

**MEMOIRE DEPOSE EN VUE DE  
L'OBTENTION DU DIPLOME D'EXPERTISE  
COMPTABLE**

**TRAITEMENT ET ANALYSE STATISTIQUES  
DE L'INFORMATION COMPTABLE ET  
FINANCIERE PRODUITE PAR LES  
COMPAGNIES D'ASSURANCE DE LA  
CONFERENCE INTERAFRICAINNE DES  
MARCHES D'ASSURANCES**

**SESSION : NOVEMBRE 2012**

**NOM ET PREMONS DU CANDIDAT : OUATTARA FANGMAN ALAIN**

**ADRESSE : CIMA BP 2750 LIBREVILLE GABON**

NOTE DE SYNTHÈSE	6
SIGLES ET ABBREVIATIONS	8
INTRODUCTION	9

<b>PREMIÈRE PARTIE : PRÉSENTATION DES MÉTHODES, DES OUTILS STATISTIQUES ET DES DONNÉES UTILISÉES</b>	<b>12</b>
--	-----------

<b>CHAPITRE I. STATISTIQUE EXPLORATOIRE ET ANALYSE DE DONNÉES</b>	<b>13</b>
1) ANALYSE EN COMPOSANTES PRINCIPALES (ACP)	14
1.1) <i>Données et caractéristiques</i>	14
1.2) <i>Recherche de sous espace principaux</i>	14
1.3) <i>Description du nuage des individus</i>	15
1.4) <i>Description du nuage des variables</i>	16
2) CLASSIFICATION	17
2.1) <i>Modèles de classification</i>	17
2.2) <i>Méthodes de partitionnement</i>	18
2.3) <i>Algorithme de classification ascendante hiérarchique</i>	18
3) AUTRES MÉTHODES D'ANALYSE DE DONNÉES	19
3.1) <i>Analyse canonique</i>	19
3.2) <i>Analyse factorielle des correspondances</i>	20
3.3) <i>Analyse des correspondances multiples</i>	21
<b>CHAPITRE II. MODELES PREDICTIFS</b>	<b>22</b>
1) REGRESSION LINEAIRE SIMPLE ET MULTIPLE	23
1.1) <i>Présentation des modèles</i>	23
1.2) <i>Estimation des coefficients de régression</i>	23
1.3) <i>Analyse de la variance et sélection des variables</i>	24
1.4) <i>Prévision et intervalle de confiance</i>	26
2) ANALYSE DISCRIMINANTE	26
2.1) <i>Matrices de variances intraclasse et interclasse</i>	27
2.2) <i>Variance d'un caractère</i>	27
2.3) <i>Facteurs et caractères discriminants</i>	28
2.4) <i>Analyse discriminante décisionnelle</i>	28
3) REGRESSION LOGISTIQUE	28
3.1) <i>Interprétation</i>	29
3.2) <i>Estimation</i>	30
3.3) <i>Test et sélection des variables</i>	31
<b>CHAPITRE III. PRÉSENTATION DE QUELQUES LOGICIELS DE TRAITEMENTS STATISTIQUES ET DES DONNÉES UTILISÉES</b>	<b>32</b>
1) LOGICIELS DE TRAITEMENTS STATISTIQUES	32
1.1) <i>Logiciels commerciaux</i>	33
1.2) <i>Logiciels libres</i>	33
1.3) <i>Logiciel  et bibliothèques utilisées</i>	33
2) PRÉSENTATION ET REFORMULATION DES DONNÉES	35
2.1) <i>Présentation des données des compagnies non vie</i>	35
2.2) <i>Présentation des données des compagnies d'assurance vie</i>	35
2.3) <i>Reformulation des données</i>	36

<b>CHAPITRE I. REALISATION ET INTERPRETATION D'UNE ANALYSE EN COMPOSANTES PRINCIPALES (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE(CAH) SUR LES DONNEES DES COMPAGNIES D'ASSURANCE NON VIE DU MARCHE</b>	<b>40</b>
1) ETUDE DU NUAGE DES INDIVIDUS A TRAVERS UNE ACP	40
1.1) <i>Représentation du nuage des individus</i>	40
1.2) <i>Pourcentage d'inertie associé à chaque axe et choix des dimensions à étudier</i>	43
1.3) <i>Qualité de la représentation des individus et contribution à la construction des axes</i>	45
2) ETUDE DU NUAGE DES VARIABLES A TRAVERS UNE ACP	47
2.1) <i>Représentation des variables</i>	47
2.2) <i>Analyse de la représentation des variables</i>	49
3) CONFIRMATION DES RESULTATS DE L'ETUDE PAR UNE CAH	51
3.1) <i>Forme du dendrogramme et nombre de classes</i>	51
3.2) <i>Description des classes par les variables initiales ou les composantes principales</i>	52
3.3) <i>Description des classes par les parangons et les individus spécifiques</i>	53
<b>CHAPITRE II. REALISATION ET INTERPRETATION D'UNE ANALYSE EN COMPOSANTES PRINCIPALES (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE(CAH) SUR LES DONNEES DES COMPAGNIES D'ASSURANCE VIE DU MARCHE</b>	<b>55</b>
1) ETUDE DU NUAGE DES INDIVIDUS A TRAVERS UNE ACP	55
1.1) <i>Représentation du nuage des individus</i>	55
1.2) <i>Pourcentage d'inertie associé à chaque axe et choix des dimensions à étudier</i>	58
1.3) <i>Qualité de la représentation des individus et contribution à la construction des axes</i>	59
2) ETUDE DU NUAGE DES VARIABLES A TRAVERS UNE ACP	61
2.1) <i>Représentation des variables</i>	61
2.2) <i>Analyse de la représentation des variables</i>	62
3) CONFIRMATION DES RESULTATS DE L'ETUDE PAR UNE CAH	64
3.1) <i>Forme du dendrogramme et nombre de classes</i>	64
3.2) <i>Description des classes par les variables initiales ou les composantes principales</i>	64
3.3) <i>Description des classes par les parangons et les individus spécifiques</i>	65
<b>CHAPITRE III. REALISATION ET INTERPRETATION D'UNE ANALYSE EN COMPOSANTE PRINCIPALE (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE(CAH) SUR LES DONNEES DES COMPAGNIES D'ASSURANCE DU GROUPE G</b>	<b>67</b>
1) ETUDE DES COMPAGNIES D'ASSURANCE NON VIE DU GROUPE G	67
1.1) <i>Etude du nuage des individus à travers une ACP</i>	68
1.2) <i>Etude du nuage des variables à travers une ACP</i>	69
1.3) <i>Confirmation des résultats par une CAH</i>	70
2) ETUDE DES COMPAGNIES D'ASSURANCE VIE DU GROUPE G	71
2.1) <i>Etude du nuage des individus à travers une ACP</i>	71
2.2) <i>Etude du nuage des variables à travers une ACP</i>	72
2.3) <i>Confirmation des résultats par une CAH</i>	74

**TROISIEME PARTIE : MISE EN ŒUVRE DES MODELES PREDICTIFS SUR LES  
DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES  
D'ASSURANCE DE L'ESPACE CIMA**

**77**

<b>CHAPITRE I. REGRESSION LINEAIRE MULTIPLE</b>	<b>78</b>
1) REGRESSION LINEAIRE SUR LES DONNEES DES COMPAGNIES D'ASSURANCE NON VIE	78
1.1) <i>Estimation des paramètres du modèle</i>	79
1.2) <i>Choix des variables</i>	82
1.3) <i>Analyse des résidus</i>	85
1.4) <i>Intérêt pour le professionnel comptable et une autorité de supervision</i>	86
2) REGRESSION LINEAIRE SUR LES DONNEES DES COMPAGNIES D'ASSURANCE VIE	89
2.1) <i>Estimation des paramètres du modèle</i>	90
2.2) <i>Choix des variables</i>	90
2.3) <i>Analyse des résidus</i>	92
2.4) <i>Intérêt pour le professionnel comptable et pour une autorité de supervision</i>	92
<b>CHAPITRE II. ANALYSE DISCRIMINANTE</b>	<b>94</b>
1) ANALYSE DISCRIMINANTE SUR LES DONNEES DES COMPAGNIES D'ASSURANCE NON VIE	94
1.1) <i>Construction du modèle</i>	94
1.2) <i>Estimation du taux de mauvais classement</i>	96
1.3) <i>Intérêt pour le professionnel comptable et pour une autorité de supervision</i>	97
2) ANALYSE DISCRIMINANTE SUR LES DONNEES DES COMPAGNIES D'ASSURANCE VIE	98
2.1) <i>Construction du modèle</i>	98
2.2) <i>Estimation du taux de mauvais classement</i>	99
2.3) <i>Intérêt pour le professionnel comptable et pour une autorité de supervision</i>	100
<b>CHAPITRE III. REGRESSION LOGISTIQUE</b>	<b>101</b>
1) REGRESSION LOGISTIQUE SUR LES DONNEES DES COMPAGNIES D'ASSURANCE NON VIE	101
1.1) <i>Construction du modèle</i>	101
1.2) <i>Estimation du taux de mauvais classement</i>	103
1.3) <i>Intérêt pour le professionnel comptable et pour une autorité de supervision</i>	104
2) REGRESSION LOGISTIQUE SUR LES DONNEES DES COMPAGNIES D'ASSURANCE VIE	106
2.1) <i>Construction du modèle</i>	107
2.2) <i>Estimation du taux de mauvais classement</i>	108
2.3) <i>Intérêt pour le professionnel comptable et pour une autorité de supervision</i>	109
<b>CONCLUSION</b>	<b>112</b>
<b>BIBLIOGRAPHIE</b>	<b>115</b>
<b>ANNEXES</b>	<b>120</b>
<b>I. ANNEXES RELATIVES A LA PRESENTATION DES METHODES, DES OUTILS STATISTIQUES ET DES DONNEES</b>	<b>121</b>
1) COMPLEMENTS SUR LES METHODES STATISTIQUES	121
1.1) <i>Analyse en composantes principales</i>	121
<i>Principe du choix du plan de principal</i>	121
<i>Données et caractéristiques</i>	122
1.2) <i>Classification</i>	124
<i>Dissimilarité et distance</i>	124
<i>Algorithme des k means</i>	124
1.3) <i>Analyse factorielle des correspondances (AFC)</i>	125
1.4) <i>Analyse des correspondances multiples</i>	126
1.5) <i>Régression linéaire multiple</i>	127
<i>Explicitation des termes de la régression linéaire</i>	127
<i>Estimation des paramètres du modèle dans le cadre d'une régression simple</i>	128
<i>Hypothèses du modèle linéaire simple</i>	129
<i>Hypothèses du modèle linéaire multiple</i>	130

2)	LOGICIELS DE DATAMINING ET DE STATISTIQUE	131
2.1)	<i>Cartographie des principaux logiciels commerciaux de datamining et de statistique</i>	131
2.1)	<i>Logiciels libres de datamining et de statistique</i>	131
3)	PRESENTATION DES DONNEES DES SOCIETES D'ASSURANCE	133
3.1)	<i>Contexte de la collecte et du traitement des données</i>	133
3.2)	<i>Présentation des données des sociétés d'assurance non vie</i>	135
3.3)	<i>Présentation des données des compagnies d'assurance vie</i>	136
4)	DONNEES BRUTES ET RESUMEES DES SOCIETES D'ASSURANCE	137
4.1)	<i>Données des sociétés d'assurance non vie</i>	137
4.2)	<i>Données résumées des sociétés d'assurance non vie</i>	141
4.3)	<i>Données des sociétés d'assurance vie</i>	142
4.4)	<i>Données résumées des sociétés d'assurance vie</i>	144
5)	DONNEES COMPLEMENTAIRES POUR L'ANALYSE DISCRIMINANTE ET LA REGRESSION LOGISTIQUE	145
5.1)	<i>Données pour les sociétés d'assurance non vie</i>	145
5.2)	<i>Données pour les sociétés d'assurance vie</i>	146

## **II. ANNEXES RELATIVES A L'ANALYSE DES DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES D'ASSURANCE**

1)	ANALYSE EN COMPOSANTES PRINCIPALES DES SOCIETES NON VIE	147
1.1)	<i>Valeurs propres</i>	148
1.2)	<i>Coordonnées, projection et contribution des individus sur les nouveaux axes</i>	148
1.3)	<i>Matrice de corrélation</i>	151
1.4)	<i>Corrélation et contribution des variables à la construction des nouveaux axes</i>	152
2)	CLASSIFICATION DES SOCIETES NON VIE	152
2.1)	<i>Représentations de la partition</i>	153
2.2)	<i>Caractérisation des classes par rapport aux variables initiales</i>	156
2.3)	<i>Caractérisation des classes par rapport aux composantes principales</i>	159
2.4)	<i>Description des classes par les parangons et les individus spécifiques</i>	160
3)	ANALYSE EN COMPOSANTES PRINCIPALES DES SOCIETES VIE	162
3.1)	<i>Valeurs propres et pourcentage d'inertie</i>	162
3.2)	<i>Coordonnées, projection et contribution des individus sur les nouveaux axes</i>	163
3.3)	<i>Matrice de corrélation</i>	165
3.4)	<i>Corrélation et contribution des variables à la construction des nouveaux axes</i>	166
4)	CLASSIFICATION DES SOCIETES VIE	167
4.1)	<i>Représentations de la partition</i>	167
4.2)	<i>Caractérisation des classes par rapport aux variables initiales</i>	170
4.3)	<i>Caractérisation des classes par rapport aux composantes principales</i>	175
4.4)	<i>Description des classes par les parangons et les individus spécifiques</i>	176
5)	ACP ET CAH DES SOCIETES NON VIE ET VIE DU GROUPE G	177
5.1)	<i>Analyse en composantes principales des sociétés non vie du groupe G</i>	178
5.2)	<i>CAH des sociétés non vie du groupe G</i>	181
	<i>Représentations de la partition</i>	181
	<i>Caractérisation des classes</i>	182
5.3)	<i>Analyse en composantes principales des sociétés vie du groupe G</i>	184
5.4)	<i>CAH des sociétés vie du groupe G</i>	187
	<i>Représentations de la partition</i>	187
	<i>Caractérisation des classes</i>	188

## **III. ANNEXES RELATIVES A LA MISE EN ŒUVRE DES MODELES PREDICTIFS SUR LES DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES D'ASSURANCE**

1)	REGRESSION LINEAIRE MULTIPLE SUR LES DONNEES DES SOCIETES NON VIE ET VIE	191
1.1)	<i>Généralités sur les provisions techniques des sociétés d'assurance</i>	191
1.2)	<i>Modèle avec PSAP comme variable à expliquer</i>	192
1.3)	<i>Modèle avec PAP comme variable à expliquer</i>	195
1.4)	<i>Modèle avec PM comme variable à expliquer</i>	196
2)	ANALYSE DISCRIMINANTE SUR LES DONNEES DES SOCIETES NON VIE ET VIE	196
2.1)	<i>Modèles pour les sociétés d'assurance non vie</i>	197
2.2)	<i>Modèles pour les sociétés d'assurance vie</i>	203

<b>INDEX DES TABLEAUX</b>	<b>206</b>
<b>INDEX DES FIGURES</b>	<b>209</b>
<b>SOMMAIRE</b>	<b>210</b>

# NOTE DE SYNTHÈSE

Les professionnels comptables appelés également **professionnels du chiffre** sont amenés dans leur pratique quotidienne à collecter, traiter, analyser et émettre un avis ou une opinion sur de nombreuses données résultant de l'activité économique. Les actuaires, les statisticiens et les économètres sont également d'autres professionnels du chiffre. Comme les professionnels comptables, ils collectent, traitent, analysent et émettent des avis sur de nombreuses données résultant notamment de l'activité économique. Dans leur travail quotidien, actuaires, statisticiens et économètres utilisent de **puissants outils**, cependant **peu démocratisés** au sein de la profession comptable. Ces outils sont indispensables pour tirer pleinement profit des **bases de données** constituées, du fait de l'explosion des capacités de stockage, par les organisations dans lesquelles interviennent les professionnels comptables.

Ce mémoire ambitionne de promouvoir au sein de la profession comptable l'utilisation d'**outils statistiques** présentant un certain niveau de complexité pour **l'analyse de l'information comptable et financière** des entreprises. A ce titre, il fournit un panorama de méthodes statistiques permettant de synthétiser et effectuer des traitements sur des données multidimensionnelles dans un objectif d'aide à la décision. Les développements pratiques du mémoire montrent comment des méthodes comme **l'analyse en composantes principales**, la **classification**, la **régression** et **l'analyse discriminante** peuvent être mises en œuvre par un professionnel comptable pour :

- **synthétiser** des **données multidimensionnelles** sur des entreprises en vue de:
  - fournir une vision synoptique et une meilleure compréhension globale des activités d'entreprises d'un secteur donnée ;
  - permettre une prise de connaissance rapide et une mise en évidence de particularités, sources ou manifestations de risques ;
- **comparer, discriminer et classifier** des entreprises en conformité avec la norme ISA 520 et la NEP 520 relatives aux **procédures analytiques**. Le positionnement d'une entreprise dans son secteur d'activité, les forces et

faiblesses mises en évidence dans ce cadre peuvent guider le professionnel comptable dans sa mission. Un tel diagnostic constitue également pour l'expert comptable une opportunité pour développer de nouvelles missions de conseil ;

- **construire des modèles**, faire des **estimations** et des **prévisions** notamment en conformité avec la norme ISA 540 et la NEP 540 relatives à l'appréciation des estimations comptables. Les modèles construits doivent aider le professionnel à forger son opinion ; ils ne peuvent en aucun cas remplacer le **jugement du professionnel**.

Les méthodes statistiques mises en œuvre à travers ce mémoire sont universelles et relativement complexes. Elles sont pleinement adoptées de façon innovante et féconde par des sciences comme la biologie ou l'économie pour donner naissance aux disciplines relativement nouvelles que sont la bio statistique et l'économétrie. Ces méthodes sont aujourd'hui implémentées dans de nombreux **logiciels libres** ou commerciaux dont une cartographie est fournie dans le mémoire. Ces logiciels permettent de **relativiser la complexité** de ces méthodes statistiques et d'en **démocratiser l'utilisation**.

En contrepartie d'un effort de prise en main, un logiciel libre comme , utilisé dans le cadre de ce mémoire, permet de mettre en œuvre la quasi totalité des méthodes statistiques utiles à un professionnel comptable. La documentation (sites internet et livres) relative à ce logiciel en permanente progression est abondante et de très bonne qualité. Elle est l'œuvre de **chercheurs** et d'**universitaires talentueux** et dévoués du monde entier.

Les données objet de l'étude sont extraites des comptes annuels des compagnies d'assurance de la Conférence Interafricaine des Marchés d'Assurances (CIMA), organisation internationale chargée de la **supervision de l'industrie des assurances** dans quatorze pays d'Afrique centrale et de l'ouest. L'étude ambitionne cependant de montrer le **caractère universel des méthodes** et **outils** mis en œuvre. Ces méthodes et outils peuvent être utilisés avec profit, notamment comme **procédures de contrôle**, par les professionnels comptables pour analyser les données des organisations dans lesquelles ils interviennent.

# SIGLES ET ABBREVIATIONS

ACP	Analyse en Composantes Principales
AFC	Analyse des Correspondances Factorielles
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CAH	Classification Ascendante Hiérarchique
CIMA	Conférence Interafricaine des Marchés d'Assurance
CNCC	Compagnie Nationale des Commissaires aux Comptes
CNRS	Centre National de Recherche Scientifique
CRAN	Comprehensive  Archive Network
CRCA	Commission Régionale de Contrôle des Assurances
ERIC	Equipe de Recherche et d'Ingénierie des Connaissances
FCFA	Francs de la Communauté Financière Africaine
IARD	Incendie Accidents Risques Divers
IC	Intervalle de Confiance
ISA	International Standards Audit
MCO	Moindre Carrés Ordinaires
NEP	Norme d'Exercice Professionnel
OEC	Ordre des Experts Comptables
PAP	Provisions pour Annulations de Primes
PM	Provisions Mathématiques
PSAP	Provisions pour Sinistre A Payer
SPSS	Statistical Package for the Social Sciences
WEKA	Waikato Environment for Knowledge Analysis

# INTRODUCTION

L'article 22 de l'ordonnance n° 45-2138 du 19 septembre 1945 portant institution de l'Ordre des Experts Comptables (OEC) prévoit que les experts comptables peuvent donner des consultations, effectuer toutes études et tous travaux d'ordre **statistique**, économique, administratif, juridique, social ou fiscal...sans pouvoir en faire l'objet principal de leur activité.

Les experts comptables exercent au-delà de leur mission principale, de nombreuses missions d'ordre économique, administratif, juridique, social ou fiscal. Par contre, les **travaux** d'ordre **statistique** effectués par les professionnels comptables sont **rares**. Ce champ de missions reste insuffisamment exploité par les experts comptables.

L'explosion des capacités de stockage informatique, le développement d'outils de calculs puissants et la constitution de grandes bases de données dans les organisations constituent cependant des **ferments** pour effectuer des études statistiques de tous genres et en particulier de la **statistique exploratoire et prédictive**.

L'information est aujourd'hui au cœur de l'activité économique et le moteur des prises de décision. Cependant, une grande partie de l'information contenue dans ces bases de données est insuffisamment exploitée en vue d'en tirer les enseignements nécessaires à une prise de **décision optimale**.

L'expert comptable est au cœur de la production et de l'analyse de l'information comptable et financière. Au regard de la masse d'informations collectées ou disponibles sur une organisation ou sur un secteur d'activité donné, il doit se procurer les moyens de la synthétiser et d'en tirer la « **substantifique moelle** ». Dans ce cadre, le mémoire mettra en évidence l'apport des techniques statistiques pour synthétiser l'information et permettre d'avoir une **vision synoptique** et une **meilleure compréhension** globale de l'activité des entreprises d'un secteur donné.

Ce travail de synthèse peut aider l'expert comptable à jouer son rôle d'information et de conseil auprès des dirigeants pour expliquer par exemple les **particularités** de leur entreprise et son positionnement dans son secteur d'activité. Ce travail de synthèse et de restitution de l'information comptable et financière constitue également une opportunité pour développer de **nouvelles missions de conseil** et pour les dirigeants un moyen de mieux communiquer vers l'actionnariat.

Dans le cadre de la documentation de ses travaux et de la prise de connaissance d'un secteur d'activité ou d'une entreprise quelconque, le professionnel comptable collecte de nombreuses informations de nature diverse. Pour l'analyse de ces informations, il utilise notamment des **procédures analytiques**. Le champ de ces procédures est relativement vaste. La norme ISA 520 précise que les procédures analytiques « vont de simples comparaisons à des analyses complexes faisant appel à des **techniques statistiques sophistiquées** ». Les techniques et les outils statistiques complexes ne font généralement pas partie de la formation première de l'expert comptable. De ce fait, les informations collectées ne sont pas toujours exploitées de façon optimale pour permettre d'orienter et de mener au mieux les missions.

L'étude permettra de montrer comment les « techniques statistiques sophistiquées » peuvent être utilisées par les professionnels comptables au titre de la procédure analytique comme technique de contrôle. Par exemple, l'analyse multidimensionnelle de données peut permettre d'apprécier les informations financières d'une entreprise à partir de leur corrélation avec d'autres informations de la même entité ou d'entités similaires.

L'objectif fondamental du mémoire est donc de montrer comment les outils statistiques peuvent contribuer de **façon innovante et féconde à la présentation et à l'analyse des informations comptables et financières** produites par les entreprises. Le mémoire ambitionne de promouvoir au sein de la profession comptable l'utilisation d'outils statistiques présentant un certain niveau de complexité pour l'analyse de l'information comptable et financière des entreprises.

D'un point de vue pratique, les exemples traités et analysés le long du mémoire permettent de partager la démarche de mise en œuvre des outils statistiques et mettent en évidence leurs **apports** pour le professionnel comptable.

Pour atteindre l'objectif visé, le mémoire est composé de trois parties.

La **première partie** présente les **fondements mathématiques, statistiques et informatiques** des outils mis en œuvre dans la deuxième et la troisième partie du mémoire. Cette partie un peu théorique, est cependant nécessaire pour bien comprendre et exploiter les outils présentés.

La **deuxième partie** est consacrée à l'application de **méthodes d'analyse de données multidimensionnelles** et de **classification** aux informations comptables et financières des compagnies d'assurance. Du fait de la nature des données, les techniques statistiques utilisées sont l'analyse en composantes principales et la classification ascendante hiérarchique. Cette partie montre comment le professionnel comptable peut utiliser des outils statistiques pour **explorer** et **synthétiser** des informations comptables et financières notamment dans la phase de prise de connaissance d'une entité ou d'un secteur d'activité donnée.

La dernière partie traite de différentes méthodes statistiques permettant de faire notamment de la **prédiction**. Les techniques statistiques mises en œuvre dans cette partie sont la **régression linéaire**, la **régression logistique** et l'**analyse discriminante**. Ces techniques sont utilisées comme **procédures analytiques** et **procédures d'estimation** pour évaluer notamment les provisions techniques ou prédire la situation financière des entreprises d'assurance.

Toutes ces différentes techniques sont mises en œuvre sur un jeu de données résultant des comptes annuels produits par les compagnies d'assurance de la Conférence Interafricaine des Marchés d'Assurance (CIMA) au titre de l'exercice 2009. A la fin de l'exercice 2009, cet espace couvrant 14 pays d'Afrique centrale et de l'ouest regroupait 90 sociétés d'assurance non vie réalisant un chiffre d'affaires de 384 milliards de FCFA (585 millions d'euros) et 45 sociétés d'assurance vie avec un chiffre d'affaires de 153 milliards de FCFA (233 millions d'euros).

**PREMIERE PARTIE : PRESENTATION  
DES METHODES, DES OUTILS  
STATISTIQUES ET DES DONNEES  
UTILISEES**

# PREMIERE PARTIE : PRESENTATION DES METHODES, DES OUTILS STATISTIQUES ET DES DONNEES UTILISEES

Cette partie introductive du mémoire présente les méthodes, les outils statistiques mis en œuvre dans la deuxième et troisième partie. Les deux premiers chapitres peuvent paraître relativement techniques. Dans ce cas, une première lecture peut être effectuée en s'en tenant simplement aux principes présentés de façon introductive avant d'aborder les aspects purement techniques et mathématiques.

Le **premier chapitre** passe en revue différentes méthodes statistiques utilisées pour explorer des données multidimensionnelles afin d'en tirer des enseignements utiles à la prise de décision. Ce chapitre présente successivement **l'analyse en composantes principales, la classification, l'analyse canonique, l'analyse des correspondances factorielles et l'analyse des correspondances multiples.**

Le **deuxième chapitre** est dédié à différentes méthodes statistiques permettant de faire de la prédiction et de discriminer des individus. Ce chapitre passe en revue successivement **la régression linéaire, l'analyse discriminante et la régression logistique.**

Au delà de la bibliographie fournie, ces deux premiers chapitres ont été rédigés en s'appuyant particulièrement sur les trois ouvrages suivants :

BOURBONNAIS Régis, Econométrie, Dunod, 8ème édition, 2011, 380 p.

SAPORTA Gilbert, Probabilités, analyse des données et statistique, Technip, 2ème édition 2006, 622 p.

TENENHAUS Michel, Statistiques : Méthodes pour décrire, expliquer et prévoir, Dunod, 2010, 679 p.

Le lecteur souhaitant approfondir des aspects techniques et mathématiques peut donc s'y référer.

L'un des intérêts du mémoire est la mise en œuvre des méthodes statistiques sur des données comptables et financières en vue d'en tirer des enseignements pour le

dirigeant ou le professionnel comptable. Ainsi, le **troisième chapitre** présente différents **outils informatiques** permettant une mise en œuvre quasi automatique de méthodes statistiques d'un certain niveau de complexité. Un point particulier est fait sur le logiciel  utilisé dans le cadre des travaux de ce mémoire. Enfin, le **jeu de données** utilisé est présenté à la fin de ce chapitre.

## **CHAPITRE I. STATISTIQUE EXPLORATOIRE ET ANALYSE DE DONNEES**

Généralement, un individu statistique est décrit par un nombre  $p$  de variables souvent élevé et non par une seule variable. Par exemple, chaque poste du bilan et du compte de résultat d'une entreprise constitue une variable permettant de décrire un aspect de son activité. L'étude séparée de chacune de ces variables à travers une statistique descriptive (moyenne, médiane, variance, écart type ...) est une phase indispensable dans le processus de dépouillement des données. Elle est cependant insuffisante.

Il faut donc analyser les données en tenant compte de leur caractère multidimensionnel. L'**analyse en composantes principales** est alors une **méthode** particulièrement **puissante** pour explorer la structure des données.

D'autres méthodes d'analyse de données multidimensionnelles existent. On peut citer notamment :

- les **méthodes de classification** qui permettent de construire une partition ou une suite de partitions emboîtées, d'un ensemble d'objets dont on connaît les distances deux à deux ;
- l'**analyse canonique** permettant de déterminer pour des individus décrits par deux groupes de variables, les liens entre ces deux groupes ;
- l'**analyse des correspondances factorielles (AFC) et l'analyse des correspondances multiples (ACM)** qui permettent d'étudier les liaisons entre variables qualitatives.

Ces différentes méthodes d'analyse de données sont présentées ci-dessous avec un accent plus particulier pour les méthodes relatives au traitement des **données quantitatives** et notamment l'analyse en composantes principales et la classification.

## 1) Analyse en composantes principales (ACP)

Lorsque les individus statistiques sont décrits par plus de trois variables, la **représentation graphique** devient **impossible**. Si on désire cependant une représentation graphique plane, on peut projeter les individus décrits par  $p$  variables donc de l'espace à  $p$  dimensions sur un plan (à deux dimensions). Il faut cependant choisir judicieusement le **plan de projection** pour que les distorsions par rapport à l'**espace originel** soient minimales. Le principe guidant le choix de ce plan est présenté sans formalisme mathématique excessif en annexe (**Principe du choix du plan principal page 121**).

L'analyse en composantes principales est une **méthode factorielle**, car elle réduit le nombre de caractères, non pas en éliminant telle ou telle variable jugée non pertinente, mais en construisant de nouveaux axes plus pertinents.

### 1.1) Données et caractéristiques

Les  $n$  individus  $x_i$  décrits par  $p$  variables sont rassemblés dans un **tableau** rectangulaire noté  $X$  à  $n$  lignes et  $p$  colonnes.

$$X = \begin{bmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_i^1 & \cdots & x_i^j & \cdots & x_i^p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix}$$

Le tableau rectangulaire  $X$  peut être assimilé à une matrice  $X_{n \times p}$  notée simplement également **matrice**  $X$ . La matrice  $X$  représente l'espace des individus. La matrice transposée de  $X$  notée  ${}^tX$  représente l'espace des variables.

On définit la **matrice de variance-covariance** notée  $V$  par le produit matriciel suivant :  $V = {}^tXDX$  et la **matrice de corrélation** notée  $R$  par le produit matriciel suivant :  $R = D_{1/s} {}^tXDXD_{1/s} = D_{1/s}VD_{1/s}$ . La **matrice**  $D$  **est diagonale**. Ces différentes matrices sont particulièrement importantes dans le cadre de l'analyse en composantes principales. Leur **construction** et leur forme **générique** sont fournies en annexe (**Données et caractéristiques page 122**).

### 1.2) Recherche de sous espaces principaux

Les données étant  $n$  points de l'espace  $\mathbb{R}^p$  des variables, l'**objectif est de réduire le nombre de variables** en projetant les points sur un sous espace de  $\mathbb{R}^p$ . Pour que ce

sous espace ait un sens, il faut que les points projetés et les points initiaux ne soient pas trop éloignés.

En notant  $p(x_i)$  la projection de l'individu  $x_i$  sur un sous espace  $\mathcal{H}$  de  $\mathbb{R}^p$ , on définit la quantité suivante  $\sum_i p_i \|x_i - p(x_i)\|^2$  à minimiser. On appelle **sous espace principal**, un sous espace de  $\mathbb{R}^p$  minimisant la quantité  $\sum_i p_i \|x_i - p(x_i)\|^2$ . La quantité  $\|x_i - p(x_i)\|$  est la distance entre le point  $x_i$  et sa projection  $p(x_i)$ .

La recherche d'un sous espace  $\mathcal{H}$  de  $\mathbb{R}^p$  minimisant la quantité  $\sum_i p_i \|x_i - p(x_i)\|^2$  est équivalente à la recherche d'un sous espace passant par l'origine du repère maximisant la quantité  $\sum_i p_i \|p(x_i)\|^2$ .

La droite  $\Delta$  maximisant  $\sum_i p_i \|p(x_i)\|^2$  a pour vecteur directeur  $u_1$ , **vecteur propre** de  $V$  associé à  $\lambda_1$ , la plus grande de ses **valeurs propres**.

Pour trouver un sous espace à  $k$  dimensions maximisant  $\sum_i p_i \|p(x_i)\|^2$ , on commence par trouver un sous espace à  $k - 1$  dimensions. Connaissant un sous espace principal  $H$  à  $k - 1$  dimensions, il suffit de trouver un sous espace principal  $H'$  à une dimension dans l'**orthogonal** de  $H$ . Le sous espace  $H \oplus H'$  est un sous espace à  $k$  dimensions.

Ainsi, trouver un sous espace à deux dimensions revient donc à trouver un sous espace à une dimension dans l'orthogonal de la droite engendrée par le vecteur directeur  $u_1$ .

Les vecteurs propres  $(u_1, u_2, \dots, u_p)$  de  $V$  rangés par ordre décroissant de leurs valeurs propres  $(\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p)$  forment une **base orthonormée**. On en conclut qu'un sous espace principal à  $k$  dimensions est exactement  $(u_1 \oplus u_2 \oplus \dots \oplus u_k)$ .

### 1.3) Description du nuage des individus

Les vecteurs propres  $(u_1, u_2, \dots, u_p)$  formant une base orthonormée de  $\mathbb{R}^p$ , ils tiennent lieu de **nouveaux axes**. Les **coordonnées des individus** dans cette **nouvelle base** s'obtiennent par **projection** sur chacun des axes.

La projection orthogonale des points sur chacun des axes principaux permet d'obtenir les **composantes principales**, vecteurs colonnes définis par le produit matriciel suivant :  $c^j = Xu^j$ . Dans la nouvelle base, l'individu  $x_i$  a donc pour coordonnées  $(c_i^1, c_i^2, \dots, c_i^p)$ .

Les composantes principales sont de **nouvelles variables, combinaisons linéaires** des **variables initiales**. On peut alors visualiser le nuage  $X$  sur le plan principal d'inertie qui est le sous espace principal de dimension 2. La qualité de la représentation du nuage  $X$  peut être étudiée sous deux angles.

Le nuage étant projeté sur une droite ou un plan, la qualité de la représentation sera d'autant meilleure que l'**inertie** de ce sous espace est **forte**. L'inertie totale du nuage valant  $(V) = p$ , on introduit les **parts expliquées** par l'axe  $u^j = \frac{\lambda_j}{p}$  et le plan formé par les axes  $u^i$  et  $u^j$  qui vaut  $\frac{\lambda_i + \lambda_j}{p}$ .

Plus le point  $x_i$  est proche du sous espace  $H$  sur lequel il est projeté, plus **pertinente** est sa **représentation**. Cette proximité est mesurée par le **cosinus carré** de l'angle de  $x_i$  et de  $H$ . On a donc  $\cos^2 \alpha = \frac{\|Projection\ de\ x_i\ sur\ H\|^2}{\|x_i\|^2}$ . Cette formule vient du fait que dans un triangle rectangle, le cosinus d'un angle est égal au côté adjacent de l'angle rapporté à l'hypoténuse.

Le cosinus carré de l'angle entre  $x_i$  et le facteur  $u^j$  est donc égal à  $\cos^2 \alpha = \frac{\|c_i^j\|^2}{\|x_i\|^2}$  et le cosinus carré de l'angle entre  $x_i$  et le plan  $u_j \oplus u_k$  est égal à  $\cos^2 \alpha = \frac{\|c_i^j\|^2 + \|c_i^k\|^2}{\|x_i\|^2}$ .

Les **individus remarquables** influent sur l'analyse. Il est donc intéressant d'évaluer leur influence sur la construction des axes. La **part d'inertie** expliquée par l'individu  $x_i$  sur l'axe  $j$  est :  $Contribution = \frac{\|Projection\ de\ x_i\ sur\ H\|^2}{n * \lambda_j} * 100$ . Cette formule fait intervenir des distances au carré et accentue le rôle des individus éloignés de l'origine.

#### 1.4) Description du nuage des variables

Les **variables initiales**  $x^1, x^2, \dots, x^p$  forment  $p$  points de l'espace  $\mathbb{R}^n$  des individus. Les  $p$  **composantes principales**  $c^1, c^2, \dots, c^p$  sont obtenues par combinaison linéaire des variables initiales. Les composantes principales peuvent être décrites par leur corrélation avec les variables initiales. La corrélation entre une composante principale  $c^j$  et une variable initiale  $x^k$  est égale à  $r(x^k, c^j) = \frac{cov(x^k, c^j)}{s(x^k)s(c^j)}$  et  $\cos^2 \theta_j^k = r^2(x^k, c^j)$ .

La **part d'inertie** expliquée par la variable  $x^k$  sur l'axe  $j$  est :  
 $Contribution = \frac{r^2(x^k, c^j)}{\lambda_j} * 100$ . La **contribution** est calculée pour chaque variable sur un axe donné. Elle permet d'identifier les variables contribuant le plus à la construction d'un axe donné.

## 2) Classification

La **classification** et l'**analyse factorielle** s'inscrivent dans une même **perspective** : l'**analyse exploratoire** d'un **tableau de données**. Elles diffèrent cependant, selon le **mode de représentation** (nuage euclidien, hiérarchie indicée ou partition).

La **combinaison** des **deux approches** permet d'obtenir une **méthodologie riche** et de disposer de plusieurs points de vue permettant de renforcer la solidité des conclusions.

Les modèles usuellement utilisés en classification sont les **partitions** et les **hiérarchies**. Dans les deux cas, les objets qui se ressemblent sont regroupés en classes. Pour les partitions, les classes sont deux à deux disjointes, pour les hiérarchies, elles peuvent être emboîtées.

### 2.1) Modèles de classification

En **classification**, toute l'information utile est contenue dans un tableau  $n \times n$  donnant les **dissemblances** entre les  $n$  individus à classer. La notion de dissemblance est plus large que celle de distance. Ces notions sont présentées de façon détaillée et formalisée en annexe (**Dissimilarité et distance page 124**).

Un sous ensemble  $K$  de  $2^X$  (ensemble des parties de  $X$ ) sera appelé système de classes sur  $X$  si et seulement si, il vérifie les trois propriétés suivantes :

- $(p_1)$ :  $X \in K$  et  $\emptyset \notin K$  ;
- $(p_2)$ : pour tout  $x \in X$ ,  $\{x\} \in K$  ;
- $(p_3)$ : pour tous  $A, B \in K$ ,  $A \cap B \neq \emptyset$  entraîne  $A \cap B \in K$ .

Par définition, une **hiérarchie** est un système de classe pour lequel, deux classes sont toujours incluses l'une dans l'autre, soit d'intersection vide. De ce fait, une hiérarchie est représentée sous la forme d'un arbre appelé plus précisément **dendrogramme** où chaque classe est représentée par un segment.

Un sous ensemble  $R = \{P_1, P_2, \dots, P_k\}$  de  $2^X$  sera appelé **recouvrement** de  $X$  si et seulement si, il vérifie les trois propriétés suivantes :

- $(p_1)$ : pour tout  $1 \leq i \leq k, P_i \neq \emptyset$  ;
- $(p_2)$ : pour tous  $1 \leq i \neq j \leq k, P_i \not\subset P_j$  et  $P_j \not\subset P_i$  ;
- $(p_3)$ :  $P_1 \cup P_2 \cup \dots \cup P_k = X$ .

Une **partition**  $P$  est un **recouvrement** tel que les classes sont donc deux à deux disjointes. Formellement, on a : pour toutes classes  $A$  et  $B$  de  $P$ :  $A \cap B = \emptyset$  si  $A \neq B$ .

## 2.2) Méthodes de partitionnement

Une partition  $P = \{C_1, C_2, \dots, C_k\}$  sur  $X$  est de meilleure qualité si l'**inertie intraclasse** est **minimisée**. La partition à  $n$  éléments possède une inertie intra classe nulle. L'objectif est donc de trouver une partition  $k < n$  qui minimise l'inertie intraclasse.

Il est irréaliste de vouloir essayer toutes les partitions possibles car le nombre de partitions possibles explose exponentiellement avec  $n$ . Les **algorithmes de partitionnement** utilisés sont tous des **heuristiques**. Ils trouvent la plupart du temps une partition acceptable, mais sans **garantie d'optimalité**.

Les principales étapes de l'**algorithme des  $k$  means** ou des centres mobiles sont fournies en annexe (**Algorithme des  $k$  means page 124** ). Cet algorithme est convergent. Il s'arrête lorsque deux itérations successives conduisent à une décroissance du critère de la somme des carrés intraclasse ou à sa stabilisation.

## 2.3) Algorithme de classification ascendante hiérarchique

L'**algorithme de CAH** est une méthode générale de construction d'une hiérarchie à partir de données décrites par une dissimilarité. Il peut être présenté comme suit :

1. A l'étape initiale, chaque individu forme une classe. Il y a donc  $n$  classes.
2. A la première étape, on utilise la matrice de dissimilarités entre les  $n$  individus et on regroupe les deux individus les plus proches. Ces deux individus forment une nouvelle classe et on obtient une nouvelle partition de  $n - 1$  classes.
3. A la deuxième étape, une nouvelle matrice de dissimilarités entre les  $n - 1$  classes est construite en utilisant le critère d'agrégation. On regroupe les deux classes les plus proches et on obtient une nouvelle partition de  $n - 2$  classes.

4. On répète le processus jusqu'à l'étape  $n - 1$  où le regroupement des classes de la partition de l'étape précédente conduit à une seule classe.

### 3) Autres méthodes d'analyse de données

Les méthodes statistiques et d'analyse de données présentées dans les paragraphes suivants ne seront pas mises en œuvre dans le cadre de l'étude effectuée. Il est cependant important de connaître leurs principes, les circonstances et le type de données auxquels elles peuvent être appliquées.

Les paragraphes suivants permettent donc de présenter successivement :

- **L'analyse canonique ;**
- **L'analyse factorielle des correspondances ;**
- **Et l'analyse des correspondances multiples.**

#### 3.1) Analyse canonique

Dans le cadre de l'**analyse canonique**, les individus sont décrits par **deux groupes de variables** (en nombre  $p$  et  $q$  respectivement). On cherche alors à déterminer les **liens** existants entre ces deux groupes de variables. Cette recherche doit permettre de savoir si les deux groupes de variables mesurent ou non les **mêmes propriétés**.

Par exemple, il peut être intéressant de savoir si les variables du bilan et celles du compte de résultat mesurent ou non les mêmes propriétés.

Le tableau des données  $X$  est généralement une concaténation de deux tableaux ( $X_1$  et  $X_2$ ) et est de la forme suivante :

$$X = \left[ \begin{array}{cccc} x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_i^1 & \cdots & x_i^j & \cdots & x_i^p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{array} \right] \left[ \begin{array}{cccc} x_1^1 & \cdots & x_1^j & \cdots & x_1^q \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_i^1 & \cdots & x_i^j & \cdots & x_i^q \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^q \end{array} \right]$$

On considère alors les deux sous espaces de  $\mathbb{R}^n$  engendrés par les colonnes de  $X_1$  et de  $X_2$  respectivement  $W_1 = \{x|x = X_1 a\}$  et  $W_2 = \{y|y = X_2 b\}$ .  $W_1$  et  $W_2$  sont les deux ensembles de variables que l'on peut constituer par combinaison linéaire des variables des deux groupes.

Si ces deux espaces sont confondus, on peut se contenter d'un seul des deux ensembles de variables car ils ont alors un même pouvoir de description. S'ils sont

orthogonaux, on conclut que les deux ensembles de variables appréhendent des phénomènes totalement différents. Ces deux cas extrêmes étant exceptionnels, les positions géométriques de  $W_1$  et  $W_2$  sont étudiées en cherchant les éléments les plus proches. Cette étude permet en particulier de connaître  $\dim(W_1 \cap W_2)$ .

Les **techniques de l'analyse canonique** se retrouvent dans d'autres méthodes comme **l'analyse des correspondances, la régression multiple, l'analyse discriminante** :

- Si la dimension de l'un des groupes de variables étudiées ( $q$  ou  $p$ ) est égale à 1, **l'analyse canonique est équivalente à la régression linéaire multiple** ;
- Si un des groupes est composé de variables indicatrices d'une variable qualitative et l'autre de variables numériques, **l'analyse canonique correspond à une analyse factorielle discriminante** ;
- Si les deux groupes de variables sont composés des indicatrices de deux variables qualitatives à  $p$  et  $q$  catégories, **l'analyse canonique est équivalente à une analyse des correspondances**.

### 3.2) Analyse factorielle des correspondances

L'**analyse factorielle des correspondances (AFC)** est un **cas particulier** de **l'analyse canonique**. Elle correspond également à une analyse en composantes principales avec une **métrique spéciale, la métrique du  $\chi^2$** .

L'objectif de l'AFC est de **résumer** et de **visualiser** un **tableau croisant deux variables qualitatives**. Un tel tableau croisé, appelé également **tableau de contingence**, se présente généralement comme suit :

	$y_1$	...	$y_j$	...	$y_l$	Somme
$x_1$	$n_{11}$	...	$n_{1j}$	...	$n_{1l}$	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_i$	$n_{i1}$	...	$n_{ij}$	...	$n_{il}$	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_k$	$n_{k1}$	...	$n_{kj}$	...	$n_{kl}$	$n_{k.}$
Somme	$n_{.1}$	...	$n_{.j}$	...	$n_{.l}$	$N$

Tableau 1: Forme générale d'un tableau de contingence

Un tableau de contingence de deux variables  $x$  et  $y$  possède autant de lignes que  $x$  a de valeurs différentes et autant de colonnes que  $y$  a de valeurs différentes. Une case  $n_{ij}$  correspond à l'effectif de la population présentant à la fois le caractère  $x_i$  et le caractère  $y_j$ .

L'AFC ne sera pas mise en œuvre dans le cadre de ce mémoire. Cependant, le champ d'application de cette technique est relativement vaste. Une possibilité d'application de l'AFC à l'épreuve du mémoire d'expertise comptable et des développements additionnels sont fournis à l'annexe suivante : **1.3) Analyse factorielle des correspondances page 125.**

### 3.3) Analyse des correspondances multiples

L'analyse des correspondances multiples (ACM) est une généralisation de l'analyse factorielle des correspondances à plus de deux variables qualitatives. Dans ce cas, le tableau de contingence simple est inopérant. Il faut soit le généraliser pour aboutir à un **tableau de contingence multiple ou tableau de Burt** soit lui substituer un **tableau disjonctif complet**.

L'objectif de l'ACM est d'étudier les **ressemblances** entre individus du point de vue de l'ensemble des variables et de dégager des **profils d'individus**. L'ACM permet également d'établir une synthèse des liaisons entre variables et d'étudier les **associations de modalités**.

Un tableau disjonctif complet se présente généralement comme suit :

	...	...	...	$X_{j1}$	...	$X_{jl}$	...	$X_{jp_j}$	...	...	...	Somme
1	...	...	...	$x_{1j1}$	...	$x_{1jl}$	...	$x_{1jp_j}$	...	...	...	$m$
⋮				⋮		⋮		⋮				⋮
$X = i$	...	...	...	$x_{ij1}$	...	$x_{ijl}$	...	$x_{ijp_j}$	...	...	...	$m$
⋮				⋮		⋮		⋮				⋮
$n$	...	...	...	$x_{nj1}$	...	$x_{njl}$	...	$x_{njp_j}$	...	...	...	$m$
Somme	...	...	...	$n_{j1}$	...	$n_{jl}$	...	$n_{jp_j}$	...	...	...	$n \times m$

Tableau 2: Forme générale d'un tableau disjonctif complet

Dans un tel tableau, on a  $m$  variables qualitatives  $x_1, \dots, x_j, \dots, x_m$  observées sur  $n$  individus. La variable  $x_j$  possède  $p_j$  modalités et le nombre total de modalités toutes variables confondues est  $p = \sum_{j=1}^m p_j$ .

Les différentes notations sont précisées ci après :

- $x_{ijl} = 1$  si l'individu  $i$  prend la modalité  $l$  de la variable  $x_j$ , sinon  $x_{ijl} = 0$ . Le tableau disjonctif complet est rempli de 0 et de 1 ;
- $X_{jl}$  : variable indicatrice de la modalité  $l$  de la variable  $x_j$  ;
- $X_j = [X_{j1} \dots \dots X_{jp_j}]$  : tableau des données binaires associées à la variable  $x_j$  ;
- $X = [X_1 \dots \dots X_m]$  : tableau disjonctif complet associé aux variables du problème.

Une possibilité d'application de l'ACM est fournie à l'annexe suivante : **1.4) Analyse des correspondances multiples page 126.**



La sagesse populaire affirme qu'une **image** vaut mieux que **mille mots** et Napoléon Bonaparte affirmait qu'un « **bon croquis** vaut mieux qu'un **long discours** ». Grâce à l'analyse factorielle, un **phénomène** peut être **synthétisé** et **représenté** sous forme d'une **image** avec une **précision infinie** dans un espace à deux ou trois dimensions. Cette représentation permet à un **œil profane** de comprendre les **caractéristiques essentielles** d'un phénomène par la localisation des **grandes masses d'individus**, la détection des **individus exceptionnels** et la mise en évidence d'éventuels **groupes isolés**.

Les méthodes factorielles sont particulièrement **efficaces** lorsque les variables étudiés sont **corrélées**. Cette **corrélation** devient cependant un **handicap** dans le cadre des **modèles prédictifs** étudiés dans le chapitre suivant.

## CHAPITRE II. MODELES PREDICTIFS

Les **techniques prédictives** s'appliquent à des problèmes variés notamment en médecine, en sciences sociales et dans de nombreux domaines de gestion des entreprises (prédiction de ventes, appétence d'un client, probabilité de faire faillite d'une entreprise ...).

Ces techniques peuvent être réparties en deux grands groupes : les **techniques de régression** (régression linéaire, régression logistique) et les **techniques de discrimination** (analyse discriminante). Dans tous les cas, l'objectif est d'estimer la

valeur d'une variable « **à expliquer ou endogène** » d'un individu en fonction des valeurs d'un certain nombre d'autres variables « **explicatives ou exogènes** » du même individu.

### 1) Régression linéaire simple et multiple

La **régression linéaire simple** est une méthode statistique permettant d'étudier la **liaison** entre **une variable à expliquer  $Y$**  et **une variable explicative  $X$** . La **régression multiple** est une extension et une généralisation du modèle de régression simple dans laquelle figurent plusieurs variables explicatives.

#### 1.1) Présentation des modèles

Lorsque les variables représentent des phénomènes observés à une même période sur plusieurs individus, cette liaison est généralement modélisée comme suit :  $Y_i = a_0 + a_1X_i + \varepsilon_i$  ou  $Y_t = a_0 + a_1X_t + \varepsilon_t$  dans le cas d'une **série temporelle**.

La relation entre la variable à expliquer et les variables explicatives est généralement **modélisée** comme suit :

$$y_i = a_0 + a_1x_{1i} + a_2x_{2i} + \dots + a_kx_{ki} + \varepsilon_i \text{ pour } i = 1, \dots, n$$

Cette écriture est détaillée en annexe (**Explicitation des termes de la régression linéaire page 127**). Elle peut également être mise sous la **forme matricielle** suivante :

$$Y_{n \times 1} = X_{n \times (k+1)} a_{(k+1) \times 1} + \varepsilon_{n \times 1} = Xa + \varepsilon$$

Les termes de cette écriture matricielle sont explicités en annexe (**Explicitation des termes de la régression linéaire page 127**).

#### 1.2) Estimation des coefficients de régression

Le vecteur  $a$ , composé des coefficients  $a_0, a_1, \dots, a_k$ , est estimé par la **méthode des moindres carrés ordinaires** consistant à minimiser la somme des carrés des erreurs :

$$\text{Min} \sum_{i=1}^n \varepsilon_i^2 = \text{Min} \varepsilon' \varepsilon = \text{Min} (Y - Xa)' (Y - Xa) = \text{Min} S \quad \text{avec } \varepsilon' \text{ transposé du vecteur } \varepsilon.$$

Cette méthode est rappelée de façon détaillée en annexe (**Estimation des paramètres du modèle dans le cadre d'une régression simple page 128**) dans le cas de la régression linéaire simple.

En **dérivant** la fonction  $S$  par rapport à  $a$ , on peut écrire :

$$\frac{\partial S}{\partial a} = -2X'Y + 2X'X\hat{a} = 0 \rightarrow \hat{a} = (X'X)^{-1}X'Y$$

Cette équation a une solution si la matrice  $X'X$  de dimension  $(k + 1, k + 1)$  est **inversible**. En cas de **colinéarité** parfaite entre deux variables explicatives, la matrice  $X'X$  est **singulière** et la méthode des moindres carrés ordinaires est inopérante.

On montre que l'estimateur  $\hat{a}$  est sans biais ( $E(\hat{a}) = a$ ) et convergent.

Les différentes hypothèses qui sous-tendent la construction d'une régression linéaire simple ou multiple sont rappelées en annexe (**Hypothèses du modèle linéaire simple et Hypothèses du modèle linéaire multiple page 130**).

Une fois les paramètres estimés, on peut calculer les **valeurs lissées** ou **ajustées** ou **prévoir de nouvelles valeurs** :  $\hat{y}_i = \hat{a}_0 + \hat{a}_1x_{1i} + \hat{a}_2x_{2i} + \dots + \hat{a}_kx_{ki}$ .

La différence entre la valeur observée et la valeur calculée est par définition le **résidu** estimé :  $\hat{\varepsilon}_i = y_i - \hat{y}_i$ .

L'**analyse des résidus** permet de se prononcer sur la **qualité de l'ajustement réalisé et du modèle construit**.

