6368822804, 6368822804, 6296449818, 3835768280, 5080583646,12202085509, 9263742261, 5124007438, 5138482035, 6368822804, 4009463447, 9365064442 , 3864717474, 6383297401, 6426721193), nrow = 1, ncol=24, dimnames = list(c("Parts"), c("Ban", "Bar", "Bum", "Gom", "Kal", "Kas", "Kim","Kin", "Kint", "Kis", "Lem", "Lim", "Lin", "Mak", "Mal", "Mas", "Mat", "Mon", "Ndj", "Nga", "Ngal", "Ngi", "Nse", "Sel"))) # Crée la matrice Part des parts des individus
15	> barplot (Part[, 1 : 24], col=c("blue", "purple", "green", "violetred1", "red", "pink", "cyan", "yellow", "violet", "blue", "purple", "green", "violetred1", "red", "pink", "cyan", "yellow", "violet", "blue", "purple", "green", "violetred1", "red", "pink"), main = "Graphique à bâtons des parts des individus", sub="Individus", col.main="red", col.sub="blue") # Crée le graphique à bâtons de la matrice Part à une ligne et 24 colonnes avec la fonction *barplot*
16	> pie (Part[, 1 : 24], col=c("blue", "purple", "green", "violetred1", "red", "pink", "cyan", "yellow", "violet", "blue", "purple", "green", "violetred1", "red", "pink", "cyan", "yellow", "violet", "blue", "purple", "green", "violetred1", "red", "pink"), main = "Graphique en camembert des parts des individus", sub="Individus", col.main="red", col.sub="blue") # Crée un diagramme en camembert de la matrice Part à une ligne et 24 colonnes avec la commande *pie*

Tableau 4.23 – Quelques commandes de logiciel R

Source : Notre conception [33]

Annexe 4. Extrait de rapport de l'Observatoire de la Dépense Publique (ODEP). Exercice 2015

« Le gouvernement provincial de Kinshasa applique convenablement les dispositions qui concernent la répartition de la rétrocession aux communes pour l'exercice 2015.

L'hôtel de ville fixe la part à rétrocéder aux communes à 138634059223 FC. Un montant de 110907247387 FC est alloué à l'investissement soit 80% de recettes et 27726811847 FC au fonctionnement soit 20% » [85].

Il y a des communes qui avaient reçu un complément (Co). Le montant alloué à chaque commune se présente de la manière suivante :

N^o	**Commune**	**Montant alloué (CDF)**
1	Bandalungwa	4450153301
2	Barumbu	2675637343 + 212572224(Co)
3	Bumbu	6086035200
4	Gombe	402038772 + 2486170796(Co)
5	Kalamu	6834659120
6	Kasa-Vubu	3063812709 + 253700328(Co)
7	Kimbanseke	18452193284
8	Kinshasa	3077676115 +250927647(Co)
9	Kintambo	2439959443 + 448250125(Co)
10	Kisenso	8123955871
11	Lemba	7860551189
12	Limete	6945566368
13	Lingwala	1954740235 + 933469332(Co)
14	Makala	4671967796
15	Maluku	3133129739 +239836922(Co)
16	Masina	11021407709
17	Matete	5988991359
18	Mont- Ngafula	3410397857 +184383299(Co)
19	Ndjili	7319878328
20	Ngaba	3784709817 +109520907(Co)
21	Ngaliema	14209991071
22	Ngiri-Ngiri	3341080828 +198246705(Co)
23	Nsele	2093374294 + 794835273(Co)
24	Selembao	7292151516

Tableau 4.24 – Montants de rétrocession avec compléments. Exercice 2015

Source : ODEP [85]

A la lumière de ce tableau, nous allons tenter d'estimer les capacités de reproduction des communes de la ville-province de Kinshasa.

1. Répartition des recettes entre une province et le pouvoir central

Soit Z, le montant total des recettes d'une province. Cette dernière en retient 40%, $40\%.Z = 40Z \div 100$ et les 60%, $60\%\ Z = 60Z \div 100$ sont envoyés au pouvoir central.