### 1.3) Analyse de la variance et sélection des variables

L'équation fondamentale d'analyse de la variance permet d'écrire que la **variabilité totale** est égale à la **variabilité expliquée par le modèle** + la **variabilité des résidus**. Cela revient à dire que la Somme des Carrés des Ecart Total ( $SCE_T$ ) est égale à la Somme des Carrés des Ecart du Modèle ( $SCE_M$ ) + la Somme des Carrés des Ecart du Résidu ( $SCE_R$ ).

$$\text{Formellement on écrit que : } \begin{cases} \sum_i^n (y_i - \bar{y})^2 = \sum_i^n (\hat{y}_i - \bar{y})^2 + \sum_i^n (y_i - \hat{y}_i)^2 \\ SCE_T = SCE_M + SCE_R \end{cases}$$

Cette équation permet de juger de la qualité de l'ajustement du modèle. Elle permet également de calculer le **coefficient de détermination**  $R^2 = \frac{SCE_M}{SCE_T} = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2}$ . Ce coefficient mesure la proportion de la variance de  $Y$  expliquée par la régression de  $X$ . Plus il est proche de 1, meilleure est la qualité d'ajustement global du modèle.

L'introduction d'une nouvelle variable dans le modèle augmente le  $R^2$  sauf en cas de colinéarité. Pour cette raison, on procède à une correction du  $R^2$  par la formule suivante :  $\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right) \left(\frac{n-1}{n-k-1}\right)$ .

Le **tableau d'analyse de la variance** se présente comme suit :

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Régression	$\sum_i^n (\hat{y}_i - \bar{y})^2$	$k$	$\sum_i^n (\hat{y}_i - \bar{y})^2 / k$
Résidu	$\sum_i^n (y_i - \hat{y}_i)^2$	$n - k - 1$	$\sum_i^n (y_i - \hat{y}_i)^2 / (n - k - 1)$
Total	$\sum_i^n (y_i - \bar{y})^2$	$n - 1$	$\sum_i^n (y_i - \bar{y})^2 / (n - 1)$

**Tableau 3: Tableau d'analyse de la variance**

Ce tableau permet d'effectuer le **test de Fischer**. Les hypothèses du test sont les suivantes :  $\begin{cases} H_0: a_1 = a_2 = \dots = a_k = 0 \\ H_1: \text{il existe au moins un des coefficients non nul} \end{cases}$

La **statistique de Fisher** calculée est la suivante :

$$F^* = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2 / k}{\sum_i^n (y_i - \hat{y}_i)^2 / (n - k - 1)} = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

L'hypothèse de normalité des erreurs implique que sous l'hypothèse  $H_0$ , la statistique  $F^*$  suit une loi de Fisher à  $k$  et  $(n - k - 1)$  degrés de liberté. La comparaison de  $F^*$  au  $F$  théorique permet de rejeter l'hypothèse  $H_0$  si  $F^* > F$ . Le rejet de l'hypothèse  $H_0$  permet de conclure que le modèle est globalement satisfaisant.

Le critère de la maximisation du coefficient de détermination  $R^2$  consiste à retenir le modèle dont le  $R^2$  est le plus élevé. Ce critère présente l'inconvénient de ne pas arbitrer entre la **perte de degrés de liberté** du modèle et la **qualité de l'ajustement** obtenu. Les critères de **Akaike AIC** ou de **Schwartz BIC** permettent de comparer des modèles impliquant un nombre différent de variables explicatives, le meilleur étant celui pour lequel  $AIC$  et  $BIC$  sont les plus bas. Ces deux critères sont définis comme suit :

$$AIC = \ln\left(\frac{SCE_R}{n}\right) + \frac{2k}{n} \text{ et } BIC = \ln\left(\frac{SCE_R}{n}\right) + \frac{k \ln(n)}{n} \text{ avec :}$$

$\ln$  = logarithme népérien

$SCE_R$  = Somme des carrés des résidus du modèle

$n$  = nombre d'observations

$k$  = nombre de variables explicatives

#### 1.4) Prévision et intervalle de confiance

Ayant une nouvelle donnée  $i + h$  pour laquelle on dispose des valeurs des variables explicatives  $x_{i+h} = (x_{1i+h}, x_{2i+h} + \dots + x_{ki+h})$ , le problème consiste à prédire  $y_{i+h}$ . Le modèle construit indique que :  $y_{i+h} = \hat{a}_0 + \hat{a}_1 x_{1i+h} + \hat{a}_2 x_{2i+h} + \dots + \hat{a}_k x_{ki+h} + \varepsilon_{i+h}$

et  $\hat{y}_{i+h} = \hat{a}_0 + \hat{a}_1 x_{1i+h} + \hat{a}_2 x_{2i+h} + \dots + \hat{a}_k x_{ki+h}$ .

Un **intervalle de confiance (IC)** de niveau  $(1 - \alpha)$  de la prévision  $y_{i+h}$  est donné par la formule suivante :

$$IC = \left[ x'_{i+h} \hat{a} \pm t_{n-k-1}^{\alpha/2} \hat{\sigma}_\varepsilon \sqrt{x'_{i+h} (X'X) x_{i+h} + 1} \right]$$

Dans cette expression :

- $x'_{i+h}$  est la transposée du vecteur  $x_{i+h}$  ;
- $\hat{a} = (\hat{a}_0, \hat{a}_1, \hat{a}_2, \dots, \hat{a}_k)$  est le vecteur des paramètres estimés du modèle ;
- $t_{n-k-1}^{\alpha/2}$  est le fractile d'ordre  $\alpha/2$  d'une loi de student de  $(n - k - 1)$  degrés de liberté ;
- $\hat{\sigma}_\varepsilon$  est l'écart type de l'erreur de prévision ;
- $(X'X)$  est le produit de la matrice  $X$  et de sa transposée.

#### 2) Analyse discriminante

L'**analyse discriminante** porte sur les **classements** effectués au sein d'une **population**. Les données étant composées de  $n$  observations réparties en  $k$  classes et décrites par  $p$  variables, on distingue deux aspects en analyse discriminante :

- La **discrimination à but descriptif** : on cherche quelles sont les combinaisons linéaires de variables permettant de séparer le mieux possible les  $k$  classes et en donner une représentation graphique rendant compte au mieux de cette séparation ;
- La **discrimination à but décisionnel** : pour un nouvel individu dont les prédicateurs sont connus, le problème est de déterminer à quelle classe il devrait être affecté.

De même qu'en analyse en composantes principales, on cherche une nouvelle base de  $\mathbb{R}^p$ . Cependant, ces nouveaux caractères ne sont plus de variance maximum mais ceux pour qui :

- Les individus d'une même classe se projettent sur des valeurs voisines,
- Deux individus de classes différentes se projettent sur des valeurs différentes.

### 2.1) Matrices de variances intraclasse et interclasse

Les données étant centrées, la **matrice de variance** du nuage est définie par le produit matriciel suivant :  $V = {}^tXDX$ .

Pour chaque classe d'individus, on peut calculer la matrice de variance  $W_k$  des caractères restreints aux éléments de la classe  $k$ . En notant  $X_k$  la matrice des individus de la classe  $k$ ,  $W_k = {}^tX_kDX_k$ . En associant à chaque classe  $k$  un poids  $P_k$  égal à la somme des poids de chaque individu de la classe ( $P_k = \sum_{x_i \in X_k} p_k$ ), on appelle **matrice intraclasse** la matrice  $W = \sum_{i=1}^q P_k W_k$ .

En notant  $g_k = (g_1^k, \dots, g_p^k)$  le centre de gravité de chaque classe, on définit la **matrice interclasse** par la matrice  $B$  dont les termes génériques sont  $b_{ij} = \sum_{k=1}^q P_k g_k^i g_k^j$ .

On démontre que  $V = \sum_{i=1}^q P_k W_k + B = W + B = {}^tXDX$ .

### 2.2) Variance d'un caractère

Le caractère associé au vecteur unitaire  $u$  de  $\mathbb{R}^p$  est défini par le produit matriciel suivant :  $c = Xu$ . Pour mesurer la norme du caractère  $c$ , on utilise la même **métrique** qu'en analyse en composantes principales : la **matrice diagonale** des poids notée  $D$ .

La norme du caractère  $c$  est alors égale à  $\|c\|$  et  $\|c\|^2 = {}^t c D c = {}^t u {}^t X D X u = {}^t u V u = {}^t u W u + {}^t u B u$ .

La **norme d'un caractère** peut ainsi se décomposer en somme de **deux variances** :

- ${}^t u W u$  **variance intraclasse**, rendant compte de la variation des valeurs de projections sur  $u$  des individus à l'intérieur d'une même classe,
- ${}^t u B u$  **variance interclasse**, rendant compte de la dispersion des projections des centres de gravité des différentes classes sur  $u$ .

Les vecteurs de la base de  $\mathbb{R}^p$  recherchés sont ceux pour lesquels  ${}^t u W u$ , la variance intraclasse est le maximum possible et  ${}^t u B u$ , la variance interclasse est le minimum possible.

### 2.3) Facteurs et caractères discriminants

A partir de l'égalité  ${}^t uVu = {}^t uWu + {}^t uBu$  on déduit que  $\frac{{}^t uVu}{{}^t uVu} = \frac{{}^t uWu}{{}^t uVu} + \frac{{}^t uBu}{{}^t uVu} = 1$ .

Le meilleur caractère possible maximise  $\frac{{}^t uBu}{{}^t uVu}$  tout en minimisant  $\frac{{}^t uWu}{{}^t uVu}$ .

Les **facteurs discriminants** sont les **vecteurs propres** de la matrice  $V^{-1}B$ . De façon similaire à l'analyse en composantes principales, on montre qu'en ordonnant les valeurs propres par ordre décroissant  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ , les vecteurs propres associés forment une base orthonormée de  $\mathbb{R}^p$  **maximisant la discrimination**.

La recherche de la minimisation de  $\frac{{}^t uWu}{{}^t uVu}$  aboutit à chercher les vecteurs propres de la matrice  $W^{-1}B$ . Ces vecteurs propres sont identiques à ceux de  $V^{-1}B$ .

### 2.4) Analyse discriminante décisionnelle

Lorsqu'il y a uniquement deux **classes d'objets**, il n'existe **qu'un seul facteur discriminant  $u$** , donné par l'équation suivante :  $u = V^{-1}(g_2 - g_1) = W^{-1}(g_2 - g_1)$ .

Par la suite, il est question de pouvoir affecter de façon pertinente tout nouvel individu  $x^*$  à une des deux classes. De part l'équation ci-dessus, le nouvel individu doit être affecté à la classe dont le centre de gravité est le plus proche au sens de la métrique  $V^{-1}$ . Cette **métrique** est appelée **distance de Mahalanobis**.

On affecte donc  $x^*$  à la classe 1 dont le centre de gravité est  $g_1$  si  ${}^t(x^* - g_1)V^{-1}(x^* - g_1) < {}^t(x^* - g_2)V^{-1}(x^* - g_2)$ . Ceci revient à se placer de part et d'autre de l'**hyperplan orthogonal** à  $u$  pour la métrique  $V^{-1}$  appelé hyperplan de fisher.

Ce critère se généralise à plus de deux classes et pour chaque nouvel individu  $x^*$ , on l'affecte à la classe  $l$  telle que  $d_{V^{-1}}(x^*, g_l) = \min_{1 \leq i \leq q} d_{V^{-1}}(x^*, g_i)$ .

## 3) Régression logistique

La **régression logistique** a pour but d'expliquer et de prédire les valeurs d'une **variable qualitative**  $Y$ , le plus **souvent binaire**, à partir de **variables explicatives**  $X = (X_1, X_2, \dots, X_p)$  **qualitatives ou quantitatives**. Dans le cadre de variables qualitatives, on utilise les variables indicatrices des différentes modalités.

On note  $\pi(x) = \text{Prob}(Y = 1|x)$  la probabilité que  $Y$  prenne la valeur 1 pour  $X = x$ . La variable aléatoire  $Y$  suit une **loi de Bernoulli** de paramètre  $p = \pi(x)$  pour  $X = x$ . La moyenne de  $\pi(x)$  et sa variance vaut  $\pi(x)[1 - \pi(x)]$ .

En régression logistique, on fait l'hypothèse que la probabilité  $\pi(x)$  s'écrit sous la forme suivante :  $\pi(x) = \text{Prob}(Y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}$

On montre que  $\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ . Cette dernière écriture permet de faire le lien avec la famille des modèles linéaires généralisés.

### 3.1) Interprétation

Le choix de la fonction logistique conduit à une expression comprise entre 0 et 1.

A titre d'exemple, considérons une seule variable explicative binaire :  $x = 1$  si les comptes d'une entreprise contiennent des anomalies significatives,  $x = 0$  sinon et  $Y = 1$  désigne le refus de certification par le commissaire aux comptes, l'événement contraire étant la certification sans réserve.

La probabilité que le commissaire aux comptes refuse la certification des comptes si ceux ci comportent des anomalies significatives est notée :  $\text{Prob}(Y = 1|x = 1) = \pi(x) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$ . La probabilité d'obtenir une certification sans réserve alors que les comptes comportent des anomalies significatives est notée :

$$\text{Prob}(Y = 0|x = 1) = 1 - \pi(x) = 1 - \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} = \frac{1}{1 + \exp(\beta_0 + \beta_1)}$$

**L'odds ou la cote** est le rapport des deux probabilités ci-dessus définies :

$$\text{Odd}(x = 1) = \frac{\text{Prob}(Y=1|x=1)}{\text{Prob}(Y=0|x=1)} = \frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0 + \beta_1).$$

On effectue des calculs similaires pour les entreprises dont les comptes ne comportent pas d'anomalies significatives :

$$x = 0 \text{ et on a : } \begin{cases} \text{Prob}(Y = 1|x = 0) = \pi(x) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \\ \text{Prob}(Y = 0|x = 0) = 1 - \pi(x) = 1 - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} = \frac{1}{1 + \exp(\beta_0)} \end{cases}$$

L'odds des deux probabilités ci-dessus définies est :

$$\text{Odd}(x = 0) = \frac{\text{Prob}(Y=1|x=0)}{\text{Prob}(Y=0|x=0)} = \frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0).$$

**L'odds-ratio** du refus de certification en fonction du fait que les comptes comportent ou non des anomalies significatives est ensuite défini par :  $OR = \frac{Odd(x=1)}{Odd(x=0)} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$ . Ce ratio s'interprète comme suit : les entreprises dont les comptes comportent des anomalies significatives présentent un risque  $\exp(\beta_1)$  fois plus élevé de ne pas voir leurs comptes certifiés que celles dont les comptes ne comportent pas d'anomalies significatives.

### 3.2) Estimation

**L'estimation des paramètres du modèle** s'effectue par la méthode du **maximum de vraisemblance** à partir d'un échantillon prélevé dans la population totale. Cette vraisemblance s'écrit comme suit :

$$L(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

$$= \prod_{i=1}^n \left( \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})} \right)^{y_i} \left( 1 - \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})} \right)^{1-y_i}$$

Dans cette écriture, on a  $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ .

On cherche  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$  maximisant la log-vraisemblance. Formellement, on écrit :  $l(\beta) = \log[L(\beta)] = \log(\prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i})$

$$l(\beta) = \sum_{i=1}^n y_i \log(\pi(x_i)) + (1 - y_i) \log[1 - \pi(x_i)]$$

$$l(\beta) = \sum_{i=1}^n y_i \log \left( \frac{\pi(x_i)}{1 - \pi(x_i)} \right) + \log[1 - \pi(x_i)]$$

En annulant les dérivés par rapport aux  $\beta_j$  de la log-vraisemblance, on aboutit au système d'équation suivant :

$$\begin{cases} \frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^n (y_i - \pi_i) = 0 \\ \frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} (y_i - \pi_i) = 0 \text{ pour } j = 1, \dots, p \end{cases}$$

Ce système n'a **pas de solution analytique** et se résout par des procédures de **calcul numérique**.

### 3.3) Test et sélection des variables

L'apport d'une variable au modèle est testé à partir de l'hypothèse  $H_0: \beta_j = 0$ .

Trois méthodes sont généralement utilisées pour tester l'apport d'une variable au modèle de régression logistique : le **test de Wald**, le **test du rapport de vraisemblances** et le **test du Score**.

Dans le cadre du **test de Wald**, on rejette l'hypothèse  $H_0$  au risque  $\alpha$  de se tromper

si la statistique de Wald vérifie la condition suivante :  $Wald = \frac{\hat{\beta}_j^2}{\text{var}(\hat{\beta}_j)} \geq \chi_{1-\alpha}^2(1)$ .

Le **test du score** consiste à calculer la statistique  $Score = U(\hat{\beta}_{H_0})' [X' \hat{\beta}_{H_0} X]^{-1} U(\hat{\beta}_{H_0})$ .

Sous l'hypothèse  $H_0$ , la statistique  $Score$  suit également asymptotiquement une loi de khi deux à un degré de liberté.

Le **test du rapport de vraisemblances** consiste à calculer la statistique  $G = -2 \log \left( \frac{\text{Vraisemblance sans la variable}}{\text{Vraisemblance avec la variable}} \right)$ . Sous l'hypothèse  $H_0$ , la statistique  $G$  suit asymptotiquement une loi de khi deux à un degré de liberté. L'hypothèse  $H_0$  est rejetée au risque  $\alpha$  lorsque la statistique  $G$  est supérieure au seuil  $\chi_{1-\alpha}^2(1)$ .

Deux autres critères dérivent du test de rapport de vraisemblances :

- Le **critère d'Akaike**  $AIC = -2 \log \mathbb{L}(\beta_j) + 2(j + 1)$ , avec  $j$  représentant le nombre de degrés de liberté d'un modèle ajusté donné ;
- Le **critère de Schwartz** (SC) appelé également Bayesian Information Criterion (BIC),  $SC = BIC = -2 \log \mathbb{L}(\beta_j) + (j + 1) \log(n)$ , avec  $n$  représentant le nombre total d'individus.

Ces deux critères permettent de comparer deux modèles, le meilleur étant celui pour lequel  $AIC$  et  $BIC$  sont les plus bas. Le critère  $BIC$  est préférable quand  $n$  est grand.



L'**appréciation des estimations comptables** est au cœur des **problématiques** rencontrées par le professionnel comptable intervenant dans une compagnie d'assurance. Ces estimations peuvent être l'objet d'**erreurs** et de **fraudes** et induire des **anomalies significatives** dans les comptes examinés.

L'utilisation appropriée des **techniques prédictives** peut aider le professionnel comptable à mettre en œuvre des diligences prévues par la **norme ISA 520** et la **NEP 520** relatives à la **procédure analytique** et de la **norme ISA 540** et la **NEP 540** relatives à l'**appréciation des estimations comptables**.

Ces techniques peuvent être également utilement mises en œuvre pour traiter des questions relatives à la **qualité des comptes annuels produits** et à la **solvabilité** des compagnies d'assurance.

De nombreux **logiciels libres** et commerciaux **facilitent** la mise en œuvre de ces **techniques** relativement **complexes**.

### CHAPITRE III. PRESENTATION DE QUELQUES LOGICIELS DE TRAITEMENTS STATISTIQUES ET DES DONNEES UTILISEES

La réalisation de ce mémoire n'aurait pas été possible sans l'utilisation d'un **puissant logiciel de statistiques** facilitant la mise en œuvre de méthodes relativement complexes. Un panorama des logiciels de statistiques et d'analyse de données est donc fourni pour permettre au lecteur intéressé de s'orienter et éventuellement de faire un choix. Une distinction est faite à ce titre entre les **logiciels commerciaux** et les **logiciels libres**. Un accent particulier est mis sur le **logiciel**  utilisé dans le cadre des travaux du présent mémoire.

En fin de chapitre, les **données** extraites des comptes annuels et des états réglementaires des sociétés d'assurance et objet de l'étude sont présentées.

#### 1) Logiciels de traitements statistiques

Le site de référence des ressources sur les **logiciels de statistique et de datamining**, [www.kdnuggets.com](http://www.kdnuggets.com), listait à la date de finalisation de ce mémoire 75 logiciels commerciaux de traitements statistiques et de datamining. Les logiciels libres de datamining référencés sont moins nombreux, on en dénombre 17.

L'examen des **critères** suivants est généralement recommandé pour comparer et choisir un logiciel de data mining :

- la **variété des techniques** de data mining et de préparation des données ;
- la **qualité des algorithmes** de data mining implémentés ;
- la **puissance de calcul** et la capacité à traiter de grands volumes ;

- le **type de données** gérées ;
- la **convivialité** du logiciel et ses aptitudes à produire des rapports synthétisant les opérations demandées et les résultats obtenus.

Ces critères doivent cependant être examinés à la lumière des **besoins** de l'utilisateur et de la variable « **prix** », ce d'autant plus que de nombreux logiciels libres n'ont rien à envier aux logiciels commerciaux.

### 1.1) Logiciels commerciaux

Le « Tableau 40: Logiciels commerciaux » en annexe fourni une **cartographie** des principaux logiciels commerciaux de statistique et de datamining. Le critère discriminant utilisé dans le cadre de cette cartographie est la capacité du logiciel à traiter de **grands volumes**.

On constate une prédominance des éditeurs anglo-saxons dans le domaine. Cependant, des éditeurs comme Complex Systems et Netral sont français. Certains éditeurs sont très spécialisés dans une technique statistique ou de datamining particulière. Les **leaders** du marché sur les gros systèmes sont les logiciels de **SAS** et **SPSS**. Ces éditeurs offrent une palette très large et très complète d'outils. En contrepartie, les prix affichés sont élevés.

### 1.2) Logiciels libres

Une liste de **logiciels libres** de datamining et de statistique est fournie en annexe (Tableau 41: Logiciels libres) avec les sites web dédiés à ces logiciels (documentation, téléchargement...). Ces logiciels sont référencés sur le site [www.kdnuggets.com](http://www.kdnuggets.com).

Nombre de ces logiciels ont **dépassé la phase d'expérimentation** et peuvent être déployés sans risque pour la réalisation d'études statistiques complexes.

### 1.3) Logiciel et bibliothèques utilisées

De nombreux logiciels libres permettent de faire de l'analyse de données et des traitements statistiques. Cependant, le **logiciel**  est devenu en une dizaine d'années d'existence, un **outil incontournable** de statistique et de visualisation des données. Son développement exceptionnel s'explique par ses trois principales qualités : il est **gratuitement** téléchargeable sur internet, très **complet** et en **essor permanent**.

Logiciel libre basé sur la coopération de nombreux spécialistes du monde entier, ce **langage** propose des **instructions de très haut niveau** orientées vers le calcul **mathématique et statistique**.

A la date de finalisation de ce mémoire à fin août 2012, la dernière version publiée le 22 juin 2012 est la version 2.15.1. Elle est disponible dans les archives du réseau **CRAN** (Comprehensive  Archive Network) dont les sites de base sont : <http://lib.stat.cmu.edu/R/CRAN/>, <http://stat.ethz.ch/CRAN/>.

D'autres sites sont associés à des **universités françaises** et au **CNRS** sur lesquels la dernière version du logiciel peut également être téléchargée. Ces sites sont les suivants : <http://cran.cict.fr/>, <http://cran.univ-lyon1.fr/>, <http://mirror.ibcp.fr/pub/CRAN/>.

Plusieurs universités proposent sur leurs sites des **cours d'initiation** au logiciel . Des cours sont notamment disponibles sur le site <http://pbil.univ-lyon1.fr>. D'autres informations peuvent être obtenues sur le site du projet qui est **www.r-project.org**.

Chaque version de base du logiciel  permet de disposer de nombreux **packages**. Cependant, certains packages doivent être téléchargés sur les sites du CRAN pour effectuer des traitements spécifiques.

Plusieurs packages différents permettent de faire de l'analyse de données et peuvent être téléchargés à partir des sites internet présentés ci dessus. Un **package** peut s'analyser comme une **bibliothèque de fonctions préprogrammées**.

Le package utilisé dans le cadre de l'analyse de données est **FactoMineR**. Il doit être installé une seule fois puis chargé à chaque utilisation. Les **méthodes classiques d'analyse de données y sont programmées** : Analyse en Composantes Principales (fonction PCA), Analyse Factorielle des Correspondances (fonction CA), Analyse Factorielle des Correspondances Multiples (fonction MCA) et construction ascendante d'une hiérarchie (fonction HCPC).

L'**analyse discriminante** est réalisée à partir de la fonction `lda` du package **MASS**. Ce package est installé par défaut dans . Il faut donc simplement le charger pour avoir accès aux fonctions nécessaires à la réalisation de l'analyse discriminante.

## 2) Présentation et reformulation des données

Les données objet de l'étude sont extraites des **états réglementaires** produits par les compagnies d'assurance à l'**autorité de supervision** du secteur des assurances de l'espace CIMA. Une distinction est effectuée entre les données des compagnies d'**assurance non vie** et celle des compagnies d'**assurance vie** pour tenir compte de certaines **variables non communes** entre ces deux catégories de sociétés.

Par convention, les éléments du passif et les charges sont des **grandeurs négatives** affectées d'un signe moins (-). Les éléments de l'actif et les produits sont des **grandeurs positives**.

Dans le tableau des données, les compagnies ont été classées par **ordre décroissant** du **chiffre d'affaires**. Pour tout couple  $(i, j)$  d'entiers naturels tel que  $(i < j)$ , le chiffre d'affaires de la compagnie  $A_i$  est supérieur au chiffre d'affaires de la compagnie  $A_j$ .

### 2.1) Présentation des données des compagnies non vie

Les données sont relatives à **90 compagnies d'assurance non vie** de l'espace CIMA. Elles sont extraites des bilans, des comptes de résultat, des tableaux de marge de solvabilité et de couverture des engagements réglementés de l'exercice 2009. **Dix neuf variables** ont été retenues. Le « Tableau 42: Variables relatives aux données des compagnies d'assurance non vie » en **annexe** permet une présentation desdites variables.

Le **tableau** comportant l'ensemble des données mesurées sur les compagnies relativement à ces variables est également joint en **annexe** (Tableau 44: Données sociétés non vie - tableau A et suivants).

Des **statistiques descriptives** en annexe (Tableau 48: Données sociétés non vie - données résumées) fournissent des **indications premières** (moyenne, médiane, quartile, minimum et maximum) sur la nature des données relatives aux sociétés d'assurance non vie.

### 2.2) Présentation des données des compagnies d'assurance vie

Les données sont relatives à **45 compagnies d'assurance vie** de l'espace CIMA. Elles sont extraites des bilans, des comptes de résultat, des tableaux de marge de solvabilité et de couverture des engagements réglementés de l'exercice 2009. **Vingt**

**deux variables** ont été retenues. Le « Tableau 43: Variables relatives aux données des compagnies d'assurance vie » en **annexe** permet une présentation desdites variables.

Le **tableau** comportant l'ensemble des données mesurées sur les compagnies relativement à ces variables est également joint en **annexe** (Tableau 49: Données sociétés vie - tableau A et suivant)

Des **statistiques descriptives** en **annexe** (Tableau 51: Données sociétés vie - données résumées) fournissent des indications premières (moyenne, médiane, quartile, minimum et maximum) sur la nature des données relatives aux sociétés d'assurance non vie.

### 2.3) Reformulation des données

Dans la plupart des **logiciels** permettant de faire de l'**analyse de données**, l'option par défaut consiste à **centrer et réduire les données**. Le **centrage** consiste à soustraire de chaque variable sa moyenne. La **réduction** consiste à diviser chaque variable par son écart type. On obtient ainsi des variables dont la moyenne est nulle et l'écart type égal à l'unité.

Cette option permet d'accorder la **même importance** à chaque **variable** quelque soit sa **dispersion**. Elle permet également de résoudre le problème pouvant résulter du fait que les variables sont exprimées dans des unités différentes. La réduction permet d'obtenir de nouvelles **variables sans unité**.

Toutes les données quantitatives utilisées dans le cadre de ce mémoire sont exprimées dans la **même unité monétaire** (en Francs de la Communauté Financière Africaine FCFA). Dans le cadre de l'analyse en composantes principales, les données sont réduites pour ne pas accorder plus d'importance aux variables présentant de **fortes variances**.

Pour permettre d'effectuer les **traitements relatifs à la régression logistique** et à l'**analyse discriminante**, deux variables supplémentaires sont prises en comptes et fournies en annexe (Tableau 52: Données sociétés non vie - ajustements et statuts et Tableau 53: Données sociétés vie - ajustements et statuts ). La variable « **Ajustement** » mesure le caractère significatif ou non des ajustements réalisés sur les comptes d'une compagnie d'assurance à la suite d'une mission de contrôle de l'autorité de régulation. La variable « **STATUT** » présente trois modalités : « **sain** », « **suivi** » et « **faillite** ». Ces trois modalités permettent de décrire respectivement les entreprises

d'assurance saines, puis celles faisant l'objet d'un suivi particulier du fait de difficultés rencontrées et enfin celles dont le retrait d'agrément est proposé à la Commission Régionale de Contrôle des Assurances.



L'**omniprésence du tableur Excel** ne favorise pas le recours systématique à d'autres **outils plus adaptés** dans le cadre des **études statistiques** et de la mise en œuvre des **procédures analytiques**. Des outils de très **bonne qualité** existent et sont parfois **gratuits**. Ils méritent une meilleure diffusion, notamment au sein de la profession comptable.

Toute étude statistique passe par l'**inventaire**, la **collecte** et la **préparation** des données. Dans le cadre des données récoltées sur les compagnies d'assurance de la zone CIMA, cette phase liminaire relativement longue a permis de mieux **calibrer** certains **objectifs de l'étude**.



La partie introductive du mémoire a permis de présenter des **méthodes classiques** et couramment mises en œuvre pour l'**exploration et l'analyse des données « multivariées »**. L'objectif de ces **méthodes descriptives ou prédictives** est **d'extraire des informations et des connaissances enfouies sous le volume des données** pour une **aide à la décision**. Une **présentation synoptique** de ces **méthodes** est fournie dans le tableau suivant :

Type	Famille	Nom	Nature des variables
Méthodes descriptives	Analyse factorielle	Analyse en composantes principales (ACP)	p variables quantitatives
		Analyse canonique	p variables quantitatives ou qualitatives et q variables quantitatives ou qualitatives
		Analyse factorielle des correspondances (AFC)	deux variables qualitatives et binaires
		Analyse des correspondances multiples (ACM)	p variables qualitatives et binaires
	Classification	Méthodes de partitionnement	p variables quantitatives ou qualitatives
		Méthodes hiérarchiques	p variables quantitatives ou qualitatives
Méthodes prédictives	Modèles paramétriques	Régression linéaire	p variables explicatives quantitatives et 1 variable à expliquer quantitative
		Analyse discriminante	p variables explicatives quantitatives et 1 variable à expliquer nominale ou binaire
		Régression logistique	p variables explicatives quantitatives et 1 variable à expliquer nominale ou binaire

**Tableau 4: Synthèse des techniques d'exploration et d'analyse de données**

Dans un contexte de **foisonnement des données et des informations**, la **maîtrise** de ces **méthodes** et de leur mise en œuvre est source de **valeur ajoutée** et d'**enrichissement** pour le professionnel comptable.

Cette **valeur ajoutée et cet enrichissement** sont mis en évidence dans les parties suivantes du mémoire à travers des données relatives aux compagnies d'assurance de la zone CIMA et l'utilisation du logiciel .

**DEUXIEME PARTIE: ANALYSE DES  
DONNEES COMPTABLES ET  
FINANCIERES DES COMPAGNIES  
D'ASSURANCE DE L'ESPACE CIMA**

## DEUXIEME PARTIE: ANALYSE DES DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES D'ASSURANCE

Dans cette **deuxième partie**, deux méthodes statistiques d'exploration et de description des données seront mises en œuvre: l'**analyse en composantes principales** et la **classification ascendante hiérarchique**. Le choix de ces méthodes résulte de la nature et de la structure des **données** étudiées qui sont **quantitatives**.

Pour permettre d'**exposer** et d'**explorer différents aspects**, ces deux méthodes sont mises en œuvre respectivement sur :

- **les données des compagnies d'assurance non vie ;**
- **les données des compagnies d'assurance vie ;**
- **et les données d'un groupe de sociétés d'assurance composé à la fois de sociétés non vie et de société vie.**

La distinction entre compagnies d'assurance non vie et vie résulte de la **spécificité** de ces **métiers** conduisant à la mesure de **variables non identiques** sur ces deux groupes.

L'étude d'un groupe de sociétés d'assurance permet l'application des méthodes sur une **structure de données proche en termes de taille** de celle avec laquelle le professionnel comptable pourrait être le plus **familier** : quelques entreprises d'un secteur donné.

Cette deuxième partie du mémoire s'appuie essentiellement sur l'ouvrage suivant :

HUSSON François, LE Sébastien, PAGES Jérôme, Analyse de données avec R, Presses Universitaires de Rennes (PUR), 2009, 224 p.

Ce livre extrêmement pédagogique explique les méthodes de l'analyse de données et la mise en œuvre avec le package FactoMineR de  développé par les auteurs.

## CHAPITRE I. REALISATION ET INTERPRETATION D'UNE ANALYSE EN COMPOSANTES PRINCIPALES (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE (CAH) SUR LES DONNEES DES COMPAGNIES D'ASSURANCE NON VIE DU MARCHE

L'analyse en composantes principales est réalisée sur l'ensemble des 90 compagnies d'assurance non vie exerçant dans l'espace CIMA. Dans ce cadre, l'étude du **nuage des individus** est suivie par celle du **nuage des variables**.

L'analyse en composantes principales est suivie d'une **classification ascendante hiérarchique** permettant d'avoir un **autre point de vue** et confirmant éventuellement les résultats de l'ACP.

### 1) Etude du nuage des individus à travers une ACP

L'étude des individus commence par une **projection et une représentation sur le plan principal**. L'analyse du **pourcentage d'inertie** associé à chaque axe permet de déterminer les dimensions et les axes intéressants pour l'étude. Enfin, **trois paramètres quantitatifs** calculés pour chaque individu permettent de conforter les analyses effectuées à partir de la représentation graphique.

#### 1.1) Représentation du nuage des individus

Les **informations initiales** sur les compagnies d'assurance non vie sont dans un fichier d'extension csv dénommé « **donnees\_non\_vie.csv** ».

Les **deux commandes** suivantes permettent dans un premier temps de **charger le package FactoMineR** puis d'**importer sur la console**  **les données** en précisant que le nom des sociétés est dans la première colonne.

```
> library (FactoMineR)
> analyse_donnees_non_vie<-read.table ("donnees_non_vie.csv", header=TRUE, sep=';', dec='.',
row.name=1)
```

Ensuite, on utilise la commande suivante :

```
> res.pca<-PCA(analyse_donnees_non_vie)
```

Cette **commande exécute l'ACP** et fournit le **graphe des individus** et le **graphe des variables**.

Les **90 sociétés d'assurance** qu'on ne pouvait pas représenter et visualiser dans un **espace à 19 dimensions** (19 variables), sont maintenant **représentées** dans un

**plan** (deux dimensions). Ce plan contenant 72,77% de l'information initiale, dont 62,45% expliqués par le **premier axe principal**, se présente comme suit :

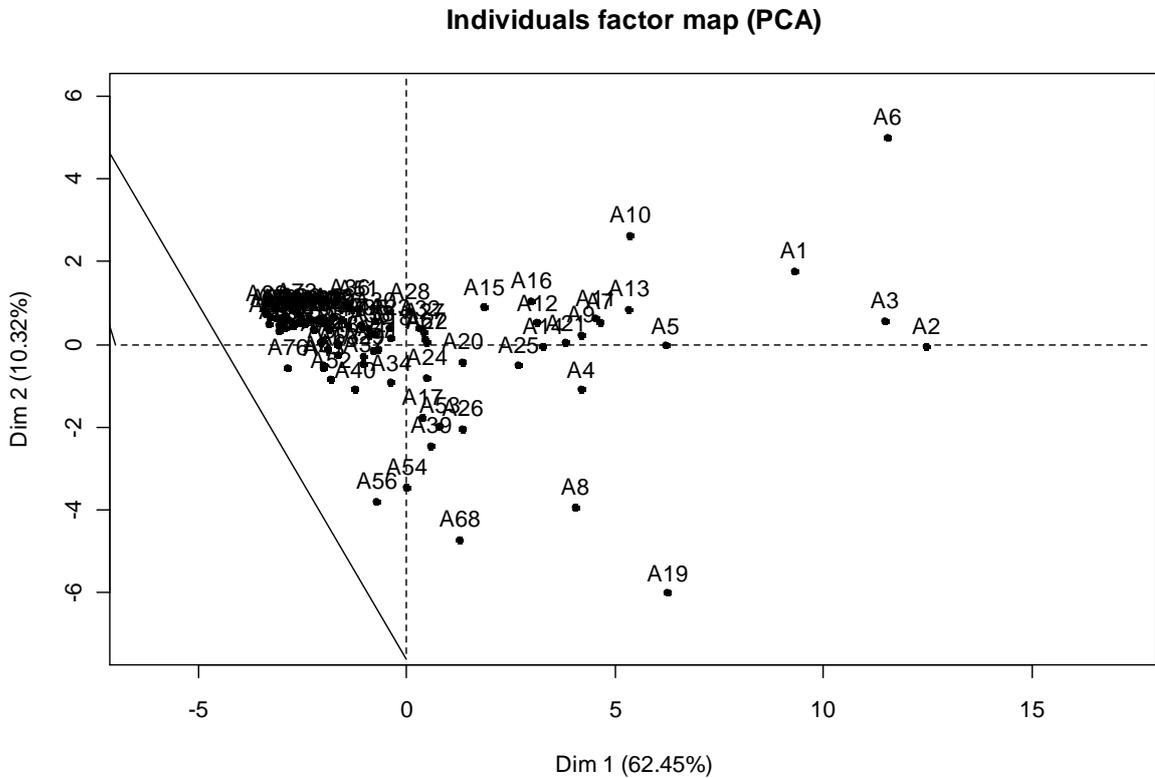


Figure 1 : Données des sociétés non vie - nuage des individus

Cette **représentation** permet d'avoir le **positionnement** de chaque compagnie d'assurance sur le **marché** par rapport aux autres, simplement à partir de **deux variables de synthèse**. Un **changement de positionnement** d'une compagnie d'un exercice à l'autre résulte forcément de **variations significatives ou de tendances inattendues**. Dans le cadre de la **surveillance prudentielle du marché des assurances**, ces changements de positionnement peuvent être des **signaux d'alertes pour l'autorité de contrôle**.

Dans le cadre de l'**analyse d'un secteur d'activité donnée** (prise de connaissance de l'entité et de son environnement) par un professionnel comptable (expert comptable ou commissaire aux comptes) et en conformité avec la **norme ISA 315 ou la NEP 315**, une ACP peut donc être utilisée pour voir le positionnement d'une entreprise dans son secteur d'activité. La **NEP 315§18** précise que dans le cadre de la phase de prise de connaissance de l'entité et d'évaluation du risque d'anomalies

significatives dans les comptes, le commissaire aux comptes peut mettre en œuvre des procédures analytiques. Ces procédures peuvent permettre au commissaire aux comptes **d'identifier des opérations ou événements inhabituels**.

L'analyse de la forme du nuage formé par les compagnies dans ce plan principal peut se faire dans un premier temps par la **détection des compagnies remarquables**. Une compagnie peut être considérée comme remarquable si elle prend des **valeurs extrêmes** sur plusieurs variables.

Le **premier axe** oppose donc des compagnies ayant une **activité importante** à des compagnies ayant un **faible niveau d'activité**.

Le **second axe** oppose les compagnies ayant une **maîtrise de leurs arriérés** de primes aux compagnies **fragilisées** par un **niveau élevé d'arriérés de primes**.

La **détection des individus remarquables** peut se faire en examinant leur **éloignement** par rapport au **centre de gravité** du nuage formé.

Les **20 compagnies** les plus éloignées du centre de gravité sont présentées dans le tableau suivant :

A6	A2	A3	A1	A19	A10	A5	A8	A68	A13	A14	A11	A21	A4	A43	A67	A56	A7	A9	A54
13,7	13,4	13,1	11,4	9,8	7,3	7,3	6,7	6,6	6,1	5,8	5,4	5,2	5,2	5,1	5,0	5,0	5,0	4,8	4,3

Tableau 5: Données sociétés non vie - éloignement du centre de gravité

Les distances étant toujours positives, deux compagnies peuvent être à une **distance équivalente** du centre de gravité et être en même temps à **des extrémités opposées**.

La **compagnie A6** est la plus éloignée du centre de gravité avec des **indicateurs de gestion positifs**. L'éloignement de la **compagnie A68** du centre de gravité se traduit par contre par des **indicateurs de gestion négatifs** (faiblesse du chiffre d'affaires et niveau élevé des arriérés de primes.)

La **représentation synoptique** du nuage formé par les compagnies d'assurance permet de constater que le marché des assurances est marqué par :

- La **présence de quelques grands acteurs** (A6, A1, A3, A2). Ces grands acteurs sont caractérisés par des **produits et des actifs importants** et une

**maîtrise de leur niveau d'arriérés de primes.** En termes de chiffre d'affaires, la compagnie A6 occupe le 6<sup>ème</sup> rang sur le marché de la CIMA. Cette position permet de constater que le classement des compagnies à partir de la seule variable chiffre d'affaires, présente des limites. Cette compagnie est plus **efficente** que les autres compagnies ayant un chiffre d'affaires plus élevé ou similaire par la **maîtrise de ses arriérés**. Elle présente également les **résultats d'exploitation** les plus **élevés** de la zone.

- La **présence d'acteurs de taille moyenne** caractérisés soit par une **maîtrise des arriérés de primes** (A10, A13), soit par un **niveau des arriérés particulièrement élevé** (A8, A19).
- La **présence de nombreux petits acteurs** formant un groupe compact, présentant une certaine homogénéité. Dans ce groupe, on constate que les compagnies n'ont pas la **taille critique** devant leur permettre une meilleure **mutualisation des risques souscrits**. Certaines compagnies de ce groupe présentent également un niveau d'arriérés particulièrement élevé (A17, A53, A39, A54, A56) qui aggrave leur fragilité.

La **non maîtrise des charges** (frais généraux et charges techniques) et des **arriérés de primes** constituent des facteurs de vulnérabilité pour les entreprises d'assurance. De nombreuses entreprises d'assurance du marché sont caractérisées par ces facteurs de vulnérabilité. Ces **facteurs de vulnérabilité influent négativement** sur la **solvabilité des entreprises d'assurance** et sur leur capacité à faire face à leurs engagements envers les assurés et bénéficiaires de contrats.

## 1.2) Pourcentage d'inertie associé à chaque axe et choix des dimensions à étudier

La commande suivante :

```
> round(res.pca$eig,2)
```

permet d'obtenir un tableau contenant la **valeur propre** associée à chaque axe (Eigen value), le **pourcentage d'inertie** (Percentage of variance) qu'elle représente dans l'analyse ainsi que le cumul de ces pourcentages (Cumulative percentage of variance).

Un extrait de ce tableau se présente comme suit :

Libellés	Eigenvalue	Percentage of variance	Cumulative percentage of variance
comp1	11,87	62,45	62,45
comp2	1,96	10,32	72,77
comp3	1,27	6,69	79,46
comp4	0,91	4,80	84,25
comp5	0,71	3,74	87,99
comp6	0,53	2,78	90,77

**Tableau 6: Données sociétés non vie - six premières valeurs propres et pourcentage d'inertie**

Le **tableau** comprenant la totalité des valeurs propres et des pourcentages d'inertie associés est fourni en **annexe** (Tableau 54: Données sociétés non vie - valeurs propres et pourcentage d'inertie).

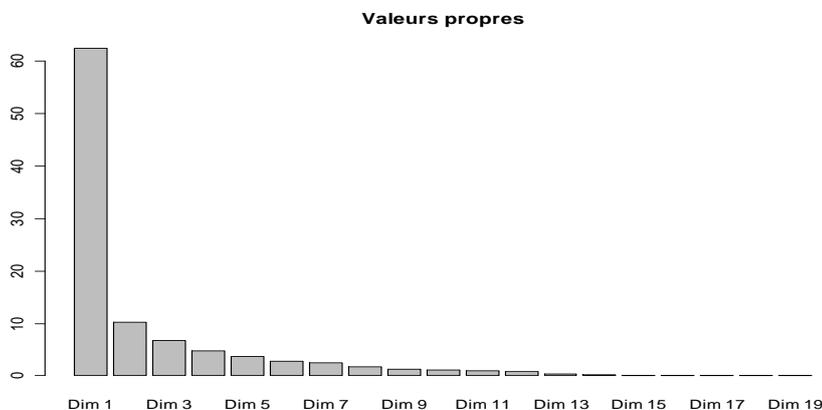
Le 1<sup>er</sup> axe exprime 62,45% de la variabilité des données contre 10,32% pour le 2<sup>ème</sup> axe. Le 1<sup>er</sup> axe exprime donc 3 fois plus de variabilité que le 2<sup>ème</sup> axe.

Généralement, on considère que la représentation est bonne si le plan principal contient plus de 80% de l'information. Le **plan principal contient 72,77% de l'information, ce qui est satisfaisant.**

Le 3<sup>ème</sup> et le 4<sup>ème</sup> axe n'expriment que 6,69% et 4,8% d'inertie. Toutefois il peut être prudent de les analyser pour s'assurer que leur interprétation ne présente pas d'intérêt particulier.

La commande suivante permet une **représentation en diagramme en barres** des différentes valeurs propres obtenues à partir de la matrice des données reformulées et notamment la matrice de corrélation (Tableau 57 : Données sociétés non vie - matrice de corrélation ).

```
> barplot(res.pca$eig[,2], main = "Valeurs propres", name=paste("Dim", 1:nrow(res.pca$eig)))
```



**Figure 2: Données sociétés non vie - éboulis des valeurs propres**

Cette représentation permet de constater une **chute importante d'inertie** entre la 1<sup>ère</sup> et la 2<sup>ème</sup> valeur propre. On constate également une chute relativement importante d'inertie entre la 2<sup>ème</sup> et la 3<sup>ème</sup> valeur propre. La 2<sup>ème</sup> valeur propre représente une fois et demie plus d'inertie que la 3<sup>ème</sup>. Après les deux premiers axes, on observe une décroissance régulière des inerties. Ces chutes d'inertie suggèrent de se **limiter aux deux premiers axes**. Cependant l'analyse du nuage des variables permettra de constater que la variable «**résultat hors exploitation**» est particulièrement **mal représentée** sur les deux premiers axes. Il peut donc être intéressant de rechercher sur l'axe 3 et éventuellement l'axe 4, les raisons de cette mauvaise représentation.

### 1.3) Qualité de la représentation des individus et contribution à la construction des axes

La commande suivante :

> round(cbind(res.pca\$ind\$coord[,1:4], res.pca\$ind\$cos2[,1:4], res.pca\$ind\$contrib[,1:4]),2)  
 permet d'avoir dans un **tableau concaténé sur les 4 premiers axes principaux** représentant 84,25% de l'inertie total :

- Les **coordonnées des compagnies** sur les nouveaux axes construits,
- Le **cosinus carré** de l'angle formé par chaque compagnie avec sa projection sur chacun des axes principaux,
- La **contribution** de chaque compagnie à la construction de chacun des axes.

Un extrait de ce tableau pour les 10 premières entreprises du marché en termes de chiffre d'affaires se présente comme suit :

	Coordonnés				Cos <sup>2</sup>				Contribution			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
A1	9,32	1,77	-1,86	-5,58	0,67	0,02	0,03	0,24	8,13	1,78	3,02	37,97
A2	12,45	-0,05	0,97	-0,79	0,86	0,00	0,01	0,00	14,52	0,00	0,82	0,76
A3	11,47	0,58	4,10	-1,53	0,77	0,00	0,10	0,01	12,33	0,19	14,72	2,85
A4	4,20	-1,09	0,34	-1,58	0,66	0,04	0,00	0,09	1,65	0,68	0,10	3,05
A5	6,22	-0,02	-2,14	0,99	0,74	0,00	0,09	0,02	3,63	0,00	4,02	1,19
A6	11,53	5,01	0,71	2,77	0,71	0,13	0,00	0,04	12,45	14,21	0,44	9,39
A7	4,64	0,53	-0,26	0,21	0,88	0,01	0,00	0,00	2,01	0,16	0,06	0,06
A8	4,06	-3,93	0,95	-0,27	0,37	0,34	0,02	0,00	1,54	8,75	0,80	0,09
A9	4,18	0,24	-1,03	-0,66	0,77	0,00	0,05	0,02	1,63	0,03	0,92	0,54
A10	5,36	2,63	-1,13	1,49	0,53	0,13	0,02	0,04	2,69	3,94	1,12	2,70

Tableau 7: Données sociétés non vie - indicateurs pour 10 sociétés sur les quatre premiers axes - extrait -

Les informations concernant toutes les compagnies sont consignées dans un **tableau en annexe** (Tableau 55: Données sociétés non vie - coordonnées et qualité de représentation des sociétés - tableau A et suivant).

Les **individus remarquables** influent sur l'analyse et il est intéressant d'examiner leur influence sur la **construction des axes**. Par ailleurs, certains individus peuvent influencer sur la construction de certains axes sans pour autant être des individus remarquables. La détection des individus qui contribuent à la construction d'un axe factoriel permet d'évaluer la **stabilité des axes**.

Les **compagnies** contribuant le plus significativement à la **construction du 1<sup>er</sup> axe** principal sont A2 avec 14,52%, A6 avec 12,45%, A3 avec 12,33% et A1 avec 8,13%. Au total, sur 90 sociétés, 4 contribuent à la construction du 1<sup>er</sup> axe principal à 47,43%. Si toutes les compagnies contribuaient de façon égalitaire à la construction de chaque axe, le pourcentage cumulé pour 4 compagnies serait de 4,4% (4/90).

Certaines compagnies réalisant un chiffre d'affaires important (A4 avec 1,65%, A5 avec 3,63%) ne contribuent que moyennement à la construction du 1<sup>er</sup> axe principal. Ce constat conforte le fait que le **chiffre d'affaires ne peut être la seule variable discriminante des compagnies**. Le chiffre d'affaires ne contribue qu'à hauteur de 7,08% à la construction du 1<sup>er</sup> axe.

Les **compagnies** contribuant le plus significativement à la **construction du 2<sup>ème</sup> axe** sont essentiellement **celles présentant un niveau d'arriérés de primes très élevé** parfois supérieur au chiffre d'affaires annuel. Il s'agit notamment des compagnies A19 avec 20,42%, A68 avec 12,65% et A8 avec 8,75%.

La compagnie A6 réalisant les meilleures performances en matière de gestion des arriérés contribue également significativement à hauteur de 14,21% à la construction du 2<sup>ème</sup> axe. **Elle est donc à l'opposé des compagnies A19, A68 et A8.**

**Quatre compagnies** ayant réalisé des **résultats hors exploitation** importants contribuent de façon significative à la **construction de l'axe 3**. Il s'agit des compagnies A3 avec une contribution de 14,72%, A14 avec une contribution de 16,25%, A43 avec une contribution de 11,31% et A67 avec une contribution de 12,53%. Au total, ces quatre compagnies contribuent à 54, 81% à la construction de l'axe 3.

La compagnie A1 contribue seule à 37,97% à la construction du 4<sup>ème</sup> axe. La 2<sup>ème</sup> compagnie contribuant à la construction de cet axe est A6 avec un taux de 9,39%. L'**interprétation** de cet axe regroupant 4,8% de l'inertie est **peu aisée**. Aucune variable de base n'est véritablement corrélée à cet axe. Les coefficients de corrélation les plus élevés enregistrés par rapport aux variables de base oscillent entre 0,3 et 0,4.

La lecture des coordonnées des compagnies permet de constater que **58 compagnies** sur **90** (soit 64%) présentent des coordonnées négatives sur le 1<sup>er</sup> axe principal. Toutes ces compagnies présentent des **signes importants de fragilité**. A ces compagnies, il convient d'ajouter **15 compagnies** présentant des coordonnées positives sur le 1<sup>er</sup> axe principal mais des coordonnées négatives sur le 2<sup>ème</sup> axe principal. Ces compagnies sont **fragilisées** par leur **niveau d'arriérés**.

Dans le cadre d'une ACP, l'étude de l'espace des individus doit être effectuée concomitamment à celle des variables.

## **2) Etude du nuage des variables à travers une ACP**

L'**étude des variables** commence par une **représentation dans un cercle de corrélation**. Ensuite, **trois paramètres quantitatifs** calculés pour chaque variable permettent de **conforter les analyses effectuées** à partir de la représentation visuelle dans le cercle de corrélation.

### **2.1) Représentation des variables**

La représentation du nuage des variables permet de visualiser rapidement les corrélations entre variables.

La représentation ci dessous permet de voir comment ces **19 variables quantitatives** sont **corrélées** aux **deux variables de synthèse** qui ont été retenues pour la représentation des compagnies d'assurance dans le plan principal.

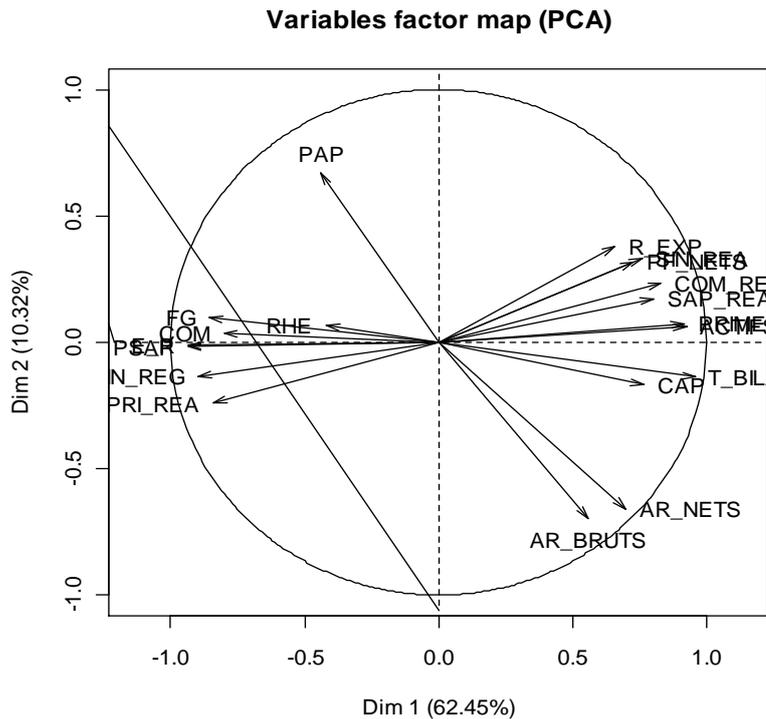


Figure 3: Données sociétés non vie - graphe des variables

La représentation met en évidence une **corrélation positive** entre la **1<sup>ère</sup> composante principale et les produits et les actifs**. Elle met en évidence également, une **corrélation négative** entre la **1<sup>ère</sup> composante principale et les charges et les passifs**.

Les **arriérés bruts**, les **arriérés nets** et la **provision pour annulations de primes** sont **peu corrélés à la 1<sup>ère</sup> composante principale**. Ces trois variables sont plutôt corrélées à la **2<sup>ème</sup> composante principale**. La provision pour annulations de primes est corrélée positivement avec la 2<sup>ème</sup> composante principale alors que les arriérés sont corrélés négativement.

Toutes les variables sont bien représentées dans le plan principal à l'exception de la variable « **résultat hors exploitation** ». La **qualité de cette représentation** est due grâce à la **proximité entre l'extrémité de la flèche et le cercle de rayon 1**. La longueur de la flèche représentant le résultat hors exploitation est à peu près égale à 1/3 du rayon du cercle.

Au delà de **l'analyse visuelle**, les coordonnées des variables, leur qualité de représentation et leur contribution à la construction des axes peuvent également être examinées à partir de **données chiffrées**.

## 2.2) Analyse de la représentation des variables

L'étude du nuage des variables se fait par l'analyse de la qualité de leur représentation et de leur contribution à la formation des nouveaux axes.

La commande suivante:

```
> round(cbind(res.pca$var$coord[,1:4], res.pca$var$cos2[,1:4], res.pca$var$contrib[,1:4]),2)
```

permet d'avoir dans un tableau concaténé sur les 4 premiers axes principaux regroupant 84,25% de l'inertie total :

- Le **coefficient de corrélation** de chaque variable initiale sur les nouveaux axes construits,
- Le **cosinus carré** de l'angle formé par chaque variable avec sa projection sur chacun des axes principaux,
- La **contribution** de chaque variable à la construction de chacun des axes.

Un extrait du tableau ainsi obtenu se présente comme suit :

Libellés	Corrélation des variables p/r axes				Corrélation des variables p/r aux axes (cos <sup>2</sup> )				Contribution des variables à la construction des axes			
	Dim 1	Dim 2	Dim 3	Dim 4	Dim 1	Dim 2	Dim 3	Dim 4	Dim 1	Dim 2	Dim 3	Dim 4
AR_BRUTS	0,56	-0,70	-0,05	0,07	0,31	0,49	0,00	0,00	2,63	<b>24,78</b>	0,23	0,52
AR_NETS	0,70	-0,66	0,10	-0,03	0,49	0,43	0,01	0,00	4,11	<b>22,17</b>	0,76	0,07
PAP	-0,44	0,67	-0,13	0,05	0,19	0,45	0,02	0,00	1,63	<b>22,87</b>	1,24	0,28
R_EXP	0,66	0,38	0,48	0,09	0,43	0,14	0,23	0,01	3,65	<b>7,38</b>	<b>18,06</b>	0,93
RHE	-0,42	0,07	-0,83	0,09	0,18	0,00	0,69	0,01	1,48	0,24	<b>54,47</b>	0,94
Total									100	100	100	100

Tableau 8: Données sociétés non vie - cordonnées et qualité de représentation des variables - extrait-

Les informations relatives à toutes les variables initiales sont consignées dans un **tableau en annexe** (Tableau 58: Données sociétés non vie - cordonnées et qualité de représentation des variables).

La plupart **des variables sont corrélées** avec le premier axe principal avec un coefficient strictement supérieur à 0,7 à l'exception des variables « arriérés bruts (0,56) », « Arriérés nets (0,7) », « PAP (0,44) », « Résultat d'exploitation (0,66) » et « résultat hors exploitation (-0,42) ».

Le **premier axe principal synthétise** donc une **part essentielle** des informations fournies par les **19 variables de base**.

On constate également que toutes les **variables fortement corrélées au 1<sup>er</sup> axe principal** contribuent dans des proportions voisines à sa construction (entre 4 et 8% pour chaque variable).

A contrario, très **peu de variables sont corrélées avec le second axe principal**. Seuls les arriérés bruts (-0,7), les arriérés nets (-0,66) et la PAP (0,67) ont des coefficients de corrélation relativement élevés avec le second axe principal. Le second axe traduit donc essentiellement la **qualité de gestion de la politique de recouvrement et de gestion des arriérés de primes**.

Cette observation est confirmée par le fait que ces trois variables contribuent à hauteur de 69,82% (24,78+22,17+22,87) à la **construction de l'axe 2**.

La **faible corrélation entre le résultat d'exploitation et le résultat hors exploitation** semble anormale. Normalement, si le résultat hors exploitation se résume à l'impôt sur société assis sur le résultat d'exploitation, les deux variables devraient être parfaitement corrélés. Ces deux variables ont cependant un coefficient de corrélation relativement faible de -0,56 (voir Tableau 57 : Données sociétés non vie - matrice de corrélation). Le résultat d'exploitation est corrélé avec le 1<sup>er</sup> axe principal avec un coefficient de 0,66 contre -0,42 pour le résultat net. Par contre, le résultat hors exploitation est corrélé à 0,83 avec le 3<sup>ème</sup> axe principal. Cette variable contribue à 54,47% à la construction du 3<sup>ème</sup> axe.

**L'axe 3** est donc intéressant à examiner puisqu'il permet d'identifier des compagnies présentant un **manque de maîtrise de leurs opérations** se traduisant par des **écritures de régularisation** alimentant les comptes de pertes et profits.

Au titre de circonstances ou d'événements augmentant le **risque de fraude** ou d'erreur, la **norme ISA 240** relative aux irrégularités et inexactitudes cite les **écritures de régularisation en nombre important et les opérations non comptabilisées selon les procédures normales**.

La **NEP 240§17** précise que lorsque le commissaire aux comptes met en œuvre des procédures analytiques lui permettant de prendre connaissance de l'entité, il apprécie si les **corrélations inhabituelles ou inattendues indiquent des risques d'anomalies significatives** résultant de fraude.

Des compagnies d'assurance du marché enregistrent dans les comptes de résultat hors exploitation de nombreuses opérations qui normalement devraient être rattachées à l'exploitation. Cette pratique traduit un **manque de maîtrise des opérations** et un **manque de fiabilité des comptes** : les opérations omises ou mal traitées au cours des exercices précédents sont corrigées par la suite par l'utilisation des comptes de pertes et profits.

A la suite de l'ACP réalisée, une CAH peut être mise en œuvre pour avoir un **autre angle d'analyse** sur les **mêmes données**.

### 3) Confirmation des résultats de l'étude par une CAH

La CAH permet de représenter les individus à travers un **arbre hiérarchique** ou **dendrogramme** avec des classes spécifiées. La **caractérisation** des classes peut être faite à partir des **variables initiales** ou à partir des **composantes principales**. Elle peut également être faite à partir des **parangons** et des **individus spécifiques**.

#### 3.1) Forme du dendrogramme et nombre de classes

La forme du **dendrogramme en annexe** (Figure 19: Données sociétés non vie - arbre hiérarchique) et le **niveau de coupure optimal** calculé par la fonction HCPC suggèrent une **partition** des sociétés d'assurance non vie en **3 classes**.

Les 3 classes peuvent être mises en évidence et représentées sur le plan principal. Cette **représentation est fournie en annexe** (Figure 20: Données sociétés non vie - représentation de la partition sur le plan principal). Le dendrogramme peut également faire l'objet de **représentation en 3 dimensions sur le plan principal** (Figure 21: Données sociétés non vie - représentation du dendrogramme en trois dimensions sur le plan principal). Ces différentes représentations d'une même réalité fournies en annexe permettent de mieux **visualiser et apprécier les classes mises en évidence**.

Les effectifs des trois différentes classes sont fournis dans le tableau suivant :

Classe	1 = noir	2 = rouge	3 = vert
Effectif	68	18	4

**Tableau 9: Données sociétés non vie - effectifs des classes résultant de la CAH**

Les différentes classes mises en évidence par l'arbre hiérarchique peuvent être caractérisées par les variables initiales et les composantes principales.

### 3.2) Description des classes par les variables initiales ou les composantes principales

Le « Tableau 59: Données sociétés non vie - description de la classe 1 par les variables initiales » en **annexe** permet de **caractériser la classe 1 par rapport aux variables initiales**. Cette classe représentée en noir sur le dendrogramme comprend 68 sociétés sur un total de 90. Pour 18 variables initiales sur un total de 19, les sociétés de la classe 1 prennent des valeurs significativement différentes de celles de la moyenne de la population.

Un extrait du tableau complet fourni en annexe est présenté ci-dessous pour 4 variables:

Libellés	Classe1_variables_quantitatives					
	v.test <sup>1</sup>	Mean in category <sup>2</sup>	Overall mean <sup>3</sup>	sd in category <sup>4</sup>	Overall sd <sup>5</sup>	p.value <sup>6</sup>
PSAP	8,23	- 1 947 199,10	- 3 801 441,41	1 409 329,37	3 737 446,81	1,89E-16
E_R	8,17	- 2 607 944,46	- 4 837 291,38	1 708 999,60	4 526 120,22	3,10E-16
T_BILAN	- 7,96	5 555 211,56	9 640 045,52	3 668 247,44	8 515 847,31	1,78E-15
ACTIFS_A	- 7,97	2 339 593,85	4 770 635,41	1 802 060,64	5 060 335,78	1,61E-15

**Tableau 10: Données sociétés non vie - description de la classe 1 par les variables initiales - extrait -**

**Une valeur test (v-test) supérieure en valeur absolue à 2 signifie que la moyenne de la classe est significativement différente de la moyenne de la population générale.** Les variables pour lesquelles la valeur test est comprise dans l'intervalle  $[-2, 2]$  ne présente pas d'intérêt particulier. Elles ne sont donc pas retenues dans l'analyse.

Un **signe positif de la valeur test** indique que la moyenne de la classe est supérieure à la moyenne générale. Un **signe négatif** de la valeur test indique que la moyenne de la classe est inférieure à la moyenne générale.

La **puissance d'un test** est la probabilité de rejeter l'hypothèse nulle  $H_0$  à raison. L'hypothèse testée ici est que la moyenne d'une classe donnée est significativement différente de la moyenne de la population générale. Pour les 4 variables figurant dans le tableau, la probabilité de rejeter l'hypothèse  $H_0$  à raison est quasiment nulle.

<sup>1</sup> v.test: valeur test

<sup>2</sup> Mean in category: moyenne prise par la variable dans la classe

<sup>3</sup> Overall mean: moyenne prise par chaque variable pour l'ensemble des individus étudiés

<sup>4</sup> Sd in category : écart type de chaque variable pour la classe

<sup>5</sup> Overall sd: écart type de chaque variable pour l'ensemble des individus étudiés

<sup>6</sup> p.value: puissance du test

Dans une **première étape**, le tableau permet de **trier** à partir de la clé **v-test** pour chaque classe les variables de la **plus « caractérisante » à la moins « caractérisante » en positif**. Pour ces variables, les individus de la classe prennent des valeurs significativement supérieures à la moyenne de l'ensemble des individus.

Dans une **seconde étape**, le tableau permet de trier pour chaque classe à partir de la clé **v-test**, les variables de la **moins « caractérisante » à la plus « caractérisante » en négatif**. Pour ces variables, les individus de la classe prennent des valeurs significativement inférieures à la moyenne de l'ensemble des individus.

Les **compagnies d'assurance de la classe 1** sont caractérisées par des provisions pour sinistres à payer (PSAP), des engagements réglementés (E\_R), des frais généraux (FG)...**significativement moins élevés par rapport à la moyenne du marché**.

Elles sont également caractérisées par des actifs admissibles en couverture des engagements réglementés (ACTIFS\_A), un total bilan (T\_BILAN), des émissions de primes (PRIMES)...**significativement moins élevés par rapport à la moyenne du marché**.

Les descriptions des autres classes par les variables initiales et les composantes principales sont fournies en annexe : **« 2.2) Caractérisation des classes par rapport aux variables initiales page 156 et 2.3) Caractérisation des classes par rapport aux composantes principales page 159 »**. Est également fournie dans cette annexe, la description de la classe 1 par les composantes principales.

A la suite de la description des classes par les variables initiales et les composantes principales, il est également possible d'effectuer une **description à partir des parangons et des individus spécifiques**.

### **3.3) Description des classes par les parangons et les individus spécifiques**

Les classes peuvent être illustrées par des individus particuliers de ces classes. A ce titre, deux types d'individus particuliers sont proposés :

- Les **parangons**, c'est-à-dire les individus les plus proches du centre de la classe

- Les **individus spécifiques** c'est-à-dire les individus les plus éloignés des centres des autres classes.

Les parangons des différentes classes sont listés dans le « Tableau 65: Données sociétés non vie - description des classes 1, 2 et 3 par les parangons » en **annexe**. Ce tableau présente pour chaque parangon la **distance le séparant du centre de la classe** à laquelle il appartient. **L'assureur A47** représente au mieux les assureurs de la **classe 1**. **L'assureur A12 et 42** sont respectivement les meilleurs parangons des **classes 2 et 3**.

Les individus spécifiques des différentes classes sont listés dans le « Tableau 66: Données sociétés non vie - description des classes 1, 2, et 3 par les individus spécifiques » en annexe. **L'assureur A88** est **spécifique** de la **classe 1** et est le plus éloigné des centres des classes 2 et 3. **L'assureur A19** est **spécifique** de la **classe 2** et est le plus éloigné des centres des **classes 1 et 3**. **L'assureur A6** est spécifique de la **classe 3** et est le plus éloigné des **classes 1 et 2**.



Les représentations des sociétés d'assurance non vie obtenues à partir de l'ACP et de la CAH permettent d'identifier globalement **trois groupes homogènes** : peu d'acteurs de **taille importante**, quelques acteurs de **taille moyenne** et un **foisonnement de petits acteurs**.

Les compagnies exerçant dans l'espace CIMA sont dans leur majorité **fragilisées par un important niveau d'arriérés de primes** et une **absence de taille critique** devant permettre de mieux couvrir les frais fixes et mutualiser les risques souscrits. La **concentration** du secteur recherchée par les autorités de supervision à travers notamment le relèvement du niveau minimum du capital social **n'a pas eu lieu**.

La **démarche** mise en œuvre pour étudier les compagnies d'assurance non vie peut également être **reproduite** sur les données des compagnies d'assurance vie en vue d'obtenir une vue synoptique du secteur et de tirer des enseignements utiles notamment pour **l'aide à la décision**.

## CHAPITRE II. REALISATION ET INTERPRETATION D'UNE ANALYSE EN COMPOSANTES PRINCIPALES (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE(CAH) SUR LES DONNEES DES COMPAGNIES D'ASSURANCE VIE DU MARCHE

Suivant la même démarche qu'au chapitre précédent, une **analyse en composantes principales** est réalisée sur l'ensemble des **45 compagnies d'assurance vie** exerçant dans l'espace CIMA. Elle est suivie d'une **classification ascendante hiérarchique** permettant d'avoir un autre point de vue et confirmant éventuellement les résultats de l'analyse en composantes principales.

### 1) Etude du nuage des individus à travers une ACP

Comme pour les compagnies non vie, **l'étude des individus** commence par une **projection et une représentation sur le plan principal**. L'analyse du **pourcentage d'inertie** associé à chaque axe permet de déterminer les **dimensions et les axes intéressants** pour l'étude. Enfin, **trois paramètres quantitatifs** calculés pour chaque individu permettent de conforter les analyses effectuées à partir de la **représentation graphique**.

#### 1.1) Représentation du nuage des individus

Les **commandes utilisées** pour importer les données et réaliser l'ACP pour les compagnies d'assurance vie sont les mêmes que celles utilisées plus haut pour les compagnies d'assurance non vie.

Le fichier initial contenant les informations sur les compagnies d'assurance vie est un fichier d'extension csv dénommé « `donnees_vie.csv` ».

Les **deux commandes** suivantes permettent dans un premier temps de **charger** le package `FactoMineR`, puis **d'importer sur la console**  **les données** en précisant que le nom des sociétés est dans la première colonne.

```
> library (FactoMineR)
> analyse_donnees_vie<-read.table ("donnees_vie.csv", header=TRUE, sep=';', dec='.',
row.name=1)
```

Ensuite, on utilise la commande suivante :

```
> res.pca<-PCA(analyse_donnees_vie)
```

Cette **commande exécute l'ACP** et fournit le **graphe des individus et le graphe des variables**.

Les 45 sociétés d'assurance vie qu'on ne pouvait pas représenter et visualiser dans un **espace à 22 dimensions** (22 variables), sont maintenant représentées dans un **plan**. Ce plan contenant **72,71% de l'information initiale** dont 60,41% pour le premier axe principal se présente comme suit :

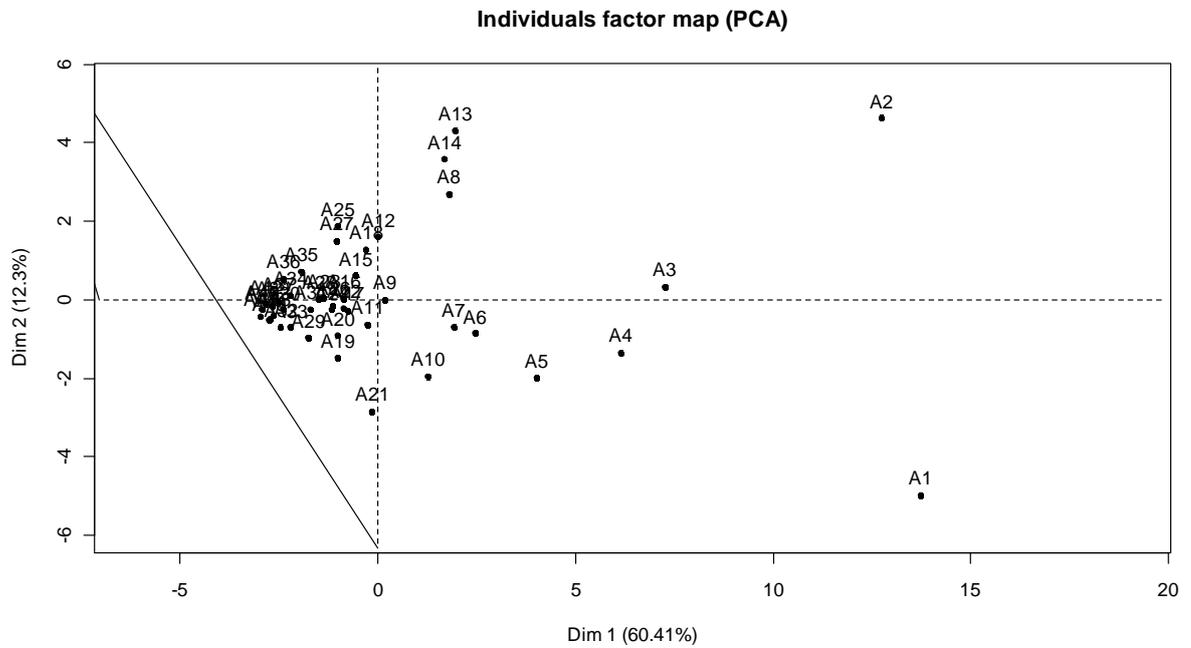


Figure 4: Données sociétés vie - nuage des individus

Les **indications générales** fournies pour la lecture du plan principal des sociétés non vie sont également valables pour les sociétés vie.

La **représentation** permet d'avoir le **positionnement** de **chaque compagnie d'assurance** sur le **marché** par rapport aux autres, simplement à partir de **deux variables de synthèse**.

Le **premier axe** oppose les compagnies ayant un niveau **d'activité important** aux compagnies ayant un **faible niveau d'activité**. Cette observation était également valable pour les compagnies d'assurance non vie.

Le **second axe** oppose les compagnies ayant recours à la **réassurance de façon intensive** aux compagnies fonctionnant quasiment **sans réassurance**. Cette observation constitue une **spécificité** des sociétés d'assurance vie par rapport aux sociétés d'assurance non vie.

La **détection** des **individus remarquables** peut se faire en examinant leur éloignement par rapport au **centre de gravité** du nuage formé.

Les **20 compagnies** les **plus éloignées** du centre de gravité sont présentées dans le tableau suivant :

A1	A2	A3	A18	A4	A21	A5	A13	A14	A15	A8	A19	A10	A6	A45	A43	A44	A42	A40
15	13,83	8,18	7,3	6,9	6,47	6,18	5,93	5,25	4,79	4,21	3,55	3,52	3,36	3,21	3,13	3,13	2,99	2,92

Tableau 11: Données sociétés vie - éloignement du centre de gravité

Les distances fournies sont toujours positives. Par conséquent, **deux compagnies** peuvent être à des **extrémités opposées** du centre de gravité et être à équidistance du centre de gravité.

Les compagnies **A1, A2, A3** sont éloignées du centre de gravité avec des **indicateurs de gestion positifs**. L'éloignement des compagnies **A45, A43, A44** du centre de gravité se traduit par contre par des **indicateurs de gestion négatifs** (faiblesse du chiffre d'affaires et niveau élevé des charges.)

La **représentation synoptique** du nuage formé par les compagnies d'assurance permet de constater que le marché des assurances est marqué par :

- La **présence de quelques grands acteurs** (A1, A2, A3, A4). Ces grands acteurs sont caractérisés par un chiffre d'affaires important. Les assureurs A2, A3, A4 protègent les risques souscrits par des **traités de réassurance** alors que l'assureur A2 n'a quasiment **pas recours à la réassurance**. Ces compagnies se ressemblent sur l'axe 1, mais sont **opposées sur l'axe 2**.
- La **présence d'acteurs de taille moyenne** caractérisés soit par des cessions importantes en réassurance (A8, A12, A13, A14,), soit par peu d'opérations de réassurance (A5, A6, A7). Ces acteurs présentent des similitudes sur l'axe 1 mais divergent sur l'axe 2.
- La **présence de nombreux petits acteurs** formant un groupe compact, présentant une certaine homogénéité. Dans ce groupe, on constate que de nombreuses compagnies n'ont pas la taille critique devant leur permettre de mutualiser les risques d'assurance et dégager la rentabilité nécessaire à la rémunération des contrats.
- Deux compagnies (A15 et A18) présentent une **spécificité par leur niveau d'arriérés de primes** particulièrement élevé.

L'**absence de taille critique** constitue un **facteur de vulnérabilité** pour de nombreuses entreprises d'assurance vie. Ce facteur de vulnérabilité influe négativement sur la **solvabilité des entreprises d'assurance vie et sur leur capacité à dégager la rentabilité nécessaire à la rémunération des contrats**.

### 1.2) Pourcentage d'inertie associé à chaque axe et choix des dimensions à étudier

La **commande** suivante permet d'obtenir un affichage de la **valeur propre** associée à chaque axe (Eigen value), le **pourcentage d'inertie** (Percentage of variance) qu'elle représente dans l'analyse ainsi que le **cumul de ces pourcentages** (Cumulative percentage of variance).

```
> round(res.pca$eig,2)
```

	Eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	13,29	60,41	60,41
comp 2	2,71	12,30	72,71
comp 3	2,17	9,84	82,55
comp 4	0,95	4,33	86,88
comp 5	0,85	3,85	90,73
comp 6	0,76	3,44	94,17

Tableau 12: Données sociétés vie - six premières valeurs propres et pourcentage d'inertie

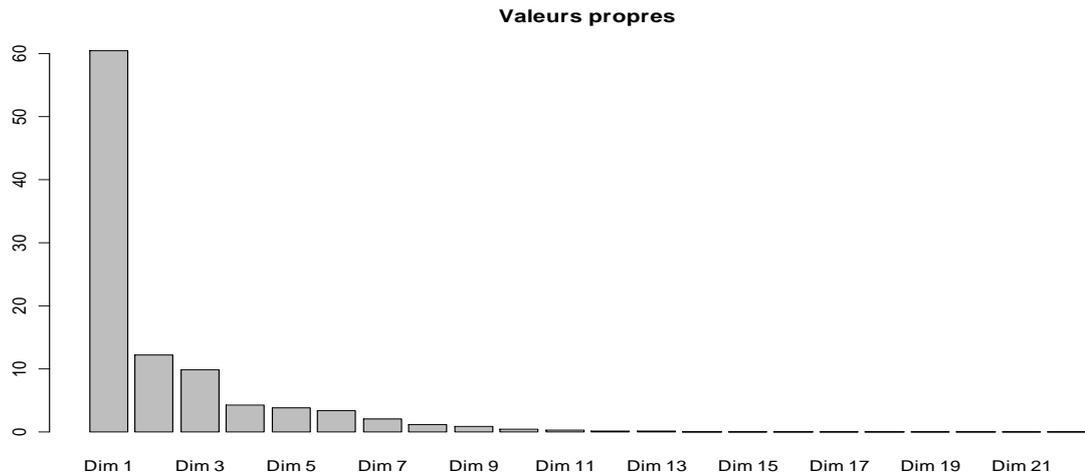
Le **tableau** comprenant la totalité des valeurs propres et des pourcentages d'inertie associés est fourni en **annexe** (Tableau 67: Données sociétés vie - valeurs propres et pourcentage d'inertie)

Le **1<sup>er</sup> axe** exprime **60,41% de la variabilité des données** contre **12,30% pour le 2<sup>ème</sup> axe**. Le 1<sup>er</sup> axe exprime à peu près 3 fois plus de variabilité que le 2<sup>ème</sup> axe. Le plan principal contient 72,71% de l'information, ce qui est satisfaisant.

Le **3<sup>ème</sup> axe** exprime **9,84%** de l'inertie total. Il n'y a pas une chute brutale d'inertie entre la 2<sup>ème</sup> et la 3<sup>ème</sup> composante principale. Le 4<sup>ème</sup> axe exprime 4,33% d'inertie. Dès cet axe, la chute de l'inertie devient régulière. L'étude peut donc se limiter aux **trois premiers axes** qui font 82,55% de l'inertie totale du nuage.

La commande suivante permet une **représentation en diagramme en barres** des différentes valeurs propres obtenues à partir de la matrice des données reformulées et notamment la matrice de corrélation (Tableau 69: Données sociétés vie - matrice de corrélation ).

```
> barplot (res.pca$eig[,2], main = "Valeurs propres", name=paste("Dim", 1:nrow(res.pca$eig)))
```



**Figure 5: Données sociétés vie - éboulis des valeurs propres**

Cette représentation permet de constater une **chute importante d'inertie entre la 1<sup>ère</sup> valeur propre et la 2<sup>ème</sup>**. La chute d'inertie entre la 2<sup>ème</sup> valeur propre et la 3<sup>ème</sup> n'est pas importante. Après les trois premiers axes, on observe une décroissance régulière des inerties. Ces chutes d'inertie suggèrent de se limiter aux trois premiers axes.

### 1.3) Qualité de la représentation des individus et contribution à la construction des axes

La commande suivante :

```
> round(cbind(res.pca$ind$coord[,1:4], res.pca$ind$cos2[,1:4], res.pca$ind$contrib[,1:4]),2)
```

permet d'avoir dans un tableau concaténé sur les 4 premiers axes principaux 84,25% de l'inertie total :

- Les **coordonnées** des compagnies sur les nouveaux axes construits,
- Le **cosinus carré** de l'angle formé par chaque compagnie avec sa projection sur chacun des axes principaux,
- La **contribution** de chaque compagnie à la construction de chacun des axes.

Un **extrait du tableau** pour les 10 premières entreprises du marché en termes de chiffre d'affaires se présente comme suit :

	Cordonnées				Cos <sup>2</sup>				Contribution			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
<b>A1</b>	13,72	-4,97	0,42	3,24	0,84	0,11	0,00	0,05	<b>31,47</b>	<b>20,32</b>	0,18	24,44
<b>A2</b>	12,72	4,63	-0,89	-1,30	0,85	0,11	0,00	0,01	<b>27,07</b>	<b>17,59</b>	0,82	3,92
<b>A3</b>	7,25	0,34	-0,82	-1,05	0,79	0,00	0,01	0,02	<b>8,79</b>	0,09	0,70	2,55
<b>A4</b>	6,15	-1,35	-1,36	-1,23	0,79	0,04	0,04	0,03	<b>6,33</b>	1,49	1,89	3,52
A5	4,00	-1,99	0,97	-1,96	0,42	0,10	0,02	0,10	2,68	3,24	0,96	8,94
A6	2,45	-0,85	-1,15	-0,72	0,53	0,06	0,12	0,05	1,00	0,59	1,35	1,21
A7	1,92	-0,69	0,34	0,90	0,44	0,06	0,01	0,10	0,62	0,39	0,12	1,90
A8	1,79	2,69	-1,13	0,17	0,18	0,41	0,07	0,00	0,54	5,95	1,32	0,07
A9	0,17	0,00	0,74	-1,39	0,01	0,00	0,09	0,34	0,01	0,00	0,56	4,54
A10	1,26	-1,95	0,35	-1,25	0,13	0,31	0,01	0,13	0,27	3,12	0,12	3,62

**Tableau 13: Données sociétés vie - coordonnées et qualité de représentation de 10 sociétés - extrait -**

Les informations concernant toutes les compagnies sont consignées dans un **tableau en annexe** (Tableau 68: Données sociétés vie - coordonnées et qualité de représentation des sociétés).

Les **compagnies A1, A2, A3 et A4** contribuent le plus à la construction de **l'axe 1** avec des contributions respectives de 31,47%, 27,07%, 8,79% et 6,33%. Au total ces 4 compagnies contribuent à hauteur de 73,66% à la construction de l'axe 1. Elles ont également des **coordonnées positives élevées** sur l'axe 1 avec un maximum de 13,72 pour la compagnie A1.

Les **compagnies A1, A2, A13 et A14** contribuent le plus à la construction de **l'axe 2** avec des contributions respectives de 20,32%, 17,59%, 15,37% et 10,54%. Les compagnies A2, A3, A13 et A14 ont des **coordonnées positives élevées sur l'axe 2**. La **compagnie A1 a des coordonnées négatives élevées sur l'axe 2**. L'axe 2 oppose donc un groupe de compagnies effectuant **d'importantes cessions en réassurance** à un autre groupe fonctionnant quasiment **sans réassurance**.

La **compagnie A18** contribue à 48,83% à la construction de **l'axe 3**. Le deuxième contributeur à la construction de cet axe est la compagnie A15 avec un taux de 21,64%. Au total, ces deux compagnies qui sont bien représentées sur le 3<sup>ème</sup> axe, contribuent à hauteur de 70,47% à sa construction. Ces compagnies sont caractérisées par un **niveau d'arriérés particulièrement élevé**.

A la suite de l'étude de l'espace des individus, l'analyse du nuage des variables est une phase indissociable de l'ACP.

## 2) Etude du nuage des variables à travers une ACP

### 2.1) Représentation des variables

La **représentation du nuage des variables** permet de visualiser rapidement les **corrélations** entre **variables**.

La représentation ci dessous permet de voir comment ces **22 variables initiales** sont **corrélées** aux **deux variables de synthèse** qui ont été retenues pour la représentation des compagnies d'assurance dans le **plan principal**.

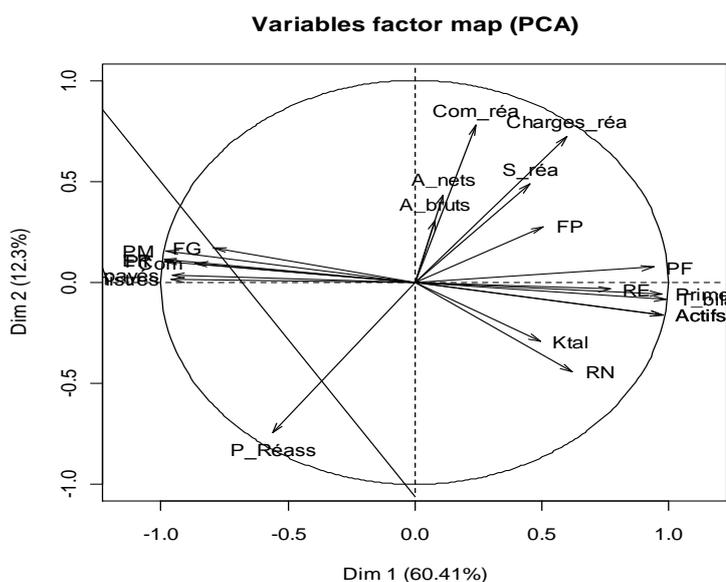


Figure 6: Données sociétés vie - graphe des variables

La représentation met en évidence une **corrélation positive** entre la **1<sup>ère</sup> composante principale** et les **produits et les actifs**. Elle met en évidence également, une **corrélation négative** entre la **1<sup>ère</sup> composante principale** et les **charges et les passifs**.

Toutes les variables liées aux **opérations de réassurances** sont **corrélées au 2<sup>ème</sup> axe principal** à l'exception de la variable « sinistres payés à la charge des réassureurs ».

Les **arriérés bruts** et les **arriérés nets** sont peu corrélés aux deux **1<sup>ères</sup> composantes principales**. Ces deux variables sont plutôt corrélées à la **3<sup>ème</sup> composante principale**.

Plusieurs variables ne sont pas bien représentées dans le plan principal. Il s'agit notamment des **variables arriérés bruts et arriérés nets** et dans une moindre mesure des **variables fonds propres et capital**.

La **qualité de cette représentation** est lue grâce à la proximité entre l'extrémité de la **flèche représentant la variable et le cercle de rayon 1**. Pour les variables citées, l'extrémité de la flèche est assez éloignée du cercle. Pour les variables arriérés, la longueur de la flèche est même inférieure à la moitié du rayon du cercle.

L'**analyse visuelle** peut être soutenue par des **données chiffrées** relatives aux coordonnées des variables, à leur qualité de représentation et à leur contribution à la construction des axes.

## 2.2) Analyse de la représentation des variables

Les **coordonnées des variables**, leur **qualité de représentation** et leur **contribution** à la construction des axes sont examinées dans les paragraphes suivants.

La commande suivante :

```
> round(cbind(res.pca$var$coord[,1:4], res.pca$var$cos2[,1:4], res.pca$var$contrib[,1:4]),2)
```

permet d'avoir dans un tableau concaténé sur les 4 premiers axes principaux 84,25% de l'inertie total :

- Le **coefficient de corrélation** de chaque variable initiale avec les nouveaux axes construits,
- Le **cosinus carré** de l'angle formé par chaque variable avec sa projection sur chacun des axes principaux,
- La **contribution** de chaque variable à la construction de chacun des axes.

Un extrait du tableau ainsi obtenu, se présente comme suit :

	Corrélation des variables p/r aux axes				Corrélation des variables p/r aux axes (cos <sup>2</sup> )				Contribution des variables à la construction des axes			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
P_Réass	<b>-0,56</b>	<b>-0,74</b>	0,07	-0,09	0,32	0,55	0,01	0,01	2,38	<b>20,34</b>	0,24	0,77
Charges_réa	<b>0,60</b>	<b>0,72</b>	-0,14	-0,07	0,36	0,52	0,02	0,00	2,71	<b>19,15</b>	0,89	0,46
RE	0,77	-0,03	0,04	0,52	0,59	0,00	0,00	0,27	4,47	0,04	0,08	<b>28,23</b>
RN	<b>0,62</b>	-0,44	0,16	0,42	0,39	0,20	0,03	0,17	2,92	7,24	1,20	<b>18,23</b>
Com_réa	<b>0,24</b>	<b>0,78</b>	-0,18	0,18	0,06	0,60	0,03	0,03	0,43	<b>22,31</b>	1,49	3,27
A_bruts	<b>0,08</b>	0,31	<b>0,92</b>	-0,07	0,01	0,09	0,84	0,00	0,04	3,46	<b>38,84</b>	0,49
A_nets	<b>0,11</b>	0,43	<b>0,85</b>	-0,06	0,01	0,19	0,72	0,00	0,09	6,92	<b>33,15</b>	0,37

**Tableau 14: Données sociétés vie - corrélation et qualité de représentation des variables - extrait -**

Les informations relatives à toutes les variables initiales sont consignées dans un **tableau en annexe** (Tableau 70: Données sociétés vie - corrélation et qualité de représentation des variables).

**Sur les 22 variables, 13 sont corrélées** avec le **premier axe principal** avec un coefficient strictement supérieur à 0,77. Les **9 variables** qui ne présentent **pas de coefficients de corrélation élevés** avec le premier axe principal sont les suivantes: « primes cédées aux réassureurs : -0,56 », « charges aux réassureurs : 0,60 », « Résultat net 0,6 », « Sinistres payés à la charge des réassureurs 0,45 », « Commissions de réassurance : 0,24 », « Arriérés bruts : 0,08 », « Arriérés nets: 0,11 », « Capital : 0,44 » et « Fonds propres : 0,51 ».

Le **premier axe principal synthétise** cependant une **part essentielle des informations** fournies par les 22 variables de base.

On constate également que toutes les **variables fortement corrélées au 1<sup>er</sup> axe principal contribuent dans des proportions quasi identiques à sa construction** (entre 4 et 8% pour chaque variable).

Les variables fortement corrélées à un axe principal sont également fortement corrélées entre elles. Les **coefficients de corrélation** entre variables d'origines sont consignés dans la **matrice de corrélation** jointe en annexe (Tableau 69: Données sociétés vie - matrice de corrélation).

Les **variables** résultant des **opérations de réassurance** sont corrélées au **2<sup>ème</sup> axe principal** avec des coefficients supérieurs à **0,72**. Seule la variable « sinistres payés à la charge des réassureurs » fait exception avec un coefficient de 0,49. Les **4 variables** traduisant les **opérations de réassurance** contribuent à hauteur de **70,53%** (20,34 + 19,15 + 8,73 + 22,31) à la construction de **l'axe 2**. On peut donc dire que le **2<sup>ème</sup> axe synthétise** essentiellement les variables relatives aux **opérations de réassurance**.

Seules **deux variables** sont fortement corrélées avec le **3<sup>ème</sup> axe principal** avec des coefficients élevés. Il s'agit des variables **arriérés bruts** (-0,92) et **arriérés nets** (-0,85). Le **3<sup>ème</sup> axe** traduit donc essentiellement la **qualité de la politique de recouvrement et de gestion des arriérés de primes**. Cette observation est confirmée par le fait que ces deux variables contribuent à hauteur de 71,99% (38,84 + 33,15) à la construction de l'axe 3. Ce constat est intéressant car en **assurance vie**, les **arriérés de primes** doivent être une **exception** : l'assuré n'a pas réglementairement d'obligation de payer la prime.

L'analyse effectuée à partir du nuage de point résultant de l'ACP peut être confirmée par une classification ascendante hiérarchique (CAH).

### 3) Confirmation des résultats de l'étude par une CAH

La **CAH** permet de représenter les individus à travers un **arbre hiérarchique ou dendrogramme** avec des **classes spécifiées**. La caractérisation des classes peut être faite à partir des **variables initiales** ou à partir des **composantes principales**. Elle peut également être faite à partir des **parangons** et des **individus spécifiques**.

#### 3.1) Forme du dendrogramme et nombre de classes

La forme du **dendrogramme en annexe** (Figure 22: Données sociétés vie - arbre hiérarchique) et le **niveau de coupure optimal** calculé par la fonction HCPC suggère une **partition** des sociétés d'assurance vie en **6 classes**.

Les **6 classes** peuvent être mises en évidence et représentées sur le **plan principal**. Cette **représentation** est fournie en **annexe** (Figure 23: Données sociétés vie - représentation de la partition sur le plan principal). Le **dendrogramme** peut également faire l'objet de **représentation en 3 dimensions sur le plan principal** (Figure 24: Données sociétés vie - représentation du dendrogramme en trois dimensions sur le plan principal).

Les **effectifs** des **six** différentes **classes** sont fournis dans le tableau suivant :

Classe	1= noir	2= rouge	3= vert	4=bleu	5=cyan	6=rose
Effectif	28	7	4	2	3	1

Tableau 15: Données société vie - effectif des classes résultant de la CAH

Les différentes **classes** mises en évidence par l'arbre hiérarchique peuvent être **caractérisées** par les **variables initiales** et les **composantes principales**.

#### 3.2) Description des classes par les variables initiales ou les composantes principales

Le «Tableau 71: Données sociétés vie - description de la classe 1 par les variables initiales» en **annexe** permet de **caractériser la classe 1** par rapport aux **variables initiales**. Cette classe représentée **en noir sur le dendrogramme** comprend **28 sociétés**. Pour **20 variables initiales** sur un total de **22**, les **sociétés de la classe 1** prennent des **valeurs significativement différentes de celles de la moyenne de la population**.

Un **extrait du tableau complet** fourni en annexe est présenté ci-dessous pour 6 variables :

Classe1_variable_quantitative						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
FG	4,96	- 390 967 117	- 681 381 343	176 018 147	498 831 392	7,19E-07
Com	4,42	- 115 010 254	- 216 545 912	85 429 316	195 372 158	9,68E-06
ER	4,23	- 3 119 944 880	- 10 242 669 996	2 886 799 934	14 346 746 312	2,37E-05
PT	4,22	- 3 067 730 958	- 10 067 134 499	2 858 991 248	14 108 339 558	2,41E-05
T_bilan	- 4,20	4 098 564 754	12 069 367 791	3 286 330 704	16 151 047 664	2,65E-05
Primes	- 4,66	1 530 716 151	3 397 430 373	1 071 181 155	3 407 911 942	3,12E-06

**Tableau 16: Données sociétés vie - description de la classe 1 par les variables initiales - extrait -**

Les compagnies d'assurance de la **classe 1** sont caractérisées par des frais généraux (FG), des commissions (Com), des engagements réglementés (E\_R), des provisions techniques (PT) ...**moins élevés par rapport à la moyenne du marché.**

Elles sont également caractérisées par des émissions de primes (PRIMES), un total bilan (T\_BILAN), ...**moins élevés par rapport à la moyenne du marché.**

La **description des 5 autres classes** par les **variables initiales et les composantes principales** est fournie en annexe « 4.2)Caractérisation des classes par rapport aux variables initiales page 170 et 4.3)Caractérisation des classes par rapport aux composantes principales page 175 », selon la même démarche que celle présentée ci-dessus. La description de la classe 1 par les composantes principales est également fournie dans les mêmes annexes.

A la suite de la description des classes par les variables initiales et les composantes principales, il est également possible d'effectuer une **description** à partir des **parangons** et des **individus spécifiques**.

### **3.3) Description des classes par les parangons et les individus spécifiques**

Les classes peuvent être illustrées par des individus particuliers de ces classes : **les parangons et les individus spécifiques.**

Les **parangons des différentes classes** sont listés dans le «Tableau 78: Données sociétés vie - description des classes 1 à 6 par les parangons » en **annexe**. Ce tableau présente pour chaque parangon la distance le séparant du centre de la classe à laquelle il appartient. **L'assureur A31** représente au mieux les assureurs de la **classe 1**. **Les assureurs A10, A8, A18 et A3** sont respectivement les meilleurs

parangons des **classes 2, 3, 4, et 5**. **L'assureur A1**, le seul individu de la **classe 6**, est le parangon de cette classe.

Les **individus spécifiques** des différentes classes sont listés dans le « Tableau 79: Données sociétés vie - description des classes 1 à 6 par les individus spécifiques » en **annexe**. **L'assureur A45** est spécifique de la **classe 1** et est le plus éloigné des centres des autres classes (2, 3, 4, 5 et 6). **L'assureur A21** est spécifique de la **classe 2** et est le plus éloigné des centres des autres classes. **L'assureur A 13** est spécifique de la **classe 3** et est le plus éloigné des autres classes. **L'assureur A18** est spécifique de la **classe 4** et est le plus éloigné des autres classes. **L'assureur A2** est spécifique de la **classe 5** et est le plus éloigné des autres classes.



Le **marché de l'assurance vie** apparaît **moins homogène** que celui de **l'assurance non vie**. Les représentations des sociétés d'assurance vie obtenues à partir de l'ACP et de la CAH permettent d'identifier globalement **6 groupes** contre 3 pour les sociétés non vie.

Les compagnies d'assurance vie exerçant dans l'espace CIMA sont dans leur **majorité fragilisées par une absence de taille critique** devant permettre de mieux couvrir les frais et dégager des résultats pour la rémunération des contrats. Le **recours ou non à la réassurance** est également un facteur important permettant de classer ces sociétés.

La démarche mise en œuvre sur les données des compagnies d'assurance vie et non vie peut également être reproduite sur les données d'un groupe de sociétés d'assurance. La **représentation synoptique** obtenue pour quelques individus peut être **plus lisible** que celle résultant de la représentation de dizaines d'individus.

### **CHAPITRE III. REALISATION ET INTERPRETATION D'UNE ANALYSE EN COMPOSANTES PRINCIPALES (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE (CAH) SUR LES DONNEES DES COMPAGNIES D'ASSURANCE DU GROUPE G**

Le professionnel comptable peut être amené à s'intéresser à la **structure et aux caractéristiques** de sociétés d'un **secteur d'activité** donné ou à un **groupe de sociétés**. Dans le cas présenté, le **groupe G** est composé de **6 sociétés d'assurance non vie et 6 sociétés d'assurance vie** exerçant dans 6 différents pays de la zone CIMA. Dans chacun des 6 pays, le groupe a constitué une société d'assurance non vie et une société d'assurance vie.

A partir du tableau initial contenant les 90 sociétés d'assurance non vie exerçant dans l'espace CIMA, les **6 sociétés du groupe étudié ont été individualisées**. Le tableau regroupant les 6 sociétés fait l'objet de l'ACP et la CAH. Un **traitement identique** a été effectué pour les **6 sociétés d'assurance vie** du groupe à partir du tableau initial des 45 sociétés d'assurance vie.

#### **1) Etude des compagnies d'assurance non vie du groupe G**

L'étude des compagnies d'assurance non vie du groupe G est effectuée suivant les trois étapes décrites ci dessous :

- **étude du nuage des individus à travers une ACP ;**
- **étude du nuage des variables à travers une ACP ;**
- **confirmation des résultats de l'ACP par une CAH.**

### 1.1) Etude du nuage des individus à travers une ACP

L'ACP effectuée permet de représenter les 6 compagnies d'assurance vie du groupe dans le **plan** ci-dessous synthétisant **91,95%** de l'information initiale.

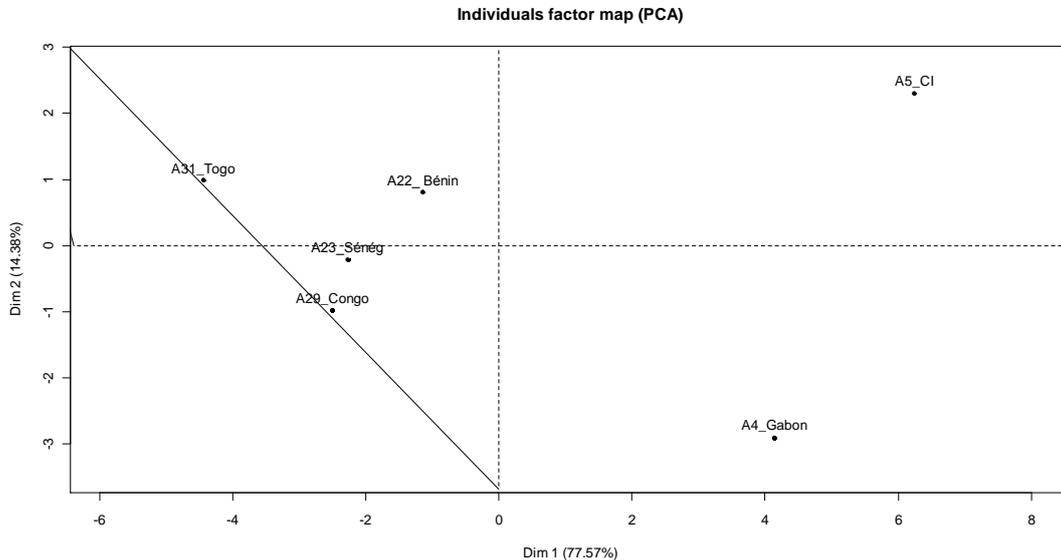


Figure 7: Données sociétés non vie groupe G - nuage des individus

Le **taux** de 91,95% résulte d'un pourcentage d'inertie de 77,57 % pour le premier axe et 14,38% pour le second. Ce taux est **excellent** dans le cadre d'une ACP.

Les **valeurs propres**, le **pourcentage d'inertie** associé à chaque axe et le **pourcentage cumulé** sont présentés en **annexe** (Tableau 80: Sociétés non vie du groupe G - valeurs propres et pourcentage d'inertie)

L'**éboulis des valeurs propres** également fourni en **annexe** (Figure 25: Données sociétés non vie groupe G - éboulis des valeurs propres) permet de constater une **chute rapide d'inertie entre l'axe 1 et l'axe 2**.

L'**axe 1** oppose les **compagnies A4 et A5** aux **autres entités du groupe**. Les sociétés A4 et A5 ont une **activité significativement plus importante** que celle des autres entités du groupe. Le « Tableau 81: Données sociétés non vie groupe G - coordonnées et qualité de représentation des sociétés » en annexe permet de chiffrer cette opposition notamment à partir des coordonnées de chaque entité sur l'axe 1.

L'**axe 2** traduit essentiellement le **résultat hors exploitation (RHE)**. Il **oppose** principalement la **compagnie A4** (coordonnée -2,9) ayant réalisé une RHE positif de 266 millions de FCFA à la **compagnie A5** (coordonnée 2,3) avec un RHE négatif de 391 millions de FCFA.

Cette première analyse permet de constater une **hétérogénéité** des sociétés du groupe et une relative **faiblesse du niveau d'activité des filiales A22, A23, A29 et A31**.

Une telle diligence mise en œuvre par un professionnel comptable dans la **phase de prise de connaissance** de sa mission permet de **situer chaque entité du groupe par rapport aux autres**. Si cette **mission est conduite par un expert comptable**, la **mise en évidence de faiblesses** au niveau de certaines entités peut faire l'objet d'une **mission de conseil** pour un **diagnostic plus approfondi** et des **recommandations adaptées**.

L'étude du nuage des individus est effectuée concomitamment avec celle du nuage des variables.

### 1.2) Etude du nuage des variables à travers une ACP

La **représentation** ci dessous permet de voir comment les **19 variables initiales** sont corrélées aux **deux variables de synthèses** retenues pour la représentation des **6 compagnies d'assurance non vie** du groupe dans le **plan principal**.

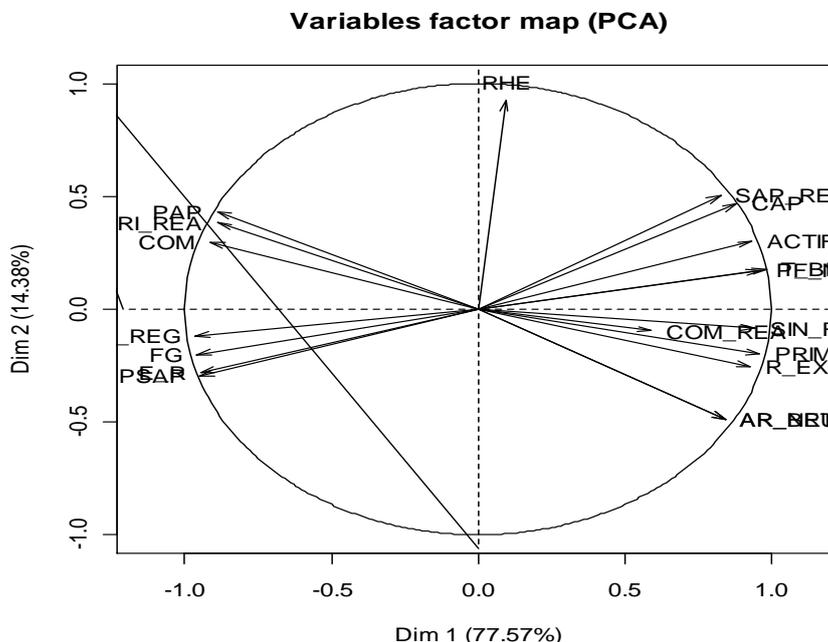


Figure 8: Données sociétés non vie du groupe G - nuage des variables

Toutes les **variables sont bien représentées** dans le **plan principal** à l'**exception** de la **variable Com\_REA** (Commissions de Réassurance). Cette variable est plutôt bien représentée sur l'**axe 3**. Elle contribue à 58,92 % à la construction de cet axe.

La **corrélation**, l'**angle** (cosinus carré) et la **contribution** de chaque variable initiale par rapport aux **4 premiers axes** sont fournis en **annexe** (Tableau 82: Données sociétés non vie groupe G - coordonnées et qualité de représentation des variables) et permettent une analyse plus détaillée.

Les quelques **variables** contribuant le plus à la construction de l'**axe 2** sont les suivantes avec leur **taux de contribution** : RHE (31,47%), SAP\_REA (9,37%), AR\_BRUTS (8,79%) et AR\_NETS (8,79%).

Les **axes 3 et 4** peuvent être intéressants à examiner même s'ils représentent **peu d'inertie**. Ils mettent en évidence des **spécificités, sources de risques dans un groupe où les procédures et traitements devraient être harmonisés**.

A la suite de l'ACP réalisée sur les données des sociétés non vie du groupe G, une CAH peut être mise en œuvre pour avoir un **autre angle d'analyse** sur les mêmes données.

### 1.3) Confirmation des résultats par une CAH

La **forme du dendrogramme** ci-dessous (Figure 9: Données sociétés non vie du groupe G - arbre hiérarchique) et le **niveau de coupure optimal** calculé par la fonction HCPC suggère une **partition** des sociétés d'assurance non vie du groupe en **3 classes**.