2. Répartition des recettes entre une province et ses ETDs

En considérant que le montant total retenu par une province est Y alors $Y = 40\%Z$. La province en retient 60%, $60\%Y = 60\%.40\%Z$, et rétrocède aux ETDs 40%, $40\%Y = 40\%.40\%Z = (40.40Z) \div (100.100)$.

Soit X le montant total rétrocédé aux ETDs alors X=40% Y=40%.40% $Z = (40.40Z) \div (100.100) = 16.Z \div 100$. Ce qui implique $X = Z.(16 \div 100)$ et donc $Z = X.(100 \div 16)$

3. Estimation des productions des communes (ETDs)

Sachant que le montant rétrocédé aux Communes de la Ville-Province de Kinshasa est $X = 144745972822 = 138634059223 + 6111913599$(Complément), on a : $Z = X.(100 \div 16) = 144745972822.\ (100 \div 16) = 904662330138$. Donc les recettes à caractère national produites par la Ville-Province de Kinshasa sont estimées à 904662330138 CDF.

A partir de ce montant nous pouvons estimer la production de chaque commune. A cet effet, nous aurions dû nous fonder sur les poids démographiques des communes mais nous ne pouvons le faire car, suivant les données fournies par ODEP, la répartition n'est même pas faite sur base des poids démographiques c'est-à-dire population. A titre d'exemple, Ngiri-Ngiri dont la population (105664) est inférieure à celle de Lingwala (123619) reçoit un montant (3539327533 CDF) supérieur à celui de Lingwala (2888209567 CDF).

C'est la raison pour laquelle, nous allons calculer la production proportionnellement au montant reçu de rétrocession. Par exemple, pour la commune de Ngiri-Ngiri, le montant reçu est 3539327533 CDF sur le montant total de 144745972822 CDF (selon nos calculs) rétrocédé.

Il faut se demander : quelle proportion du montant total rétrocédé représente le montant reçu par Ngiri-Ngiri ? Il suffit de faire :$3539327533 \div 144745972822 = 0,024451993$ et par la suite multiplier cette proportion par le montant total de contribution de la Ville/Province de Kinshasa pour avoir le montant estimé de la production (capacité de production) de la commune de Ngiri-Ngiri : $0,024451993$x$904662330138 = 22120796964$.

Donc nous estimons la capacité de production de la commune de Ngiri-Ngiri à 22120796964 CDF. Nous ferons de même pour les autres communes en prenant $\frac{X_i}{X}Z$ avec X_i le montant de la rétrocession reçu par la commune, X le montant total rétrocédé et Z la production total de la Ville/Province de Kinshasa.

Les montants estimés des capacités de production des communes de la ville-province de Kinshasa pour l'exercice 2015 se présentent comme suit :

N^o	**Commune**	**Montant estimé de la production (CDF)**
1	Bandalungwa	27813458131
2	Barumbu	18051309794
3	Bumbu	38037720000
4	Gombe	18051309800
5	Kalamu	42716619500
6	Kasa-Vubu	20734456481
7	Kimbanseke	115326208025
8	Kinshasa	20803773513
9	Kintambo	18051309800
10	Kisenso	50774724194
11	Lemba	49128444931
12	Limete	43409789800
13	Lingwala	18051309794
14	Makala	29199798725
15	Maluku	21081041631
16	Masina	68883798181
17	Matete	37431195994
18	Mont- Ngafula	22467382225
19	Ndjili	45749239550
20	Ngaba	24338942025
21	Ngaliema	88812444194
22	Ngiri-Ngiri	22120797081
23	Nsele	18051309794
24	Selembao	45575946975
Total		**904662330138**

Tableau 4.25 – Montants estimés de capacité de production des communes

Source : Notre conception à partir des données de l'ODEP. Exercice 2015 [85]

Annexe 5. Table de Test Z (Table de la fonction de répartition inverse de la loi normale)

$Z = Z_\alpha = F^{-1}(1 - \frac{\alpha}{2})$ (Il suffit de remplacer α ou $\frac{\alpha}{2}$ (Niveau de signification) par sa valeur) (Cfr 2.2.6.3.)