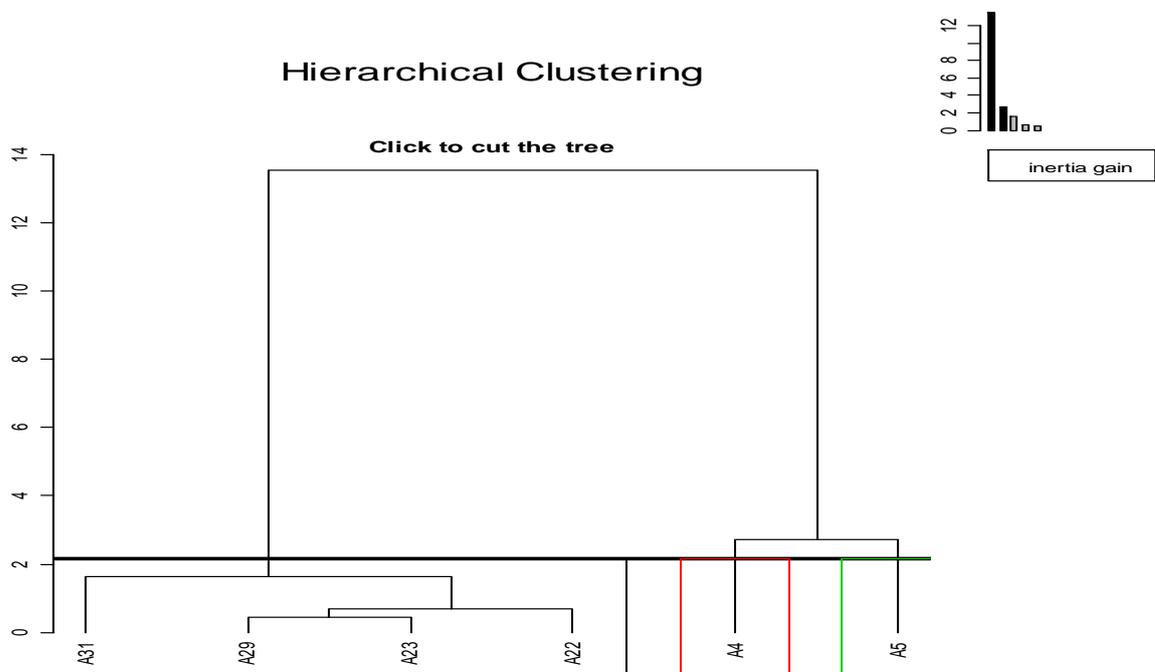


Figure 9: Données sociétés non vie du groupe G - arbre hiérarchique

Les **3 classes** peuvent être mises en évidence et représentées sur le **plan principal**. Cette **représentation** est fournie en **annexe** (Figure 26: Données sociétés non vie groupe G - représentation de la partition sur le plan principal). Le **dendrogramme** ci-dessus peut également faire l'objet de **représentation en 3 dimensions sur le plan principal**. Cette **représentation** est également fournie en **annexe** (Figure 27: Données sociétés non vie du groupe G - représentation du dendrogramme en trois dimensions sur le plan principal).

La démarche mise en œuvre pour l'analyse des données des sociétés non vie du groupe G peut être reproduite à partir des données des sociétés vie du groupe.

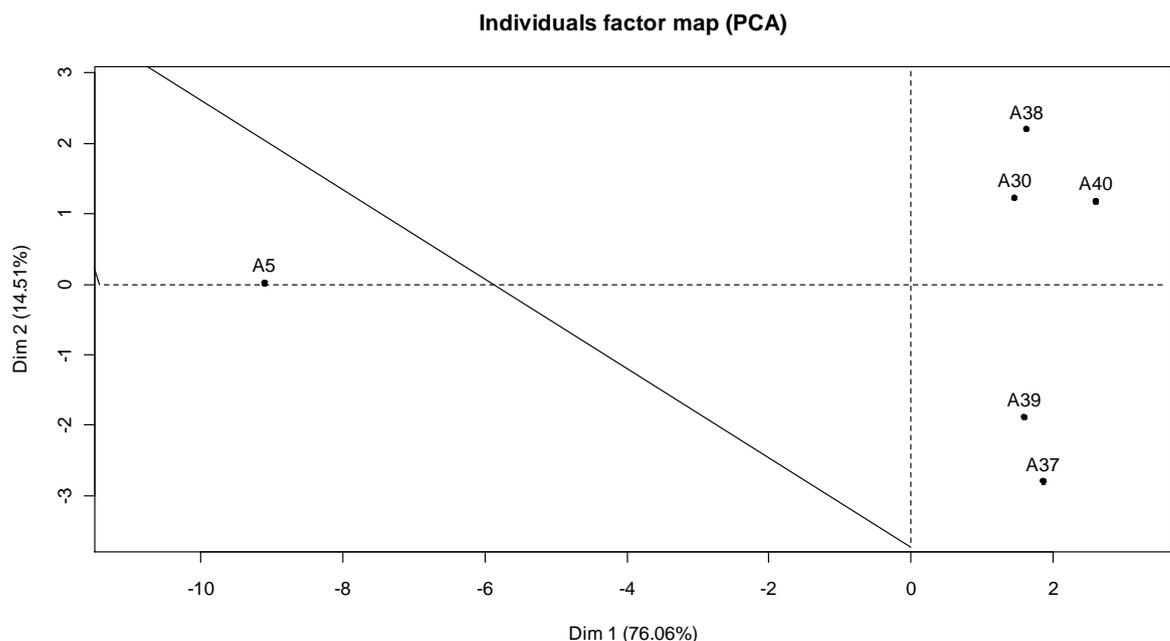
## 2) Etude des compagnies d'assurance vie du groupe G

L'étude des compagnies d'assurance vie du Groupe G est effectuée suivant les trois étapes suivantes :

- **Etude du nuage des individus à travers une ACP ;**
- **Etude du nuage des variables à travers une ACP ;**
- **Confirmation des résultats de l'ACP par une CAH.**

### 2.1) Etude du nuage des individus à travers une ACP

L'ACP effectuée permet de représenter les 6 compagnies d'assurance vie du groupe dans le **plan** ci-dessous synthétisant **90,57%** de l'information initiale.



**Figure 10: Données sociétés vie groupe G - nuage des individus**

Le taux de 90,57% résulte d'une inertie de 76,06 % pour le premier axe et 14,51% pour le second. Ce **taux est excellent** dans le cadre d'une ACP.

Les **valeurs propres**, le **pourcentage d'inertie** associé à chaque axe et le **pourcentage cumulé** sont présentés en **annexe** (Tableau 84: Données sociétés vie groupe G - valeurs propres et pourcentage d'inertie). **L'éboulis des valeurs propres** est également présenté en **annexe** (Figure 28: Données sociétés vie du groupe G - éboulis des valeurs propres).

**L'axe 1** oppose la **compagnie A5** à toutes les **autres compagnies du groupe**. Le Tableau 85: Données sociétés vie groupe G - coordonnées et qualité de représentation des sociétés permet de constater que l'assureur A5 est très bien représenté sur l'axe 1 (coordonnées : -9,11, projection : 1, contribution : 82,64%) par rapport aux autres compagnies.

**L'assureur A5** est cependant mal représenté sur **l'axe 2** (coordonnées : 0,03, projection : 0, contribution : 0,01%). Cet axe oppose les **compagnies A37 et A39** aux **compagnies A30, A38 et A40**.

L'étude du nuage des individus est soutenue par celle du nuage des variables réalisée dans les paragraphes ci-dessous.

## **2.2) Etude du nuage des variables à travers une ACP**

La **représentation** ci dessous permet de voir comment les **19 variables quantitatives** sont corrélées aux **deux variables de synthèses** retenues pour la représentation des **6 compagnies d'assurance vie du groupe** dans le plan principal.

### Variables factor map (PCA)

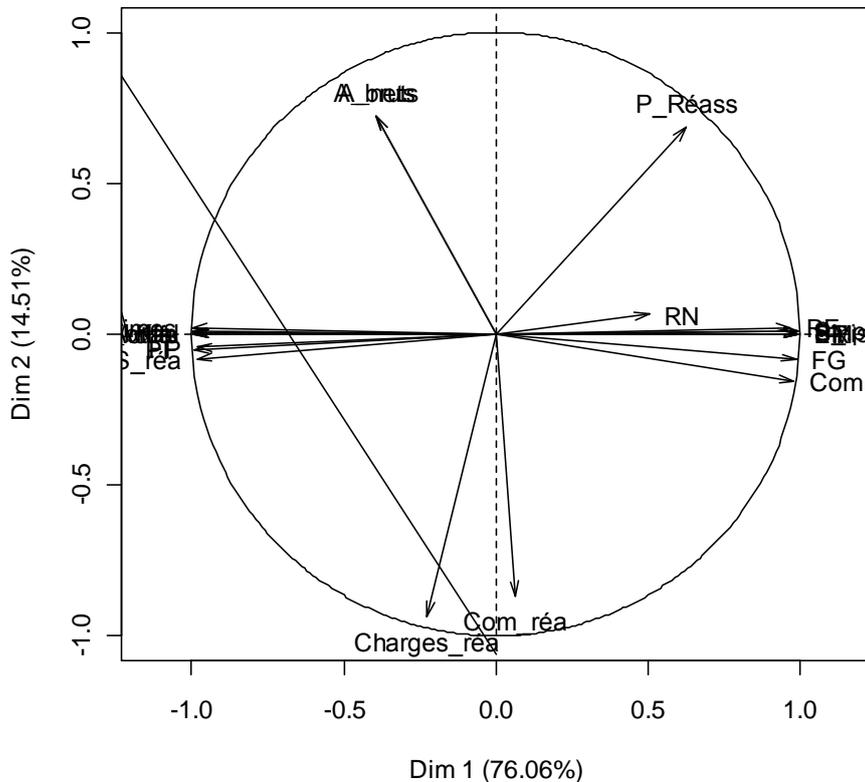


Figure 11: Données sociétés vie groupe G - nuage des variables

La **corrélation**, l'**angle** (cosinus carré) et la **contribution** de chaque variable initiale par rapport aux **4 premiers axes** sont fournis en annexe (Tableau 86: Données sociétés vie groupe G - corrélation et qualité de représentation des variables) et permettent une analyse plus détaillée.

Toutes les **variables sont bien représentées** dans le **plan principal** à l'**exception** de la variable **résultat net (RN)**. Cette variable contribue à **80,69%** à la construction du **4<sup>ème</sup> axe**, non représenté ici. Le 4<sup>ème</sup> axe représente 4,10% de l'inertie totale. La représentativité est donc faible. Il permet cependant d'attirer l'attention sur l'**absence de corrélation** entre le **résultat net** par rapport au **résultat d'exploitation**.

Le **1<sup>er</sup> axe** représente le **niveau d'activité** des compagnies. Les **opérations de réassurance et les arriérés** contribuent pour **98,5%** à la construction de l'**axe 2**.

A la suite de l'ACP réalisée sur les données des sociétés vie du groupe, une CAH peut être mise en œuvre pour avoir un **autre angle d'analyse** sur les mêmes données.

### 2.3) Confirmation des résultats par une CAH

La forme du dendrogramme et le niveau de coupure optimal calculé par la fonction HCPC suggère une **partition** des sociétés d'assurance vie du groupe **en 3 classes**.

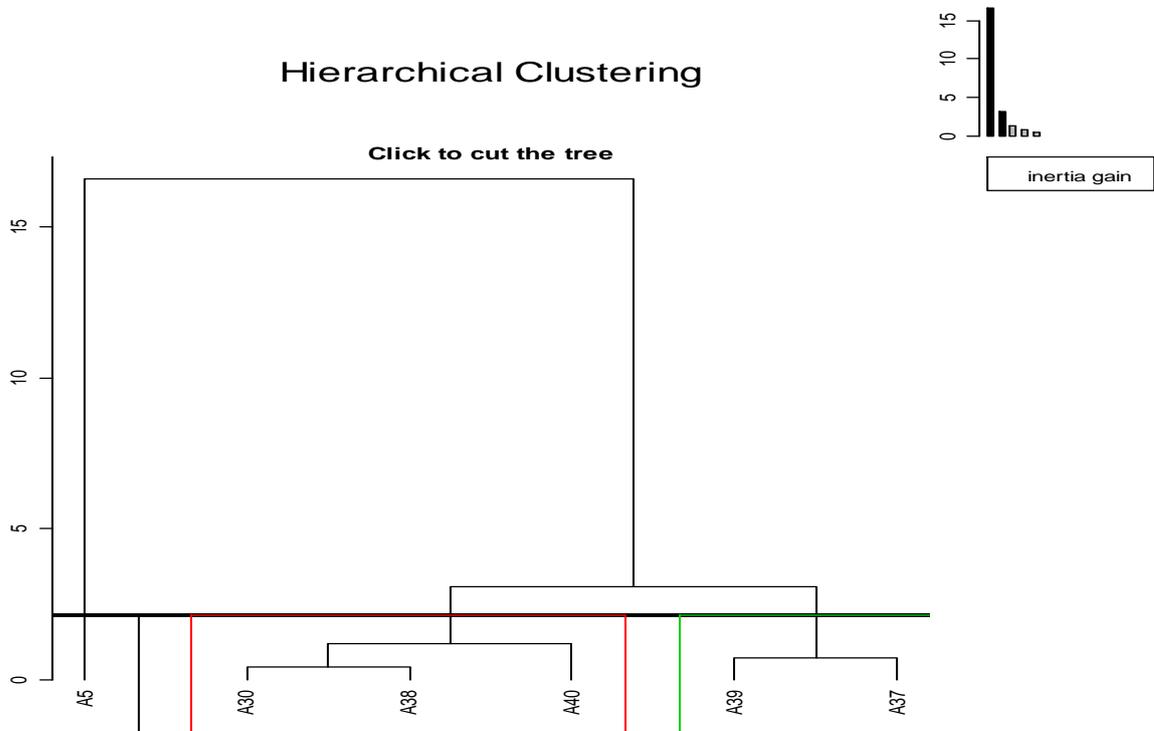


Figure 12: Données sociétés vie du groupe G - arbre hiérarchique

Les **3 classes** peuvent être mises en évidence et représentées sur le **plan principal**. Cette **représentation** est fournie en **annexe** (Figure 29: Données sociétés vie du groupe G - représentation de la partition sur le plan principal). Le **dendrogramme** ci-dessus peut également faire l'objet de **représentation en 3 dimensions sur le plan principal**. Cette **représentation** est également fournie en **annexe** (Figure 30: Données sociétés vie groupe G - représentation du dendrogramme en trois dimensions dans le plan principal).



L'**ACP** et la **CAH** réalisées sur le groupe G mettent en évidence une **hétérogénéité au sein du groupe**. La majorité des sociétés du groupe peine à atteindre la **taille critique nécessaire à une bonne mutualisation des risques d'assurance**. Cette situation est plus **alarmante** au niveau **des sociétés vie** où une seule émerge. Au niveau de l'activité non vie du groupe, des sociétés semblent avoir des **politiques de suivi et de recouvrement des arriérés de primes complètement opposées**.

Les **bonnes pratiques** en la matière enregistrées au niveau de A5 ne semblent **pas diffusées** et mises en œuvre dans les autres entités du groupe.

La **représentation synoptique** du groupe obtenue et les particularités mises en évidence peuvent être plus **intelligibles** pour les **dirigeants** et le **lecteur profane** des états financiers que les **comptes individuels** et **consolidés du groupe**.



Les travaux effectués dans le cadre de l'ACP et de la CAH peuvent être résumés dans le tableau suivant :

Méthodes	Données	Objectifs	Résultats obtenus	Autres objectifs possibles
ACP	Données comptables et financières extraites des comptes annuels des compagnies d'assurance vie (22 variables) et non vie (19 variables) de la Conférence Interafricaine des Marchés d'Assurance	Explorer sans a priori un tableau de données en vue d'en tirer des enseignements utiles à la prise de décision. Obtenir une représentation synoptique des acteurs du marché. Identifier les principales caractéristiques du secteur et mettre en évidence des particularités, sources de risques.	Résultat satisfaisant. Deux variables de synthèse permettent de résumer à plus de 72% l'essentiel de l'information contenue dans les tableaux initiaux. Le marché est dans son ensemble très atomisé avec de nombreux petits acteurs qui peinent à atteindre la taille critique nécessaire à une meilleure couverture des frais fixes et une mutualisation des risques d'assurance. Les assureurs non vie sont handicapés par le niveau élevé des arriérés de primes	Reproduire l'exercice d'année en année et examiner comment la représentation (carte ou dendrogramme) des individus évolue. Enrichir à partir de cette méthode le rapport annuel produit sur le marché.
CAH		Mise en évidence d'une hétérogénéité avec 3 classes au niveau des compagnies d'assurance non vie et 6 classes au niveau des compagnies d'assurance vie.		

Tableau 17: Synthèse des travaux effectués dans le cadre de l'ACP et de la CAH

Dans le cadre de l'ACP, l'information relative aux **90 sociétés d'assurance non vie** de l'espace CIMA mesurées à travers **19 variables** est **synthétisée à 72,77% par deux variables** permettant une représentation synoptique de ces sociétés dans un plan. Pour les **45 sociétés d'assurance vie**, le **plan principal** contient **72,71%** de l'information initiale mesurée à travers **22 variables d'origine**.

L'ACP et la CAH ont permis de mettre en évidence les **caractéristiques principales** des compagnies d'assurance vie et non vie de l'espace CIMA (**absence de taille critique, non maîtrise des arriérés de primes**) et d'en effectuer une classification en groupes relativement homogènes. Des **axes non représentés** mettent en évidence des **particularités** au niveau de certaines compagnies d'assurance : **résultat hors exploitation élevé, importance des cessions en réassurances...** . Ces particularités, sources de **risques potentiels**, doivent faire l'objet de **vigilance et de diligence accrues** dans le cadre des missions mises en œuvre par un professionnel comptable et les autorités de contrôle de la CIMA.

**TROISIEME PARTIE : MISE EN ŒUVRE  
DES MODELES PREDICTIFS SUR LES  
DONNEES COMPTABLES ET  
FINANCIERES DES COMPAGNIES  
D'ASSURANCE DE L'ESPACE CIMA**

# TROISIEME PARTIE : MISE EN ŒUVRE DES MODELES PREDICTIFS SUR LES DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES D'ASSURANCE DE L'ESPACE CIMA

Dans cette **dernière partie** du mémoire, la **régression linéaire multiple**, l'**analyse discriminante** et la **régression logistique** sont mises en œuvre dans un **objectif décisionnel** et de **prédiction**.

La **régression linéaire multiple** est utilisée pour **estimer les provisions techniques** des compagnies d'assurance. Pour les compagnies d'assurance non vie, les deux provisions techniques objet de l'étude sont les provisions pour sinistres à payer et les provisions pour annulations de primes. Dans le cas des compagnies d'assurance vie, l'étude portera sur les provisions mathématiques.

L'**analyse discriminante** est mise en œuvre pour **discriminer les entreprises saines des entreprises en situation précaire** faisant l'objet de suivi par la CRCA et des **entreprises en faillite** dont les agréments devraient être retirés.

La **régression logistique** est utilisée pour estimer la **probabilité que les comptes d'une entreprise contiennent des anomalies significatives**. Cette étude part d'un échantillon de sociétés ayant fait l'objet de contrôle, construit un modèle et l'applique aux autres sociétés non contrôlées.

Cette troisième partie du mémoire s'appuie essentiellement sur l'ouvrage suivant :

CORNILLON Pierre-André et autres, Statistiques avec R, Presses Universitaires de Rennes (PUR), 3ème édition, 2012, 296 p.

La qualité, le succès et l'intérêt de ce livre, qui est en quelques années à sa 3<sup>ème</sup> édition, ne se démentent pas. Certains de ces auteurs développent des bibliothèques sur .

## CHAPITRE I. REGRESSION LINEAIRE MULTIPLE

Les différents contrôles effectués par les commissaires contrôleurs de la CIMA sur les entreprises d'assurance permettent de constater le **manque de fiabilité** des **comptes** arrêtés.

Les **ajustements** les plus importants portent généralement pour les compagnies non vie sur les **provisions pour sinistres à payer et les provisions pour annulations de primes**. Pour les entités vie, les ajustements les plus importants portent sur les **provisions mathématiques**.

Une **sous évaluation** de ces provisions permet aux compagnies de présenter une situation financière **en apparence conforme à la réglementation** alors qu'en réalité les ratios minimum de couverture des engagements réglementés et de marge de solvabilité ne sont pas respectés.

La **norme ISA 540** et la **NEP 540** précisent que pour réunir des éléments probants sur le caractère raisonnable d'une **estimation comptable**, le professionnel comptable peut utiliser une **estimation indépendante** pour la comparer avec celle de la direction.

L'approche développée dans les paragraphes suivants vise donc à mettre en place un **outil** simple d'utilisation permettant au professionnel comptable de faire une **estimation indépendante** des **principales provisions** constituées par une compagnie d'assurance afin de vérifier la vraisemblance des évaluations effectuées par la direction.

### 1) Régression linéaire sur les données des compagnies d'assurance non vie

Dans le cadre de la **régression multiple** une étape principale consiste à **estimer les paramètres du modèle** à partir des données disponibles. Ce **modèle** est généralement **complexe** car il fait intervenir toutes les variables. Il faut donc construire un **autre modèle** en **choisissant les variables les plus pertinentes**. La qualité de ce deuxième modèle ainsi construit est ensuite testée par **l'analyse des résidus**. Si le **modèle est satisfaisant**, il peut être enfin utilisé dans le cadre d'un **exercice de prédiction**.

### 1.1) Estimation des paramètres du modèle

Dans le cadre des sociétés d'assurance non vie et en fonction des anomalies identifiées par les commissaires contrôleurs, **deux variables à expliquer** seront étudiées. Le **1<sup>er</sup> modèle** aura comme **variable à expliquer** la **provision pour sinistre à payer « PSAP »**. Le **2<sup>nd</sup> modèle** aura comme **variable à expliquer** la **provision pour annulations de primes « PAP »**.

#### **Modèle avec « PSAP » comme variable à expliquer**

La fonction `lm` permet d'ajuster un modèle linéaire. La **régression de la variable à expliquer PSAP** en fonction des **autres variables explicatives** fournit les résultats suivants :

```
Call:
lm(formula = PSAP ~ ., data = ADREG)

Residuals:
    Min       1Q   Median       3Q      Max
-2436684 -374206   34149   408386 1501983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.265e+05  3.050e+05  -0.743  0.4641
PRIMES       5.697e-01  6.652e-01   0.856  0.3993
SIN_REG     -2.497e-01  1.069e+00  -0.234  0.8170
AR_BRUTS    4.248e-01  1.829e-01  2.323  0.0280 *
AR_NETS    -6.348e-01  3.184e-01  -1.994  0.0563 .
PF_NETS     -9.754e-01  1.280e+00  -0.762  0.4525
COM         1.567e+00  9.814e-01   1.597  0.1220
FG         1.287e+00  6.802e-01   1.892  0.0693 .
PRI_REA     1.507e-01  5.848e-01   0.258  0.7987
SIN_REA    -1.594e+00  1.341e+00  -1.189  0.2447
SAP_REA   -6.971e-01  4.015e-01  -1.736  0.0939 .
COM_REA     -6.189e-01  2.729e+00  -0.227  0.8223
PAP         -4.562e-01  6.977e-01  -0.654  0.5187
CAP         1.487e-01  1.767e-01   0.842  0.4074
R_EXP       4.288e-02  7.092e-01   0.060  0.9522
RHE        -3.007e-01  1.169e+00  -0.257  0.7990
ACTIFS_A  -5.327e-01  8.202e-02  -6.494  5.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 890300 on 27 degrees of freedom
Multiple R-squared:  0.9688,    Adjusted R-squared:  0.9503
F-statistic: 52.43 on 16 and 27 DF,  p-value: 3.607e-16
```

Tableau 18: Données sociétés non vie - régression avec PSAP comme variable à expliquer

La fonction de régression linéaire utilisée `lm` retourne une **liste d'objets et d'informations**. Les objets et les informations affichés dans un premier temps sur la console sont les suivants :

- L'objet `call` contient la **fonction** et la **formule** de régression linéaire utilisées,

- **L'objet `residual`** contient tous les **résidus**  $\hat{\varepsilon}_i = y_i - \hat{y}_i$  calculés après avoir déterminé les paramètres du modèle. Cependant, la fonction `summary` permet d'avoir pour les résidus le minimum, le 1<sup>er</sup> quartile, la médiane, le 2<sup>ème</sup> quartile et le maximum.
- **L'objet `coefficients`** se présente sous la forme d'une **matrice**. Pour chaque **paramètre**, **4 colonnes** permettent d'avoir les **informations suivantes** :
  - **L'estimation du paramètre** (colonne `Estimate`) par la méthode des moindres carrés,
  - **L'estimation de son écart type** (colonne `std. Error`),
  - La **valeur observée** de la **statistique de test d'hypothèse**  $H_0: \beta_i = 0$  contre  $H_1: \beta_i \neq 0$  (colonne `t value`). Cette valeur est obtenue en divisant chaque paramètre par son écart type.
  - La **probabilité critique** ( $\Pr(> |t|)$ ) donnant pour la **statistique de test de student sous  $H_0$** , la probabilité de dépasser la valeur estimée (colonne `Pr(>|t|)`). Il s'agit d'un test bilatéral.
- La **codification des tests** (`Signif. codes`) se révélant significatifs est fournie **sous la matrice des coefficients**. Lorsqu'elles sont testées une par une, les variables significatives sont les actifs admis en couverture des engagements réglementés `ACTIFS_A ***`, les arriérés bruts `AR_BRUTS *`, les arriérés nets `AR_NETS .`, les frais généraux `FG .` et les provisions pour sinistres à payer cédées aux réassureurs `SAP_REA..`. Cependant, la régression étant multiple et les variables explicatives non orthogonales, ces **tests doivent être utilisés prudemment**. En effet, le test sur un coefficient dans le modèle revient à tester la significativité d'une variable alors que les autres variables sont dans le modèle. Il importe donc d'utiliser **d'autres procédures de choix de modèles**.
- Le **résumé de l'étape d'estimation** fournit l'**erreur type résiduel**  $\sigma$  égale à 890300, le nombre de **degrés de liberté** associé : 27 `degrees of freedom` ( $n - k - 1 = (44 - 16 - 1)$ ).
- Le **coefficient de détermination** permettant de juger de la qualité de l'ajustement du modèle est fourni. Ce coefficient est le suivant : `Multiple R-`

squared: 0.9688. Pour mémoire, ce coefficient est obtenu par la formule suivante :  $R^2 = \frac{SCE_M}{SCE_T} = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2}$ . Ce coefficient mesure la proportion de la variance de  $Y$  expliquée par la régression de  $X$ . Plus il est proche de 1, meilleure est la qualité d'ajustement global du modèle. La statistique  $\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right) \left(\frac{n-1}{n-k-1}\right)$  est aussi fournie. Elle est la suivante : Adjusted R-squared: 0.9503.

- La **dernière ligne** présente le résultat du **test de comparaison** entre le **modèle utilisé** et le **modèle n'utilisant que la constance comme variable explicative**. Le test est significatif car les variables explicatives apportent de l'information sur la variable à expliquer. La statistique  $F^*$  suit une loi de Fisher à  $k$  et  $(n - k - 1)$  degrés de liberté. La comparaison de  $F^*$  au  $F$  théorique permet de rejeter l'hypothèse  $H_0$  si  $F^* > F$ . Le rejet de l'hypothèse  $H_0$  permet de conclure que le modèle est globalement satisfaisant. Dans l'étude effectuée, la statistique de fisher se présente comme suit : F-statistic: 52.43 on 16 and 27 DF, p-value: 3.607e-16. On peut lire dans des tables de fisher les valeurs suivantes :  $F_{0,01}(16; 27) = 2,75$  et  $F_{0,05}(16; 27) = 2,4$ . L'hypothèse  $H_0$  est donc rejetée car la valeur de la statistique de test de fisher est très élevée et la probabilité que  $F_\alpha(16; 27)$  dépasse 52,43 est pratiquement nulle.

### Modèle avec « PAP » comme variable à expliquer

Avec la fonction `lm`, la régression de la **variable à expliquer** PAP en fonction des **autres variables explicatives** fournit les résultats suivants :

```
Call:
lm(formula = PAP ~ ., data = ADREG)

Residuals:
    Min       1Q   Median       3Q      Max
-527870 -128599   4752  125836  525093

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.584e+05  9.093e+04   1.742  0.092423 .
PRIMES      -1.087e-01  1.080e-01  -1.007  0.322678
SIN_REG     -4.340e-01  1.811e-01  -2.397  0.023464 *
AR_BRUTS    3.874e-02  6.006e-02   0.645  0.524208
AR_NETS     -3.469e-01  8.368e-02  -4.145  0.000284 ***
PF_NETS     -3.079e-01  3.972e-01  -0.775  0.444802
COM         5.273e-01  2.727e-01   1.933  0.063348 .
FG          2.976e-02  1.650e-01   0.180  0.858194
PRI_REA     1.132e-01  1.663e-01   0.681  0.501666
SIN_REA     -1.123e-01  3.047e-01  -0.369  0.715188
SAP_REA     -1.370e-01  1.055e-01  -1.299  0.204692
COM_REA     1.441e+00  6.809e-01   2.116  0.043413 *
PSAP        -4.800e-02  5.931e-02  -0.809  0.425148
CAP         5.001e-02  5.204e-02   0.961  0.344713
```

R_EXP	1.380e-01	1.558e-01	0.886	0.383055
RHE	2.420e-01	3.591e-01	0.674	0.505818
ACTIFS_A	1.166e-03	3.995e-02	0.029	0.976931
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 277600 on 28 degrees of freedom				
Multiple R-squared: 0.8606, Adjusted R-squared: 0.7809				
F-statistic: 10.8 on 16 and 28 DF, p-value: 4.603e-08				

**Tableau 19: Données sociétés non vie - régression avec PAP comme variable à expliquer**

Lorsqu'elles sont testées une par une, les variables significatives sont les arriérés nets `AR_NETS` \*\*\*, les sinistres réglés `SIN_REG` \*, les commissions de réassurance `COM_REA` \* et les commissions `COM`.

Les **modèles construits ci-dessus** avec toutes les variables explicatives sont relativement **complexes**. Des **procédures** permettent de sélectionner les variables pertinentes afin de **réduire la complexité du modèle**.

### 1.2) Choix des variables

Le **choix des variables** peut se faire **pas à pas**: on enlève la variable la moins significative, soit le résultat d'exploitation `R_EXP`, puis on recalcule les estimations et ainsi de suite. Le logiciel  possède cependant un **package** `leaps` traitant du choix automatique des variables. La **fonction** `regsubsets` de ce package retourne, pour différents critères, le **meilleur modèle** à 1 variable explicative, 2 variables explicatives, ... à `nvmax` variables explicatives.

#### ***Modèle avec « PSAP » comme variable à expliquer***

Pour le **modèle avec « PSAP » comme variable à expliquer**, la **représentation graphique** suivante permet d'analyser les résultats de la fonction `regsubsets` utilisée pour déterminer le **meilleur modèle** pour `nvmax=17` variables explicatives :

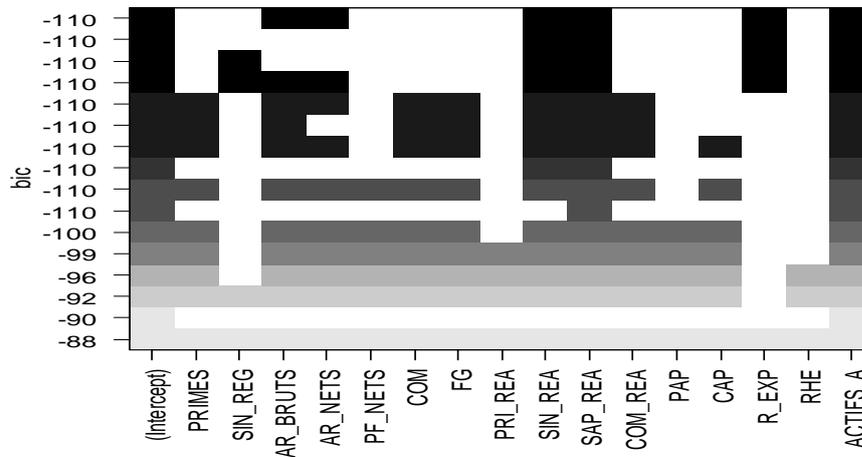


Figure 13: Données sociétés non vie - modèle PSAP - choix des variables avec BIC

Le critère est optimum pour la ligne en haut du graphique. Pour le critère BIC, le modèle avec les 6 variables AR\_BRUTS+AR\_NETS+SIN\_REA+SAP\_REA+R\_EXP+ACTIFS\_A, peut être retenu. L'ajustement du modèle avec les 6 variables sélectionnées donne les résultats suivants :

```
Call:
lm(formula = PSAP ~ AR_BRUTS + AR_NETS + SIN_REA + SAP_REA +
R_EXP + ACTIFS_A, data = ADREG)

Residuals

Min       1Q   Median       3Q      Max
-3046091 -323531   67944   513724  1281384

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.199e+05  2.052e+05  -2.046  0.047917 *
AR_BRUTS     3.760e-01  1.525e-01   2.465  0.018471 *
AR_NETS     -4.446e-01  1.682e-01  -2.643  0.011965 *
SIN_REA    -1.340e+00  3.582e-01  -3.741  0.000619 ***
SAP_REA     -7.020e-01  2.063e-01  -3.403  0.001614 **
R_EXP        1.041e+00  3.015e-01   3.451  0.001412 **
ACTIFS_A    -5.782e-01  4.961e-02 -11.655  6.02e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 866600 on 37 degrees of freedom
Multiple R-squared:  0.9595,    Adjusted R-squared:  0.9529
F-statistic: 146.1 on 6 and 37 DF,  p-value: < 2.2e-16
```

Tableau 20: Données sociétés non vie - régression ajustée avec PSAP comme variable à expliquer

La fonction regsubsets a permis de passer de 16 variables explicatives à 6 variables explicatives tout en gardant le même pouvoir explicatif du modèle. Pour le modèle à 16 variables explicatives, le coefficient de détermination R-squared: 0.9688 est comparable à celui obtenu pour le modèle à 6 variables R-squared: 0.9595.

### Modèle avec « PAP » comme variable à expliquer

Pour le modèle avec « PAP » comme variable à expliquer, la représentation graphique suivante permet d'analyser les résultats de la fonction `regsubsets` utilisée pour déterminer le meilleur modèle pour  $n_{vmax}=17$  variables explicatives:

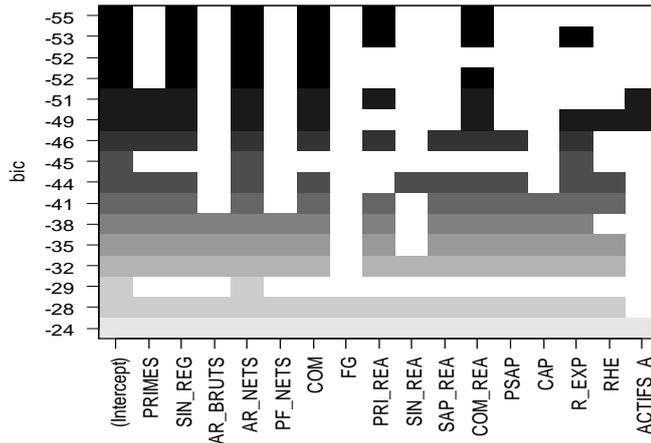


Figure 14: Données sociétés non vie - modèle PAP - choix de variables avec BIC

Le critère est optimum pour la ligne en haut du graphique. Pour le critère BIC, le modèle avec les 5 variables `SIN_REG+AR_NETS+COM+PRI_REA+COM_REA`, peut être retenu. L'ajustement du modèle avec les 5 variables sélectionnées donne les résultats suivants :

```
Call:
lm(formula = PAP ~ SIN_REG + AR_NETS + COM + PRI_REA + COM_REA,
    data = ADREG)

Residuals:
    Min       1Q   Median       3Q      Max
-590232.3 -107320.6  -640.4  169121.2  668022.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.728e+05  6.381e+04   2.708  0.00999 **
SIN_REG      -3.647e-01  5.812e-02  -6.275  2.15e-07 ***
AR_NETS      -3.185e-01  2.787e-02 -11.431  5.08e-14 ***
COM           7.952e-01  1.344e-01   5.915  6.81e-07 ***
PRI_REA       2.484e-01  1.010e-01   2.458  0.01851 *
COM_REA       1.618e+00  5.708e-01   2.834  0.00725 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 266500 on 39 degrees of freedom
Multiple R-squared:  0.821,    Adjusted R-squared:  0.798
F-statistic: 35.77 on 5 and 39 DF,  p-value: 1.453e-13
```

Tableau 21: Données sociétés non vie - régression ajustée avec PAP comme variable à expliquer

La fonction `regsubsets` a permis de passer de **16 variables explicatives à 5 variables explicatives** tout en gardant le **même pouvoir explicatif du modèle**. Pour le modèle à 16 variables explicatives, le coefficient de détermination  $R\text{-squared: } 0.8606$  est comparable à celui obtenu pour le modèle à 6 variables  $R\text{-squared: } 0.821$ .

La construction de modèles de régression linéaire repose sur plusieurs hypothèses présentées en annexe (**Hypothèses du modèle linéaire multiple page 130**). **L'analyse des résidus** constitue une des principales méthodes permettant de s'assurer du respect de ces hypothèses.

### 1.3) Analyse des résidus

Les **résidus** sont obtenus par la **fonction** `residuals`. Ces résidus ne sont pas de même variance (hétéroscédasticité). Une **transformation** est effectuée sur ces résidus pour obtenir des **résidus « studentisés »** ayant la même variance. La **figure** ci-dessous permet une **visualisation** de ces résidus « studentisés ».

#### ***Modèle avec « PSAP » comme variable à expliquer***

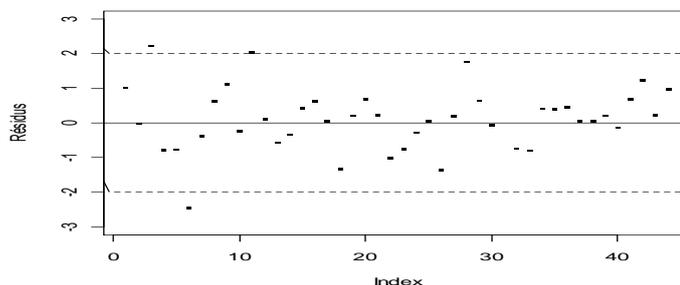


Figure 15: Données sociétés non vie - modèle PSAP - représentation des résidus

L'**analyse des résidus** est la **principale méthode** utilisée pour valider les **hypothèses** associées à un **modèle de régression**. Si les **tests statistiques** montrent le **caractère significatif des paramètres du modèle** et si le **coefficient de détermination** est **élevé**, il est alors possible de développer de **bonnes estimations** en utilisant l'équation estimée de la régression.

Si l'analyse des résidus indique qu'une ou plusieurs **hypothèses sont contestables**, il convient d'être **prudent** dans l'**interprétation des résultats** de la régression.

La standardisation des résidus est obtenue en divisant chaque résidu par son écart type. Ce traitement permet de valider l'hypothèse de normalité des erreurs en

s'assurant que **95%** des **résidus standardisés** sont compris dans l'**intervalle**  $[-2, 2]$ . C'est ici le cas où **deux résidus sur 45** se trouvent à l'extérieur de cet intervalle.

### **Modèle avec « PAP » comme variable à expliquer**

Avec le modèle ayant la « PAP » comme variable à expliquer, on constate que **3 résidus sur 45** ne sont pas compris dans l'intervalle  $[-2, 2]$ . La figure suivante permet d'illustrer ce constat :

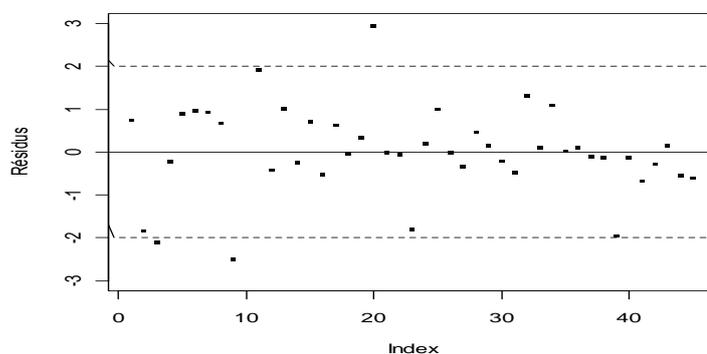


Figure 16: Données sociétés non vie - modèle PAP - représentation des résidus

**Aucune autocorrélation** n'est mise en évidence à travers le **nuage des résidus**. Les hypothèses associées au modèle de régression ne semblent pas remises en cause.

Les modèles construits et validés peuvent être utilisés dans le cadre **d'exercices de prédiction** et pour faire des **estimations**.

#### **1.4) Intérêt pour le professionnel comptable et une autorité de supervision**

La **correcte évaluation** des **provisions techniques** est essentielle pour juger de la **solvabilité** d'une compagnie d'assurance. Pour les compagnies d'assurance non vie de la zone CIMA, les provisions pour sinistres à payer (PSAP) et les provisions pour annulations de primes (PAP) sont les deux principales provisions techniques pour lesquelles des sous évaluations sont couramment rencontrées. Elles sont des **estimations comptables** au sens de la **norme ISA 540** et de la **NEP 540**.

En conformité avec ces normes relatives à l'appréciation des estimations comptables, le professionnel comptable peut **utiliser un modèle de régression**

**linéaire pour effectuer une estimation et la comparer avec l'estimation établie par la direction** d'une entreprise.

Les **modèles** sont **construits** avec un **échantillon** de la population composé des compagnies d'assurance non vie ayant un **numéro impair (exemple A1)**. Ils sont par la suite **testés** sur les compagnies ayant un **numéro pair (exemple A2)**. Ils auraient pu également être construits avec toutes les compagnies d'assurance non vie pour un exercice  $n$  et être testés sur ces mêmes compagnies pour l'exercice  $n + 1$ .

### **Modèle avec « PSAP » comme variable à expliquer**

Le « Tableau 89: Données sociétés non vie - prédiction des PSAP » en **annexe** donne pour **chaque compagnie** ayant un numéro pair, **l'estimation effectuée** avec le modèle pour obtenir le montant des PSAP avec un intervalle de confiance à 95%. Un extrait de ce tableau se présente comme suit :

Libellés	Fit <sup>7</sup>	lwr <sup>8</sup>	upr <sup>9</sup>	Test1 <sup>10</sup>	Test2 <sup>11</sup>	PSAP	Ecart	En %
A2	- 22 350 430	- 25 961 018	- 18 739 841	NON	INFERIEUR	-12 267 624	-10 082 806	82%
A6	- 25 040 795	- 30 475 402	- 19 606 187	NON	INFERIEUR	-14 716 191	-10 324 604	70%
A8	- 3 943 728	- 6 960 217	- 927 239	NON	SUPERIEUR	-7 933 254	3 989 526	-50%
A38	- 1 783 167	- 3 604 716	38 382	NON	SUPERIEUR	-4 899 929	3 116 762	-64%

**Tableau 22: Données sociétés non vie - prédiction des PSAP - extrait**

L'extrait permet de constater que sur 45 entreprises, **l'estimation effectuée pour 4 entreprises** n'est **pas comprise dans l'intervalle de confiance défini**. La classification ascendante hiérarchique a montré que les assureurs **A2 et A6** présentent des **caractéristiques exceptionnelles** par rapport au reste de la population. Il est généralement recommandé **d'exclure les individus extrêmes** avant de construire un modèle car ils **influent négativement** sur la **robustesse du modèle construit**.

Les assureurs **A8 et A38** ont dans leur bilan des PSAP largement supérieures en valeurs absolues aux **provisions prédites par le modèle**. Ce constat pourrait permettre de conclure que les provisions constituées par ces deux entités sont

<sup>7</sup> Fit : estimation effectuée à partir du modèle retenu

<sup>8</sup> Lwr : lower pour la valeur de borne inférieure de l'intervalle

<sup>9</sup> Upr : upper pour la valeur de la borne supérieure de l'intervalle

<sup>10</sup> Test1 : test permettant de savoir si la valeur affichée par la société est comprise dans l'intervalle estimé par le modèle, la réponse est binaire (oui/non).

<sup>11</sup> Test2 : test permettant de comparer la valeur estimée par le modèle de celle de la société, la réponse est binaire (inférieur/supérieur)

prudentes. Une analyse plus poussée lors des contrôles opérés dans ces deux compagnies d'assurance permet de faire les constats suivants :

- L'assureur **A38** est une mutuelle ayant **transféré l'essentiel de son portefeuille** à une société anonyme. Il gère essentiellement le run off (liquidation du passé) mais ne manifeste **aucun empressement à indemniser effectivement les assurés**. De nombreux sinistres sont **prescrits** mais **maintenus dans le stock de PSAP**.
- L'assureur **A8** a une **organisation inadaptée** ne lui permettant pas de procéder à un **correct inventaire de ses PSAP**. Par ailleurs, il s'est forgé la **réputation** de ne **pas honorer ses engagements** envers les assurés et bénéficiaires de contrats en cas de sinistres. Enfin, comparativement aux assureurs A2 et A6, l'assureur A8 est parmi les entreprises ayant le **niveau d'arriérés** le plus **élevé** du marché. **L'insuffisance de trésorerie** a une **incidence nette** sur le **paiement des sinistres et le niveau des PSAP**.

Au delà de ces **4 cas extrêmes**, on constate que sur les 45 assureurs, une nette majorité de **29 compagnies** d'assurance présente dans leur bilan un montant de **PSAP inférieur** à celui **prédit par le modèle**, avec parfois des **écarts significatifs**.

Au niveau d'une autorité de supervision, un **critère** pour sélectionner les entreprises à contrôler serait de choisir en **priorité les assureurs** dont le montant des **PSAP** au **bilan** présente un **écart significatif** par rapport au montant des **PSAP** estimé à partir du **modèle**.

Le professionnel comptable qui met en œuvre une telle technique d'estimation doit par la suite **analyser de façon plus approfondie les écarts** mis en évidence notamment lorsque ceux-ci sont significatifs.

La démarche mise en œuvre pour l'estimation des « PSAP » est reproduite dans les paragraphes suivants pour l'estimation des « PAP ».

#### ***Modèle avec « PAP » comme variable à expliquer***

Le « Tableau 90: Données sociétés non vie - prédiction des PAP » en **annexe** donne pour chaque **compagnie** ayant un **numéro pair**, **l'estimation effectuée** avec le modèle pour obtenir le montant des PAP avec un intervalle de confiance à 95%. Un extrait de ce tableau se présente comme suit :

Libellés	fit	lwr	upr	Test1	Test2	PAP	Ecart	En %
A8	- 2 279 195	- 2 935 013	- 1 623 377	NON	INFERIEUR	- 702 216	- 1 576 979	225%
A14	- 970 289	- 1 539 382	- 401 196	NON	INFERIEUR	- 98 669	- 871 620	883%
A24	- 766 996	- 1 325 877	- 208 115	NON	INFERIEUR	- 146 697	- 620 299	423%
A52	- 999 463	- 1 572 259	- 426 668	NON	INFERIEUR	-	- 999 463	
A68	- 1 802 284	- 2 428 434	- 1 176 134	NON	INFERIEUR	- 663 912	- 1 138 372	171%
A2	- 51 108	- 1 002 583	900 367	NON	SUPERIEUR	- 2 141 000	2 089 892	-98%
A6	2 633 636	1 498 592	3 768 681	NON	SUPERIEUR	- 433 174	3 066 810	-708%
A40	- 140 484	- 696 675	415 708	NON	SUPERIEUR	- 1 100 000	959 516	-87%
A56	- 1 294 236	- 1 880 339	- 708 133	NON	SUPERIEUR	- 2 863 107	1 568 871	-55%

**Tableau 23: Données sociétés non vie - prédiction des PAP - extrait**

L'extrait permet de constater que sur 45 entreprises, **l'estimation effectuée pour 9 entreprises n'est pas comprises dans l'intervalle de confiance défini**. Les provisions pour annulations de primes constituées pour les assureurs **A8, A14, A24 et A52** sont **largement inférieures à celles évaluées par le modèle**. Ces sociétés présentent un niveau d'arriérés de primes particulièrement élevé comparativement à leur chiffre d'affaires et à la provision pour annulations de primes constituée. Une **diligence particulière** devrait être mise en œuvre sur ce point en cas de contrôle ou d'audit de ces entités.

Les assureurs **A2, A40 et A56** ont dans leur bilan des **PAP supérieures** en valeurs absolues aux **provisions prédites par le modèle**. Ce constat pourrait permettre de conclure que les provisions constituées par ces trois entités sont prudentes. Cependant une **analyse plus poussée** est nécessaire dans le cadre d'un **contrôle sur place ou d'une mission d'audit**.

La démarche mise en œuvre pour l'estimation des « PSAP » et des « PAP » pour les compagnies d'assurance non vie peut être reproduite au titre des « PM » pour les compagnies d'assurance vie.

## **2) Régression linéaire sur les données des compagnies d'assurance vie**

Comme pour les compagnies d'assurance non vie, un premier **modèle** relativement **complexe** est construit à partir de **toutes les variables**. Un second modèle relativement **moins complexe** est construit en choisissant les **variables les plus pertinentes**. La **qualité de ce deuxième modèle** ainsi construit est ensuite testée par **l'analyse des résidus**. Ce modèle est ensuite utilisé dans le cadre d'un **exercice de prédiction et d'estimation**.

## 2.1) Estimation des paramètres du modèle

Pour les compagnies d'assurance vie, la régression de la **variable à expliquer PM** en **fonction des autres variables explicatives** fournit les résultats suivants :

```
Call:
lm(formula = PM ~ ., data = ADREG)

Residuals:
    Min       1Q   Median       3Q      Max
-579727623 -72285323 -15813176  75489295 455878284

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.393e+08  2.578e+08   1.316  0.2362
Primes      -9.216e-01  3.172e-01  -2.906  0.0271 *
PF          -4.642e+00  2.145e+00  -2.164  0.0737 .
Com         -5.461e+00  1.804e+00  -3.027  0.0232 *
FG           1.920e+00  9.833e-01   1.953  0.0987 .
P_Réass     -6.134e+00  2.946e+00  -2.083  0.0824 .
Charges_réa  1.972e+00  4.417e+00   0.446  0.6709
RE          -2.444e-01  1.876e+00  -0.130  0.9006
RN          -7.797e-01  1.122e+00  -0.695  0.5130
S_.payés    4.260e-01  2.437e-01   1.748  0.1310
S_réa       -5.574e+00  4.886e+00  -1.141  0.2975
Com_réa     -2.464e+01  1.030e+01  -2.392  0.0539 .
A_bruts     1.552e+00  2.435e+00   0.637  0.5474
A_nets     -1.503e+00  2.391e+00  -0.629  0.5526
Ktal        8.405e-02  6.704e-01   0.125  0.9043
FP           2.398e-01  4.184e-01   0.573  0.5874
Actifs     -4.646e-01  1.340e-01  -3.466  0.0134 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 406500000 on 6 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9993
F-statistic: 1957 on 16 and 6 DF,  p-value: 8.424e-10
```

**Tableau 24: Données sociétés vie - régression avec PM comme variable à expliquer**

Lorsqu'elles sont testées une par une, les **variables significatives** sont les actifs Actifs \*, les primes Primes \* , les commissions com \* , etc... . Les limites de ces tests ont été exposées plus haut.

## 2.2) Choix des variables

La **représentation graphique** suivante permet d'analyser les résultats de la **fonction regsubsets** utilisée pour déterminer le **meilleur modèle** pour  $nvmax=17$  variables explicatives :

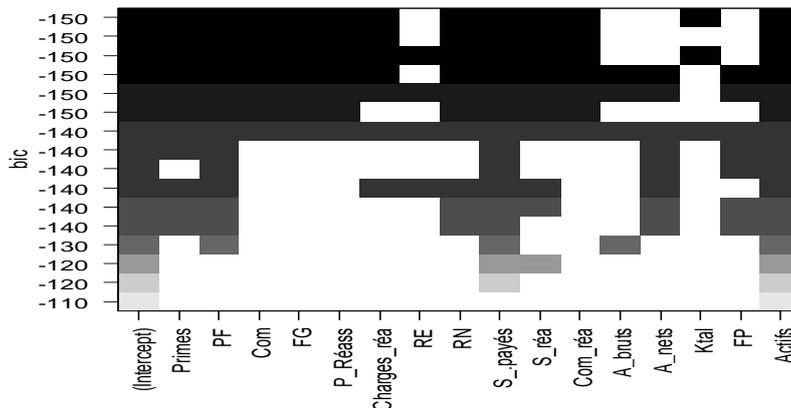


Figure 17 : Données sociétés vie - choix des variables avec BIC

Le critère est optimum pour la ligne en haut du graphique. Pour le critère BIC, le modèle permet d'éliminer les 4 variables suivantes: RE, A\_bruts, A\_nets et FP. Ces 4 variables retirées du jeu de données initial permettent d'aboutir à un nouveau jeu de données à partir duquel la régression est effectuée. L'ajustement du modèle avec les 12 variables restantes donne les résultats suivants :

```
Call:
lm(formula = PM ~ ., data = ADREG_2)

Residuals:
    Min       1Q   Median       3Q      Max
-499571268 -125404661 -21835561  77401991 499113284

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.622e+08  1.973e+08   1.835 0.096312 .
Primes       -9.128e-01  1.665e-01  -5.482 0.000268 ***
PF           -5.159e+00  9.432e-01  -5.470 0.000273 ***
Com          -5.253e+00  1.299e+00  -4.043 0.002350 **
FG            2.309e+00  5.718e-01   4.038 0.002369 **
P_Réass      -6.739e+00  1.692e+00  -3.983 0.002590 **
Charges_réa  3.513e+00  1.881e+00   1.868 0.091345 .
RN           -7.278e-01  1.290e-01  -5.642 0.000215 ***
S_.payés     5.429e-01  1.615e-01   3.361 0.007230 **
S_réa        -7.327e+00  1.827e+00  -4.011 0.002476 **
Com_réa      -2.705e+01  7.400e+00  -3.656 0.004422 **
Ktal         4.285e-01  3.121e-01   1.373 0.199689
Actifs      -4.270e-01  5.088e-02  -8.392 7.72e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 336800000 on 10 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9995
F-statistic: 3802 on 12 and 10 DF,  p-value: < 2.2e-16
```

Tableau 25: Données sociétés vie - régression ajustée avec PM comme variable à expliquer

La fonction `regsubsets` a permis de réduire le nombre de variables explicatives tout en gardant le même pouvoir explicatif du modèle. Pour le modèle à 16 variables

explicatives, le **coefficient de détermination**  $R\text{-squared: } 0.998$  est quasi identique à celui obtenu pour 12 variables.

L'**analyse des résidus** effectuée dans les paragraphes suivants permet de s'assurer du respect des hypothèses de la régression linéaire multiple.

### 2.3) Analyse des résidus

Pour le modèle construit avec la « **PM** » comme **variable à expliquer**, on constate que **tous les résidus sont compris dans l'intervalle  $[-2, 2]$** . La « Figure 18: Données sociétés vie - représentation des résidus » suivante permet d'illustrer ce constat :

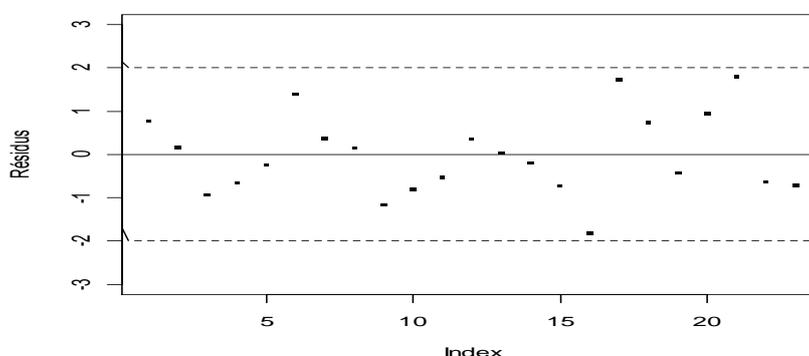


Figure 18: Données sociétés vie - représentation des résidus

L'examen du nuage des résidus ne remet pas en cause les hypothèses associées au modèle de régression.

Les **modèles construits et validés** peuvent être utilisés pour faire de la **prédiction et de l'estimation**.

### 2.4) Intérêt pour le professionnel comptable et pour une autorité de supervision

Les **indications générales** fournies en introduction du point « 1.4) Intérêt pour le professionnel comptable et une autorité de supervision page 86 » restent valables pour l'estimation des PM.

Pour être factuel, le « Tableau 91: Données sociétés vie - prédiction des PM » en **annexe** donne pour chaque compagnie ayant un **numéro pair**, l'**estimation effectuée** avec le **modèle** pour obtenir le **montant des PM** avec un **intervalle de confiance à 95%**.

Un extrait de ce tableau se présente comme suit :

Libellés	Fit	Lwr	Upr	Test1	Test2	PM	Ecart	En %
A4	-37 581 496	-39 733 523	-35 429 469	NON	INFERIEUR	-32 085 330	-5 496 166	17%
A8	-16 032 881	-17 783 196	-14 282 566	NON	INFERIEUR	-11 877 989	-4 154 892	35%
A14	-15 213 212	-17 581 568	-12 844 856	NON	INFERIEUR	-11 527 536	-3 685 676	32%
A16	-8 730 869	-10 154 471	-7 307 268	NON	INFERIEUR	-5 495 963	-3 234 907	59%
A22	-7 049 539	-7 996 808	-6 102 270	NON	INFERIEUR	-5 562 564	-1 486 975	27%
A26	-7 155 891	-8 283 421	-6 028 360	NON	INFERIEUR	-4 404 668	-2 751 222	62%

**Tableau 26: Données sociétés non vie - prédiction de PM - extrait**

L'extrait permet de constater que sur 22 entreprises, l'estimation effectuée **pour 9** entreprises n'est pas comprises dans l'intervalle de confiance défini. Les **provisions mathématiques constituées** pour ces assureurs sont **largement inférieures** à celles évaluées par le modèle. Une **diligence particulière** devrait être mise en œuvre sur ce point en cas d'audit ou de contrôle de ces entités.



En conformité avec la **norme ISA 540** et pour réunir des éléments probants sur le **caractère raisonnable des estimations** effectuées par les compagnies d'assurance au titre de différentes **provisions techniques**, des régressions linéaires multiples ont été mises en œuvre. Pour chaque type de provision, **l'estimation indépendante** obtenue **comparée à celle des compagnies d'assurance** permet de mettre en évidence des **écarts**. Le professionnel comptable doit faire preuve de **vigilance et de diligences accrues** lorsque les écarts dégagés s'avèrent significatifs.

Si la régression linéaire multiple a permis de faire des estimations sur les provisions techniques des compagnies d'assurance, une **analyse discriminante** peut permettre de discriminer ces sociétés selon qu'elles sont **saines**, en **difficulté** ou en **faillite**.

## CHAPITRE II. ANALYSE DISCRIMINANTE

La situation de quelques sociétés vie et non vie examinées par la CRCA (Tableau 52: Données sociétés non vie - ajustements et statuts et Tableau 53: Données sociétés vie - ajustements et statuts) permet d'identifier trois statuts : les sociétés **saines**, les sociétés en difficultés faisant l'objet de **suivi** et les sociétés en **faillite**.

L'**objectif** des paragraphes suivants est donc de construire pour les compagnies d'assurance vie et non vie, des modèles permettant de **discriminer les sociétés saines, des sociétés devant faire l'objet de suivi et des sociétés en faillite**.

### 1) Analyse discriminante sur les données des compagnies d'assurance non vie

Des modèles sont construits à partir d'un **échantillon de 24 sociétés** dont les dossiers ont fait l'objet d'examen par la CRCA. La fonction `greedy.wilks()` du package `k1ar` et une validation croisée permettent de tester la qualité de chaque modèle. Enfin, le modèle retenu est utilisé dans le cadre d'un exercice de prédiction.

#### 1.1) Construction du modèle

La fonction `lda()` de la librairie `MASS` permet de réaliser l'**analyse discriminante linéaire**. La librairie `MASS` est installée par défaut dans . Il suffit donc de la charger. La fonction `lda()` utilise une formule comme la fonction `lm` pour la régression linéaire.

Les **probabilités a priori** choisies pour l'analyse doivent être spécifiées par l'utilisateur. En l'absence d'information sur ces probabilités, deux stratégies peuvent être envisagées :

- Les **probabilités** sont choisies **égales**, c'est-à-dire  $\pi_0 = \pi_1 = \pi_2 = \frac{1}{3}$  dans les exemples étudiés ;
- Les **probabilités** sont estimées par la **proportion d'observations** dans chaque groupe.

Le choix des probabilités a priori est spécifié dans l'argument `prior` de la fonction `lda`. Si rien n'est spécifié, la deuxième stratégie est utilisée par défaut.

### **Modèle avec toutes les variables explicatives et modèle avec 6 variables explicatives**

Pour le modèle contenant toutes les variables explicatives et le modèle avec 6 variables explicatives, les sorties du logiciel sont relativement longues. Elles sont donc communiquées et analysées en **annexe (Modèle avec toutes les variables explicatives page 197 et Modèle avec 6 variables explicatives page 198 )**.

### **Modèle sélectionné**

La fonction `greedy.wilks()` du package `klAR` peut être utilisée pour détecter les **variables prédictives pertinentes**. La fonction `greedy.wilks()` permet une sélection pas à pas de variables pertinentes. Le modèle initial est défini en commençant par la variable qui sépare le mieux les groupes. Le modèle est ensuite étendu en incluant d'autres variables en fonction du **critère Wilk lambda**.

La mise en œuvre de la fonction `greedy.wilks()` fournit les résultats suivants :

```
> Library (klAR)

> variables_pertinentes <- greedy.wilks(STATUT ~.,
data=donnees_analyse_discriminante_non_vie)

> variables_pertinentes
Formula containing included variables:

STATUT ~ PF_NETS

Values calculated in each step of the selection procedure:

      vars wilks.lambda F.statistics.overall p.value.overall F.statistics.diff
1 PF_NETS      0.450293           12.20776      0.0003427298           12.20776
  p.value.diff
1 0.0003427298
```

**Tableau 27: Données sociétés non vie - analyse discriminante - sélection des variables**

La fonction `greedy.wilks()` détecte une **seule variable pertinente**, les **produits financiers nets PF\_NETS**. Cette variable est utilisée pour construire le modèle d'analyse discriminante suivant à partir de la fonction `lda()`:

```
> modele_non_vie3 <- lda(STATUT ~ PF_NETS,
data=donnees_analyse_discriminante_non_vie)

> modele_non_vie3
Call:
lda(STATUT ~ PF_NETS, data = donnees_analyse_discriminante_non_vie)

Prior probabilities of groups:
 Faillite      Sain      Suivi
0.3043478 0.3043478 0.3913043

Group means:
      PF_NETS
Faillite  -40.14286
Sain     361094.14286
Suivi    57592.44444
```

```

Coefficients of linear discriminants:
                LD1
PF_NETS 6.728772e-06

```

**Tableau 28: Données sociétés non vie - analyse discriminante - modèle sélectionné**

Les sorties du logiciel fournissent les **probabilités a priori du modèle**, le **centre de gravité** de chacun des deux groupes et les **coefficients des variables canoniques**.

Les sociétés en faillite présentent en moyenne des **produits financiers nets négatifs** contre des **produits financiers nets positifs** et importants pour les sociétés saines. Les sociétés faisant l'objet de suivi sont entre ces deux extrêmes. La moyenne des produits financiers de ce groupe est de 57592.44444 KFCFA.

## 1.2) Estimation du taux de mauvais classement

### **Modèle avec 1 variable explicative**

Pour la variable explicative retenue, les labels prédits par le modèle et les taux de mauvais classement sont obtenus avec la commande :

```

> prevision3 <- lda(STATUT ~ PF_NETS,
data=donnees_analyse_discriminante_non_vie, CV=TRUE)$class
> table(prevision3, donnees_analyse_discriminante_non_vie$STATUT)

prevision3 Faillite Sain Suivi
Faillite   0     0     0
Sain       0     5     0
Suivi      7     2     9

```

**Tableau 29: Données sociétés non vie - analyse discriminante - classement**

Le modèle n'arrive pas à détecter les **sociétés en faillite** puisque que le **taux de mauvais classement** sur cette catégorie est de 100% (7/7). Cependant, il **détecte** bien les sociétés faisant l'**objet d'un suivi** (9/9). Le taux de mauvais classement à ce niveau est nul. Le modèle traduit une situation rencontrée sur le marché : de nombreuses sociétés faisant l'objet d'un suivi devraient normalement faire l'objet de retrait d'agrément si les règles édictées par le code des assurances étaient strictement respectées.

Une estimation du **taux de mauvais classement global** est obtenue par la commande suivante:

```

> sum(prevision3!= donnees_analyse_discriminante_non_vie$STATUT)/nrow
(donnees_analyse_discriminante_non_vie)
[1] 0.3913043

```

Du point de vue des erreurs de classement, le modèle avec les **6 variables explicative** PF\_NETS+ COM+ FG+ PRI\_REA+COM\_REA+RHE donne de **meilleurs résultats** que le modèle avec toutes les variables explicatives et le modèle avec une seule variable.

Le modèle à une **seule variable** est cependant retenu du fait de sa simplicité et de sa performance sur le critère de minimisation de la **statistique du lambda de Wilks** noté  $\Lambda$ . Cette statistique est utilisée pour mesurer le pouvoir discriminant des variables explicatives.

### 1.3) Intérêt pour le professionnel comptable et pour une autorité de supervision

Les **modèles construits** peuvent être utilisés dans un **contexte de prévision** lorsqu'ils sont **jugés satisfaisants** et dans de nombreux contextes. Il s'agit là d'un vaste **champ de recherche** pour valoriser les bases de données constituées par les organisations.

Dans le cas étudié, la qualité des modèles établis est sujette à celle des données. Or, les **données** produites par les compagnies d'assurance de la zone CIMA sont entachées de **nombreuses irrégularités**. La frontière entre les sociétés faisant l'objet de suivi et les sociétés en faillite n'est pas nette. Du fait notamment des **difficultés** résultant des **procédures de liquidation** des sociétés en faillite, la CRCA effectue le suivi de nombreuses sociétés qui devraient être normalement admises à la liquidation.

**L'analyse discriminante** effectuée permet cependant de montrer que l'un des critères optimum pour discriminer les sociétés d'assurance non vie est le niveau des **produits financiers nets**. Elle permet également d'ouvrir une réflexion au sein du Secrétariat Général de la CIMA sur la collecte et la fiabilisation des données relatives au suivi des compagnies d'assurance et leur discrimination.

Pour **prédire le label** des **autres sociétés** ne faisant pas l'objet de suivi au cours de la période, les données de ces sociétés sont mises dans un tableau ayant la même structure que celle des données précédemment étudiées.

La fonction `predict` permet d'affecter un groupe à chacune des sociétés n'ayant pas encore fait l'objet de contrôle par la CRCA. Même s'ils ne présentent pas d'intérêt véritable, les résultats obtenus sont fournis en annexe (200**Prédiction avec le modèle à une variable explicative** page 200).

## 2) Analyse discriminante sur les données des compagnies d'assurance vie

Comme pour les compagnies d'assurance non vie, des modèles sont construits à partir d'un **échantillon de 10 sociétés** dont les dossiers ont fait l'objet d'examen par la CRCA. La fonction `greedy.wilks()` du package `k1aR` et une **validation croisée** permettent de tester la qualité de chaque modèle. Enfin, le **modèle retenu** est utilisé dans le cadre d'un **exercice de prédiction**.

### 2.1) Construction du modèle

Le modèle prenant en compte toutes les variables explicatives est fourni en annexe. En tenant compte du **taux de mauvais classement de 60%** résultant de la **validation croisée**, ce modèle fait à peine mieux qu'un **modèle aléatoire**. Plusieurs modèles peuvent être conçus et comparés en fonction du taux de mauvais classement.

La fonction `greedy.wilks()` du package `k1aR` peut être utilisée pour détecter les variables **prédictives pertinentes**. La mise en œuvre de cette fonction pour les sociétés vie examinées fournit les résultats suivants :

```
> library(k1aR)
> test_vie <- greedy.wilks(STATUT ~., data=donnees_analyse_discriminante_vie)
> test_vie

Formula containing included variables:
STATUT ~ Kta1 + A_nets

values calculated in each step of the selection procedure:
      vars wilks.lambda F.statistics.overall p.value.overall F.statistics.diff
p.value.diff
1  Kta1    0.3917872          5.433420          0.03764229          5.433420
0.03764229
2  A_nets  0.2095343          3.553808          0.03913021          2.609401
0.14231247
```

Tableau 30: Données sociétés vie - analyse discriminante - sélection des variables

La fonction `greedy.wilks()` détecte deux **variables pertinentes**, le capital et les actifs nets `Kta1 + A_nets`. Ces variables sont utilisées pour construire le modèle d'analyse discriminante suivant à partir de la fonction `lda()` :

```
> modele_vie2 <- lda(STATUT ~ Kta1 + A_nets,
data=donnees_analyse_discriminante_non_vie)
> modele_vie2
Call:
lda(STATUT ~ Kta1 + A_nets, data = donnees_analyse_discriminante_non_vie)

Prior probabilities of groups:
```

```

Faillite      Sain      Suivi
      0.2      0.4      0.4

Group means:
      Kta1      A_nets
Faillite 112500000 1267214661
Sain      581250000 405123161
Suivi     500000000 532775084

Coefficients of linear discriminants:
      LD1      LD2
Kta1   -9.185139e-09 -2.130986e-09
A_nets  1.987103e-09  1.722762e-09

Proportion of trace:
      LD1      LD2
0.9867 0.0133

```

**Tableau 31: Données sociétés vie - analyse discriminante - modèle sélectionné**

Les résultats de l'analyse peuvent faire l'objet des commentaires suivants :

- L'objet `ca11` contient la formule de l'analyse discriminante ;
- La probabilité d'appartenance à chaque groupe est fournie `Prior probabilities of groups` ;
- Le centre de gravité de chacun des trois groupes est fourni dans l'objet `Group means` ;
- Les coefficients des deux variables canoniques sont fournis de sorte que la variance intraclasse est égale à 1 = 0.9867+ 0.0133.

## 2.2) Estimation du taux de mauvais classement

La fonction `lda` permet d'estimer le **taux de mauvais classement** par **validation croisée** en ajoutant l'argument `CV=TRUE` lors de l'appel de la fonction. On obtient les **labels prédits** par le modèle et les taux de mauvais classement avec les commandes successives suivantes :

```

> prevision_vie2 <- lda(STATUT ~ Kta1 + A_nets,
data=donnees_analyse_discriminante_vie, CV=TRUE)$class

> table(prevision_vie2, donnees_analyse_discriminante_vie$STATUT)

prevision_vie2 Faillite Sain Suivi
  Faillite      1      0      0
  Sain          1      3      2
  Suivi         0      1      2

```

**Tableau 32: Données sociétés vie - analyse discriminante - classement**

Les taux de mauvais classement sont de **50%** pour les sociétés faisant l'objet d'un suivi et pour les sociétés en faillite. Ils sont de **25%** pour les sociétés saines.

Une **estimation du taux de mauvais classement global** peut être obtenue par la commande suivante:

```

> sum(prevision_vie2!= donnees_analyse_discriminante_non_vie$STATUT)/nrow
(donnees_analyse_discriminante_non_vie)

```

[1] 0.4

Avec un taux d'erreurs de classement de 40%, le modèle avec les 2 variables explicatives  $K_{ta1} + A_{nets}$  donne de meilleurs résultats que le modèle avec toutes les variables explicatives. Ce taux reste cependant élevé.

### 2.3) Intérêt pour le professionnel comptable et pour une autorité de supervision

Les indications générales au point « 1.3) Intérêt pour le professionnel comptable et pour une autorité de supervision page 97 » restent valables pour cette section.

Comme pour les compagnies d'assurance non vie, les différents modèles construits pour les compagnies vie ne fournissent pas des taux de bons classements satisfaisants. Ils peuvent difficilement être utilisés pour faire de la prédiction. Cependant, les prédictions effectuées sont fournies en annexe (**Prédiction avec le modèle sélectionné**  $K_{ta1} + A_{nets}$  page 204)



Pour tous les modèles construits à partir des données des compagnies d'assurance de la zone CIMA, les taux de mauvais classement obtenus dans le cadre de l'analyse discriminante sont élevés. Les prédictions pouvant en résulter doivent être examinées avec réserve. Les taux de mauvais classement et le mauvais pouvoir de prédiction des modèles sont imputables essentiellement à la qualité des données et aux incertitudes sur la variable à expliquer. Les travaux montrent donc la nécessité de réfléchir à deux problématiques : le traitement et la fiabilisation des données recueillies sur les sociétés et la fixation de critères plus précis relatifs aux trois classes de sociétés établies.

Les modèles retenus résultent d'une procédure de sélection pas à pas des variables explicatives les plus pertinentes. Une autre démarche consistant à construire les modèles à partir des composantes principales permet d'éviter la problématique de colinéarité des variables explicatives initiales.

Les informations ayant permis d'effectuer l'analyse discriminante pourraient également permettre la mise en œuvre d'une régression logistique. Une démarche similaire, mais cependant différente, va être développée dans les paragraphes suivants pour permettre de présenter et de partager la mise en œuvre d'autres outils.

### CHAPITRE III. REGRESSION LOGISTIQUE

L'**effectif de la brigade de contrôle** de la CIMA et les **contraintes de budget** ne permettent pas d'effectuer au moins chaque année un **contrôle de chaque entreprise** du marché. Par exemple au titre de l'exercice clos au 31/12/2009, une **vingtaine d'assureurs non vie** ont fait l'objet de contrôle pour une **dizaine d'assureurs vie**. Les quelques contrôles effectués au titre de chaque exercice permettent de constater que les comptes produits par les compagnies d'assurance présentent **d'importantes anomalies**.

Un **tableau en annexe** (Tableau 52: Données sociétés non vie - ajustements et statuts et Tableau 53: Données sociétés vie - ajustements et statuts) donne pour les compagnies contrôlées sur la période, les ajustements effectués et leur incidence globale sur le niveau des fonds propres. Par hypothèse, les ajustements effectués sont significatifs si le **total des ajustements** rapporté aux **fonds propres** est supérieur au taux de **10%**.

L'objectif des paragraphes suivants est de mettre en évidence un outil simple d'utilisation permettant de déterminer la **probabilité** que les comptes d'une compagnie d'assurance contiennent des **anomalies significatives**. Plus cette probabilité sera élevée et proche de 1, plus urgent il sera nécessaire d'être vigilant en cas de mission d'audit ou de contrôle.

#### 1) Régression logistique sur les données des compagnies d'assurance non vie

Le modèle sera construit à partir des sociétés ayant fait l'objet de contrôle au cours de l'année et testé sur les autres sociétés.

##### 1.1) Construction du modèle

La régression logistique appartient à la famille des modèles linéaires généralisés. Un ajustement de ces modèles sur  est réalisé par la **fonction glm**. L'utilisation de cette fonction est similaire à celle de la fonction lm. Le modèle est écrit avec une formule du type :

```
> modele_complet_Y<-glm(Y~., data=donnees_societes_controlees,family=binomial)
```

Il faut également spécifier à travers l'écriture du modèle une **famille de lois de probabilité**. Dans le cadre de la régression logistique, il s'agit de la famille

**binomiale.** Le `modele_complet_Y` comprenant toutes les variables explicatives se présente comme suit :

```
> modele_complet_Y
Call: glm(formula = Ajustement ~ ., family = binomial, data =
donnees_societes_controlees)

Coefficients:
(Intercept)      PRIMES      SIN_REG      AR_BRUTS      AR_NETS      PF_NETS
 3.111e+01  2.187e-05  1.351e-05  2.828e-05 -4.257e-05  1.778e-04
      COM      FG      PRI_REA      SIN_REA      SAP_REA      COM_REA
 7.557e-05  2.461e-05  7.010e-05  1.654e-05 -2.128e-05  2.256e-04
      PSAP      PAP      CAP      T_BILAN      R_EXP      RHE
-7.563e-06  4.425e-06 -2.121e-05  2.385e-05 -3.928e-05 -7.292e-05
      E_R      ACTIFS_A
 2.923e-05 -6.011e-06

Degrees of Freedom: 19 Total (i.e. Null); 0 Residual
Null Deviance:      16.91
Residual Deviance: 3.154e-10  AIC: 40
```

**Tableau 33: Données sociétés non vie - régression logistique - modèle brut**

La fonction de régression linéaire utilisée `glm` retourne une **liste d'objets et d'informations**. Les objets et les informations affichés dans un premier temps sur la console sont présentés comme suit:

- **L'objet `call`** contient la fonction et la formule de régression logistique utilisées ;
- **Les coefficients** sont des paramètres estimés par la méthode du maximum de vraisemblance ;
- **Le nombre de degrés de libertés** `Degrees of Freedom` ;
- La `Null deviance: 16.91` correspond à la **déviante du modèle trivial** réduit à la seule constante  $a_0$ . On montre que la déviante triviale  $D_0 = -2 \times LL_0 = -2 \times \left[ n \times \ln(1 - \hat{p}) + n_+ \times \left( \frac{\hat{p}}{1 - \hat{p}} \right) \right]$ . Dans cette expression  $n_+$  est le nombre de sociétés présentant des anomalies significatives dans leurs comptes sur un nombre  $n$  de sociétés contrôlées. La proportion de sociétés présentant des anomalies significatives s'en déduit et est notée  $\hat{p} = \frac{n_+}{n}$  ;
- La `Residual deviance: 3.154e-10` est pour la régression logistique l'équivalent de la **somme des carrés des résidus** de la régression linéaire multiple. L'objectif est de minimiser cette déviante ;
- `L'AIC` est également un **critère** de mesure de la **déviante à minimiser**.

D'autres objets non affichés directement sur la console peuvent être intéressants à examiner. On peut s'aider de la **commande** `help(glm)` pour avoir une description des autres objets générés par la fonction.

Le modèle complet comprend **19 variables explicatives**. La fonction `step` permet de sélectionner un modèle à l'aide d'une procédure pas à pas basée sur la minimisation du critère **AIC** (Akaike Information Criterion). Les caractéristiques du **modèle sélectionné** se présentent comme suit :

```
> modele_selectionne

Call: glm(formula = Ajustement ~ PF_NETS + PRI_REA + CAP, family = binomial,
data = donnees_societes_controlees)

Coefficients:
(Intercept)      PF_NETS      PRI_REA          CAP
  4.556e+02   -1.057e-03   -1.587e-04   -1.632e-04

Degrees of Freedom: 19 Total (i.e. Null); 16 Residual
Null Deviance:      16.91
Residual Deviance: 6.588e-09    AIC: 8
```

**Tableau 34: Données sociétés non vie - régression logistique - modèle sélectionné**

La sélection pas à pas effectuée à partir de la fonction `step` est une **procédure descendante**. A chaque étape, la variable dont le retrait du modèle conduit à la diminution la plus grande du critère AIC est éliminée. Le processus prend fin lorsque toutes les variables sont retirées ou lorsque le retrait d'aucune variable ne permet de diminuer le critère AIC. Sur 19 variables initiales, le **modèle sélectionné** conserve in fine **3 variables** qui sont les produits financiers nets `PF_NETS`, les primes cédées en réassurance `PRI_REA` et les capitaux propres `CAP`. Sur la base du critère AIC, la fonction `step` a permis de passer d'un modèle à **19 variables avec  $AIC = 40$**  à un modèle à **3 variables avec  $AIC = 8$** .

## 1.2) Estimation du taux de mauvais classement

Le modèle a été construit à partir d'un **échantillon de 20 sociétés** dont les comptes de l'exercice 2009 ont fait l'objet d'un audit par la brigade de contrôle de la CIMA. La fonction `predict` peut être utilisée pour **estimer le taux de mal classés du modèle**. Ce taux est égal à la **proportion d'erreurs** commises par le modèle lorsqu'il est appliqué à l'échantillon de base.

Dans ce cadre, on calcule d'abord les **probabilités prédites** pour chaque société de l'échantillon. La commande suivante permet d'effectuer ce calcul :

```
> previsions_proba_controlees<-predict(modele_selectionne,
newdata=donnees_societes_controlees, type="response")
```

On compare ensuite les **probabilités obtenues à 0.5 afin d'obtenir le label prédit** pour chaque société. La commande suivante permet d'effectuer cette comparaison :

```
> prevision_label<-as.numeric(previsions_proba_controlees > 0.5)
```

Il est alors possible d'établir un **tableau de contingence** entre les **labels prédits et les vrais labels**. La commande permettant d'effectuer cette opération et sa résultante sont présentées comme suit :

```
> table(donnees_societes_controlees$Ajustement, prevision_label)


```

prevision_label
  0  1
0  3  0
1  0 17
```


```

On obtient le **taux de mal classés** du modèle à l'aide de la commande suivante :

```
> mal_classes<-
sum(prevision_label!=donnees_societes_controlees$Ajustement)/nrow(donnees_socie
tes_controlees)
> mal_classes
[1] 0
```

Le taux de mal classés est nul. Toutes les sociétés de l'échantillon sont bien classées par le modèle. Le taux ainsi obtenu est généralement optimiste puisque le même échantillon est utilisé pour construire le modèle et pour estimer le taux de mal classé. Des **techniques de validation croisée** permettent de faire des estimations plus précises. Pour cela, il faut au préalable charger le package `boot` et créer une fonction de coût ayant comme paramètres les valeurs observées des ajustements et les probabilités prédites.

```
> library(boot)
> cout<-function (Ajustements, prevision_proba_controlees)
return(mean(abs(Ajustements - prevision_proba_controlees)>0.5))
> cv.glm(donnees_societes_controlees, modele_selectionne, cout)$delta[1]
  1
0.2
```

Le **taux de mal classés** obtenu par la **validation croisée est de 20%**. Il est nettement plus important que le **taux optimiste** obtenu avec la fonction `predict`.

### 1.3) Intérêt pour le professionnel comptable et pour une autorité de supervision

Le modèle a été construit à partir d'un **échantillon de 20 sociétés** dont les comptes de l'exercice 2009 ont fait l'objet d'un examen par la brigade de contrôle de la CIMA. Le modèle logistique construit est par la suite utilisé pour prédire la probabilité que les comptes de chacune des **70 autres sociétés** contiennent des **anomalies**

**significatives.** Les données sur ces 70 sociétés ont la même structure que celles relatives aux 20 sociétés utilisées pour la construction du modèle. Ces données sont extraites du tableau initial relatif aux sociétés non vie à l'aide de la commande suivante:

```
> donnees_societes_non_contrrolees<-
donnees_non_vie[c("A1","A3","A6","A7","A8","A9","A10","A13","A14","A15",
"A17","A19","A20","A21","A22","A24","A25","A26","A27","A28","A29","A30",
"A31","A32","A33","A34","A36","A37","A38","A39","A40","A41",
"A43","A44","A45","A46","A47","A48","A49","A50","A51","A54","A55",
"A57","A58","A59","A60","A61","A62","A63","A64","A65","A66","A67","A68",
"A70","A73","A74","A75","A76","A77","A80","A81","A83","A84","A85","A86",
"A87","A88","A90"),1:19]
```

La fonction `predict` permet de **prédire les probabilités**  $P(Y = 1/X = x)$  pour chacune de ces 70 sociétés. La commande relative à cette fonction est la suivante :

```
> prevision_du_modele<- predict (modele_selectionne,
newdata=donnees_societes_non_contrrolees, type="response")
```

Les **prévisions** obtenues peuvent être récupérées dans un fichier d'extension csv pour des travaux de mise en forme avec la commande suivante :

```
> write.table(prevision_du_modele,"prevision_du_modele_logistique.csv",sep=";")
```

A la suite des travaux de mise en forme, ces probabilités sont consignées dans le tableau suivant :

Sociétés	A1	A3	A6	A7	A8	A9	A10	A13	A14	A15	A17	A19	A20	A21	A22	A24	A25	A26
Ajustements	1	1	0	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0
Sociétés	A27	A28	A29	A30	A31	A32	A33	A34	A36	A37	A38	A39	A40	A41	A43	A44	A45	A46
Ajustements	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sociétés	A47	A48	A49	A50	A51	A54	A55	A57	A58	A59	A60	A61	A62	A63	A64	A65	A66	A67
Ajustements	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0
Sociétés	A68	A70	A73	A74	A75	A76	A77	A80	A81	A83	A84	A85	A86	A87	A88	A90		
Ajustements	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0		

**Tableau 35: Données sociétés non vie - prédiction d'anomalies**

Le modèle prévoit que les **comptes de 58 sociétés** sur les **70 listées** contiennent des **anomalies significatives**. Les comptes de **12 sociétés** sur un total de 70 sont identifiés comme ne comportant pas d'anomalies significatives. A ce stade, un modèle de **régression linéaire multiple** peut être utilisé pour évaluer pour chacune des sociétés le **montant des anomalies** pouvant affecter les comptes. La démarche suivrait la même logique que celle présentée au chapitre relatif à la régression multiple pour l'estimation des provisions techniques des sociétés vie et non vie.

Les **résultats obtenus** sont globalement **conformes aux attentes**. Les comptes de seulement 3 sociétés contrôlées sur 20 ne comportent pas d'anomalies significatives,

soit un taux de 15%. Le modèle prévoit que les comptes de 12 sociétés non contrôlées sur 70 ne comportent pas d'anomalies significatives, soit un taux de 17%.

Les résultats obtenus doivent cependant être confrontés à la réalité du terrain au cours des contrôles ultérieurs effectués.

L'idée initiale était de faire un lien entre l'opinion émise par les commissaires aux comptes et les comptes annuels des compagnies d'assurance. La difficulté de ce lien réside dans la **pratique très hétérogène du commissariat aux comptes** dans l'espace CIMA et dans l'absence de formation de nombreux commissaires aux comptes sur les spécificités et les risques des compagnies d'assurance. Le risque de non détection d'anomalies significatives est particulièrement élevé.

**L'option finalement retenue** et mise en œuvre est une **modélisation** uniquement à partir des résultats des contrôles effectués par les **commissaires contrôleurs de la CIMA**. Les méthodologies, les résultats obtenus et les opinions émises sont plus homogènes et permettent de construire un modèle plus robuste.

La CIMA a mis en place une **procédure d'agrément des commissaires aux comptes des compagnies d'assurance**. Disposant d'un outil permettant d'estimer la probabilité que les comptes annuels comportent des anomalies significatives et de quantifier ces anomalies à travers une régression linéaire multiple, elle peut attirer la vigilance des commissaires aux comptes sur ces risques. Cet outil doit cependant être utilisé avec prudence.

Dans un environnement où la pratique des professionnels comptables est harmonisée et de haut niveau, on pourrait construire à partir d'un échantillon un modèle permettant de prédire l'opinion qui sera émise sur les comptes d'une entreprise audité.

## **2) Régression logistique sur les données des compagnies d'assurance Vie**

Comme pour les compagnies d'assurance non vie, le modèle est construit à partir des **sociétés** ayant fait **l'objet de contrôle** sur les comptes produits au titre de l'exercice 2009 et **testé sur les autres sociétés**.

## 2.1) Construction du modèle

Le modèle est écrit avec une formule du type :

```
> modele_complet_Y<-glm(Y~., data=donnees_societes_controlees,family=binomial)
```

Le `modele_complet` comprenant toutes les variables explicatives se présente comme suit :

```
> modele_complet

Call:  glm(formula = Ajustement ~ ., family = binomial, data =
donnees_societes_controlees)

Coefficients:
(Intercept)      Primes          PF      Sinistres          Com          FG
3.275e+01  1.014e-08 -4.753e-08  8.371e-09  1.353e-08  2.690e-08
P_Réass  Charges_réa          RE          RN      S_.payés      S_réa
-8.562e-09  5.199e-09          NA -1.468e-09 -6.954e-09          NA
Com_réa          PM      A_bruts      A_nets      Ktal          FP
NA          NA          NA          NA          NA          NA
PT      T_bilan      Actifs      ER      Actifs.
NA          NA          NA          NA          NA

Degrees of Freedom: 9 Total (i.e. Null); 0 Residual
Null Deviance:      6.502
Residual Deviance: 4.287e-10  AIC: 20
```

**Tableau 36: Données sociétés vie - régression logistique - modèle brut**

Le modèle complet comprend **22 variables explicatives**. La fonction `step` permet de **sélectionner un modèle** à l'aide d'une procédure pas à pas basée sur la minimisation du critère AIC (Akaike Information Criterion). Les caractéristiques du modèle sélectionné se présentent comme suit :

```
> modele_selectionne

Call:  glm(formula = Ajustement ~ Sinistres, family = binomial, data =
donnees_societes_controlees)

Coefficients:
(Intercept)      Sinistres
 1.227e+02  1.686e-08

Degrees of Freedom: 9 Total (i.e. Null);  8 Residual
Null Deviance:      6.502
Residual Deviance: 3.761e-10  AIC: 4
```

**Tableau 37: Données sociétés vie - régression logistique - modèle sélectionné**

Sur **22 variables initiales**, le modèle sélectionné conserve in fine **une variable** qui est la charge de sinistres `sinistres`. Sur la base du critère AIC, la fonction `step` a permis de passer d'un modèle à 22 variables avec  $AIC = 20$  à un modèle à 1 variable avec  $AIC = 4$ .

Une autre façon de tester la qualité du modèle sélectionné consiste à évaluer le **taux de mauvais classement**.

## 2.2) Estimation du taux de mauvais classement

Le modèle a été construit à partir d'un **échantillon de 10 sociétés** dont les comptes de l'exercice 2009 ont fait l'objet d'un **audit par la brigade de contrôle de la CIMA**. La fonction `predict` peut être utilisée pour estimer le taux de mal classés du modèle. Ce taux est égal à la **proportion d'erreurs** commises par le modèle lorsqu'il est appliqué à l'échantillon de base.

Dans ce cadre, on calcule d'abord les **probabilités prédites** pour chaque société de l'échantillon. La commande suivante permet d'effectuer ce calcul :

```
> previsions_proba_controlees <- predict (modele_selectionne,
newdata=donnees_societes_controlees, type="response")
```

On compare ensuite les **probabilités obtenues à 0.5 afin d'obtenir le label prédit** pour chaque société. La commande suivante permet d'effectuer cette comparaison :

```
> prevision_label<-as.numeric(previsions_proba_controlees > 0.5)
```

Il est alors possible d'établir un **tableau de contingence entre les labels prédits et les vrais labels**. La commande permettant d'effectuer cette opération et sa résultante sont présentées comme suit :

```
> table(donnees_societes_controlees$Ajustement, prevision_label)
  prevision_label
0 1
0 1 0
1 0 9
```

On obtient le **taux de mal classés** du modèle à l'aide de la commande suivante :

```
> mal_classes<-
sum(prevision_label!=donnees_societes_controlees$Ajustement)/nrow(donnees_socie
tes_controlees)
> mal_classes
[1] 0
```

Le taux de mal classés est nul. Toutes les sociétés de l'échantillon sont bien classées par le modèle. Le taux ainsi obtenu est généralement optimiste puisque le même échantillon est utilisé pour construire le modèle et pour estimer le taux de mal

classé. Des **techniques de validation croisée** permettent de faire des estimations plus précises. Pour cela, il faut au préalable charger le package `boot` et créer une fonction de coût ayant comme paramètres les valeurs observées des ajustements et les probabilités prédites.

```
> library(boot)

> cout<-function (Ajustements, prevision_proba_controlees)
  return(mean(abs(Ajustements - prevision_proba_controlees)>0.5))

> cv.glm(donnees_societes_controlees, modele_selectionne, cout)$delta[1]
  1
  0
```

Le **taux de mal classés** obtenu par la validation croisée est également nul. Le modèle sélectionné semble **satisfaisant**. Il peut donc être utilisé dans le cadre d'un exercice de **prédiction**.

### 2.3) Intérêt pour le professionnel comptable et pour une autorité de supervision

Les **indications générales** au point « 1.3) Intérêt pour le professionnel comptable et pour une autorité de supervision page 104» restent valables pour cette section.

Le modèle a été construit à partir d'un **échantillon de 10 sociétés** dont les comptes de l'exercice 2009 ont fait l'objet d'un audit par la brigade de contrôle de la CIMA. Le modèle logistique construit est par la suite utilisé pour prédire la probabilité que les comptes de chacune des 35 autres sociétés contiennent des anomalies significatives. Les données sur ces 35 sociétés ont la même structure que celles relatives aux 10 sociétés utilisées pour la construction du modèle. Ces données sont extraites du tableau initial relatif aux sociétés non vie à l'aide de la commande suivante :

```
> donnees_societes_non_controlees<-
  donnees_vie[c("A1","A2","A3","A5","A7","A10", "A11", "A12", "A13","A14", "A16",
  "A17","A18","A19","A21","A22","A24", "A25", "A27","A29","A30", "A31","A33",
  "A34","A35","A36","A37","A38", "A39", "A40", "A41","A42",
  "A43","A44","A45"),1:22]
```

La fonction `predict` permet de **prédire les probabilités**  $P(Y = 1/X = x)$  pour chacune de ces 35 sociétés. La commande relative à cette fonction est la suivante :

```
> prevision_du_modele<- predict (modele_selectionne,
  newdata=donnees_societes_non_controlees, type="response")
```

Les **prévisions obtenues** peuvent être récupérées dans un **fichier d'extension csv** pour des travaux de mise en forme avec la commande suivante :

```
write.table(previson_du_modele,"previson_du_modele_logistique.csv",sep=";")
```

A la suite des travaux de mise en forme, ces **probabilités** sont consignées dans le tableau suivant :

Sociétés	A1	A2	A3	A5	A7	A10	A11	A12	A13	A14	A16	A17	A18	A19	A21	A22	A24	A25
Ajustements	0,00	0,00	0,00	0,69	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sociétés	A27	A29	A30	A31	A33	A34	A35	A36	A37	A38	A39	A40	A41	A42	A43	A44	A45	
Ajustements	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

**Tableau 38: Données sociétés vie - prédiction d'anomalies**

Le modèle prévoit que les **comptes de 32 sociétés** sur les 35 listées contiennent des **anomalies significatives**. Les comptes de 3 sociétés sur un total de 35 sont identifiés comme ne comportant pas d'anomalies significatives. A ce stade, un **modèle de régression linéaire multiple** peut être utilisé pour évaluer pour chacune des sociétés le **montant des anomalies** pouvant affecter les comptes. La démarche suivrait la même logique que celle présentée au chapitre relatif à la régression multiple pour l'estimation des provisions techniques des sociétés vie et non vie.

Les **résultats obtenus** sont globalement **conformes aux attentes**. Les comptes de seulement 1 société contrôlée sur 10 ne comportent pas d'anomalies significatives, soit un taux de 10%. Le modèle prévoit que les comptes de 3 sociétés non contrôlés sur 35 ne comportent pas d'anomalies significatives, soit un taux de 9%.

Les résultats obtenus doivent cependant être confrontés à la réalité du terrain au cours des contrôles ultérieurs effectués.



A partir d'un **échantillon** de sociétés ayant fait l'**objet d'un contrôle**, une **régression logistique** a permis d'estimer la **probabilité** que les comptes annuels d'une entité non contrôlée présentent **des anomalies significatives**. Le professionnel comptable doit faire preuve de **vigilance et de diligences accrues** dans le cadre de ses travaux sur les comptes d'entités identifiés comme présentant des anomalies significatives.



La **régression linéaire multiple** et la **régression logistique** aboutissent à des **résultats satisfaisants** utilisés par la suite dans le cadre d'un exercice de **prédiction**. Même si l'**analyse discriminante** fournit des **résultats peu satisfaisants** du fait de la qualité des données utilisées, son **utilité** dans le cadre de l'exercice effectué est **démontrée**.

Pour la mise en œuvre des **normes de travail** des professionnels comptables, la **régression linéaire multiple** trouve une application plus immédiate que la régression logistique ou l'analyse discriminante. Ces deux dernières techniques peuvent cependant être utilisées par l'expert comptable non seulement dans le cadre de travaux liés aux comptes mais également dans le cadre de **travaux de valorisation des immenses bases de données** constituées par les entreprises dans lesquels il intervient : identification des prospects le plus susceptibles de devenir client, recherche des facteurs de risque notamment en assurance, prévention des impayés, utilisation du score de risque pour proposer ou refuser un crédit,...

Les **travaux** résultant de la mise en œuvre des **méthodes prédictives** peuvent être **résumés** dans le tableau suivant :

Données Additionnelles	Méthodes mises en œuvre	Objectifs	Résultats obtenus	Autres objectifs possibles
Aucune.	Régression linéaire multiple.	Estimation des provisions techniques des compagnies d'assurances en conformité avec les normes ISA 520, 540 et les NEP 520 et 540.	Résultat satisfaisant. Le modèle obtenu est utilisé pour effectuer de la prédiction et identifier notamment les entreprises dont les provisions techniques ne seraient pas suffisantes.	Estimation du montant des ajustements à réaliser sur les provisions techniques des compagnies d'assurance.
Statut des sociétés examinés par la CRCA.	Analyse discriminante.	Discriminer les entreprises d'assurance selon les 3 statuts suivants: "sain", "suivi" et "faillite".	Le pouvoir discriminant du modèle obtenu n'est pas satisfaisant. L'utilité d'une telle modélisation est cependant prouvée.	Discriminer les compagnies dont les comptes contiennent des anomalies significatives et de celles dont les comptes sont réguliers et sincères.
Anomalies identifiées et quantifiées sur les comptes annuels examinés par les contrôleurs de la CIMA.	Régression logistique.	Estimer la probabilité que les comptes annuels d'une compagnie d'assurance contiennent des anomalies significatives.	Résultat satisfaisant. Le modèle obtenu est utilisé pour effectuer de la prédiction et identifier les entreprises dont les comptes comportent des anomalies significatives	Discriminer les entreprises d'assurance selon les 3 statuts suivants: "sain" = 1, "suivi" = 2 et "faillite" = 3 dans le cadre d'une régression logistique multinomiale.

**Tableau 39: Synthèse des travaux effectués dans le cadre de la mise en œuvre des méthodes prédictives**

# CONCLUSION

L'analyse factorielle et les techniques de classification permettent d'explorer sans a priori un **tableau de données** et d'en effectuer une **représentation synthétique** sous forme d'un nuage euclidien, d'une hiérarchie ou d'une partition. Une telle représentation synoptique aide à comprendre rapidement les **caractéristiques essentielles d'un phénomène** par la localisation des grandes masses d'individus, la détection des individus exceptionnels et la mise en évidence d'éventuels groupes isolés. Le bilan, le compte de résultat, le tableau de financement... constituent autant de tableaux de données pouvant faire l'objet d'analyse factorielle ou de classification. Pour les professionnels comptables, le **champ d'application** de ces techniques est donc particulièrement **vaste**. A cet égard, tout **investissement réalisé** pour en **assurer la maîtrise** devrait pouvoir être rapidement **rentabilisé** en termes de **développement de missions nouvelles** pour l'expert comptable et de **qualité de documentation** des travaux pour le commissaire aux comptes.

Les **techniques prédictives** peuvent être utilisées par le professionnel comptable notamment dans le cadre de la mise en œuvre de **diligences** prévues par **la norme ISA 520 et la NEP 520** relatives aux procédures analytiques et de **la norme ISA 540 et la NEP 540** relatives à l'appréciation des estimations comptables. Dans ce sens, le mémoire constitue une **voie d'approfondissement** de la note d'information **NI VIII de la CNCC** relative aux procédures analytiques.

L'**appréciation des estimations comptables** est au cœur des problématiques rencontrées par les professionnels comptables et en particulier les commissaires aux comptes intervenant dans le secteur des assurances. Ces derniers doivent, conformément à la **NEP 705 § 8**, **justifier de leurs appréciations** portant sur les estimations comptables importantes. La correcte mise en œuvre de **techniques prédictives** peut aider à la justification des appréciations et éventuellement permettre la détection d'erreurs et de fraudes ayant un impact significatif sur les comptes.



Les **techniques factorielles et de classification** mises en œuvre à travers le mémoire permettent de constater que les compagnies exerçant dans l'espace CIMA sont dans leur majorité **fragilisées** par un **important niveau d'arriérés de primes** et une **absence de taille critique** devant permettre de mieux couvrir les frais fixes et mutualiser les risques souscrits. La concentration du secteur recherchée par les autorités de supervision à travers notamment le relèvement du niveau minimum du capital social ne s'est pas opérée.

Le marché de **l'assurance vie** apparaît **plus hétérogène** que celui de **l'assurance non vie**. Les représentations des sociétés d'assurance vie obtenues à partir de la classification mettent en évidence globalement 6 groupes contre 3 pour les sociétés non vie.

Des axes non représentés dans le cadre de l'ACP mettent en évidence des **particularités** au niveau de certaines compagnies d'assurance : résultat hors exploitation élevé, importance des cessions en réassurances... . Ces particularités, sources de **risques potentiels**, doivent faire l'objet de **vigilance et de diligence accrues** dans le cadre des missions mises en œuvre par un professionnel comptable et les commissaires contrôleurs de la CIMA.



En conformité avec la **norme ISA 540** et pour réunir des éléments probants sur le **caractère raisonnable des estimations** effectuées par les compagnies d'assurance au titre de différentes provisions techniques, des régression linéaires multiples ont été mises en œuvre. Pour chaque type de provision, l'estimation indépendante obtenue comparée à celle des compagnies d'assurance permet de mettre en évidence des écarts devant faire l'objet de diligences accrues lorsqu'ils sont significatifs.

La **régression linéaire multiple et la régression** logistique aboutissent à des **résultats satisfaisants** utilisés pour estimer le niveau des provisions techniques et la probabilité d'existence d'anomalies significatives dans les comptes des compagnies d'assurance. **L'analyse discriminante** fournit des **résultats peu**

**satisfaisants** du fait notamment de la qualité des données utilisées. Son utilité dans le cadre de l'exercice effectué est cependant confirmée.

La **régression linéaire multiple** trouve une **application plus immédiate** que la **régression logistique ou l'analyse discriminante** dans le cadre de la mise en œuvre des normes de travail des professionnels comptables.

Cependant, toutes ces techniques peuvent être utilisées par l'expert comptable dans le cadre de missions d'aide à la valorisation des bases de données constituées par les entreprises clientes. Ces missions rentrent dans le cadre des travaux d'ordre **statistique** prévus par l'article 22 de l'ordonnance n° 45-2138 du 19 septembre 1945 portant institution de l'ordre des experts comptables et devraient constituer un **axe** important de **développement des cabinets d'expertise comptable**. Il s'agit également d'une occasion d'appeler à la profession d'autres talents et notamment des **actuares**, des **statisticiens**, des **économètres** et des **informaticiens** pour répondre aux **besoins de compétence** et de **pluridisciplinarité** exigés par les normes professionnelles et la satisfaction de la clientèle.

# BIBLIOGRAPHIE

## Ouvrages

1. AZAÏS Jean-Marc et BARDET Jean-Marc, Le modèle linéaire par l'exemple - Régressions, analyse de la variance et plans d'expériences illustrées avec R, SAS, et Splus, Cours et exercices corrigés- Master, écoles d'ingénieurs, Dunod, 2006, 326 p.
2. BARDOS Mireille, RASSON Jean-Paul, Analyse discriminante : Application au risque et scoring financier, Dunod, 2001, 224 p.
3. BERTRAND Frédéric, MAUMY-BERTRAND Myriam, Initiation à la statistique avec R, Cours, exemples, exercices et problèmes corrigés - Licence 3, Master 1, écoles d'ingénieurs, Dunod, 2010, 384 p.
4. BLONDEAU Jacques, PARTRAT Christian, La réassurance : Approche technique - Collection Assurance Audit Actuariat, Economica, 2003, 642 pages
5. BOURBONNAIS Régis, Econométrie, Dunod, 8ème édition, 2011, 380 p.
6. CARRICANO Manu, POUJOL Fanny, BERTRANDIAS Laurent, Analyse de données avec SPSS, Pearson Education, 2ème éd., 2010, 234Pp
7. CLARKSON Douglas B. , FRALEY Chris, GU Charles C. , RAMSAY James O. S+ Functional Data Analysis User's Guide, User's Manual for Windows, Springer, 2005, 192 p.
8. CODY Ron, SMITH Jeffrey K., Applied Statistics and the SAS Programming Language; Prentice Hall, Pearson Education, 5ème édition, 2005, 572 pages
9. CORNILLON Pierre-André, MATZNER-LOBER Eric, Régression avec R, Springer, 2010, 242 p.
10. CORNILLON Pierre-André et autres, Statistiques avec R, Presses Universitaires de Rennes (PUR), 3ème édition, 2012, 296 p.

11. COUILBAULT François, ELIASHBERG Constant, Les grands principes de l'assurance, L'argus de l'assurance, 9ème édition, 2009, 388 p.
12. CRUCIANU Michel, ASSELIN DE BEAUVILLE Jean-Pierre, BONE Romuald, Méthodes factorielles pour l'analyse des données : méthodes linéaires et extensions non-linéaires, Hermès - Lavoisier, 2004, 288 p.
13. DENGLOS Grégory, Introduction à l'économétrie, Cours et exercices, PUF, 2009, 238 p.
14. DENUIT Michel, CHARPENTIER Arthur, Mathématiques de l'assurance non-vie : Tome I : Principe fondamentaux de théorie du risque, Michel Denuit, Arthur Charpentier ; Editeur : Economica, 2004, 464 pages
15. DENUIT Michel, CHARPENTIER Arthur, Mathématiques de l'assurance non-vie : Tome II : Tarification et provisionnement, Economica, 2005, 569 p.
16. ESCOFIER Brigitte, PAGES Jérôme, Analyses factorielles simples et multiples ; Objectifs, méthodes et interprétation - Cours et études de cas - Masters et écoles d'ingénieur; Dunod, 4ème édition, 2008, 318 p.
17. EVERITT Brian S. An R and S-Plus Companion to Multivariate Analysis, Springer, 2005, 235 p.
18. EVERITT Brian, LANDAU Sabine, A Handbook of Statistical Analyses Using SPSS; Chapman and Hall / CRC, 2004, 354 p.
19. FOUCART Thierry, L'analyse des données : Mode d'emploi, Presses Universitaires de Rennes (PUR), 1997, 188 p.
20. GEORGIN Jean-Pierre, Analyse interactive des données (ACP, AFC) avec Excel 2000 : Théorie et pratique, Presses Universitaires de Rennes (PUR), 2ème édition 2007, 270 p.
21. GHIL Michael, ROUX Jean, Mathématiques appliquées aux sciences de la vie et de la planète, Licence 3, Master, Ecoles d'ingénieurs - Cours et

exercices corrigés, Dunod, 390 p., 2010

22. HAENNI AMO Anne-Valère, KHAW Christophe, LULA Jonela, ROYER Daniel, Mathématiques de base - Sciences économiques et gestion ; Exercices corrigés avec rappels de théorie, Economica, 3ème édition, 2010, 492 p.
23. HUSSON François, LE Sébastien, PAGES Jérôme, Analyse de données avec R, Presses Universitaires de Rennes (PUR), 2009, 224 p.
24. JALLAIS Sophie, OZGÜR Gün, Introduction à l'algèbre linéaire, PUF, 2011, 292 p.
25. LAFAYE DE MICHEAUX Pierre, DROUILHET Rémy, LIQUET Benoît, Le logiciel R ; Maîtriser le langage - Effectuer des analyses statistiques, Springer, 2010, 486 p.
26. LUZI Michel, Assurance IARD : Interprétation des chiffres, Economica, 2007, 492 p.
27. NAKACHE Jean-Pierre, CONFAIS Josiane, Approche pragmatique de la classification : Arbres hiérarchiques – Partitionnements, Technip, 2004, 262 p.
28. OUELLET Fernando, BAILLARGEON Gérald, Analyse de données avec SPSS pour windows versions 10.0 et 11.0 : Applications en gestion, assurance qualité et en sciences humaines, SMG, 2004, 190 p.
29. PAGES Jérôme, Statistique générale pour utilisateurs - Tome 1, Méthodologie, Presses Universitaires de Rennes (PUR), 2ème édition, 2010, 264 p.
30. PETAUTON Pierre, Théorie de l'assurance dommages, Dunod, 1ère édition, 2000, 182 p.
31. PETAUTON Pierre, Théorie et pratique de l'assurance vie, Dunod, 3ème

édition, 2004, 252 p.