Tableau 4.26 – Table de fonction de répartition inverse de la loi normale

α	**0,05**	**0,01**	**0,001**
α/2	**0,025**	**0,005**	**0,0005**
Z	1,96	2,58	3,29

Source : [79]

Annexe 6. Table de fonction de répartition de la loi normale centrée réduite

Tableau 4.27 – Table de fonction de répartition de la loi normale centrée réduite
(Probabilité $F(Z)$ de trouver une valeur inférieure à Z)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de z

z	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9
F(z)	0,998650	0,999032	0,999313	0,999517	0,999663	0,999767	0,999841	0,999892	0,999928	0,999952
z	4,0	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9
F(z)	0,999968	0,999979	0,999987	0,999991	0,999995	0,999997	0,999998	0,999999	0,999999	1,000000

Nota. La table donne F(z) pour z positif. Pour z négatif, il faut prendre le complément à l'unité de la valeur lue dans la table. Exemple : F(-1,37) = 1 - F(1,37) =1 - 0,9147 = 0,0853.

Source : [79]

$F(Z)$: probabilité de trouver une valeur inférieure à Z (probabilité cumulée), égale à Z_i=$\frac{X_i-\overline{X}}{\sigma}$. L'emploi de cette table exige par conséquent la standardisation préalable de la valeur de X dont on veut connaître la probabilité cumulée, Z se lit dans la première colonne pour sa partie cumulée et sa première décimale, la deuxième décimale se trouvant dans la première ligne.

Exemples de lecture : (1) Si Z est donné. Pour $Z = 0,92$; on a $F(Z) = 0,8212$. Pour $Z = -0,92$; on a $F(Z) = F(-0,92) = 1 - F(0,92) = 1 - 0,8212 = 0,1788$. (2) Si $F(Z)$ est donné. Pour $F(Z) = 0,975$; $Z = F^{-1}(0,975) = 1,96$ $(= 1,9 + 0,06)$

Annexe 7. Table de la loi du Khi-deux (loi de Pearson)

(Les valeurs de χ^2 ayant la probabilité P d'être dépassée)

Tableau 4.28 – Table de la loi Khi deux

ν	P = 0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,00004	0,0002	0,001	0,0039	0,0158	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,39	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672

Nota. ν est le nombre de degrés de liberté.
Pour $\nu > 30$, on peut admettre que la quantité $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ suit la loi normale centrée réduite

Source : [79]

Annexe 8. Variables caractérisant les classes

Tableau 4.29 – Tableau des variables caractérisant les classes

```
Link between the cluster variable and the quantitative variables
================================================================
                Eta2      P-value
SUPERFICIE 0.9871243 2.357660e-16
POPULATION 0.9456561 9.458492e-11
PRODUCTION 0.9073095 1.093230e-08

Description of each cluster by quantitative variables
=====================================================
$`1`
              v.test Mean in category  Overall mean sd in category Overall sd     p.value
PRODUCTION -2.355200       -0.7491997 -8.890458e-18     0.06648241   0.978945 0.018512745
POPULATION -3.012745       -0.9583678 -1.243218e-17     0.14352214   0.978945 0.002588963

$`2`
NULL

$`3`
NULL

$`4`
             v.test Mean in category Overall mean sd in category Overall sd      p.value
SUPERFICIE 4.760645         4.660409 3.729655e-17              0   0.978945 1.929754e-06
$`5`
             v.test Mean in category  Overall mean sd in category Overall sd    p.value
PRODUCTION 2.490695         1.686208 -8.890458e-18      0.4082709   0.978945 0.01274936

$`6`
             v.test Mean in category  Overall mean sd in category Overall sd     p.value
PRODUCTION 3.249248         3.180835 -8.890458e-18              0   0.978945 0.001157107
POPULATION 2.996721         2.933625 -1.243218e-17              0   0.978945 0.002729000
```

Source : Tableau réalisé à partir des résultats de la classification