32. SAPORTA Gilbert, Probabilités, analyse des données et statistique, Technip, 2ème édition 2006, 622 p.
33. SIMONET Guy, La comptabilité des entreprises d'assurance, L'argus, 5ème édition, 382 p.
34. STAFFORD Jean, BODSON Paul, L'analyse multivariée avec SPSS, Presses de l'Université du Québec, 2005, 260 p.
35. TENENHAUS Michel, Statistiques : Méthodes pour décrire, expliquer et prévoir, Dunod, 2010, 679 p.
36. TOSETTI Alain, BEHAR Thomas, FROMENTEAU Michel, MENART Stéphane Assurance- Comptabilité – Réglementation- Actuariat, Economica, 2ème édition, 2011, 344 p.
37. TUFFERY Stéphane Data mining et statistique décisionnelle, L'intelligence des données - 3e édition revue et augmentée, Technip, 3ème édition, 2010, 706 p.
38. VENABLES William, RIPLEY Brian D. , Modern applied statistics with S, Springer, 4ème édition, 2002, 496 p.
39. YEATMAN Jérôme, Manuel international de l'assurance, Economica, 2ème édition 2005, 403 p.
40. ZAJDENWEBER Daniel, Economie et gestion de l'assurance, Economica, 2006, 172 p.

### **Mémoires**

1. BOIVIN-CHAMPEAUX Bertrand, Les provisions techniques en assurance non vie : principes de détermination et d'évaluation, proposition de solutions pratiques, Bibliothèque, 1999, 149 p

2. MORELLI Francine, L'analyse et le contrôle des provisions techniques d'une compagnie d'assurance IARD eu égard à son environnement réglementaire et fiscal, Bibliothèque, 1987, 141

### **Sites internet**

- <http://cedric.cnam.fr/~saporta>** Site du professeur SAPORTA proposant de nombreuses ressources statistiques
- <http://eric.univ-lyon2.fr>** Site de l'Equipe de Recherche en Ingénierie des Connaissances (ERIC)
- <http://factominer.free.fr>** Site dédié au package FactoMineR destiné à l'analyse exploratoire multidimensionnelle de données
- <http://pbil.univ-lyon1.fr>** Site de l'Université de Lyon1 permettant d'accéder à des cours de statistiques et ressources relatives au logiciel R
- <http://r.developpez.com/>** Site proposant de nombreuses ressources relatives au logiciel R
- [www.agrocampus-ouest.fr/math](http://www.agrocampus-ouest.fr/math)** Site du laboratoire de mathématiques appliquées d'Agrocampus. Des packages du logiciel R sont développés par les membres de ce laboratoire et notamment le package FactoMineR utilisé dans le mémoire pour l'analyse de données
- [www.r-project.org](http://www.r-project.org)** Site officiel pour se procurer le logiciel R et les documents d'aide
- [www.rstat.ouvaton.org](http://www.rstat.ouvaton.org)** Site proposant des supports de formations aux statistiques appliquées avec le logiciel R

# **ANNEXES**

# ANNEXES

	Page
<b>I. ANNEXES RELATIVES A LA PRESENTATION DES METHODES, DES OUTILS STATISTIQUES ET DES DONNEES</b>	121
1. Compléments sur les méthodes statistiques	121
2. Logiciels de datamining et de statistique	131
3. Présentation des données des sociétés d'assurance	133
4. Données brutes et résumées des sociétés d'assurance	137
5. Données complémentaires pour l'analyse discriminante et la régression logistique	145
<b>II. ANNEXES RELATIVES A L'ANALYSE DES DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES D'ASSURANCE</b>	146
1. Analyse en composantes principales des sociétés non vie	146
2. Classification des sociétés non vie	152
3. Analyse en composantes principales des sociétés vie	162
4. Classification des sociétés vie	167
5. ACP et CAH des sociétés non vie et vie du groupe G	177
<b>III. ANNEXES RELATIVES A LA MISE EN ŒUVRE DES MODELES PREDICTIFS SUR LES DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES D'ASSURANCE</b>	191
1. Régression linéaire multiple sur les données des sociétés non vie et vie	191
2. Analyse discriminante sur les données des sociétés non vie et vie	196

## I. ANNEXES RELATIVES A LA PRESENTATION DES METHODES, DES OUTILS STATISTIQUES ET DES DONNEES

Ces annexes relatives à la présentation des méthodes, des outils statistiques et des données sont composées des 5 éléments suivants :

- Complément sur les méthodes statistiques ;
- Logiciels de datamining et de statistique ;
- Présentation des données des sociétés d'assurance ;
- Données brutes et résumées des sociétés d'assurance ;
- Données complémentaires pour l'analyse discriminante et la régression logistique.

### 1) Compléments sur les méthodes statistiques

#### 1.1) Analyse en composantes principales

##### *Principe du choix du plan principal*

Le principe de l'analyse en composantes principales est exposé dans la première partie suivante du mémoire : **CHAPITRE I.1) Analyse en composantes principales (ACP) page 13**. Les paragraphes ci dessous expliquent sans formalisme mathématique excessif comment le plan principal est judicieusement choisi.

On considère  $x_i$  et  $x_j$  deux éléments (par exemple deux entreprises d'assurance) de  $X$  (matrice ou tableau des données) et  $d(x_i, x_j)$  la distance de l'un à l'autre dans  $R^p$ . En projetant ces éléments sur un plan, la distance entre les deux projections  $d(p(x_i), p(x_j))$  est plus petite que  $d(x_i, x_j)$ . On se fixera comme critère de choix de plan, celui qui maximise la moyenne des carrés des distances entre les projections.

On peut déterminer un plan par deux droites  $D_1$  et  $D_2$  orthogonales entre elles. De part la relation de Pythagore, la distance au carré entre les deux projetés sur ce plan est égale à la somme des deux distances au carré des projections des points sur les deux droites :  $d^2(p(x_i), p(x_j)) = d^2(\alpha_i, \alpha_j) + d^2(\beta_i, \beta_j)$  avec  $\alpha_k$  et  $\beta_k$  les projetés de  $x_k$  ( $1 \leq k \leq n$ ) sur  $D_1$  et  $D_2$  respectivement ( $n$  étant le nombre d'individus statistiques)

Le plan maximisant la moyenne des carrés des distances entre les projections, appelé plan principal peut donc être déterminé itérativement. On commence par

chercher la droite  $D_1$  maximisant la moyenne des  $d^2(\alpha_i, \alpha_j)$  puis une droite  $D_2$  orthogonale à  $D_1$  maximisant la moyenne des  $d^2(\beta_i, \beta_j)$ . On peut continuer le processus et trouver  $p$  droites orthogonales entre elles formant une nouvelle base de  $\mathbb{R}^p$ , appelées axes principaux.

La meilleure représentation des données dans une dimension  $q \leq p$  est alors une représentation de  $X$  sur les  $q$  premiers axes. Ceci est la méthode de l'analyse en composantes principales : remplacer la base canonique  $\mathbb{R}^p$  par une base formée des axes principaux représentant mieux les données et permettre ainsi de réduire l'espace de représentation aux  $q$  axes les plus représentatifs.

### ***Données et caractéristiques***

Les vecteurs propres de la matrice de corrélation associés aux valeurs propres classées par ordre décroissant constituent les composantes principales. Le calcul matriciel, la diagonalisation des matrices, le calcul des valeurs propres et des vecteurs, les notions de matrice variance-covariance et de matrice de corrélation sont importantes en analyse de données et particulièrement en analyse en composantes principales. Ces notions ont été brièvement présentées dans la partie suivante du mémoire : **1.1) Données et caractéristiques page 14.**

Les paragraphes suivants décrivent de façon plus détaillée le passage d'une matrice initiale à une matrice variance covariance puis à une matrice de corrélation.

A l'étape initiale, les  $n$  individus  $x_i$  décrits par  $p$  variables sont rassemblés dans un tableau rectangulaire noté  $X$  à  $n$  lignes et  $p$  colonnes.

$$X = \begin{bmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{bmatrix}$$

Le tableau rectangulaire  $X$  peut être assimilé à une matrice  $X_{n \times p}$  noté simplement également matrice  $X$ . La matrice  $X$  représente l'espace des individus. La matrice transposée de  $X$  notée  ${}^tX$  représente l'espace des variables.

$x_i^j$  est la valeur prise par la variable  $n^\circ j$  sur le  $i^{\text{ème}}$  individu.

Le centre de gravité  $g$  ou point moyen est un individu généralement fictif décrit par les moyennes des différentes variables.

$$g = (\bar{x}^1, \dots, \bar{x}^j, \dots, \bar{x}^p)$$

Une variable est dite centrée si sa moyenne est nulle. L'opération de centrage revient à déplacer le centre du repère vers  $g$  et donc à retirer sa moyenne à chaque variable :  $x^j - \bar{x}^j$ .

A chaque individu  $x_i$  ( $1 \leq i \leq n$ ) est associé un poids  $p_i$  tel que  $\sum_{i=1}^n p_i = 1$ . Ces poids sont rassemblés dans une matrice diagonale =  $diag(p_1, \dots, p_i, \dots, p_n)$ .

Une métrique est une matrice permettant de définir un produit scalaire et donc des distances entre des individus et entre variables. La matrice diagonale  $D$  est une métrique.

On définit la matrice de variance-covariance notée  $V$  par le produit matriciel suivant :  $V = {}^tXDX$ .  $V$  est une matrice carrée contenant à l'intersection de la ligne  $i$  et de la colonne  $j$ , la covariance de ces variables notée  $s_{ij} = cov(x^i, x^j)$ . Cette matrice est symétrique et sa diagonale contient la variance des différentes variables  $s_i^2$ . La somme des variances, appelée également inertie est notée  $I = trace(V)$ . L'inertie exprime la dispersion du nuage de points.

On définit également la matrice diagonale  $D_{1/s} = diag(\frac{1}{s_1}, \dots, \frac{1}{s_j}, \dots, \frac{1}{s_p})$ , ou  $s_j$  désigne l'écart type de la variable  $j$ . La matrice diagonale  $D_{1/s}$  est également une métrique.

Enfin, on définit la matrice de corrélation notée  $R$  par le produit matriciel suivant :

$$R = D_{1/s} {}^tXDXD_{1/s} = D_{1/s}VD_{1/s} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

Dans cette matrice, le terme générique  $r(x^i, x^j) = \frac{cov(x^i, x^j)}{s(x^i)s(x^j)}$  désigne le coefficient de corrélation entre deux variables quelconques. Chaque élément de la diagonale est égal à 1 car toute variable est parfaitement corrélée à elle-même.

Les vecteurs propres de la matrice de corrélation classés par ordre décroissant des valeurs propres constituent les composantes principales, nouvelles variables de synthèse.

## 1.2) Classification

### *Dissimilarité et distance*

La notion de dissimilarité est essentielle en classification. Elle est introduite dans le corps du mémoire dans la section suivante : **2) Classification page 17**. Elle est mieux explicitée dans cette annexe.

Une dissimilarité sur un ensemble d'objets  $X$  est une fonction  $d$  de  $X \times X$  dans l'ensemble des réels positifs  $\mathbb{R}^+$  telle que les propriétés ci-dessous soient satisfaites :

- $(P_1): d(x, y) \geq 0$  pour tous  $x, y \in X$  (positivité)
- $(P_2): d(x, x) = 0$  pour tous  $x \in X$
- $(P_3): d(x, y) = d(y, x)$  pour tous  $x, y \in X$  (symétrie)

Une dissimilarité est dite *propre* lorsque la 4<sup>ème</sup> propriété suivante est vérifiée :

- $(P_4): d(x, y) = 0 \Rightarrow x = y$  pour tous  $x, y \in X$

Une dissimilarité *propre*  $d$  sur  $X$  est appelé une *distance*, si elle satisfait l'inégalité triangulaire :

- $(P_5): d(x, y) \leq d(x, z) + d(z, y)$  pour tous  $x, y, z \in X$

### *Algorithme des k means*

L'algorithme des *k means* ou des centres mobiles est introduit à la section suivante du mémoire : **2.2) Méthodes de partitionnement page 18** . Il permet de partitionner des données euclidiennes.

Les paragraphes suivants présentent les principales étapes de l'algorithme des *k means* ou des centres mobiles :

1. A l'étape initiale, on choisit au mieux  $k$  centres de classe provisoires formés des points individus, appelés aussi noyaux. Les  $k$  noyaux  $\{x(1), \dots, x(k)\}$  permettent de construire une classification initiale  $\{C_1^0, C_2^0, \dots, C_k^0\}$  des  $n$  individus en affectant chaque individu  $j$  à la classe dont il est le plus proche ( $d(x_j, x(i))$  minimum). On calcule ensuite les centres de gravité :

$\{g_1^0, g_2^0, \dots, g_k^0\}$  des classes  $\{C_1^0, C_2^0, \dots, C_k^0\}$ . Ces centres de gravité  $\{g_1^0, g_2^0, \dots, g_k^0\}$  deviennent les nouveaux centres de classe.

2. A l'étape 2, les nouveaux centres de classe  $\{g_1^0, g_2^0, \dots, g_k^0\}$  permettent d'obtenir une nouvelle partition  $\{C_1^1, C_2^1, \dots, C_k^1\}$  en affectant chaque individu  $j$  à la classe dont il est le plus proche ( $d(x_j, g_i^0)$  minimum). On calcule ensuite les centres de gravité  $\{g_1^1, g_2^1, \dots, g_k^1\}$  des classes  $\{C_1^1, C_2^1, \dots, C_k^1\}$ . Ces centres de gravité  $\{g_1^1, g_2^1, \dots, g_k^1\}$  deviennent les nouveaux centres de classe.
3. A l'étape  $m$ , les centres de classe  $\{g_1^{m-1}, g_2^{m-1}, \dots, g_k^{m-1}\}$  obtenus à l'étape précédente permettent d'obtenir une nouvelle partition  $\{C_1^m, C_2^m, \dots, C_k^m\}$  en affectant chaque individu  $j$  à la classe dont il est le plus proche ( $d(x_j, g_i^{m-1})$  minimum). On calcule ensuite les centres de gravité  $\{g_1^m, g_2^m, \dots, g_k^m\}$  des classes  $\{C_1^m, C_2^m, \dots, C_k^m\}$ . Ces centres de gravité  $\{g_1^m, g_2^m, \dots, g_k^m\}$  deviennent les nouveaux centres de classe.

L'algorithme est convergent. Il s'arrête lorsque deux itérations successives conduisent à une décroissance du critère de la somme des carrés intraclasse ou à sa stabilisation.

### 1.3) Analyse factorielle des correspondances (AFC)

L'AFC a été brièvement présentée dans la section suivante du mémoire : **3.2) Analyse factorielle des correspondances page 20** . Elle n'y a pas fait l'objet d'un développement important car elle n'est pas mise en œuvre. Elle a cependant un champ d'application relativement vaste comme le montre l'exemple ci-dessous.

Dans le cadre du mémoire d'expertise comptable, tout candidat doit préciser la discipline (15 au total numérotée de 1 à 15) et le secteur d'activité (22 au total numéroté de A à V) dans lesquels s'inscrit son sujet. En retenant pour chaque étudiant une discipline (variable  $x$ ) et un secteur d'activité (variable  $y$ ), on obtient un tableau de contingence à 15 lignes et 22 colonnes. A l'intersection de la variable ligne  $x_i$  = "audit et commissariat aux comptes" et de la variable colonne  $y_i$  = "assurances et mutuelles", on a l'effectif des candidats dont le sujet porte sur le commissariat aux comptes dans le secteur des assurances et mutuelles. L'effectif total  $N$  correspondra au nombre total de candidats inscrits à la session à l'épreuve du mémoire. L'étude d'un tel tableau de contingence pourrait être riche d'enseignement.

Le test d'indépendance du  $\chi^2$  permet, par le calcul de la statistique suivante  $\chi^2 = \sum_{ij} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$ , de savoir s'il est légitime de rechercher une structure entre les deux variables. La statistique calculée sera proche de 0 si les deux variables sont indépendantes et grande dans le cas contraire.

Les effectifs marginaux du tableau de contingence permettent de définir deux matrices diagonales  $D_L = (n_{1.}, \dots, n_{i.}, \dots, n_{l.})$  et  $D_K = (n_{.1}, \dots, n_{.j}, \dots, n_{.k})$ .

Le tableau de contingence noté  $X$  correspond à priori à deux nuages de points :

- En ligne,  $L$  points dans  $\mathbb{R}^k$ , les  $n_{ij}$  formant les coordonnées du point  $x_i$ ,
- En colonne,  $K$  points dans  $\mathbb{R}^l$ , les  $n_{ij}$  formant les coordonnées du point  $y_j$ .

A partir de la matrice  $X$  et des matrices diagonales  $D_L$  et  $D_K$ , on peut caractériser de façon matricielle les profils lignes et les profils colonnes du tableau de contingence. Ainsi, la matrice  $X_L = D_L^{-1}X$  caractérise les profils lignes et la matrice  $X_K = D_K^{-1}X$  caractérise les profils colonnes.

On définit deux métriques appelées métrique de  $\chi^2$  telles que  $M_L = nD_L^{-1}$  pour l'analyse en ligne et  $M_K = nD_K^{-1}$  pour l'analyse en colonne. Ces métriques permettent de calculer des distances respectivement entre les lignes et les colonnes.

Dans le cadre d'une analyse en composantes principales en ligne et en colonne, on montre que les facteurs propres sont les vecteurs propres des matrices  $M_L V_L$  et  $M_K V_K$  avec :

- $V_L = {}^t X_L \left(\frac{1}{n} D_L\right) X_L$  pour l'analyse en ligne,
- $V_K = {}^t X_K \left(\frac{1}{n} D_K\right) X_K$  pour l'analyse en colonne.

#### 1.4) Analyse des correspondances multiples

L'ACM est une généralisation de l'AFC. Elle a été brièvement présentée dans la section suivante du mémoire : **3.3) Analyse des correspondances multiples page 21**. Comme l'AFC, l'ACM a également un champ d'application relativement vaste comme le montre l'exemple ci-dessous.

Relativement au mémoire d'expertise comptable, les candidats doivent préciser la discipline et le secteur d'activité dans lesquels s'inscrit leur sujet. La discipline et le

secteur d'activité constituent deux variables qualitatives. Une troisième variable qualitative pourrait être la note du mémoire déclinée selon les 5 modalités suivantes : médiocre, passable, moyen, bien, excellent. Avec trois variables, il est impossible d'établir un tableau de contingence simple. Un tableau disjonctif complet des trois variables serait composé de 42 colonnes (15 disciplines +22 secteurs d'activités+5 niveaux de notes) et d'environ 1500 lignes (hypothèse retenue pour le nombre de candidats au sujet de mémoire).

### 1.5) Régression linéaire multiple

#### *Explicitation des termes de la régression linéaire*

La relation entre la variable à expliquer et les variables explicatives est généralement modélisée comme suit :

$$y_i = a_0 + a_1x_{1i} + a_2x_{2i} + \dots + a_kx_{ki} + \varepsilon_i \text{ pour } i = 1, \dots, n$$

Cette écriture est fournie dans le corps du mémoire à la section suivante : **1.1) Présentation des modèles page 23**. Elle peut paraître non explicite. Elle est donc détaillée comme suit :

$$y_1 = a_0 + a_1x_{11} + a_2x_{21} + \dots + a_kx_{k1} + \varepsilon_1$$

$$y_2 = a_0 + a_1x_{12} + a_2x_{22} + \dots + a_kx_{k2} + \varepsilon_2$$

...

$$y_i = a_0 + a_1x_{1i} + a_2x_{2i} + \dots + a_kx_{ki} + \varepsilon_i$$

...

$$y_n = a_0 + a_1x_{1n} + a_2x_{2n} + \dots + a_kx_{kn} + \varepsilon_n$$

Cette écriture détaillée est équivalente à l'écriture matricielle suivante présentée dans le corps du mémoire à la section **1.1) Présentation des modèles page 23**:

$$Y_{n \times 1} = X_{n \times (k+1)} a_{(k+1) \times 1} + \varepsilon_{n \times 1} = Xa + \varepsilon$$

Dans cette écriture matricielle, on a :

$$Y_{n \times 1} = Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}; X_{n \times (k+1)} = X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1i} & x_{2i} & \cdots & x_{ki} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}; a_{(k+1) \times 1} = a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{pmatrix};$$

$$\varepsilon_{n \times 1} = \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

La première colonne de la matrice  $X$ , composée de 1 correspond au coefficient  $a_0$ , le terme constant.

Un tel modèle peut permettre par exemple d'étudier dans une compagnie d'assurance le lien entre le nombre de sinistres survenus  $Y_i$  et le nombre de contrats souscrits  $X_i$ . Le nombre de sinistres survenus ne dépend pas exclusivement du nombre de contrats souscrits. D'autres facteurs peuvent également influencer sur le nombre de sinistres survenus comme par exemple la qualité des risques souscrits. Ces facteurs non explicites sont regroupés dans le terme  $\varepsilon_i$ . Ce terme mesure la différence entre les valeurs réellement observées de  $Y_i$ , nombre de sinistres, et les valeurs observées si la relation définie était rigoureusement exacte. Le terme  $\varepsilon_i$  regroupe trois types d'erreurs :

- une erreur de spécification résultant du fait que la seule variable explicative  $X_i$  n'est pas suffisante pour rendre compte de la totalité du phénomène à expliquer ;
- une erreur de mesure résultant du fait que les données ne représentent pas exactement le phénomène ;
- une erreur de fluctuation d'échantillonnage résultant du fait que d'un échantillon à l'autre, les observations sont légèrement différentes.

#### ***Estimation des paramètres du modèle dans le cadre d'une régression simple***

Les estimateurs des coefficients  $a_0$  et  $a_1$  notés respectivement  $\hat{a}_0$  et  $\hat{a}_1$  sont des variables aléatoires de mêmes lois de probabilité que  $\varepsilon_i$ . Les caractéristiques de moyennes et d'écart type de ces coefficients permettent de construire des tests de validité du modèle estimé.

Les estimateurs des coefficients  $a_0$  et  $a_1$  sont obtenus en minimisant la distance au carré entre chaque observation et la droite, d'où le nom d'estimateur des moindres carrés ordinaires (MCO). Cette méthode est présentée brièvement dans la section suivante du mémoire : **1.1) Présentation des modèles page 23**. Elle est rappelée de façon plus détaillée ci-dessous pour l'estimation des coefficients d'une régression linéaire simple.

La résolution analytique est la suivante :

$$\text{Min} \sum_{i=1}^{i=n} \varepsilon_i^2 = \text{Min} \sum_{i=1}^{i=n} (y_i - a_0 - a_1 x_i)^2 = \text{Min } S$$

En opérant par dérivation par rapport à  $a_0$  et  $a_1$  afin de trouver le minimum de cette fonction, on obtient les résultats suivants :

$$\frac{\delta S}{\delta a_0} = 0 \text{ et } \frac{\delta S}{\delta a_1} = 0$$

Sommant par rapport à  $i$ , il vient :

$$\sum_i x_i y_i - \hat{a}_0 \sum_i x_i - \hat{a}_1 \sum_i x_i^2 = 0$$

$$\sum_i y_i - n \hat{a}_0 - \hat{a}_1 \sum_i x_i = 0$$

Ces équations impliquent que :

$$\hat{a}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}$$

On montre que les estimateurs ainsi déterminés sont sans biais et convergents.

#### ***Hypothèses du modèle linéaire simple***

Les modèles linéaires simple et multiple sont présentés au chapitre suivant du mémoire : **1.1) Présentation des modèles page 23**. Ces modèles se fondent sur des hypothèses qui sont rappelées comme suit pour le modèle linéaire simple:

- $H1$ : le modèle est linéaire en  $x_i$  ;
- $H2$ : les valeurs de  $x_i$  sont observées sans erreur ( $x_i$  non aléatoire) ;

- $H3: E(\varepsilon_i) = 0$ , l'espérance mathématique de l'erreur est nulle, en moyenne le modèle est bien spécifié et donc l'erreur moyenne est nulle ;
- $H4: E(\varepsilon_i^2) = \sigma_\varepsilon^2$ , la variance de l'erreur est constante, cette hypothèse s'appelle également hypothèse d'homoscédasticité ;
- $H5: E(\varepsilon_i, \varepsilon_{i'}) = 0$  si  $i \neq i'$ , les erreurs sont non corrélées (ou encore indépendantes) ;
- $H6: Cov(x_i, \varepsilon_i) = 0$ , l'erreur est indépendante de la variable explicative.

### ***Hypothèses du modèle linéaire multiple***

Les modèles linéaires simple et multiple sont présentés au chapitre suivant du mémoire : **1.1) Présentation des modèles page 23**. Ces modèles se fondent sur des hypothèses qui sont rappelées comme suit pour le modèle linéaire multiple :

- $H1$ : le modèle est linéaire en  $X$  ;
- $H2$ : les valeurs de  $x_{j,i}$  sont observées sans erreur ;
- $H3: E(\varepsilon_i) = 0$ , l'espérance mathématique de l'erreur est nulle, en moyenne le modèle est bien spécifié et donc l'erreur moyenne est nulle ;
- $H4: E(\varepsilon_i^2) = \sigma_\varepsilon^2$ , la variance de l'erreur est constante, cette hypothèse s'appelle également hypothèse d'homoscédasticité ;
- $H5: E(\varepsilon_i, \varepsilon_{i'}) = 0$  si  $i \neq i'$ , les erreurs sont non corrélées (ou encore indépendantes) ;
- $H6: Cov(x_i, \varepsilon_i) = 0$ , l'erreur est indépendante de la variable explicative ;
- $H7$ : l'absence de colinéarité entre les variables explicatives implique que la matrice  $X'X$  est régulière et inversible ;
- $H8: X'X/n$  tend vers une matrice finie non singulière ;
- $H9: n > k + 1$ , le nombre d'observations est supérieur au nombre des séries explicatives.

## 2) Logiciels de datamining et de statistique

### 2.1) Cartographie des principaux logiciels commerciaux de datamining et de statistique

La cartographie des logiciels commerciaux a été introduite à la section **1.1) Logiciels commerciaux page 33** du mémoire. Cette cartographie est fournie ci après :

Volume de données	Produit	Editeur	Site internet
Faible: dizaine de milliers d'enregistrements	NeuralWorks Predict	Neuralware	www.neuralware.com
	NeuroOne	Netral	www.netral.com
	Wizwhy	Wizsoft	www.wizsoft.com
	DataLab	Complex Systems	www.complex-systems.fr
Moyen: centaine de milliers d'enregistrements	Alice	Isoft	www.alice-soft.com
	KnowledgeSEEKER	Angoss	www.angoss.com
	KnowledgeSTUDIO	Angoss	www.angoss.com
	C5.0	RuleQuest Research	www.rulequest.com
	Data Mining Suite	Salford Systems	www.salford-systems.com
	CART	Salford Systems	www.salford-systems.com
	Polyanalyste	Megaputer	www.megaputer.com
	JMP	SAS	www.sas.com
S-PLUS	TIBCO Software	www.tibco.com	
Elevé: millions d'enregistrements	KXEN	KXEN	www.kxen.com
	Oracle Data Mining	Oracle	www.oracle.com
	SPAD	Coheris SPAD	www.coheris.fr
	IBM SPSS Statistics	SPSS (groupe IBM)	www.ibm.com
	IBM SPSS Modeler	SPSS (groupe IBM)	www.ibm.com
	Statistica Data Miner	Statsoft	www.statsoft.fr
	Insighful Miner	TIBCO Software	www.tibco.com
	SAS/STAT	SAS	www.sas.com
	Entreprise Miner	SAS	www.sas.com

**Tableau 40: Logiciels commerciaux**

Le critère discriminant utilisé dans le cadre de cette cartographie est la capacité du logiciel à traiter de grands volumes.

### 2.1) Logiciels libres de datamining et de statistique

La présentation des logiciels libre de datamining et de statistique a été introduite à la section suivante du mémoire : **1.2) Logiciels libres page 33**.

Les paragraphes suivants développent la brève présentation effectuée dans le corps du mémoire et insistent sur quelques logiciels libres ayant acquis une certaine notoriété :

- RapidMiner

RapidMiner est un logiciel open source d'analyse de données et de datamining développé par la société Rapid-I. Cette société propose également une version commerciale de son logiciel avec un service d'assistance. D'autres solutions pour des traitements spécifiques sont également proposées par cette société.

Le logiciel est téléchargeable à l'adresse suivante <http://rapid-i.com/content/view/281/225/lang,en/>.

- Tanagra

Tanagra est un outil open source de datamining développé essentiellement par le Professeur Ricco Rakotomalala du laboratoire ERIC (Equipe de Recherche et d'Ingénierie des Connaissances) de l'université Lyon 2. Le logiciel implémente des méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

Le logiciel est téléchargeable à l'adresse suivante <http://chirouble.univ-lyon2.fr/~ricco/tanagra/>.

- WEKA

WEKA est un outil open source développé par l'université de Waikato en Nouvelle Zélande. Le logiciel dispose d'outils permettant de faire du traitement de données, de la classification, de la régression, du clustering et des règles d'association.

Le logiciel est téléchargeable à l'adresse suivante <http://www.cs.waikato.ac.nz/ml/weka/>.

Le tableau suivant fourni une liste de logiciels libres de statistiques et de datamining et les sites internet dédiés à ces logiciels :

Logiciel	Site internet
Adam	<a href="http://datamining.itsc.uah.edu/adam/">http://datamining.itsc.uah.edu/adam/</a>
Alphaminer	<a href="http://www.eti.hku.hk/alphaminer/">http://www.eti.hku.hk/alphaminer/</a>
Databionic ESOM Tools	<a href="http://databionic-esom.sourceforge.net/">http://databionic-esom.sourceforge.net/</a>
ELKI	<a href="http://elki.dbs.ifi.lmu.de">http://elki.dbs.ifi.lmu.de</a>
Gnome Data Mining Tools	<a href="http://www.togaware.com/datamining/gdatamine/">http://www.togaware.com/datamining/gdatamine/</a>
IBM Intelligent Miner	<a href="http://www.developer.ibm.com/university/scholars/">http://www.developer.ibm.com/university/scholars/</a>
KEEL	<a href="http://www.keel.es/">http://www.keel.es/</a>
KNIME	<a href="http://www.knime.org">http://www.knime.org</a>
Mining Mart	<a href="http://mmart.cs.uni-dortmund.de/">http://mmart.cs.uni-dortmund.de/</a>
MLC++	<a href="http://www.sgi.com/tech/mlc/">http://www.sgi.com/tech/mlc/</a>
Machine Learning in Java (MLJ)	<a href="http://www.kddresearch.org/Groups/Machine-Learning/MLJ/">http://www.kddresearch.org/Groups/Machine-Learning/MLJ/</a>
Orange	<a href="http://magix.fri.uni-lj.si/orange/">http://magix.fri.uni-lj.si/orange/</a>
RapidMiner	<a href="http://rapid-i.com/">http://rapid-i.com/</a>
Star Probe	<a href="http://www.roselabdb.com/starprobe.htm">http://www.roselabdb.com/starprobe.htm</a>
TANAGRA	<a href="http://chirouble.univ-lyon2.fr/ricco~/tanagra/">http://chirouble.univ-lyon2.fr/ricco~/tanagra/</a>
Weka	<a href="http://www.cs.waikato.ac.nz/ml/weka/">www.cs.waikato.ac.nz/ml/weka/</a>

**Tableau 41: Logiciels libres**

### **3) Présentation des données des sociétés d'assurance**

La structure des données objet de l'étude a été présentée brièvement dans le corps du mémoire à la section suivante : **2) Présentation et reformulation des données page 35.**

Les paragraphes suivants fournissent des informations complémentaires sur la collecte et la structure des données.

#### **3.1) Contexte de la collecte et du traitement des données**

La Conférence Interafricaine des Marchés d'Assurances (CIMA) est une organisation internationale, mise en place dans le cadre d'un Traité, chargée de l'intégration, du développement et de la supervision de l'industrie des assurances dans quatorze (14) pays d'Afrique centrale et d'Afrique de l'ouest.

Dans le cadre de son activité de supervision, le Secrétariat Général de la CIMA reçoit annuellement les états financiers de toutes les compagnies d'assurance exerçant sur ce marché. A la fin de l'exercice 2009, cet espace regroupait 90 sociétés d'assurance non vie réalisant un chiffre d'affaires de 384 milliards de FCFA

(585 millions d'euros) et 45 sociétés d'assurance vie avec un chiffre d'affaires de 153 milliards de FCFA (233 millions d'euros).

L'exploitation des données produites par les entreprises du marché fait l'objet d'un rapport annuel produit par le Secrétariat Général de la CIMA. Ce rapport d'environ 80 pages, hors annexes, fournit des informations sur l'évolution du marché de l'assurance et la solvabilité globale des entreprises exerçant dans ce secteur.

Aucune méthode d'analyse multidimensionnelle des données n'est cependant utilisée dans le cadre de l'établissement du rapport annuel et dans le cadre de l'activité de supervision des commissaires contrôleurs de la CIMA. Ces méthodes permettent d'étudier simultanément plusieurs individus statistiques décrits par plusieurs variables qualitatives ou quantitatives.

L'observation effectuée sur le rapport annuel produit par le Secrétariat Général de la CIMA peut s'étendre aux rapports annuels établis par de nombreuses autorités de supervisions similaires et aux rapports annuels des groupes de sociétés.

### 3.2) Présentation des données des sociétés d'assurance non vie

Les données sont relatives à 90 compagnies d'assurance non vie de l'espace CIMA. Elles sont extraites des bilans, des comptes de résultat, des tableaux de marge de solvabilité et de couverture des engagements réglementés de l'exercice 2009. Dix neuf variables ont été retenues.

Le tableau suivant présente lesdites variables :

Nom de la variable	Observation
PRIMES	V1 Primes émises
SIN_REG	V2 Sinistres réglés
AR_BRUTS	V3 Arriérés de primes bruts
AR_NETS	V4 Arriérés de primes nets de provision pour dépréciation
PF_NETS	V5 Produits financiers nets des charges financières
COM	V6 Commissions payées aux intermédiaires
FG	V7 Frais généraux
PRI_REA	V8 Primes cédées aux réassureurs
SIN_REA	V9 Sinistres réglés par les réassureurs
SAP_REA	V10 Provisions pour sinistres à payer à la charge des réassureurs
COM_REA	V11 Commissions reçues des réassureurs
PSAP	V12 Provision pour sinistres à payer
PAP	V13 Provision pour annulations de primes
CAP	V14 Capitaux propres
T_BILAN	V15 Total bilan
REXP	V16 Résultat d'exploitation
RHE	V17 Résultat hors exploitation
E_R	V18 Total des engagements réglementés
ACTIFS_A	V19 Total des actifs admis en couverture des engagements réglementés

Tableau 42: Variables relatives aux données des compagnies d'assurance non vie

Pour les sociétés d'assurance non vie, les données sont fournies dans l'annexe au point suivant : **4.1) Données des sociétés d'assurance non vie page 137.**

### 3.3) Présentation des données des compagnies d'assurance vie

Les données sont relatives à 45 compagnies d'assurance vie de l'espace CIMA. Elles sont extraites des bilans, des comptes de résultat, des tableaux de marge de solvabilité et de couverture des engagements réglementés de l'exercice 2009. Vingt deux variables ont été retenues.

Le tableau suivant présente lesdites variables :

Variable		Observation
Primes	V1	Primes émises
PF	V2	Produits financiers
Sinistres	V3	Charge de sinistres
Com	V4	Charge de commissions
FG	V5	Frais généraux
P_Réass	V6	Primes cédées aux réassureurs
Charges_réa	V7	Charges cédées aux réassureurs
RE	V8	Résultat d'exploitation
RN	V9	Résultat net
S_payés	V10	Sinistres payés
S_réa	V11	Sinistres payés à la charge des réassureurs
Com_réa	V12	Commissions de réassurance
PM	V13	Provisions mathématiques
A_bruts	V14	Arriérés bruts
A_nets	V15	Arriérés nets
Ktal	V16	Capital
FP	V17	Fonds propres
PT	V18	Provisions techniques
T_bilan	V19	Total bilan
Actifs	V20	Actifs admis en couverture des engagements réglementés
ER	V21	Engagements réglementés
Actifs.	V22	Actifs admis en couverture des engagements réglementés et avances sur polices

Tableau 43: Variables relatives aux données des compagnies d'assurance vie

Pour les sociétés d'assurance vie, les données sont fournies dans l'annexe au point suivant : **4.3) Données des sociétés d'assurance vie page 142.**

#### 4) Données brutes et résumées des sociétés d'assurance

Les tableaux suivants contiennent les données (brutes et résumées) des sociétés d'assurance vie et non vie, objet de l'étude. Sur ces données, sont mises en œuvre dans le mémoire différentes techniques statistiques : analyse en composantes principales, classification ascendante hiérarchique, régression linéaire multiple, analyse discriminante et régression logistique.

##### 4.1) Données des sociétés d'assurance non vie

Les données sont consignées dans 4 tableaux notés A, B, C et D. Elles sont en milliers de FCFA.

	PRIMES	SIN_REG	AR_BRUTS	AR_NETS	PF_NETS	COM	FG	PRI_REA	SIN_REA	SAP_REA
A1	22 953 749	- 5 473 068	3 944 098	3 944 098	555 721	- 2 816 568	- 4 207 550	- 13 479 084	1 448 102	2 042 948
A2	20 607 464	- 12 733 605	7 296 400	6 714 295	592 948	- 2 898 092	- 2 453 512	- 8 629 554	6 031 713	2 629 641
A3	17 434 533	- 5 804 252	8 243 079	8 055 763	558 418	- 269 493	- 4 265 856	- 11 842 713	3 226 474	1 915 865
A4	13 563 400	- 4 372 709	5 089 826	5 089 826	202 904	- 2 191 043	- 2 487 910	- 2 913 095	614 144	1 302 052
A5	12 339 581	- 5 448 815	3 768 065	3 768 065	295 609	- 1 715 587	- 3 466 112	- 2 561 794	732 142	4 766 006
A6	11 222 631	- 9 741 082	2 771 628	2 649 691	1 285 403	- 1 300 489	- 3 412 248	- 6 028 270	8 329 518	7 969 502
A7	11 094 279	- 4 032 970	4 213 876	3 275 426	412 961	- 1 215 482	- 2 845 414	- 3 299 805	801 488	2 261 592
A8	10 003 516	- 3 781 421	14 972 844	8 334 936	- 75 947	- 1 682 185	- 1 570 257	- 847 148	460 935	570 058
A9	9 918 871	- 3 970 235	4 080 685	4 047 296	86 240	- 1 643 771	- 1 702 922	- 3 215 065	1 051 425	2 291 117
A10	9 585 203	- 3 890 774	1 993 054	1 982 415	1 679 842	- 1 014 298	- 2 542 027	- 5 219 137	1 738 921	1 536 330
A11	8 395 102	- 2 176 176	2 112 543	2 050 381	487 990	- 1 228 930	- 2 737 797	- 4 145 234	613 165	1 734 555
A12	8 019 471	- 2 818 028	2 836 151	2 570 604	435 583	- 1 403 885	- 1 694 446	- 2 509 460	556 423	1 268 026
A13	8 007 531	- 5 615 192	4 388 698	4 388 698	742 040	- 1 412 089	- 1 481 893	- 1 877 202	2 188 024	3 350 654
A14	7 716 600	- 1 962 396	4 380 832	3 978 418	345 149	- 838 012	- 2 812 987	- 3 437 659	390 860	4 230 244
A15	7 566 368	- 3 127 812	2 337 270	2 337 270	118 024	- 1 199 129	- 1 708 390	- 3 173 440	1 159 699	1 581 504
A16	7 053 666	- 1 570 753	1 985 726	1 889 669	658 356	- 789 176	- 2 083 225	- 3 614 091	371 566	1 713 452
A17	6 973 806	- 1 833 479	3 666 027	3 666 027	33 135	- 1 239 132	- 1 537 093	- 1 630 246	260 548	-
A18	6 873 578	- 2 880 741	2 288 888	2 288 888	29 136	- 955 560	- 1 545 839	- 994 393	110 384	264 354
A19	6 570 761	- 1 617 579	9 556 511	9 472 077	513 215	- 1 402 298	- 2 717 089	- 2 202 552	296 435	3 250 618
A20	6 522 954	- 2 462 251	2 840 610	2 840 610	141 992	- 617 604	- 2 125 671	- 2 281 377	457 516	1 177 489
A21	6 509 817	- 2 034 752	4 622 297	2 808 174	426 382	- 905 526	- 2 079 936	- 986 112	293 461	1 563 235
A22	5 919 064	- 2 211 430	2 149 696	2 149 696	128 793	- 723 395	- 1 336 876	- 1 358 539	24 583	505 513
A23	5 689 676	- 1 864 282	1 696 109	1 696 109	21 760	- 780 777	- 1 284 929	- 1 259 603	240 593	417 599
A24	5 618 950	- 1 874 452	4 438 647	4 438 647	34 670	- 617 221	- 1 725 363	- 1 913 822	234 493	1 136 606
A25	5 388 824	- 1 423 323	4 504 717	4 504 717	576 880	- 380 061	- 1 985 296	- 1 688 449	222 364	1 155 063
A26	4 675 904	- 3 181 876	5 876 588	5 876 588	111 022	- 320 343	- 2 346 418	- 339 904	742 848	704 620
A27	4 513 681	- 549 905	2 497 678	2 497 678	131 448	- 631 002	- 1 014 287	- 2 862 380	240 206	1 258 688
A28	4 382 285	- 1 967 594	1 703 418	1 558 464	301 293	- 480 412	- 875 316	- 1 944 098	748 371	519 358
A29	4 157 667	- 856 384	2 944 420	2 944 420	4 885	- 442 879	- 939 165	- 1 889 721	137 942	260 625
A30	3 859 737	- 1 371 450	1 504 153	1 504 153	146 521	- 429 517	- 976 826	- 1 092 665	233 523	524 587
A31	3 851 817	- 1 549 475	740 624	740 624	2 176	- 481 441	- 1 167 038	- 561 670	26 785	654 969
A32	3 723 087	- 986 945	2 172 900	1 723 014	256 574	- 440 487	- 1 547 787	- 1 151 300	157 228	1 138 209
A33	3 651 181	- 1 087 383	2 189 966	2 189 966	55 828	- 526 944	- 763 158	- 1 396 887	503 553	380 397
A34	3 566 074	- 1 275 262	2 616 725	2 616 725	164 287	- 887 025	- 1 370 730	- 246 981	4 308	23 300
A35	3 514 568	- 1 071 501	1 266 810	1 266 810	93 462	- 397 734	- 1 355 188	- 535 339	76 109	334 983
A36	3 389 258	- 422 053	1 222 060	1 222 060	94 980	- 262 578	- 728 519	- 1 432 655	12 663	58 330
A37	3 367 512	- 1 163 555	2 738 720	2 738 720	28 979	- 500 702	- 1 083 073	- 612 483	195 897	232 699
A38	3 329 285	- 463 018	1 220 666	1 220 666	- 31 989	- 387 842	- 1 175 649	- 308 910	2 157	269 576
A39	3 295 107	- 1 214 372	4 995 925	4 995 925	- 225 551	- 367 663	- 1 994 004	- 508 904	21 075	629 877
A40	3 240 685	- 1 732 987	2 464 641	2 464 641	16 612	- 219 043	- 1 162 572	- 247 538	84 784	124 356
A41	3 228 090	- 1 412 393	2 087 070	2 087 070	80 337	- 600 921	- 825 609	- 975 565	450 318	578 594
A42	3 195 868	- 526 698	1 293 629	1 211 714	10 000	- 572 682	- 1 151 448	- 280 544	-	53 902
A43	3 167 525	- 1 045 317	2 438 577	871 394	- 9 178	- 408 061	- 3 132 081	- 477 708	102 458	529 086
A44	3 024 035	- 1 349 873	753 417	753 417	76 773	- 470 468	- 1 008 908	- 295 743	123 307	332 671
A45	3 023 067	- 786 759	582 250	8690	54 770	-	- 1 624 483	- 25911	-	-

Tableau 44: Données sociétés non vie - tableau A

	COM_REA	PSAP	PAP	CAP	T_BILAN	R_EXP	RHE	E_R	ACTIFS_A
A1	2 480 468	- 11 239 154	- 558 439	2 569 111	21 918 288	404 199	- 212 891	- 13 183 575	14 496 407
A2	1 062 620	- 12 267 624	- 2 141 000	7 199 215	40 336 606	1 836 767	- 330 874	- 14 085 894	23 644 338
A3	1 902 341	- 15 054 016	- 511 276	5 306 494	37 040 927	2 597 011	- 1 593 425	- 19 364 782	19 335 010
A4	399 128	- 5 995 145	- 1 038 363	2 007 836	18 803 851	999 625	- 390 971	- 8 142 276	7 613 778
A5	480 209	- 10 445 603	- 717 961	4 984 632	28 808 423	952 217	266 363	- 13 575 883	13 240 336
A6	1 411 771	- 14 716 191	- 433 174	3 269 293	29 184 177	2 595 807	- 518 199	- 18 246 610	18 037 547
A7	570 436	- 8 371 736	- 316 635	3 916 314	19 884 400	1 362 937	- 96 606	- 10 940 825	10 741 991
A8	230 068	- 7 933 254	- 702 216	3 327 935	18 662 789	1 001 281	- 442 660	- 9 280 694	9 464 666
A9	822 405	- 10 525 466	- 236 000	1 787 283	19 651 730	623 308	- 67 272	- 12 307 097	11 800 732
A10	778 558	- 9 981 689	- 286 457	2 786 058	19 828 295	1 006 605	132 939	- 11 368 185	13 560 683
A11	718 987	- 11 262 695	- 288 826	5 639 675	22 542 904	408 070	- 163 734	- 12 892 904	12 580 224
A12	435 761	- 9 365 585	- 258 078	2 536 671	16 751 242	601 854	- 99 025	- 11 154 334	11 966 910
A13	321 594	- 11 831 624	- 47 695	4 180 047	22 037 735	1 017 161	- 252 514	- 13 192 879	11 433 650
A14	574 143	- 9 735 175	- 98 669	1 795 113	22 325 407	- 427 559	785 437	- 11 182 684	10 588 015
A15	507 868	- 4 305 729	- 178 189	1 104 431	15 317 471	951 141	- 156 882	- 5 961 231	4 434 024
A16	531 149	- 9 994 856	- 301 495	1 917 360	18 870 703	418 876	213 796	- 11 052 035	14 163 212
A17	373 694	- 1 533 894	- 1 702 367	1 098 470	7 215 723	200 153	- 63 112	- 3 316 503	2 806 013
A18	236 706	- 2 362 919	-	497 993	6 632 806	174 918	- 16 465	- 3 430 180	3 097 064
A19	561 869	- 10 183 289	- 3 574 768	6 216 534	28 151 742	- 761 400	- 1 144 692	- 12 111 456	9 584 649
A20	407 822	- 6 264 136	- 761 638	1 425 761	13 582 422	- 5 384	249 001	- 7 744 155	7 257 350
A21	226 360	- 10 143 327	- 91 538	3 221 086	24 896 811	845 761	- 34 660	- 17 172 230	18 823 250
A22	445 891	- 4 548 846	- 305 716	1 880 847	10 635 093	140 092	28 203	- 6 311 348	6 694 587
A23	351 930	- 2 254 367	- 159 356	977 844	6 583 148	442 972	- 138 215	- 3 699 148	2 850 363
A24	467 774	- 3 201 612	- 146 697	2 142 465	9 369 002	137 953	- 60 418	- 5 310 991	-
A25	467 522	- 8 700 913	- 457 250	2 537 231	17 172 300	533 905	- 140 625	- 11 097 730	12 067 435
A26	91 635	- 4 686 841	- 762 618	2 568 462	15 789 945	82 680	140 050	- 6 052 521	6 709 428
A27	920 301	- 3 413 338	- 191 808	2 177 972	11 613 684	174 729	- 61 320	- 4 294 581	2 618 998
A28	352 450	- 3 194 085	- 323 309	1 159 753	8 594 073	647 194	- 181 208	- 4 678 763	4 432 732
A29	452 847	- 1 197 724	- 359 000	456 733	7 046 987	278 416	- 19 367	- 1 525 496	1 470 271
A30	211 359	- 3 000 005	- 104 500	1 003 612	6 224 296	448 516	- 148 198	- 3 885 561	3 521 827
A31	87 072	- 1 039 850	- 15 166	602 181	3 266 243	- 19 617	- 10 695	- 1 421 849	985 727
A32	197 772	- 4 537 375	- 232 260	1 519 144	11 651 916	675 941	- 217 077	- 5 780 447	6 019 184
A33	312 222	- 2 020 284	- 220 041	630 899	6 375 559	454 428	- 212 557	- 2 762 007	1 786 112
A34	84 970	- 2 848 770	- 904 895	1 637 303	7 978 207	254 580	- 83 363	- 3 962 963	4 107 655
A35	123 230	- 4 150 807	- 100 000	900 991	6 910 949	296 341	- 38 204	- 4 417 756	4 343 027
A36	380 398	- 660 348	-	837 679	4 663 015	904 526	- 319 365	- 1 831 769	2 185 733
A37	191 025	- 1 924 446	- 424 880	903 430	5 780 850	253 574	- 107 742	- 2 596 228	2 320 567
A38	51 537	- 4 899 929	- 112 882	1 128 114	9 197 480	625 445	- 444 478	- 5 952 604	3 006 417
A39	123 745	- 4 061 274	- 526 358	4 483 450	18 669 401	- 401 050	- 236 006	- 6 067 359	5 083 821
A40	46 665	- 1 543 251	- 1 100 000	1 071 344	6 197 565	57 755	- 29 506	- 1 942 936	2 702 985
A41	223 374	- 3 846 840	- 336 304	1 629 576	8 618 773	- 125 267	204 118	- 4 474 546	4 041 516
A42	58 536	- 875 876	- 231 375	524 677	2 883 822	103 495	- 42 973	- 1 070 634	1 148 154
A43	103 226	- 3 405 648	- 15 617	- 1 911 341	7 238 032	- 1 888 245	364 705	- 4 330 821	688 806
A44	83 432	- 2 890 000	- 285 000	1 044 415	5 963 654	298 366	- 143 645	- 3 621 536	5 791 273
A45	-	- 3 745 407	-	566 846	5 068 416	- 245 480	404 538	- 4 332 339	4 577 604

**Tableau 45 : Données sociétés non vie - tableau B**

	PRIMES	SIN_REG	AR_BRUTS	AR_NETS	PF_NETS	COM	FG	PRI_REA	SIN_REA	SAP_REA
A46	2 956 490	- 1 022 755	819 703	819 703	81 625	- 211 050	- 511 737	- 1 143 079	167 515	501 040
A47	2 837 870	- 931 320	1 279 679	1 279 679	80 319	- 403 223	- 677 264	- 804 064	64 891	671 522
A48	2 650 789	- 967 157	2 335 274	1 307 689	- 64 053	- 137 918	- 1 024 942	- 259 982	68 794	242 454
A49	2 571 255	- 1 303 117	2 220 524	1 624 486	65 362	- 460 113	- 797 687	- 101 891	-	71 708
A50	2 562 973	- 1 012 164	1 938 663	1 914 430	30 571	- 333 230	- 729 356	- 313 554	196 487	171 982
A51	2 504 685	- 1 095 094	356 394	288 569	218 772	- 607 716	- 863 816	- 194 603	85 187	136 517
A52	2 338 273	- 533 354	3 723 381	3 723 381	- 89 101	- 197 696	- 1 090 949	- 94 500	-	-
A53	2 333 498	- 1 714 532	5 334 293	5 334 293	80 679	- 415 024	- 1 670 976	- 298 382	70 710	210 164
A54	2 320 294	- 868 347	7 278 571	7 278 571	228 585	- 273 488	- 637 700	- 259 136	357 450	384 593
A55	2 263 441	- 858 343	1 075 723	1 042 223	9 718	- 378 196	- 551 870	- 215 271	71 335	106 732
A56	2 232 687	- 911 065	4 937 220	4 937 220	- 45 078	- 276 569	- 985 120	- 26 568	-	846 156
A57	2 207 020	- 906 125	5 036 236	875 963	329 602	- 338 558	- 906 938	- 142 242	11 723	177 113
A58	2 204 984	- 39 623	932 891	899 983	201	- 399 409	- 937 745	- 267 955	1 304	20 026
A59	2 077 153	- 605 296	882 581	882 581	119 291	- 253 698	- 611 755	- 656 492	40 782	134 968
A60	2 028 566	- 396 572	629 452	601 555	60 124	- 153 965	- 534 052	- 675 703	90 045	218 139
A61	1 984 753	- 878 300	5 195 363	815 422	- 44 480	- 278 366	- 821 014	- 195 851	11 461	232 595
A62	1 972 286	- 66 264	544 994	544 994	- 2 208	- 330 965	- 588 298	- 383 309	-	27 222
A63	1 823 460	- 632 564	3 679 120	2 136 759	- 65 901	- 266 670	- 804 181	- 119 802	53 631	573 564
A64	1 809 621	- 508 177	874 418	874 418	23 596	- 314 619	- 746 774	- 250 634	128 915	557 052
A65	1 684 423	- 274 665	570 039	570 039	137 761	- 168 758	- 1 022 636	- 95 564	-	330 973
A66	1 620 199	- 387 440	1 027 101	935 047	71 419	- 152 490	- 610 012	- 604 535	12 814	261 595
A67	1 617 825	- 886 429	2 433 981	2 408 057	564 627	- 160 065	- 466 107	- 97 059	79 088	579 428
A68	1 597 301	- 883 932	16 877 393	6 709 747	184 868	- 202 162	- 2 578 055	- 78 198	76 588	578 336
A69	1 427 810	- 603 666	829 897	829 897	- 24 468	- 179 677	- 536 032	- 128 485	3 881	16 363
A70	1 423 541	- 177 705	581 479	581 479	27 602	- 312 936	- 437 299	- 319 373	27 599	626 922
A71	1 375 826	- 428 707	984 061	984 061	33 191	- 135 446	- 311 613	- 228 337	41 236	95 165
A72	1 303 841	- 697 993	1 099 400	687 440	72 139	- 136 881	- 802 132	- 66 234	3 495	107 423
A73	1 280 267	- 356 664	343 251	343 251	78 043	- 220 511	- 342 115	- 271 868	62 894	146 050
A74	1 244 303	- 21 438	566 942	566 942	35 700	- 200 219	- 432 495	- 103 627	-	-
A75	1 237 448	- 237 033	566 952	566 952	16 177	- 172 783	- 404 515	- 372 624	31 075	133 249
A76	1 173 679	- 304 699	1 623 101	1 383 726	- 177 174	- 135 911	- 915 565	- 118 009	-	5 691
A77	1 092 940	- 913 581	644 124	644 124	141 298	- 77 528	- 572 035	- 239 540	410 308	77 243
A78	1 070 286	- 173 228	451 143	341 727	7 034	- 142 497	- 703 585	- 159 082	6 002	107 219
A79	1 044 940	- 169 668	321 055	321 055	16 306	- 89 630	- 321 972	- 168 224	-	100 000
A80	1 013 109	- 109 029	698 361	698 361	- 11 043	- 114 869	- 416 428	- 334 487	11 615	157 726
A81	977 695	- 177 816	102 385	102 385	18 909	- 110 554	- 384 060	- 161 170	5 078	123 595
A82	782 568	- 20 996	113 973	91 359	28 194	- 24 483	- 293 712	- 230 869	4 213	113 800
A83	752 273	- 15 587	192 932	192 932	- 1 786	- 117 736	- 265 745	- 275 250	1 462	17 231
A84	665 053	- 92 839	536 333	536 333	2 279	- 64 305	- 266 554	- 152 840	11 267	45 808
A85	605 993	- 89 651	738 608	738 608	- 2 079	- 6 576	- 348 169	- 218 538	5 562	12 426
A86	591 413	- 26 190	235 106	235 106	- 26 413	- 83 263	- 405 538	- 60 870	-	-
A87	386 666	- 98 119	520 528	520 528	11 841	- 16 880	- 214 731	- 105 478	25 544	129 941
A88	265 159	-	-	-	1 652	- 6 279	- 203 227	- 78 553	-	-
A89	245 951	- 28 603	114 034	114 034	- 10 045	- 3 983	- 257 737	- 57 107	13 018	4 792
A90	166 475	- 11 559	-	-	201 157	- 26 267	- 60 861	- 19 611	-	5 201

Tableau 46: Données sociétés non vie - tableau C

	COM_REA	PSAP	PAP	CAP	T_BILAN	R_EXP	RHE	E_R	ACTIFS_A
A46	77 198	- 2 211 896	- 90 000	649 852	4 255 678	158 769	- 62 550	- 2 732 912	2 291 837
A47	161 063	- 2 663 875	- 98 343	677 617	5 946 290	191 531	- 51 082	- 3 131 260	2 665 314
A48	58 513	- 2 310 425	- 5 600	1 097 041	6 249 685	30 175	- 1 859	- 3 947 876	3 584 829
A49	-	- 969 402	-	537 504	3 553 336	45 183	492	- 1 028 264	1 214 224
A50	105 570	- 1 198 764	- 324 470	1 157 849	1 150 959	100 354	39 409	- 1 983 080	1 424 460
A51	8 378	- 4 768 977	- 38 436	645 586	6 575 518	140 517	- 10 793	- 5 368 348	5 478 253
A52	-	- 1 127 721	-	814 272	5 253 617	246 671	- 135 372	- 1 438 158	1 112 757
A53	219 867	- 4 246 400	- 1 030 824	2 430 615	12 079 999	856 707	- 398 636	- 5 344 384	4 603 535
A54	69 242	- 3 236 573	- 1 917 594	866 937	11 262 376	132 410	51 782	- 3 812 582	2 425 100
A55	47 001	- 1 203 033	-	1 112 406	3 443 084	47 372	66 953	- 1 723 212	1 338 404
A56	-	- 1 939 503	- 2 863 107	1 069 196	8 300 108	- 372 471	55 655	- 2 487 712	2 258 608
A57	25 325	- 1 564 505	-	1 008 347	13 270 929	177 598	- 242 347	- 3 897 533	6 643 307
A58	32 136	- 425 118	-	229 555	2 335 982	369 081	- 80 938	- 728 317	950 553
A59	146 157	- 1 118 822	- 70 000	537 328	3 401 303	145 168	- 46 478	- 1 507 967	2 052 145
A60	182 630	- 1 408 974	- 117 179	883 129	3 921 149	201 926	- 131 668	- 1 947 592	2 488 159
A61	34 309	- 2 436 862	- 107 630	797 707	7 823 991	- 387 999	308 059	- 3 503 526	5 141 918
A62	76 080	- 475 621	-	500 000	2 151 973	- 174 716	-	- 1 242 606	882 566
A63	23 305	- 4 098 715	- 142 576	763 592	5 915 573	368 787	4 275	- 3 993 033	1 687 110
A64	-	- 2 732 516	- 207 045	564 514	5 239 097	154 404	- 2 233	- 3 255 679	3 113 956
A65	16 085	- 4 490 485	- 10 000	779 234	6 732 803	29 743	- 16 099	- 4 915 856	5 559 749
A66	119 194	- 2 232 799	- 2 118	800 000	4 683 138	145 860	- 78 808	- 2 633 502	2 546 051
A67	85 153	- 2 295 244	-	6 576 230	15 509 576	1 104 325	- 857 798	- 2 628 298	2 967 042
A68	12 453	- 6 716 140	- 663 912	1 837 229	18 291 543	- 906 162	457 184	- 7 906 021	8 721 170
A69	23 705	- 206 149	-	180 517	1 386 939	- 18 542	- 34 799	- 524 686	71 283
A70	69 841	- 1 330 222	- 52 677	585 233	3 635 225	129 111	- 53 921	- 1 569 564	1 586 725
A71	50 546	- 470 004	- 38 479	558 513	2 499 503	349 648	448	- 1 003 059	979 450
A72	-	- 2 717 813	-	659 299	6 161 831	458 042	- 16 425	- 3 726 454	5 011 093
A73	51 766	- 847 880	8 729	569 819	2 279 636	141 867	- 19 182	- 1 219 470	1 276 797
A74	26 704	- 309 992	-	500 000	1 492 309	- 19 856	- 12 554	- 624 425	488 829
A75	79 147	- 711 642	- 58 760	323 140	1 802 810	81 979	- 14 593	- 862 305	428 534
A76	898	- 1 710 892	-	744 622	3 935 227	- 1 031 852	- 50 688	- 2 026 740	178 822
A77	81 386	- 2 668 139	- 172 530	595 640	4 793 484	9 631	32 759	- 2 809 931	2 860 187
A78	26 415	- 760 047	-	249 402	1 440 545	- 230 055	36 585	- 1 014 313	357 527
A79	-	- 734 936	- 15 000	336 020	1 537 247	15 093	- 7 012	- 899 518	740 672
A80	66 206	- 568 300	- 50 655	500 000	1 496 778	107 381	- 254 311	- 665 802	335 608
A81	28 397	- 1 314 975	- 54 846	238 489	1 724 485	- 205 413	- 5 979	- 1 750 843	786 717
A82	63 555	- 188 461	-	1 012 493	2 390 949	23 957	4 259	- 1 520 702	1 566 298
A83	64 014	- 146 321	- 15 025	600 000	1 388 884	- 16 492	-	- 349 835	457 486
A84	27 070	- 181 797	-	412 665	4 377 835	- 59 927	-	- 379 466	188 339
A85	12 656	- 190 647	- 30 369	366 161	2 178 774	- 188 741	- 58 236	- 586 428	952 243
A86	7 098	- 40 278	-	500 000	1 269 569	- 235 923	-	- 271 321	140 257
A87	21 342	- 854 102	- 109 538	278 580	1 836 821	1 964	- 61 329	- 921 293	793 100
A88	27 494	-	-	272 836	990 498	- 11 671	-	- 86 862	375 471
A89	17 571	- 23 961	-	741 860	941 683	- 3 767	- 2 000	- 126 558	673 481
A90	4 540	- 119 856	- 92	4 195 270	4 758 496	226 146	3 114	- 130 586	133 150

**Tableau 47: Données sociétés non vie - tableau D**

#### 4.2) Données résumées des sociétés d'assurance non vie

L'un des objectifs du mémoire est de montrer comment les techniques statistiques d'analyses des données « multivariées » peuvent être utilisées avec profit par un professionnel comptable. Même si ces techniques permettent d'analyser les données en tenant compte de leur caractère multidimensionnel, il peut être intéressant d'examiner également des éléments de statistique descriptive. Quelques uns de ces éléments sont fournis pour les compagnies d'assurance non vie dans le tableau suivant :

Libellés	Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
PRIMES	166 475	1 602 432	2 989 779	4 267 032	5 671 995	22 953 749
SIN_REG	- 12 733 605	- 1 856 581	- 949 239	- 1 568 598	- 402 942	-
AR_BRUTS	-	739 112	1 989 390	2 608 169	3 712 316	16 877 393
AR_NETS	-	708 423	1 531 309	2 196 885	2 832 501	9 472 077
PF_NETS	- 225 551	7 705	68 391	156 244	197 085	1 679 842
COM	- 2 898 092	- 627 653	- 372 930	- 546 230	- 169 764	-
FG	- 4 265 856	- 1 688 579	- 980 973	- 1 256 867	- 576 101	- 60 861
PRI_REA	- 13 479 084	- 1 580 848	- 356 264	- 1 327 148	- 194 915	- 19 611
SIN_REA	- 4 213	7 318	76 349	421 812	342 196	8 329 518
SAP_REA	- 113 800	101 683	300 275	756 389	810 772	7 969 502
COM_REA	-	27 720	86 113	256 810	368 383	2 480 468
PSAP	- 15 054 016	- 4 652 342	- 2 550 369	- 3 801 441	- 1 059 593	-
PAP	- 3 574 768	- 324 180	- 108 584	- 331 532	- 2 988	8 729
CAP	- 1 911 341	560 013	940 637	1 493 530	1 908 232	7 199 215
T_BILAN	941 683	3 470 647	6 475 539	9 640 046	14 883 709	40 336 606
R_EXP	- 1 888 245	- 2 334	150 132	258 471	447 130	2 597 011
RHE	- 1 593 425	- 140 023	- 36 502	- 78 628	336	785 437
E_R	- 19 364 782	- 6 029 699	- 3 562 531	- 4 837 291	- 1 511 151	- 86 862
ACTIFS_A	-	1 121 606	2 828 188	4 770 635	6 487 276	23 644 338

**Tableau 48: Données sociétés non vie - données résumées**

Le tableau présente pour chaque variable le minimum et le maximum observés, le 1<sup>er</sup> quartile, la médiane, la moyenne et le 3<sup>ème</sup> quartile. La comparaison pour chaque individu des valeurs prises avec les indicateurs fournis dans le tableau peut constituer un premier niveau d'analyse.

### 4.3) Données des sociétés d'assurance vie

Les données des sociétés d'assurance vie sont consignées dans 2 tableaux notés A et B. Elles sont en milliers de FCFA.

	Primes	PF	Sinistres	Com	FG	P_réa	Charges_réa	RE	RN	S_payés	S_réa
A1	15 464 713	2 458 369	-12 340 927	-745 646	-2 150 008	-170 264	0	2 516 237	4 052 080	-8 616 431	0
A2	14 094 980	3 129 480	-13 011 533	-827 021	-2 014 639	-1 616 689	1 517 044	1 271 621	401 813	-9 644 781	272 746
A3	9 789 984	2 043 483	-9 792 585	-377 565	-860 602	-276 912	553 473	1 079 276	566 826	-5 202 608	552 354
A4	8 708 852	2 035 861	-8 671 542	-652 745	-938 799	-43 288	265 525	703 864	691 780	-8 203 492	265 525
A5	8 341 051	608 371	-7 233 171	-457 084	-1 848 692	-128 895	44 857	-673 563	-180 935	-3 785 538	35 290
A6	7 162 413	529 954	-5 926 617	-553 807	-876 978	-119 582	90 078	305 462	416 354	-3 475 417	8 495
A7	6 266 730	819 601	-4 223 685	-240 652	-1 082 675	-449 671	-29 325	1 060 323	300 000	-1 837 178	848
A8	6 133 942	583 494	-5 108 546	-352 861	-701 773	-608 633	430 729	376 352	224 464	-2 390 547	95 122
A9	5 866 911	204 151	-5 391 885	-278 402	-670 772	-93 306	52 971	-310 333	-232 672	-1 084 050	3 709
A10	4 849 956	-11 443	-2 031 618	-517 311	-1 814 323	-84 170	566	391 657	335 255	-1 915 626	566
A11	4 380 566	134 910	-3 826 470	-104 151	-443 615	-62 800	33 182	111 624	103 401	-2 945 595	21 303
A12	3 980 004	95 897	-3 036 549	-207 298	-479 700	-384 969	319 134	286 519	234 307	-1 693 253	58 831
A13	3 790 598	635 173	-2 179 665	-170 350	-695 185	-984 805	739 798	1 135 565	436 486	-1 987 911	175 507
A14	3 462 296	614 355	-2 426 759	-246 341	-832 058	-709 556	426 244	288 180	375 065	-1 874 132	634 568
A15	3 290 671	39 686	-1 815 518	-83 416	-1 162 221	-140 354	74 149	202 999	205 776	-953 517	27 821
A16	3 219 775	442 089	-2 777 553	-164 132	-526 009	-173 973	39 888	60 085	49 300	-1 907 152	43 770
A17	3 168 154	159 122	-2 482 864	-158 046	-474 074	-59 452	145 414	298 254	255 661	-670 367	141 701
A18	3 130 321	-155 550	-1 384 585	-174 364	-1 229 470	-50 606	21 123	156 870	52 725	-1 200 599	21 123
A19	3 017 658	239 797	-2 224 560	-435 634	-1 179 055	0	0	-581 795	0	-2 149 501	0
A20	2 997 225	205 984	-2 425 619	-164 559	-474 205	-22 069	-7 703	109 055	143 197	-942 258	0
A21	2 673 235	-14 401	-533 041	-203 783	-1 178 131	-43 769	-13 701	686 408	2 480 488	-1 658 442	0
A22	2 669 894	311 293	-1 750 328	-255 692	-732 467	-142 468	58 143	158 373	182 261	-1 038 957	25 000
A23	2 513 206	37 329	-1 198 850	-378 501	-702 432	-74 968	25 967	221 751	117 161	-222 308	5 568
A24	2 512 273	80 397	-1 518 272	-209 337	-554 094	-174 421	5 516	142 061	134 087	-940 502	18 861
A25	2 465 807	115 694	-1 709 275	-251 549	-395 182	-440 855	277 486	62 126	35 737	-756 229	82 234
A26	1 907 231	159 618	-725 637	-53 823	-758 680	-71 934	66 938	523 713	119 595	-931 866	104 739
A27	1 904 259	68 805	-1 391 063	-84 408	-538 337	-347 593	169 529	-218 807	-217 178	-665 061	91 916
A28	1 632 223	151 101	-968 076	-179 387	-474 960	-140 861	39 896	59 936	43 963	-632 703	3 730
A29	1 296 324	154 250	-821 623	-145 219	-434 805	-7 400	0	41 526	54 014	-388 785	0
A30	1 279 306	9 873	-828 550	-140 514	-322 642	-64 456	13 093	-53 891	-62 555	-61 482	0
A31	1 210 456	219 689	-988 538	-93 884	-269 435	-57 863	43 813	64 239	107 181	-1 067 681	7 459
A32	1 170 973	60 708	-783 377	-118 319	-427 817	-15 335	0	-113 168	-125 168	-312 839	0
A33	1 150 953	70 685	-729 384	-141 851	-292 758	-11 364	5 141	51 422	95 426	-421 303	0
A34	1 056 466	66 566	-568 638	-101 765	-193 673	-124 856	96 133	230 234	174 351	-79 793	0
A35	1 007 394	87 790	-833 185	-68 949	-228 805	-54 848	181 337	90 734	88 536	-424 376	181 337
A36	994 114	-19 367	-460 126	-51 911	-267 024	-168 575	45 372	72 484	31 481	-107 290	0
A37	892 871	48 939	-643 955	-34 602	-271 362	-116 610	80 492	-44 227	-42 110	-85 524	0
A38	855 581	-402	-344 730	-114 367	-546 091	-8 454	4 358	-154 105	-159 552	-41 466	0
A39	690 964	9 425	-374 690	-84 565	-402 190	-69 274	46 773	-183 556	-189 587	-131 464	6 617
A40	591 271	3 577	-236 349	-45 745	-287 947	-11 966	500	13 341	13 348	-17 608	0
A41	502 257	-1 190	-96 262	-38 399	-361 282	-15 512	2 402	-7 986	-7 986	-1 562	0
A42	312 336	14 632	-153 729	-11 681	-178 509	-64 089	34 946	-46 094	-48 436	-6 113	0
A43	228 259	5 191	-115 577	-11 543	-191 048	-58 634	33 150	-110 202	-111 342	-15 998	5 252
A44	183 151	-2 231	10 411	-6 461	-168 360	-12 024	-3 358	1 128	1 128	-3 370	0
A45	66 760	9 728	-37 527	-6 926	-29 274	-34 922	36 171	4 009	2 466	-25 631	19 198

Tableau 49: Données sociétés vie - tableau A

	Com_réa	PM	A_bruts	A_nets	Ktal	FP	PT	T_bilan	Actifs	ER	Actifs+
A1	0	-65 403 682	100 105	100 105	2 000 000	5 745 735	-65 517 940	78 515 129	75 779 527	-67 363 426	77 175 449
A2	154 617	-48 092 231	1 077 118	1 077 118	825 000	3 391 231	-56 493 546	62 586 288	56 868 336	-57 108 692	57 432 930
A3	1 119	-40 891 447	530 049	530 049	1 000 000	1 274 544	-41 891 545	45 062 838	43 822 295	-42 297 839	44 109 402
A4	0	-32 085 330	16 206	16 206	1 000 000	1 423 770	-33 074 873	36 829 512	36 307 998	-33 230 636	36 307 998
A5	9 567	-26 552 241	239 888	239 888	2 250 000	1 598 504	-27 811 850	37 191 912	32 318 167	-27 911 058	32 441 102
A6	59 106	-19 885 955	38 583	38 583	500 000	897 175	-20 989 292	22 970 789	22 111 204	-21 094 491	22 111 204
A7	-4 325	-16 006 353	458 244	458 244	1 000 000	1 865 234	-16 201 908	19 032 283	17 347 570	-16 435 726	17 451 161
A8	214 589	-11 877 989	389 738	388 415	500 000	697 330	-14 193 941	16 200 080	13 783 378	-14 302 195	14 213 693
A9	22 337	-12 109 481	1 063 237	1 061 659	500 000	416 521	-12 902 594	15 529 724	13 835 578	-14 690 216	13 835 578
A10	0	-14 190 010	117 749	117 749	1 000 000	-1 168 678	-16 505 653	16 968 943	18 906 366	-16 505 653	18 906 366
A11	11 302	-10 973 818	247 263	247 263	500 000	663 584	-10 997 293	12 503 211	11 559 456	-11 169 552	11 806 719
A12	119 688	-8 317 261	575 411	575 411	500 000	623 600	-8 757 702	10 500 142	8 280 009	-8 926 715	9 258 078
A13	186 520	-6 872 527	1 008 922	1 007 614	1 500 000	4 845 835	-8 346 858	17 981 281	4 620 186	-7 904 824	4 673 100
A14	125 194	-11 527 536	1 018 298	824 938	500 000	762 920	-13 465 942	17 103 844	13 549 386	-13 846 320	13 596 665
A15	0	-5 465 276	2 499 441	2 499 441	1 500 000	836 842	-5 849 557	7 792 445	3 236 070	-5 970 262	3 856 799
A16	51 341	-5 495 963	0	0	500 000	781 588	-5 630 411	6 932 067	5 309 808	-5 710 544	5 309 808
A17	3 713	-7 375 196	73 379	73 379	500 000	815 983	-7 508 125	8 963 416	6 728 239	-7 605 469	8 307 493
A18	0	-6 200 872	4 775 416	3 206 453	1 350 000	1 723 392	-7 101 080	9 740 078	6 690 903	-7 339 731	8 268 282
A19	0	-7 414 983	1 939	1 939	550 000	-3 019 683	-8 428 991	6 010 124	5 398 223	-8 644 906	5 398 223
A20	0	-8 137 911	253 813	16 888	500 000	704 987	-8 159 997	9 454 955	8 613 194	-8 341 496	8 637 516
A21	0	-8 218 882	1 639 966	281 748	2 000 000	-6 576 767	-9 591 316	8 346 325	6 750 748	-9 989 005	6 750 748
A22	41 207	-5 562 564	69 041	69 041	500 000	750 428	-5 991 229	7 738 910	5 632 197	-6 018 925	5 701 239
A23	20 399	-2 784 109	961 767	961 767	500 000	538 995	-3 007 927	4 262 741	2 653 758	-3 112 904	2 792 963
A24	8 341	-4 162 999	618 907	618 907	1 000 000	1 031 411	-4 387 976	5 916 134	4 214 913	-4 387 976	4 431 421
A25	110 177	-3 221 778	627 014	627 014	580 000	559 728	-3 304 276	4 681 990	3 110 601	-3 328 831	3 255 974
A26	36 025	-4 404 668	34 989	34 989	750 000	1 247 601	-4 663 864	7 844 953	5 131 996	-4 663 864	5 184 176
A27	114 357	-6 008 455	567 015	567 015	1 000 000	1 193 613	-6 113 854	7 883 972	6 681 175	-6 155 314	6 757 398
A28	54 620	-4 980 470	138 071	138 071	500 000	645 569	-5 216 681	6 460 914	5 508 698	-5 284 983	5 508 698
A29	0	-2 850 651	86 545	83 010	1 350 000	1 153 252	-2 873 078	4 198 298	2 876 003	-2 910 899	3 018 535
A30	13 093	-1 420 974	266 007	266 007	500 000	289 855	-1 498 177	1 988 221	1 507 629	-1 538 058	1 559 948
A31	14 034	-4 902 515	371 981	371 981	500 000	683 656	-5 196 310	6 458 528	4 834 256	-5 191 296	4 834 256
A32	0	-2 346 992	102 592	69 092	500 000	-601 529	-2 398 269	2 087 687	1 767 586	-2 545 335	1 767 586
A33	5 339	-1 429 863	135 868	64 868	800 000	461 648	-1 529 502	2 311 478	2 117 689	-1 609 484	2 117 689
A34	53 027	-1 128 982	0	0	500 000	581 713	-1 174 915	2 144 002	1 895 433	-1 129 017	1 895 433
A35	0	-2 246 901	932 372	932 372	500 000	568 586	-2 304 273	3 068 223	2 037 149	-2 408 600	2 433 084
A36	54 126	-730 148	624 829	624 829	600 000	372 642	-871 739	1 565 596	595 902	-912 424	859 421
A37	24 897	-1 293 392	110 685	110 685	500 000	213 913	-1 312 097	1 730 460	1 409 393	-1 346 916	1 409 393
A38	4 358	-439 241	253 812	253 812	500 000	245 406	-468 855	1 133 179	299 401	-501 786	299 401
A39	17 630	-243 226	66 944	66 944	500 000	500 000	-352 167	1 057 872	286 016	-382 694	352 910
A40	0	-352 903	124 156	124 156	500 000	293 072	-370 502	1 000 595	727 536	-407 145	764 918
A41	2 402	-115 051	166 619	166 619	600 000	600 000	-115 677	798 364	472 407	-130 288	475 531
A42	16 022	-147 616	192 790	192 790	500 000	500 000	-153 729	808 280	370 955	-289 200	378 336
A43	11 727	-145 386	16 350	16 350	500 000	392 659	-145 386	627 414	328 115	-148 810	344 465
A44	2 405	-101 958	186 164	186 164	500 000	280 144	-101 958	485 157	120 342	-101 958	120 342
A45	10 177	-22 570	23 642	23 642	500 000	508 006	-48 201	653 196	393 801	-24 689	393 801

Tableau 50: Données sociétés vie - tableau B

#### 4.4) Données résumées des sociétés d'assurance vie

L'un des objectifs du mémoire est de montrer comment les techniques statistiques d'analyses des données « multivariées » peuvent être utilisées avec profit par un professionnel comptable. Même si ces techniques permettent d'analyser les données en tenant compte de leur caractère multidimensionnel, il peut être intéressant d'examiner également des éléments de statistique descriptive. Quelques uns de ces éléments sont fournis pour les compagnies d'assurance vie dans le tableau suivant :

Libellés	Minimum	1er Quartile	Médiane	Moyenne	3ème Quartile	Maximum
Primes	66 760	1 056 000	2 513 000	3 397 000	3 980 000	15 460 000
PF	-155 550	9 873	95 897	365 789	311 293	3 129 480
Sinistres	-13 010 000	-2 778 000	-1 391 000	-2 580 000	-644 000	10 410
Com	-827 021	-255 692	-164 132	-216 546	-84 408	-6 461
FG	-2 150 000	-860 600	-526 000	-681 400	-322 600	-29 270
P_Réass	-1 617 000	-170 300	-74 970	-189 200	-43 770	0
Charges_réa	-29 325	4 358	39 896	132 606	96 133	1 517 044
RE	-673 563	-7 986	90 734	228 438	298 254	2 516 237
RN	-232 672	0	95 426	247 738	234 307	4 052 080
S_payés	-9 645 000	-1 907 000	-940 500	-1 612 000	-131 500	-1 562
S_réa	0	0	6 617	64 693	58 831	634 568
Com_réa	-4 325	0	11 302	34 870	51 341	214 589
PM	-65 400 000	-10 970 000	-5 465 000	-9 425 000	-1 421 000	-22 570
A_bruts	0	86 550	239 900	506 700	618 900	4 775 000
A_nets	0	69 040	192 800	429 600	575 400	3 206 000
Ktal	500 000	500 000	500 000	781 200	1 000 000	2 250 000
FP	-6 577 000	416 500	645 600	695 900	1 031 000	5 746 000
PT	-65 520 000	-11 000 000	-5 630 000	-10 070 000	-1 498 000	-48 200
T_bilan	485 200	2 088 000	6 932 000	12 070 000	15 530 000	78 520 000
Actifs	120 300	1 768 000	5 132 000	10 370 000	11 560 000	75 780 000
ER	-67 360 000	-11 170 000	-5 711 000	-10 240 000	-1 538 000	-24 690
Actifs.	120 300	1 768 000	5 184 000	10 590 000	11 810 000	77 180 000

**Tableau 51: Données sociétés vie - données résumées**

Le tableau présente pour chaque variable le minimum et le maximum observés, le 1<sup>er</sup> quartile, la médiane, la moyenne et le 3<sup>ème</sup> quartile. La comparaison pour chaque individu des valeurs prises avec les indicateurs fournis dans le tableau peut constituer un premier niveau d'analyse.

## 5) Données complémentaires pour l'analyse discriminante et la régression logistique

### 5.1) Données pour les sociétés d'assurance non vie

Le tableau ci-dessous donne pour les compagnies contrôlées sur la période, les ajustements effectués et leur incidence globale sur le niveau des fonds propres. Par hypothèse, les ajustements effectués sont significatifs si le total des ajustements rapporté aux fonds propres est supérieur au taux de 10%.

Le tableau donne également le statut de quelques sociétés dont la situation financière a été examinée par la CRCA. Trois statuts sont identifiés: les sociétés saines, les sociétés en difficultés faisant l'objet de suivi et les sociétés en faillite.

Société	Fonds propres	Fonds propres corrigés	Correction en %	Ajustements	STATUT
A2	7584	6263	0,17	1	Sain
A4	3654	3067	0,16	1	Suivi
A5	6646	6364	0,04	0	Sain
A11	5900	5853	0,01	0	Sain
A12	3057	3057	0	0	Sain
A16	1669	1282	0,23	1	Sain
A18	1266	-824	1,65	1	Suivi
A23	1394	909	0,35	1	Suivi
A35	1195	204	0,83	1	Suivi
A37	1454	1051	0,28	1	Sain
A42	480	-1834	4,82	1	Faillite
A45	-372	-1576	-3,24	1	Faillite
A49	633	-881	2,39	1	Suivi
A52	1013	-2060	3,03	1	Faillite
A53	2833	2141	0,24	1	Suivi
A56	1352	-1695	2,25	1	Faillite
A69	102	-748	8,33	1	Suivi
A71	806	57	0,93	1	Suivi
A72	1163	-1644	2,41	1	Faillite
A78	-521	-1531	-1,94	1	Faillite
A79	307	19	0,94	1	Suivi
A82	1131	697	0,38	1	Sain
A89	418	216	0,48	1	Faillite

Tableau 52: Données sociétés non vie - ajustements et statuts

Un ajustement significatif est codé « 1 » et un ajustement non significatif codé « 0 ». Cette information figure dans la colonne « Ajustement » du tableau. Elle doit être concaténée au tableau des données brutes avant de réaliser la régression logistique.

## 5.2) Données pour les sociétés d'assurance vie

Les indications générales fournies sur le « Tableau 52: Données sociétés non vie - ajustements et statuts » ci-dessus sont également valables pour le tableau suivant :

Société	Fonds propres initiaux	Fonds propres corrigés	Observations	Ajustements	STATUT
A20	804	645	20%	1	Sain
A15	-34	-1930	-5576%	1	Faillite
A23	165	-246	249%	1	Suivi
A6	890	101	89%	1	Suivi
A8	1004	748	25%	1	Sain
A28	724	239	67%	1	Sain
A32	-616	-694	-13%	1	Suivi
A9	774	-170	122%	1	Suivi
A26	1098	-1827	266%	1	Faillite
A4	2474	2474	0%	0	Sain

**Tableau 53: Données sociétés vie - ajustements et statuts**

Les informations dans ces tableaux sont annoncées notamment dans le corps du mémoire dans les parties relatives à la mise en œuvre des modèles prédictifs (**Chapitre I Analyse discriminante page 94 et Chapitre III Régression logistique page 101**).

## **II. ANNEXES RELATIVES A L'ANALYSE DES DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES D'ASSURANCE**

Ces annexes relatives aux méthodes descriptives sont composées des différents documents suivants :

- Annexes relatives à l'analyse en composantes principales des sociétés non vie ;
- Annexes relatives à la classification ascendante hiérarchique des sociétés non vie ;
- Annexes relatives à l'analyse en composantes principales des sociétés vie ;
- Annexes relatives à la classification ascendante hiérarchique des sociétés vie ;
- Annexes relatives à l'ACP et à la CAH des sociétés vie et non vie du group G.

### **1) Analyse en composantes principales des sociétés non vie**

Cette annexe relative à l'ACP des sociétés non vie est composée des 4 points suivants :

- Le tableau complet des valeurs propres ;
- Les tableaux complets des coordonnées, des projections et des contributions des individus sur les 4 premiers axes ;
- La matrice de corrélation ;
- Le tableau des corrélations et des coordonnées des variables initiales par rapport aux 4 premières composantes principales.

Des extraits de ces tableaux ont été fournis dans le corps du rapport relatif à ces points.

### 1.1) Valeurs propres

Le tableau suivant contient la valeur propre associée à chaque axe (Eigen value), le pourcentage d'inertie (Percentage of variance) qu'elle représente dans l'analyse ainsi que le cumul de ces pourcentages (Cumulative percentage of variance).

Libellés	Eigenvalue	Percentage of variance	Cumulative percentage of variance
comp 1	11,87	62,45	62,45
comp 2	1,96	10,32	72,77
comp 3	1,27	6,69	79,46
comp 4	0,91	4,8	84,25
comp 5	0,71	3,74	87,99
comp 6	0,53	2,78	90,77
comp 7	0,47	2,47	93,24
comp 8	0,33	1,76	95
comp 9	0,25	1,3	96,3
comp 10	0,21	1,08	97,38
comp 11	0,18	0,94	98,32
comp 12	0,15	0,76	99,08
comp 13	0,06	0,34	99,42
comp 14	0,04	0,2	99,62
comp 15	0,02	0,13	99,75
comp 16	0,02	0,11	99,86
comp 17	0,01	0,07	99,93
comp 18	0,01	0,05	99,98
comp 19	0	0,02	100

**Tableau 54: Données sociétés non vie - valeurs propres et pourcentage d'inertie**

Ce tableau a fait l'objet d'analyse dans le corps du mémoire et un extrait y a été produit (Tableau 6: Données sociétés non vie - six premières valeurs propres et pourcentage d'inertie ).

### 1.2) Coordonnées, projection et contribution des individus sur les nouveaux axes

Les deux tableaux A et B suivants relatifs aux 4 premières composantes principales représentant 84,25% de l'inertie totale fournissent 3 types d'informations :

- Les coordonnées des compagnies sur les nouveaux axes construits,
- Le cosinus carré de l'angle formé par chaque compagnie avec sa projection sur chacun des axes principaux,
- La contribution de chaque compagnie à la construction de chacun des axes.

	Coordonnées				Cos <sup>2</sup>				Contribution			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
A1	9,32	1,77	-1,86	-5,58	0,67	0,02	0,03	0,24	8,13	1,78	3,02	37,97
A2	12,45	-0,05	0,97	-0,79	0,86	0,00	0,01	0,00	14,52	0,00	0,82	0,76
A3	11,47	0,58	4,10	-1,53	0,77	0,00	0,10	0,01	12,33	0,19	14,72	2,85
A4	4,20	-1,09	0,34	-1,58	0,66	0,04	0,00	0,09	1,65	0,68	0,10	3,05
A5	6,22	-0,02	-2,14	0,99	0,74	0,00	0,09	0,02	3,63	0,00	4,02	1,19
A6	11,53	5,01	0,71	2,77	0,71	0,13	0,00	0,04	12,45	14,21	0,44	9,39
A7	4,64	0,53	-0,26	0,21	0,88	0,01	0,00	0,00	2,01	0,16	0,06	0,06
A8	4,06	-3,93	0,95	-0,27	0,37	0,34	0,02	0,00	1,54	8,75	0,80	0,09
A9	4,18	0,24	-1,03	-0,66	0,77	0,00	0,05	0,02	1,63	0,03	0,92	0,54
A10	5,36	2,63	-1,13	1,49	0,53	0,13	0,02	0,04	2,69	3,94	1,12	2,70
A11	4,54	0,62	-0,61	0,48	0,72	0,01	0,01	0,01	1,93	0,22	0,33	0,28
A12	3,13	0,53	-0,64	0,29	0,75	0,02	0,03	0,01	0,91	0,16	0,36	0,10
A13	5,33	0,85	0,05	2,11	0,77	0,02	0,00	0,12	2,66	0,41	0,00	5,43
A14	3,26	-0,04	-4,31	0,87	0,31	0,00	0,55	0,02	0,99	0,00	16,25	0,92
A15	1,85	0,90	0,16	-0,91	0,51	0,12	0,00	0,12	0,32	0,46	0,02	1,01
A16	3,00	1,05	-1,77	0,84	0,52	0,06	0,18	0,04	0,84	0,63	2,75	0,86
A17	0,37	-1,76	0,15	-1,61	0,01	0,33	0,00	0,28	0,01	1,75	0,02	3,16
A18	-0,37	0,16	-0,46	-1,08	0,04	0,01	0,07	0,37	0,01	0,02	0,19	1,41
A19	6,25	-6,00	1,85	0,73	0,41	0,38	0,04	0,01	3,66	20,42	3,00	0,65
A20	1,35	-0,42	-1,55	-0,22	0,34	0,03	0,45	0,01	0,17	0,10	2,09	0,06
A21	3,82	0,04	-1,03	2,02	0,54	0,00	0,04	0,15	1,37	0,00	0,92	4,98
A22	0,47	0,05	-0,52	-0,37	0,16	0,00	0,20	0,10	0,02	0,00	0,24	0,16
A23	-0,43	0,41	0,26	-0,91	0,11	0,10	0,04	0,49	0,02	0,09	0,06	1,01
A24	0,47	-0,80	-0,11	-0,69	0,06	0,16	0,00	0,12	0,02	0,36	0,01	0,58
A25	2,69	-0,50	-0,27	1,24	0,58	0,02	0,01	0,12	0,68	0,14	0,06	1,88
A26	1,33	-2,04	-0,72	0,70	0,18	0,43	0,05	0,05	0,17	2,35	0,45	0,60
A27	0,41	0,29	0,05	-0,90	0,04	0,02	0,00	0,19	0,02	0,05	0,00	1,00
A28	0,08	0,80	0,70	-0,08	0,00	0,41	0,31	0,00	0,00	0,36	0,42	0,01
A29	-1,02	-0,29	0,20	-1,06	0,32	0,03	0,01	0,35	0,10	0,05	0,04	1,38
A30	-0,77	0,57	0,42	-0,10	0,49	0,27	0,14	0,01	0,06	0,18	0,15	0,01
A31	-1,85	0,55	-0,24	-0,50	0,71	0,06	0,01	0,05	0,32	0,17	0,05	0,31
A32	0,31	0,39	0,39	0,53	0,06	0,10	0,10	0,18	0,01	0,09	0,13	0,35
A33	-0,90	0,23	0,78	-0,60	0,37	0,02	0,27	0,16	0,08	0,03	0,53	0,44
A34	-0,38	-0,92	0,18	-0,12	0,06	0,34	0,01	0,01	0,01	0,48	0,03	0,02
A35	-0,92	0,41	-0,17	0,08	0,56	0,11	0,02	0,00	0,08	0,10	0,02	0,01
A36	-1,38	0,88	1,49	-0,62	0,31	0,13	0,37	0,06	0,18	0,44	1,93	0,47
A37	-1,05	-0,47	0,36	-0,41	0,59	0,12	0,07	0,09	0,10	0,13	0,12	0,20
A38	-0,80	0,26	1,15	0,09	0,14	0,01	0,29	0,00	0,06	0,04	1,16	0,01
A39	0,59	-2,46	0,26	0,34	0,03	0,51	0,01	0,01	0,03	3,42	0,06	0,14
A40	-1,26	-1,10	0,25	-0,16	0,35	0,27	0,01	0,01	0,15	0,68	0,05	0,03
A41	-0,68	-0,12	-0,78	0,07	0,24	0,01	0,33	0,00	0,04	0,01	0,54	0,01
A42	-2,02	0,08	0,06	-0,56	0,83	0,00	0,00	0,06	0,38	0,00	0,00	0,38
A43	-1,99	-0,57	-3,60	-0,90	0,15	0,01	0,51	0,03	0,37	0,19	11,31	0,98
A44	-1,11	0,45	0,29	0,10	0,54	0,09	0,04	0,00	0,11	0,12	0,07	0,01
A45	-1,99	0,59	-1,70	0,37	0,45	0,04	0,33	0,02	0,37	0,20	2,54	0,16

Tableau 55: Données sociétés non vie - coordonnées et qualité de représentation des sociétés - tableau A

	Coordonnés				Cos <sup>2</sup>				Contribution			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
A46	-1,77	0,67	0,23	-0,04	0,80	0,12	0,01	0,00	0,29	0,26	0,04	0,00
A47	-1,42	0,46	0,06	-0,05	0,84	0,09	0,00	0,00	0,19	0,12	0,00	0,00
A48	-1,64	-0,01	-0,15	0,12	0,80	0,00	0,01	0,00	0,25	0,00	0,02	0,02
A49	-2,07	0,05	-0,01	-0,19	0,87	0,00	0,00	0,01	0,40	0,00	0,00	0,04
A50	-1,89	-0,13	0,13	-0,17	0,85	0,00	0,00	0,01	0,33	0,01	0,01	0,04
A51	-1,19	0,84	-0,38	0,45	0,37	0,19	0,04	0,05	0,13	0,40	0,13	0,24
A52	-1,81	-0,85	0,56	-0,18	0,55	0,12	0,05	0,01	0,31	0,41	0,28	0,04
A53	0,79	-1,97	1,54	0,31	0,07	0,47	0,29	0,01	0,06	2,20	2,07	0,12
A54	0,00	-3,44	0,33	0,53	0,00	0,65	0,01	0,02	0,00	6,69	0,09	0,34
A55	-2,24	0,35	-0,03	-0,08	0,91	0,02	0,00	0,00	0,47	0,07	0,00	0,01
A56	-0,73	-3,78	0,04	-0,01	0,02	0,58	0,00	0,00	0,05	8,10	0,00	0,00
A57	-0,81	-0,14	0,40	0,81	0,14	0,00	0,03	0,14	0,06	0,01	0,14	0,79
A58	-2,43	0,49	0,40	-0,39	0,84	0,03	0,02	0,02	0,55	0,14	0,14	0,18
A59	-2,12	0,59	0,26	-0,15	0,89	0,07	0,01	0,00	0,42	0,20	0,06	0,03
A60	-2,05	0,63	0,58	-0,10	0,83	0,08	0,07	0,00	0,39	0,23	0,29	0,01
A61	-1,98	-0,49	-1,58	0,08	0,44	0,03	0,28	0,00	0,37	0,14	2,18	0,01
A62	-2,72	0,45	-0,04	-0,37	0,93	0,03	0,00	0,02	0,69	0,12	0,00	0,16
A63	-1,67	-0,27	-0,20	0,10	0,52	0,01	0,01	0,00	0,26	0,04	0,04	0,01
A64	-1,84	0,36	-0,07	0,25	0,87	0,03	0,00	0,02	0,32	0,07	0,00	0,08
A65	-1,54	0,59	-0,35	0,67	0,57	0,08	0,03	0,11	0,22	0,20	0,10	0,55
A66	-1,99	0,52	0,29	0,12	0,89	0,06	0,02	0,00	0,37	0,15	0,07	0,02
A67	0,43	0,10	3,79	1,94	0,01	0,00	0,58	0,15	0,02	0,01	12,53	4,59
A68	1,28	-4,72	-2,73	1,56	0,04	0,50	0,17	0,05	0,15	12,65	6,53	2,96
A69	-2,91	0,39	0,23	-0,28	0,95	0,02	0,01	0,01	0,79	0,09	0,05	0,10
A70	-2,35	0,60	0,24	0,04	0,89	0,06	0,01	0,00	0,52	0,21	0,05	0,00
A71	-2,60	0,53	0,48	0,03	0,89	0,04	0,03	0,00	0,63	0,16	0,20	0,00
A72	-1,78	0,62	0,16	0,53	0,70	0,08	0,01	0,06	0,30	0,22	0,02	0,35
A73	-2,64	0,78	0,30	0,04	0,90	0,08	0,01	0,00	0,65	0,34	0,08	0,00
A74	-2,94	0,51	0,23	-0,10	0,95	0,03	0,01	0,00	0,81	0,14	0,04	0,01
A75	-2,78	0,57	0,27	-0,17	0,93	0,04	0,01	0,00	0,73	0,19	0,06	0,04
A76	-2,85	-0,57	-0,54	-0,35	0,71	0,03	0,03	0,01	0,76	0,18	0,25	0,15
A77	-2,04	0,59	-0,03	0,40	0,81	0,07	0,00	0,03	0,39	0,20	0,00	0,19
A78	-2,98	0,49	-0,19	-0,14	0,95	0,03	0,00	0,00	0,83	0,14	0,03	0,02
A79	-3,00	0,63	0,23	0,01	0,94	0,04	0,01	0,00	0,84	0,23	0,04	0,00
A80	-2,75	0,45	0,96	-0,20	0,84	0,02	0,10	0,00	0,71	0,11	0,80	0,05
A81	-3,03	0,65	-0,09	-0,08	0,92	0,04	0,00	0,00	0,86	0,24	0,01	0,01
A82	-2,94	0,70	0,34	0,11	0,90	0,05	0,01	0,00	0,81	0,28	0,10	0,01
A83	-3,11	0,65	0,30	-0,12	0,93	0,04	0,01	0,00	0,91	0,24	0,08	0,02
A84	-3,06	0,47	0,25	0,00	0,94	0,02	0,01	0,00	0,88	0,13	0,05	0,00
A85	-3,08	0,32	0,29	-0,03	0,95	0,01	0,01	0,00	0,89	0,06	0,08	0,00
A86	-3,29	0,48	0,12	-0,11	0,95	0,02	0,00	0,00	1,01	0,13	0,01	0,01
A87	-3,01	0,47	0,42	0,09	0,93	0,02	0,02	0,00	0,85	0,13	0,16	0,01
A88	-3,38	0,74	0,32	-0,02	0,93	0,04	0,01	0,00	1,07	0,31	0,09	0,00
A89	-3,27	0,65	0,38	0,07	0,93	0,04	0,01	0,00	1,00	0,24	0,13	0,01
A90	-2,55	0,67	1,08	0,98	0,47	0,03	0,08	0,07	0,61	0,25	1,01	1,18

**Tableau 56: Données sociétés non vie - coordonnées et qualité de représentation des sociétés - tableau B**

Ces tableaux ont fait l'objet d'analyse dans le corps du mémoire et un extrait y a été produit (Tableau 7: Données sociétés non vie - indicateurs pour 10 sociétés sur les quatre premiers axes - extrait - ).

### 1.3) Matrice de corrélation

Lorsque les variables sont parfaitement corrélées, l'analyse en composantes principales est particulièrement efficace pour effectuer la synthèse de ces variables. Cette corrélation devient un handicap important dans le cadre des techniques prédictives (régression linéaire, analyse discriminante, régression logistique).

Un examen de la matrice de corrélation permet dans ce cas d'identifier les variables trop corrélées. On peut faire l'option dans le cadre de l'utilisation de techniques prédictives d'ignorer certaines de ces variables colinéaires.

La matrice de corrélation des variables mesurées sur les compagnies d'assurance non vie se présente comme suit :

		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19
PRIMES	V1	1,00	0,87	0,44	0,59	0,56	0,89	0,82	0,90	0,64	0,64	0,87	0,81	0,35	0,61	0,83	0,60	0,33	0,81	0,82
SIN_REG	V2		1,00	0,42	0,54	0,61	0,80	0,70	0,74	0,87	0,74	0,70	0,78	0,35	0,63	0,82	0,68	0,29	0,77	0,81
AR_BRUTS	V3			1,00	0,88	0,19	0,41	0,55	0,28	0,23	0,31	0,29	0,52	0,54	0,47	0,64	0,15	0,22	0,53	0,50
AR_NETS	V4				1,00	0,28	0,54	0,63	0,44	0,34	0,43	0,45	0,62	0,72	0,63	0,74	0,29	0,40	0,62	0,56
PF_NETS	V5					1,00	0,47	0,56	0,60	0,64	0,67	0,59	0,73	0,18	0,56	0,67	0,56	0,24	0,71	0,73
COM	V6						1,00	0,67	0,68	0,53	0,60	0,67	0,70	0,41	0,56	0,72	0,44	0,19	0,69	0,70
FG	V7							1,00	0,74	0,52	0,70	0,74	0,85	0,35	0,56	0,84	0,35	0,20	0,86	0,79
PRI_REA	V8								1,00	0,66	0,58	<b>0,97</b>	0,74	0,23	0,53	0,73	0,57	0,38	0,73	0,74
SIN_REA	V9									1,00	0,75	0,64	0,65	0,23	0,48	0,65	0,67	0,34	0,63	0,67
SAP_REA	V10										1,00	0,62	0,80	0,28	0,57	0,77	0,50	0,17	0,79	0,72
COM_REA	V11											1,00	0,73	0,22	0,51	0,70	0,55	0,40	0,73	0,70
PSAP	V12												1,00	0,32	0,69	<b>0,93</b>	0,54	0,28	<b>0,99</b>	<b>0,94</b>
PAP	V13													1,00	0,43	0,46	0,05	0,28	0,31	0,32
CAP	V14														1,00	0,82	0,54	0,51	0,69	0,70
T_BILAN	V15															1,00	0,56	0,38	<b>0,94</b>	<b>0,93</b>
R_EXP	V16																1,00	0,56	0,56	0,60
RHE	V17																	1,00	0,30	0,26
E_R	V18																		1,00	<b>0,95</b>
ACTIFS_A	V19																			1,00

Tableau 57 : Données sociétés non vie - matrice de corrélation

Les coefficients de corrélation les plus élevés sont mis en gras pour permettre une éventuelle élimination de ces variables dans le cadre de la mise en œuvre des méthodes prédictives.

### 1.4) Corrélation et contribution des variables à la construction des nouveaux axes

Les coordonnées des variables, leur qualité de représentation et leur contribution à la construction des axes sont examinées à travers les paramètres suivants :

- Le coefficient de corrélation de chaque variable initiale avec les nouveaux axes construits,
- Le cosinus carré de l'angle formé par chaque variable avec sa projection sur chacun des axes principaux,
- La contribution de chaque variable à la construction de chacun des axes.

Libellés	Corrélation des variables p/r axes				Corrélation des variables p/r aux axes (cos <sup>2</sup> )				Contribution des variables à la construction des axes			
	Dim 1	Dim 2	Dim 3	Dim 4	Dim 1	Dim 2	Dim 3	Dim 4	Dim 1	Dim 2	Dim 3	Dim 4
PRIMES	0,92	0,07	-0,08	-0,34	0,84	0,01	0,01	0,12	7,08	0,27	0,48	12,88
SIN_REG	-0,90	-0,14	0,00	0,04	0,80	0,02	0,00	0,00	6,78	0,96	0,00	0,17
AR_BRUTS	0,56	-0,70	-0,05	0,07	0,31	0,49	0,00	0,00	2,63	<b>24,78</b>	0,23	0,52
AR_NETS	0,70	-0,66	0,10	-0,03	0,49	0,43	0,01	0,00	4,11	<b>22,17</b>	0,76	0,07
PF_NETS	0,73	0,32	-0,02	0,35	0,53	0,10	0,00	0,12	4,43	5,19	0,02	13,11
COM	-0,80	0,03	0,17	0,29	0,64	0,00	0,03	0,08	5,39	0,06	2,16	9,32
FG	-0,86	0,10	0,30	0,11	0,73	0,01	0,09	0,01	6,16	0,48	6,88	1,22
PRI_REA	-0,84	-0,24	-0,01	0,40	0,71	0,06	0,00	0,16	5,96	2,99	0,01	17,57
SIN_REA	0,76	0,33	0,13	0,12	0,57	0,11	0,02	0,02	4,85	5,67	1,36	1,66
SAP_REA	0,80	0,17	-0,18	0,27	0,64	0,03	0,03	0,07	5,42	1,53	2,55	8,05
COM_REA	0,83	0,23	0,01	-0,39	0,69	0,05	0,00	0,16	5,79	2,76	0,01	17,04
PSAP	-0,94	-0,02	0,17	-0,16	0,88	0,00	0,03	0,02	7,39	0,02	2,33	2,69
PAP	-0,44	0,67	-0,13	0,05	0,19	0,45	0,02	0,00	1,63	<b>22,87</b>	1,24	0,28
CAP	0,77	-0,17	0,28	0,23	0,59	0,03	0,08	0,05	4,94	1,45	6,01	5,82
T_BILAN	0,96	-0,14	-0,04	0,14	0,92	0,02	0,00	0,02	7,73	0,97	0,14	2,16
R_EXP	0,66	0,38	0,48	0,09	0,43	0,14	0,23	0,01	3,65	<b>7,38</b>	<b>18,06</b>	0,93
RHE	-0,42	0,07	-0,83	0,09	0,18	0,00	0,69	0,01	1,48	0,24	<b>54,47</b>	0,94
E_R	-0,94	-0,01	0,16	-0,16	0,88	0,00	0,02	0,03	7,38	0,01	1,90	2,77
ACTIFS_A	0,92	0,06	-0,13	0,16	0,86	0,00	0,02	0,03	7,21	0,20	1,40	2,81
Total									100	100	100	100

Tableau 58: Données sociétés non vie - coordonnées et qualité de représentation des variables

Ce tableau a fait l'objet d'analyse dans le corps du mémoire et un extrait y a été produit (Tableau 8: Données sociétés non vie - coordonnées et qualité de représentation des variables - extrait- ).

### 2) Classification des sociétés non vie

Cette annexe relative à la classification des sociétés non vie est composée des points suivants :

- Représentations des partitions obtenues à la suite de la classification selon différentes modalités ;
- Caractérisation des classes par rapport aux variables initiales ;
- Caractérisation des classes par rapport aux composantes principales ;
- Description des classes par les parangons et les individus spécifiques.

## 2.1) Représentations des partitions

La forme du dendrogramme et le niveau de coupure optimal calculé par la fonction HCPC suggère une partition des sociétés d'assurance vie en 3 classes.

Les 3 classes peuvent être mises en évidence et représentée sur le plan principal. Le dendrogramme peut également faire l'objet de représentation en 3 dimensions sur le plan principal.

Le dendrogramme mettant en évidence les 3 classes se présente comme suit :

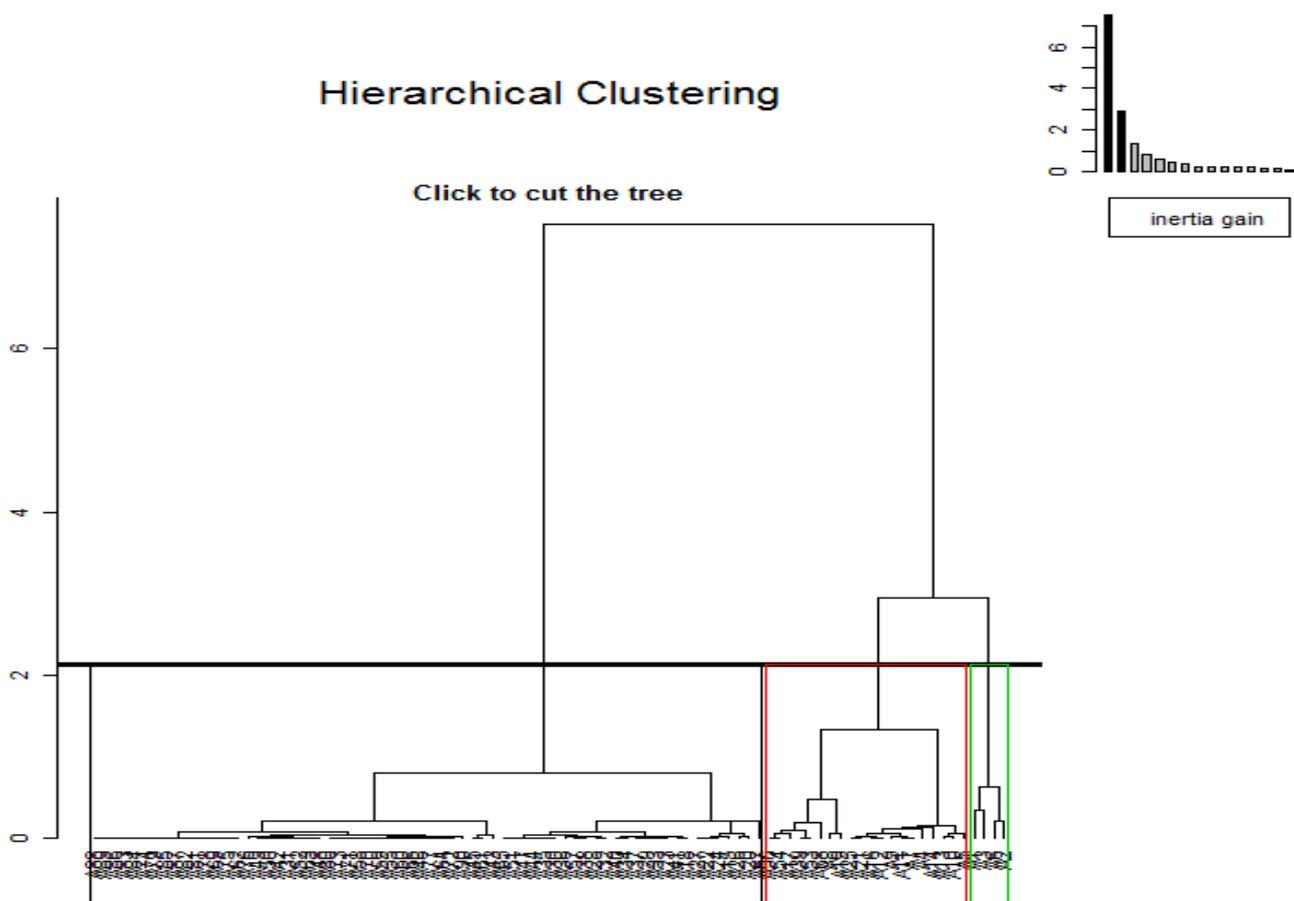


Figure 19: Données sociétés non vie - arbre hiérarchique

A la suite de la mise en évidence des différentes classes à partir du dendrogramme, une visualisation peut également être effectuée sur le plan principal.

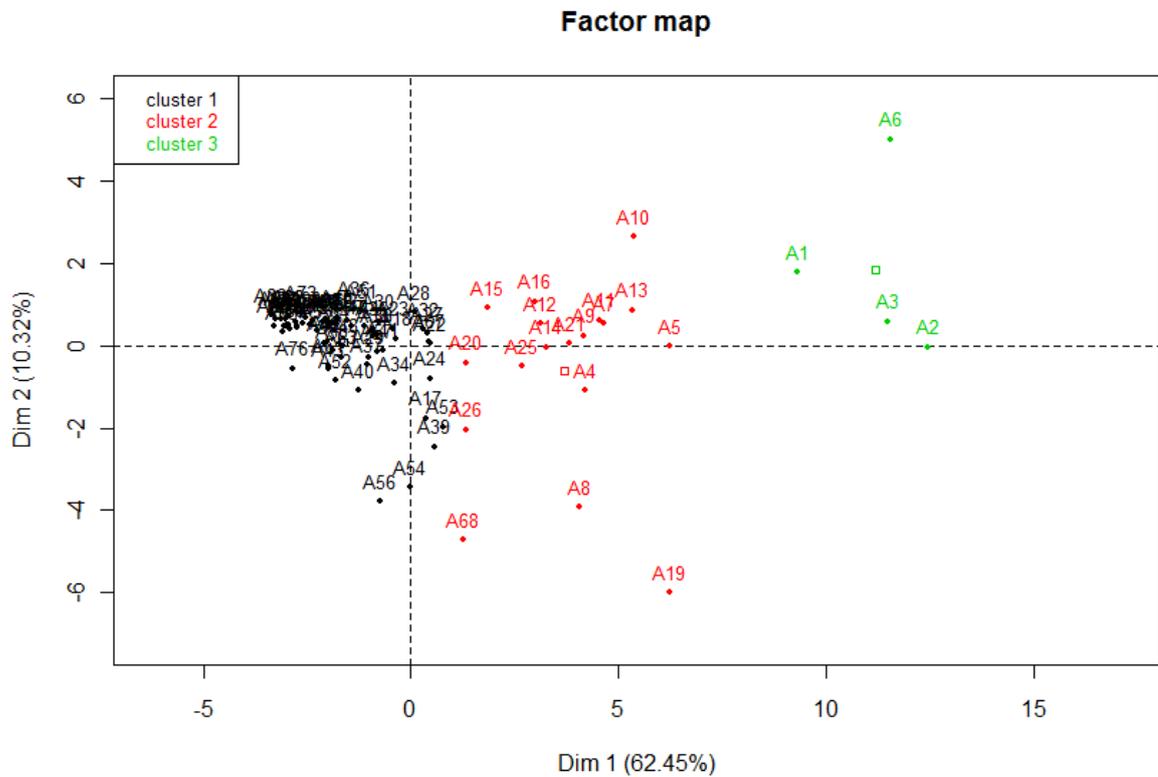


Figure 20: Données sociétés non vie - représentation de la partition sur le plan principal

La fonction HCPC permet également de représenter le dendrogramme en trois dimensions sur le plan principal. Cette représentation est la suivante :

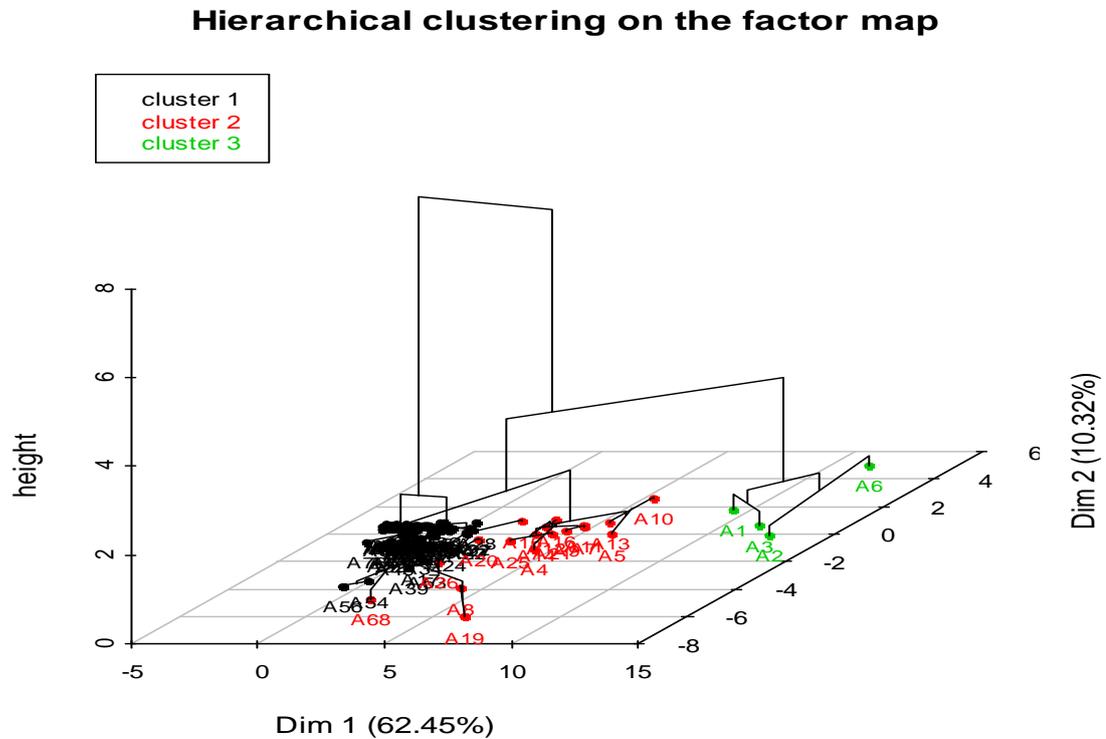


Figure 21: Données sociétés non vie - représentation du dendrogramme en trois dimensions sur le plan principal

Ces trois représentations ont été annoncées et analysées dans le corps du mémoire à la section suivante : **3.1) Forme du dendrogramme et nombre de classes** page 51.

## 2.2) Caractérisation des classes par rapport aux variables initiales

Les différentes classes mises en évidence peuvent être caractérisées par les variables initiales. Les principes de cette caractérisation par les variables initiales ont été exposés dans le corps du mémoire (Description des classes par les variables initiales ou les composantes principales page 52).

Le tableau complet permettant de caractériser la classe 1 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe1_variables_quantitatives						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
PSAP	8,23	- 1 947 199,10	- 3 801 441,41	1 409 329,37	3 737 446,81	1,89E-16
E_R	8,17	- 2 607 944,46	- 4 837 291,38	1 708 999,60	4 526 120,22	3,10E-16
FG	7,32	- 850 191,03	- 1 256 866,68	508 796,53	920 915,14	2,40E-13
COM	6,55	- 319 330,68	- 546 230,10	243 377,10	574 267,55	5,63E-11
SIN_REG	6,51	- 780 158,56	- 1 568 597,59	626 833,40	2 010 179,17	7,75E-11
PRI_REA	5,99	- 515 793,49	- 1 327 147,78	574 070,04	2 248 241,09	2,16E-09
PAP	2,92	- 226 642,57	- 331 532,13	483 525,24	596 769,34	3,55E-03
R_EXP	- 4,15	104 760,44	258 471,22	399 225,73	614 578,39	3,35E-05
SIN_REA	- 4,80	90 577,06	421 811,68	139 331,37	1 145 672,50	1,62E-06
AR_BRUTS	- 5,30	1 706 504,15	2 608 168,59	1 551 064,81	2 823 093,76	1,18E-07
COM_REA	- 5,76	119 378,13	256 809,78	156 061,36	395 454,11	8,21E-09
AR_NETS	- 5,77	1 476 924,71	2 196 884,93	1 420 700,43	2 067 842,46	7,71E-09
CAP	- 5,91	915 939,57	1 493 529,58	1 118 562,19	1 619 981,14	3,35E-09
PF_NETS	- 6,03	54 799,15	156 243,80	111 988,45	279 050,41	1,64E-09
SAP_REA	- 6,67	271 795,10	756 388,60	292 308,71	1 205 594,99	2,62E-11
PRIMES	- 7,09	2 460 079,18	4 267 032,33	1 527 065,59	4 228 920,82	1,37E-12
T_BILAN	- 7,96	5 555 211,56	9 640 045,52	3 668 247,44	8 515 847,31	1,78E-15
ACTIFS_A	- 7,97	2 339 593,85	4 770 635,41	1 802 060,64	5 060 335,78	1,61E-15

**Tableau 59: Données sociétés non vie - description de la classe 1 par les variables initiales**

Un extrait de ce tableau a été fourni et analysé dans le corps du mémoire (Tableau 10: Données sociétés non vie - description de la classe 1 par les variables initiales - extrait -)

Le tableau complet permettant de caractériser la classe 2 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe2_variables_quantitatives						
	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
T_BILAN	5,78	20 076 095,17	9 640 045,52	4 020 112,15	8 515 847,31	7,44E-09
ACTIFS_A	5,64	10 819 527,94	4 770 635,41	3 200 560,43	5 060 335,78	1,72E-08
SAP_REA	4,66	1 946 386,17	756 388,60	1 178 634,43	1 205 594,99	3,22E-06
AR_NETS	4,61	4 218 050,94	2 196 884,93	2 118 915,67	2 067 842,46	4,02E-06
AR_BRUTS	4,59	5 357 649,22	2 608 168,59	4 128 438,32	2 823 093,76	4,35E-06
CAP	4,35	2 988 314,33	1 493 529,58	1 429 880,53	1 619 981,14	1,35E-05
PF_NETS	4,25	407 950,56	156 243,80	374 771,48	279 050,41	2,09E-05
PRIMES	4,20	8 029 397,17	4 267 032,33	2 750 548,28	4 228 920,82	2,71E-05
COM_REA	2,33	452 109,28	256 809,78	212 368,89	395 454,11	1,98E-02
PAP	- 2,11	- 599 017,11	- 331 532,13	772 785,86	596 769,34	3,45E-02
PRI_REA	- 2,39	- 2 466 095,67	- 1 327 147,78	1 306 330,23	2 248 241,09	1,69E-02
SIN_REG	- 3,41	- 3 020 610,78	- 1 568 597,59	1 335 441,89	2 010 179,17	6,56E-04
COM	- 4,71	- 1 120 087,83	- 546 230,10	521 376,96	574 267,55	2,43E-06
FG	- 5,22	- 2 275 880,28	- 1 256 866,68	524 342,78	920 915,14	1,79E-07
E_R	- 6,14	- 10 729 730,00	- 4 837 291,38	2 749 984,83	4 526 120,22	8,20E-10
PSAP	- 6,17	- 8 691 289,06	- 3 801 441,41	2 175 187,96	3 737 446,81	6,77E-10

**Tableau 60: Données sociétés non vie - description de la classe 2 par les variables initiales**

Les individus de la classe 2 sont représentés en rouge sur le dendrogramme. Cette classe est composée de 18 sociétés sur un total de 90. Sur 16 variables initiales, les sociétés de la classe 2 prennent des valeurs significativement différentes de celles de la moyenne des individus.

Les compagnies d'assurance de la classe 2 sont caractérisées par un total bilan (T\_BILAN), des actifs admissibles en représentation des engagements réglementés (ACTIFS\_A) ... plus élevés par rapport à la moyenne du marché. Elles sont également caractérisées par des provisions pour sinistres à payer (PSAP), des engagements réglementés (E\_R), des frais généraux (FG)... plus élevés par rapport à la moyenne du marché.

Le tableau complet permettant de caractériser la classe 3 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe3_variables_quantitatives						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
SIN_REA	7,70	4 758 951,75	421 811,68	2 630 545,51	1 145 672,50	1,34E-14
COM_REA	7,50	1 714 300,00	256 809,78	533 522,13	395 454,11	6,45E-14
PRIMES	6,63	18 054 594,25	4 267 032,33	4 403 948,77	4 228 920,82	3,28E-11
ACTIFS_A	5,67	18 878 325,50	4 770 635,41	3 272 291,02	5 060 335,78	1,41E-08
T_BILAN	5,37	32 119 999,50	9 640 045,52	7 148 813,21	8 515 847,31	7,84E-08
R_EXP	5,30	1 858 446,00	258 471,22	895 053,60	614 578,39	1,18E-07
SAP_REA	4,87	3 639 489,00	756 388,60	2 514 390,01	1 205 594,99	1,14E-06
PF_NETS	4,32	748 122,50	156 243,80	310 546,13	279 050,41	1,59E-05
CAP	3,88	4 586 028,25	1 493 529,58	1 813 116,30	1 619 981,14	1,03E-04
AR_NETS	3,09	5 340 961,75	2 196 884,93	2 147 693,92	2 067 842,46	1,98E-03
AR_BRUTS	2,13	5 563 801,25	2 608 168,59	2 269 366,53	2 823 093,76	3,32E-02
PAP	- 1,98	- 910 972,25	- 331 532,13	711 564,51	596 769,34	4,82E-02
RHE	- 4,16	- 663 847,25	- 78 627,89	547 622,34	286 406,57	3,22E-05
COM	- 4,52	- 1 821 160,50	- 546 230,10	1 098 793,47	574 267,55	6,27E-06
E_R	- 5,12	- 16 220 215,25	- 4 837 291,38	2 634 913,99	4 526 120,22	3,11E-07
FG	- 5,14	- 3 584 791,50	- 1 256 866,68	735 058,09	920 915,14	2,70E-07
PSAP	- 5,18	- 13 319 246,25	- 3 801 441,41	1 611 953,31	3 737 446,81	2,20E-07
SIN_REG	- 6,95	- 8 438 001,75	- 1 568 597,59	2 994 899,64	2 010 179,17	3,58E-12
PRI_REA	- 7,84	- 9 994 905,25	- 1 327 147,78	2 878 900,58	2 248 241,09	4,36E-15

**Tableau 61: Données sociétés non vie - description de la classe 3 par les variables initiales**

Les individus de la classe 3 sont représentés sur le dendrogramme en vert. Cette classe est composée de 4 sociétés sur un total de 90. Sur 19 variables initiales, les sociétés de la classe 3 prennent des valeurs significativement différentes de celles de la moyenne des individus.

Les compagnies d'assurance de la classe 3 sont caractérisées par des provisions pour sinistres payés cédés aux réassureurs (SIN\_REA), des commissions de réassurance (COM\_REA), des émissions de primes (PRIMES)... plus élevées par rapport à la moyenne du marché. Elles sont également caractérisées par des primes cédées aux réassureurs (PRI\_REA), des sinistres réglés (SIN\_REG), des provisions pour sinistres à payer (PSAP) plus élevées en valeur absolue par rapport à la moyenne du marché.

### 2.3) Caractérisation des classes par rapport aux composantes principales

Le raisonnement effectué ci-dessus s'applique également à la caractérisation des classes par les composantes principales. La caractérisation de la classe 1 par rapport aux composantes principales est effectuée dans le tableau suivant :

Classe1_variables_quantitatives						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.3	2,36	0,16	1,14E-18	0,82	1,13	1,82E-02
Dim.1	- 7,88	- 1,64	-9,63E-17	1,16	3,44	3,36E-15

**Tableau 62: Données sociétés non vie - description de la classe 1 par les composantes principales**

Les compagnies d'assurance de la classe 1 prennent des valeurs significativement plus faibles que la moyenne des individus sur le 1<sup>er</sup> axe principal. Elles prennent par contre des valeurs significativement plus élevées que la moyenne des individus sur le 3<sup>ème</sup> axe.

L'inertie totale expliquée par le 1<sup>er</sup> axe est de 62,45% contre 6,69% pour le 3<sup>ème</sup> axe (voir Tableau 6: Données sociétés non vie - six premières valeurs propres et pourcentage d'inertie). Le résultat d'exploitation (R\_EXP) et le résultat hors exploitation (RHE) contribue pour 73% à la construction de l'axe 3 (voir Tableau 8: Données sociétés non vie - coordonnées et qualité de représentation des variables)

La caractérisation de la classe 2 par rapport aux composantes principales est effectuée dans le tableau suivant :

Classe2_variables_quantitatives						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.1	5,06	3,69	0,00	1,54	3,44	0,00
Dim.4	2,71	0,55	0,00	0,97	0,95	0,01
Dim.2	-2,13	-0,63	0,00	2,15	1,40	0,03
Dim.3	-3,45	-0,82	0,00	1,36	1,13	0,00

**Tableau 63: Données sociétés non vie - description de la classe 2 par les composantes principales**

Les compagnies d'assurance de la classe 2 ont des coordonnées significativement plus élevées que la moyenne des individus sur le 1<sup>er</sup> et le 4<sup>ème</sup> axe principal. Elles ont des coordonnées significativement plus faibles que la moyenne des individus sur le 2<sup>ème</sup> et le 3<sup>ème</sup> axe principal.

La caractérisation de la classe 3 par rapport aux composantes principales est effectuée dans le tableau suivant :

Classe3_variables_quantitatives						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.1	6,61	11,19	0,00	1,15	3,44	0,00
Dim.2	2,65	1,83	0,00	1,95	1,40	0,01
Dim.4	-2,73	-1,28	0,00	2,97	0,95	0,01

**Tableau 64: Données sociétés non vie - description de la classe 3 par les composantes principales**

Les compagnies d'assurance de la classe 3 ont des coordonnées significativement plus élevées que la moyenne des individus sur les deux premiers axes principaux. Elles ont également des coordonnées significativement plus faibles que la moyenne des individus sur le 4<sup>ème</sup> axe principal. Cet axe qui traduit 4,8% de l'inertie totale résume essentiellement les opérations de réassurance et l'assureur A1 de la classe 3 contribue pour 38% à la construction de cet axe.

#### **2.4) Description des classes par les parangons et les individus spécifiques**

A la suite de la description des classes, par les variables initiales et les composantes principales, l'examen des parangons et des individus spécifiques permet de compléter l'analyse.

Les parangons des différentes classes sont listés dans le tableau suivant :

Parangons de la classe 1					
Individus	A47	A48	A50	A63	A49
Distance/centre de classe	0,46	0,51	0,52	0,52	0,53
Parangons de la classe 2					
Individus	A12	A9	A7	A25	A11
Distance/centre de classe	1,34	1,61	1,64	1,71	1,97
Parangons de la classe 3					
Individus	A2	A3	A6	A1	
Distance/centre de classe	3,92	5,20	5,72	5,73	

**Tableau 65: Données sociétés non vie - description des classes 1, 2 et 3 par les parangons**

Le tableau présente pour chaque parangon la distance le séparant du centre de la classe à laquelle il appartient. L'assureur A47 représente au mieux les assureurs de la classe 1. Les assureurs A12 et A2 sont respectivement les meilleurs parangons des classes 2 et 3.

Les individus spécifiques des différentes classes sont listés dans le tableau suivant :

Individus spécifiques de la classe 1					
Individus	A88	A89	A86	A83	A85
Distances	7,34	7,21	7,18	7,07	6,97
Individus spécifiques de la classe 2					
Individus	A19	A8	A10	A13	A14
Distances	9,52	7,01	6,94	6,91	6,71
Individus spécifiques de la classe 3					
Individus	A6	A3	A2	A1	
Distances	10,58	9,99	9,94	8,75	

**Tableau 66: Données sociétés non vie - description des classes 1, 2, et 3 par les individus spécifiques**

Les tableaux détaillés relatifs à la description des classes par les parangons et les individus spécifiques n'ont pas été fournis dans le corps du mémoire. Cependant une analyse y a été effectuée au point suivant : **(3.3) Description des classes par les parangons et les individus spécifiques page 53)**

### 3) Analyse en composantes principales des sociétés vie

Cette annexe relative à l'ACP des sociétés vie est constituée de façon similaire à celle des sociétés non vie. Elle est composée des 4 points suivants :

- Le tableau complet des valeurs propres ;
- Les tableaux complets des coordonnées, des projections et des contributions des individus sur les 4 premiers axes ;
- La matrice de corrélation ;
- Le tableau des corrélations et des coordonnées des variables initiales par rapport aux 4 premières composantes principales.

Des extraits de ces tableaux ont été fournis dans le corps du mémoire relatif à ces points.

#### 3.1) Valeurs propres et pourcentage d'inertie

Le tableau contient la valeur propre associée à chaque axe (Eigen value), le pourcentage d'inertie (Percentage of variance) qu'elle représente dans l'analyse ainsi que le cumul de ces pourcentages (Cumulative percentage of variance).

Libellés	Eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	13,29	60,41	60,41
comp 2	2,71	12,3	72,71
comp 3	2,17	9,84	82,55
comp 4	0,95	4,33	86,88
comp 5	0,85	3,85	90,73
comp 6	0,76	3,44	94,17
comp 7	0,45	2,06	96,24
comp 8	0,27	1,23	97,47
comp 9	0,2	0,92	98,39
comp 10	0,12	0,56	98,95
comp 11	0,07	0,32	99,27
comp 12	0,05	0,23	99,5
comp 13	0,04	0,18	99,68
comp 14	0,03	0,13	99,81
comp 15	0,02	0,11	99,92
comp 16	0,01	0,07	99,98
comp 17	0	0,01	99,99
comp 18	0	0	100
comp 19	0	0	100
comp 20	0	0	100
comp 21	0	0	100
comp 22	0	0	100

**Tableau 67: Données sociétés vie - valeurs propres et pourcentage d'inertie**

Un extrait de ce tableau a été fourni et analysé dans le corps du mémoire (Tableau 12: Données sociétés vie - six premières valeurs propres et pourcentage d'inertie )

### **3.2) Coordonnées, projection et contribution des individus sur les nouveaux axes**

Le tableau suivant relatif aux 4 premières composantes principales représentant 86,88% de l'inertie totale fournit 3 types d'informations :

- Les coordonnées des compagnies sur les nouveaux axes construits,
- Le cosinus carré de l'angle formé par chaque compagnie avec sa projection sur chacun des axes principaux,
- La contribution de chaque compagnie à la construction de chacun des axes.

	Coordonnées				Cos <sup>2</sup>				Contribution			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
A1	13,72	-4,97	0,42	3,24	0,84	0,11	0,00	0,05	31,47	20,32	0,18	24,44
A2	12,72	4,63	-0,89	-1,30	0,85	0,11	0,00	0,01	27,07	17,59	0,82	3,92
A3	7,25	0,34	-0,82	-1,05	0,79	0,00	0,01	0,02	8,79	0,09	0,70	2,55
A4	6,15	-1,35	-1,36	-1,23	0,79	0,04	0,04	0,03	6,33	1,49	1,89	3,52
A5	4,00	-1,99	0,97	-1,96	0,42	0,10	0,02	0,10	2,68	3,24	0,96	8,94
A6	2,45	-0,85	-1,15	-0,72	0,53	0,06	0,12	0,05	1,00	0,59	1,35	1,21
A7	1,92	-0,69	0,34	0,90	0,44	0,06	0,01	0,10	0,62	0,39	0,12	1,90
A8	1,79	2,69	-1,13	0,17	0,18	0,41	0,07	0,00	0,54	5,95	1,32	0,07
A9	0,17	0,00	0,74	-1,39	0,01	0,00	0,09	0,34	0,01	0,00	0,56	4,54
A10	1,26	-1,95	0,35	-1,25	0,13	0,31	0,01	0,13	0,27	3,12	0,12	3,62
A11	-0,26	-0,63	-0,61	-0,21	0,03	0,15	0,14	0,02	0,01	0,33	0,38	0,10
A12	-0,01	1,60	-0,38	0,30	0,00	0,58	0,03	0,02	0,00	2,11	0,15	0,21
A13	1,96	4,33	0,72	2,93	0,11	0,53	0,01	0,24	0,64	15,37	0,53	19,99
A14	1,70	3,58	-0,23	-0,59	0,10	0,47	0,00	0,01	0,48	10,54	0,06	0,80
A15	-0,56	0,62	4,59	-0,06	0,01	0,02	0,92	0,00	0,05	0,32	21,64	0,01
A16	-0,87	0,03	-1,10	0,00	0,30	0,00	0,49	0,00	0,13	0,00	1,25	0,00
A17	-0,77	-0,27	-0,86	0,10	0,22	0,03	0,28	0,00	0,10	0,06	0,76	0,02
A18	-0,30	1,28	6,90	-0,44	0,00	0,03	0,89	0,00	0,01	1,36	48,83	0,46
A19	-1,02	-1,48	-0,52	-2,60	0,08	0,17	0,02	0,54	0,17	1,80	0,27	15,80
A20	-1,01	-0,90	-0,66	-0,02	0,36	0,29	0,16	0,00	0,17	0,67	0,45	0,00
A21	-0,15	-2,85	2,91	-0,02	0,00	0,19	0,20	0,00	0,00	6,65	8,68	0,00
A22	-0,87	-0,21	-0,78	0,03	0,38	0,02	0,30	0,00	0,13	0,04	0,63	0,00
A23	-1,36	0,05	0,89	-0,29	0,36	0,00	0,16	0,02	0,31	0,00	0,82	0,19
A24	-1,17	-0,25	0,64	0,24	0,55	0,03	0,17	0,02	0,23	0,05	0,42	0,14
A25	-1,03	1,88	-0,20	0,11	0,18	0,62	0,01	0,00	0,18	2,90	0,04	0,03
A26	-1,13	-0,15	-0,55	0,83	0,34	0,01	0,08	0,18	0,21	0,02	0,31	1,60
A27	-1,03	1,51	0,10	0,20	0,18	0,39	0,00	0,01	0,18	1,86	0,01	0,09
A28	-1,49	0,02	-0,75	0,20	0,69	0,00	0,17	0,01	0,37	0,00	0,57	0,09
A29	-1,75	-0,96	0,05	0,56	0,46	0,14	0,00	0,05	0,51	0,76	0,00	0,73
A30	-2,39	-0,25	-0,41	-0,04	0,93	0,01	0,03	0,00	0,96	0,05	0,17	0,00
A31	-1,72	-0,24	-0,34	0,24	0,77	0,01	0,03	0,02	0,49	0,05	0,12	0,14
A32	-2,45	-0,68	-0,63	-0,45	0,83	0,06	0,05	0,03	1,00	0,38	0,41	0,48
A33	-2,20	-0,70	-0,42	0,30	0,82	0,08	0,03	0,02	0,81	0,40	0,18	0,21
A34	-2,21	0,13	-1,01	0,79	0,73	0,00	0,15	0,09	0,82	0,01	1,04	1,45
A35	-1,96	0,72	0,54	-0,06	0,53	0,07	0,04	0,00	0,64	0,42	0,30	0,01
A36	-2,39	0,55	0,19	0,49	0,87	0,05	0,01	0,04	0,96	0,25	0,04	0,55
A37	-2,52	-0,02	-0,73	0,22	0,88	0,00	0,08	0,01	1,06	0,00	0,55	0,12
A38	-2,64	-0,40	-0,28	-0,26	0,92	0,02	0,01	0,01	1,16	0,13	0,08	0,16
A39	-2,67	-0,15	-0,74	-0,06	0,90	0,00	0,07	0,00	1,20	0,02	0,56	0,01
A40	-2,76	-0,52	-0,57	0,26	0,89	0,03	0,04	0,01	1,27	0,22	0,33	0,16
A41	-2,74	-0,47	-0,39	0,34	0,91	0,03	0,02	0,01	1,25	0,18	0,15	0,28
A42	-2,85	-0,11	-0,57	0,38	0,91	0,00	0,04	0,02	1,36	0,01	0,33	0,34
A43	-2,93	-0,24	-0,86	0,26	0,88	0,01	0,08	0,01	1,44	0,05	0,76	0,15
A44	-2,97	-0,41	-0,51	0,37	0,90	0,02	0,03	0,01	1,47	0,14	0,26	0,32
A45	-2,94	-0,25	-0,91	0,53	0,84	0,01	0,08	0,03	1,45	0,05	0,84	0,65

**Tableau 68: Données sociétés vie - coordonnées et qualité de représentation des sociétés**

Un extrait de ce tableau a été fourni et analysé dans le corps du mémoire (Tableau 13: Données sociétés vie - coordonnées et qualité de représentation de 10 sociétés - extrait - )

### 3.3) Matrice de corrélation

Lorsque les variables sont parfaitement corrélées, l'analyse en composantes principales est particulièrement efficace pour effectuer la synthèse de ces variables. Cette corrélation devient un handicap important dans le cadre des techniques prédictives (régression linéaire, analyse discriminante, régression logistique).

Un examen de la matrice de corrélation permet dans ce cas d'identifier les variables trop corrélées. On peut faire l'option dans le cadre de l'utilisation de techniques prédictives d'ignorer certaines de ces variables colinéaires.

La matrice de corrélation des variables mesurées sur les compagnies d'assurance vie se présente comme suit :

		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22
Primes	V1	1,00	0,89	0,98	0,89	0,81	0,50	0,52	0,70	0,57	0,94	0,37	0,20	0,97	0,07	0,11	0,47	0,47	0,98	0,97	0,97	0,98	0,97
PF	V2		1,00	0,92	0,78	0,61	0,58	0,67	0,73	0,50	0,94	0,52	0,22	0,92	0,06	0,00	0,31	0,52	0,93	0,92	0,91	0,93	0,91
Sinistres	V3			1,00	0,86	0,71	0,49	0,56	0,63	0,48	0,95	0,42	0,20	0,96	0,00	0,06	0,38	0,48	0,97	0,96	0,96	0,97	0,96
Com	V4				1,00	0,81	0,43	0,46	0,53	0,48	0,87	0,28	0,19	0,85	0,00	0,01	0,36	0,27	0,86	0,84	0,85	0,86	0,85
FG	V5					1,00	0,37	0,32	0,51	0,53	0,73	0,17	0,05	0,77	0,28	0,26	0,67	0,20	0,78	0,79	0,77	0,78	0,77
PRéass	V6						1,00	0,92	0,45	0,08	0,49	0,50	0,79	0,41	0,16	0,26	0,10	0,44	0,47	0,49	0,41	0,46	0,41
Charges_réa	V7							1,00	0,42	0,04	0,58	0,62	0,68	0,47	0,14	0,23	0,04	0,39	0,52	0,53	0,46	0,52	0,46
RE	V8								1,00	0,79	0,68	0,32	0,19	0,74	0,08	0,08	0,41	0,50	0,73	0,74	0,72	0,73	0,73
RN	V9									1,00	0,56	0,07	0,06	0,65	0,06	0,06	0,57	0,12	0,63	0,63	0,64	0,63	0,64
S_payés	V10										1,00	0,42	0,19	0,94	0,01	0,03	0,42	0,42	0,95	0,94	0,94	0,95	0,94
S_réa	V11											1,00	0,35	0,40	0,11	0,16	0,02	0,21	0,42	0,41	0,38	0,41	0,38
Com_réa	V12												1,00	0,10	0,05	0,13	0,10	0,30	0,14	0,17	0,10	0,13	0,10
PM	V13													1,00	0,01	0,03	0,48	0,47	1,00	0,99	1,00	1,00	1,00
A_bruts	V14														1,00	0,94	0,36	0,01	0,02	0,04	0,01	0,03	0,00
A_nets	V15															1,00	0,28	0,20	0,05	0,07	0,01	0,05	0,02
Ktal	V16																1,00	0,15	0,47	0,52	0,47	0,47	0,48
FP	V17																	1,00	0,46	0,53	0,48	0,45	0,48
PT	V18																		1,00	0,99	1,00	1,00	0,99
T_bilan	V19																			1,00	0,99	0,99	0,99
Actifs	V20																				1,00	1,00	1,00
ER	V21																					1,00	1,00
Actifs.	V22																						1,00

Tableau 69: Données sociétés vie - matrice de corrélation

Au regard des risques de colinéarité, certaines variables trop corrélées peuvent être éliminées dans le cadre de la mise en œuvre des méthodes prédictives.

### 3.4) Corrélation et contribution des variables à la construction des nouveaux axes

Les coordonnées des variables, leur qualité de représentation et leur contribution à la construction des 4 premiers axes sont examinées à travers les paramètres suivants :

- Le coefficient de corrélation de chaque variable initiale avec les nouveaux axes construits,
- Le cosinus carré de l'angle formé par chaque variable avec sa projection sur chacun des axes principaux,
- La contribution de chaque variable à la construction de chacun des axes.

Ces éléments sont consignés dans le tableau suivant :

	Corrélation des variables p/r aux axes				Corrélation des variables p/r aux axes (cos <sup>2</sup> )				Contribution des variables à la construction des axes			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
Primes	0,98	-0,06	0,01	-0,09	0,95	0,00	0,00	0,01	7,17	0,12	0,00	0,78
PF	0,94	0,08	-0,19	-0,01	0,88	0,01	0,04	0,00	6,66	0,21	1,68	0,02
Sinistres	-0,96	0,01	0,10	0,15	0,92	0,00	0,01	0,02	6,91	0,01	0,43	2,25
Com	-0,87	0,10	0,05	0,28	0,75	0,01	0,00	0,08	5,64	0,33	0,13	8,50
FG	-0,79	0,17	-0,35	0,21	0,63	0,03	0,12	0,04	4,72	1,08	5,56	4,53
P_Réass	<b>-0,56</b>	<b>-0,74</b>	0,07	-0,09	0,32	0,55	0,01	0,01	2,38	<b>20,34</b>	0,24	0,77
Charges_réa	<b>0,60</b>	<b>0,72</b>	-0,14	-0,07	0,36	0,52	0,02	0,00	2,71	<b>19,15</b>	0,89	0,46
RE	0,77	-0,03	0,04	0,52	0,59	0,00	0,00	0,27	4,47	0,04	0,08	<b>28,23</b>
RN	<b>0,62</b>	-0,44	0,16	0,42	0,39	0,20	0,03	0,17	2,92	7,24	1,20	<b>18,23</b>
S_payés	-0,96	0,04	0,08	0,12	0,91	0,00	0,01	0,01	6,88	0,05	0,28	1,45
S_réa	<b>0,45</b>	<b>0,49</b>	-0,12	-0,18	0,21	0,24	0,02	0,03	1,55	<b>8,73</b>	0,72	3,36
Com_réa	<b>0,24</b>	<b>0,78</b>	-0,18	0,18	0,06	0,60	0,03	0,03	0,43	<b>22,31</b>	1,49	3,27
PM	-0,98	0,16	0,04	0,03	0,96	0,02	0,00	0,00	7,23	0,92	0,06	0,07
A_bruts	<b>0,08</b>	0,31	<b>0,92</b>	-0,07	0,01	0,09	0,84	0,00	0,04	3,46	<b>38,84</b>	0,49
A_nets	<b>0,11</b>	0,43	<b>0,85</b>	-0,06	0,01	0,19	0,72	0,00	0,09	6,92	<b>33,15</b>	0,37
Ktal	<b>0,49</b>	-0,29	0,56	0,16	0,24	0,09	0,32	0,02	1,84	3,18	14,70	2,57
FP	<b>0,51</b>	0,28	-0,07	0,48	0,26	0,08	0,00	0,23	1,93	2,83	0,22	<b>23,86</b>
PT	-0,99	0,11	0,03	0,05	0,98	0,01	0,00	0,00	7,35	0,42	0,06	0,32
T_bilan	0,99	-0,08	-0,01	0,01	0,98	0,01	0,00	0,00	7,38	0,26	0,01	0,02
Actifs	0,98	-0,16	-0,05	-0,03	0,95	0,03	0,00	0,00	7,17	0,98	0,13	0,10
ER	-0,99	0,11	0,03	0,05	0,97	0,01	0,00	0,00	7,33	0,46	0,04	0,31
Actifs.	0,98	-0,16	-0,04	-0,03	0,95	0,03	0,00	0,00	7,18	0,95	0,08	0,07

Tableau 70: Données sociétés vie - corrélation et qualité de représentation des variables

Ce tableau a fait l'objet d'analyse dans le corps du mémoire et un extrait y a été produit (Tableau 14: Données sociétés vie - corrélation et qualité de représentation des variables - extrait - ).

#### 4) Classification des sociétés vie

Cette annexe relative à la classification des sociétés vie est composée des points suivants :

- Représentations des partitions obtenues à la suite de la classification selon différentes modalités ;
- Caractérisation des classes par rapport aux variables initiales ;
- Caractérisation des classes par rapport aux composantes principales ;
- Description des classes par les parangons et les individus spécifiques.

##### 4.1) Représentations de la partition

La forme du dendrogramme et le niveau de coupure optimal calculé par la fonction HCPC suggère une partition des sociétés d'assurance vie en 6 classes.

Les 6 classes peuvent être mises en évidence et représentées sur le plan principal. Le dendrogramme peut également faire l'objet de représentation en 3 dimensions sur le plan principal.

Le dendrogramme mettant en évidence les 6 classes se présente comme suit :

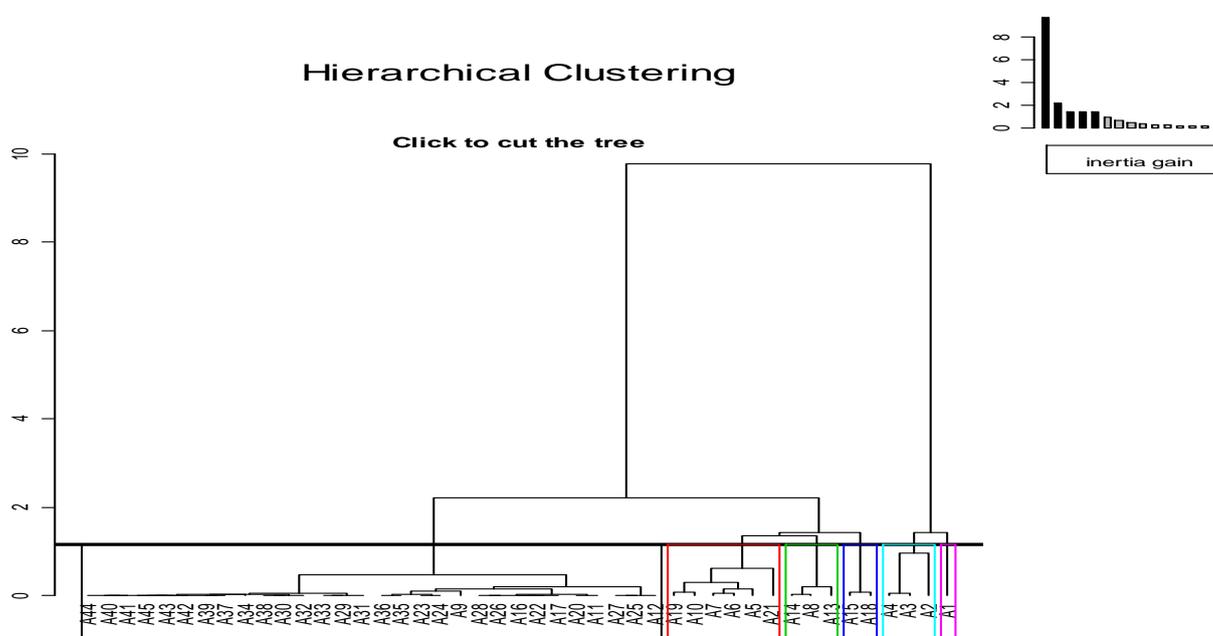


Figure 22: Données sociétés vie - arbre hiérarchique

A la suite de la mise en évidence des différentes classes à partir du dendrogramme, une visualisation peut également être effectuée sur le plan principal.

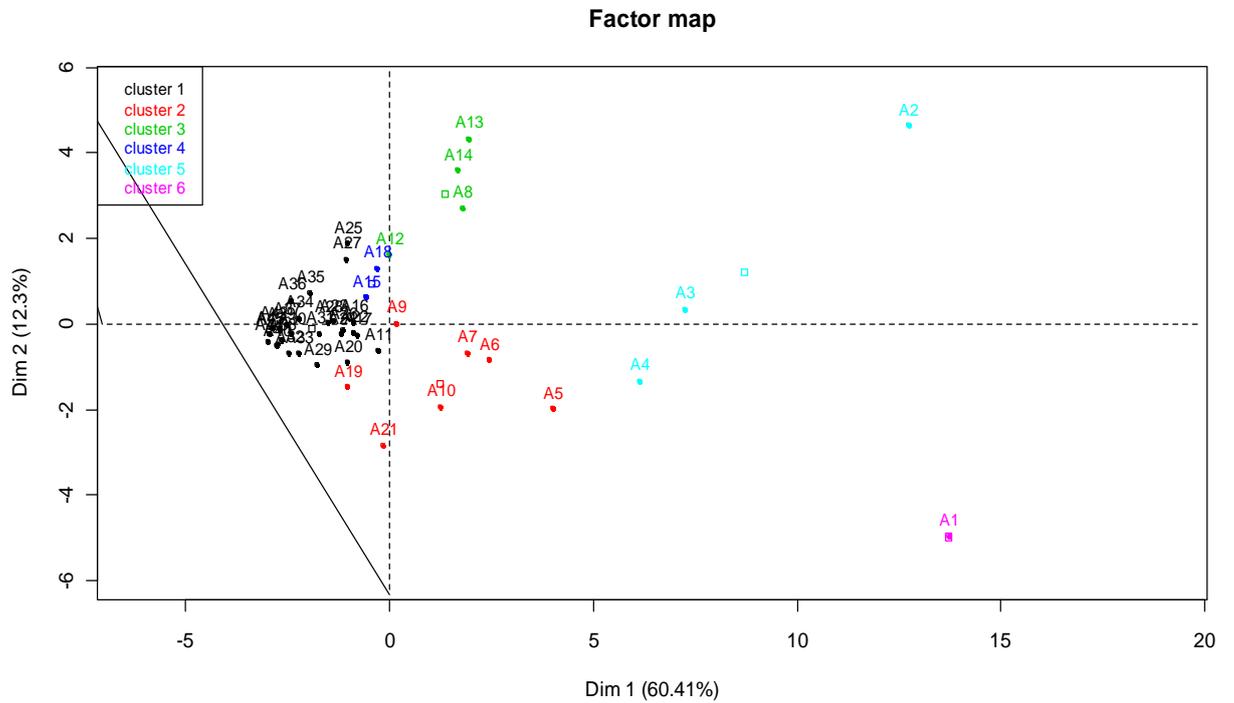


Figure 23: Données sociétés vie - représentation de la partition sur le plan principal

La fonction HCPC permet également de représenter le dendrogramme en trois dimensions sur le plan principal. Cette représentation est la suivante :

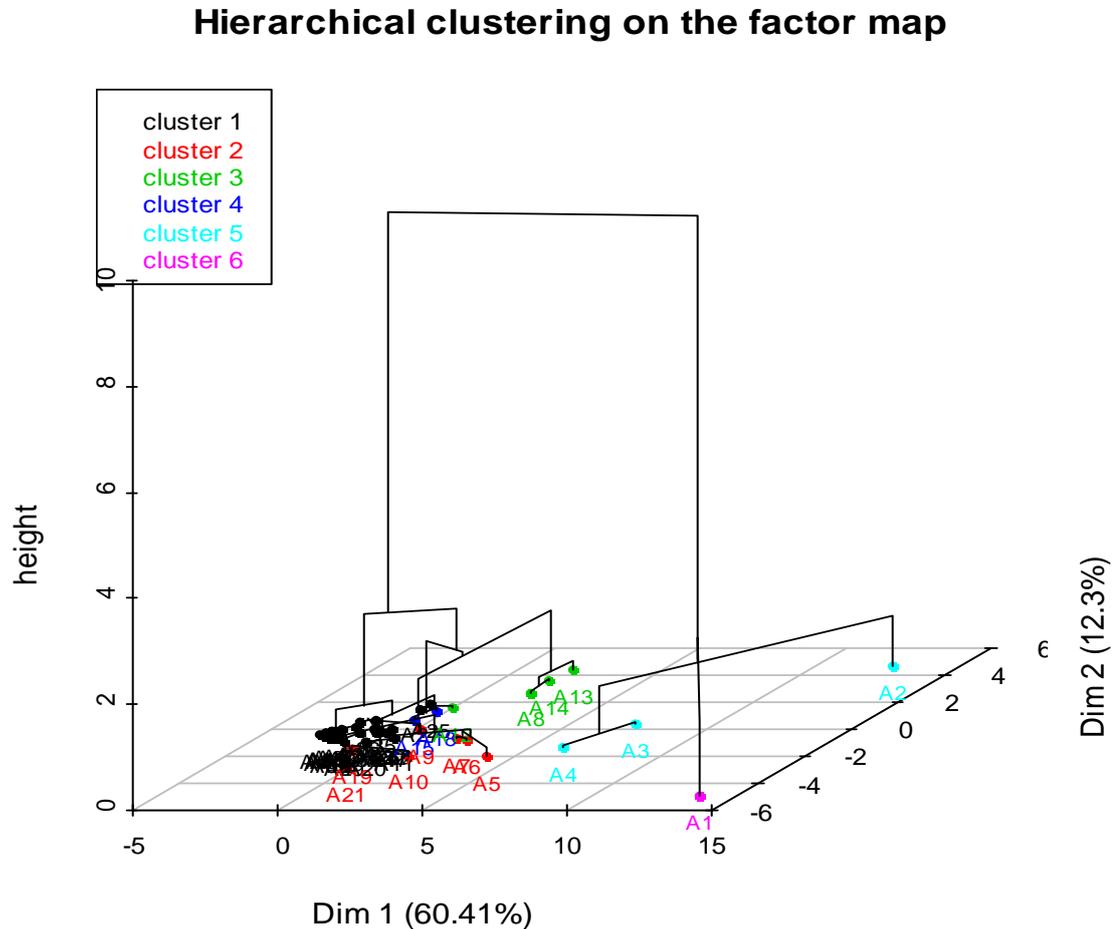


Figure 24: Données sociétés vie - représentation du dendrogramme en trois dimensions sur le plan principal

Ces trois représentations ont été annoncées et analysées dans le corps du mémoire à la section suivante : **3.1) Forme du dendrogramme et nombre de classes page 64.**

## 4.2) Caractérisation des classes par rapport aux variables initiales

Les classes peuvent être décrites à partir des variables initiales ou à partir des composantes principales.

Le tableau complet permettant de caractériser la classe 1 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe1_variable_quantitative							
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value	
FG	4,96	- 390 967 117	- 681 381 343	176 018 147	498 831 392	7,19E-07	
Com	4,42	- 115 010 254	- 216 545 912	85 429 316	195 372 158	9,68E-06	
ER	4,23	- 3 119 944 880	- 10 242 669 996	2 886 799 934	14 346 746 312	2,37E-05	
PT	4,22	- 3 067 730 958	- 10 067 134 499	2 858 991 248	14 108 339 558	2,41E-05	
Sinistres	4,22	- 1 027 851 241	- 2 580 280 465	914 978 755	3 130 478 168	2,43E-05	
S_payés	4,12	- 530 188 714	- 1 611 517 959	651 160 446	2 235 590 167	3,83E-05	
PM	4,11	- 2 965 225 029	- 9 425 296 831	2 818 782 724	13 374 123 532	3,92E-05	
P_Réass	2,78	- 93 127 682	- 189 178 751	99 035 092	294 136 377	5,44E-03	
S_réa	- 2,41	27 095 808	64 693 051	47 236 885	132 767 966	1,59E-02	
A_nets	- 2,52	246 702 000	429 627 029	273 348 136	617 113 831	1,16E-02	
Charges_réa	- 2,58	52 663 598	132 605 506	65 674 644	263 383 051	9,77E-03	
A_bruts	- 2,61	259 022 007	506 731 699	267 423 012	809 251 612	9,17E-03	
RN	- 2,67	28 156 395	247 737 575	114 467 978	700 871 701	7,65E-03	
RE	- 2,91	49 430 795	228 437 908	150 851 253	524 060 483	3,64E-03	
PF	- 3,38	93 007 340	365 788 501	104 635 271	687 411 080	7,30E-04	
Ktal	- 3,49	595 714 286	781 222 222	202 209 731	453 030 727	4,90E-04	
Actifs	- 3,98	3 113 701 714	10 373 102 037	2 867 803 937	15 528 829 103	6,90E-05	
Actifs.	- 3,99	3 239 944 747	10 589 005 119	2 969 709 111	15 677 863 451	6,59E-05	
T_bilan	- 4,20	4 098 564 754	12 069 367 791	3 286 330 704	16 151 047 664	2,65E-05	
Primes	- 4,66	1 530 716 151	3 397 430 373	1 071 181 155	3 407 911 942	3,12E-06	

**Tableau 71: Données sociétés vie - description de la classe 1 par les variables initiales**

Un extrait de ce tableau a été fourni (Tableau 16: Données sociétés vie - description de la classe 1 par les variables initiales - extrait - ) et analysé **(3.2) Description des classes par les variables initiales ou les composantes principales page 64** ) dans le corps du mémoire. Cet extrait et l'analyse effectuée sont prolongés dans les paragraphes suivants pour les autres classes restantes.

Le tableau complet permettant de caractériser la classe 2 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe2_variable_quantitative						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Ktal	2,09	1 114 285 714	781 222 222	673 325 588	453 030 727	3,63E-02
Com	-2,44	- 383 810 380	- 216 545 912	130 269 374	195 372 158	1,48E-02
FP	-2,61	- 855 384 950	695 866 984	2 819 044 928	1 689 036 418	8,93E-03
FG	-3,16	- 1 235 803 715	- 681 381 343	412 020 658	498 831 392	1,55E-03

**Tableau 72: Données sociétés vie - description de la classe 2 par les variables initiales**

Sur les 4 variables du tableau, les sociétés de la classe 2 prennent des valeurs significativement différentes de la moyenne du marché. Cette classe représentée en rouge sur le dendrogramme comprend 7 sociétés. Les compagnies d'assurance de cette classe sont caractérisées par un niveau de capitalisation (Ktal) plus élevé par rapport à la moyenne du marché. Elles sont également caractérisées par un niveau de frais généraux (FG) plus élevé par rapport à la moyenne du marché et un niveau de fonds propres (FP) plus faible par rapport à la moyenne du marché.

Il s'agit essentiellement d'entreprises ayant consommé une part significative de leurs fonds propres par une absence de rentabilité et des frais généraux excessifs. Cette observation doit cependant être nuancée car l'écart type de 2 819 millions de FCFA au sein de cette classe sur la variable FP est significatif.

Plusieurs entreprises de cette classe font l'objet d'un suivi particulier de la Commission Régionale de Contrôle des Assurances du fait de leur fragilité.

Le tableau complet permettant de caractériser la classe 3 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe3_variable_quantitative						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Com_réa	4,95	161 497 673	34 869 610	40 344 924	52 954 594	7,26E-07
S_réa	2,75	241 006 824	64 693 051	231 111 892	132 767 966	5,93E-03
Charges_réa	2,72	478 976 380	132 605 506	157 071 770	263 383 051	6,44E-03
P_Réass	-3,40	- 671 990 704	- 189 178 751	215 441 212	294 136 377	6,72E-04

**Tableau 73: Données sociétés vie - description de la classe 3 par les variables initiales**

Cette classe représentée en vert sur le dendrogramme comprend 4 sociétés. Les compagnies d'assurance de la classe 3 sont caractérisées par un niveau

d'opérations de réassurance (Com\_rea, S\_réa, Charges\_réa, P\_réass) plus élevé en valeur absolue par rapport à la moyenne du marché.

Sur les 4 compagnies composant cette classe, 3 appartiennent au même groupe avec les mêmes politiques de produit et de réassurance. Une diligence particulière résultant de ce constat est de s'assurer dans le cadre d'un contrôle de l'adéquation du programme de réassurance avec l'activité de l'entreprise et de sa correcte application pour l'établissement des comptes de réassurance.

Le tableau complet permettant de caractériser la classe 4 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe4_variable_quantitative						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
A_nets	5,62	2 852 946 762	429 627 029	353 506 199	617 113 831	1,94E-08
A_bruts	5,53	3 637 428 514	506 731 699	1 137 987 951	809 251 612	3,12E-08
Ktal	2,03	1 425 000 000	781 222 222	75 000 000	453 030 727	4,21E-02

**Tableau 74: Données sociétés vie - description de la classe 4 par les variables initiales**

Les deux compagnies d'assurance de la classe 4 sont caractérisées par un niveau d'arriérés de primes très élevé par rapport à la moyenne du marché.

Les contrôles opérés au sein de ces entités ont confirmé le risque de non recouvrement pesant sur ces arriérés et l'insuffisance des provisions constituées pour faire face à ce risque.

Le tableau complet permettant de caractériser la classe 5 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe5_variable_quantitative						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
PF	5,25	2 402 941 294	365 788 501	513 749 609	687 411 080	1,49E-07
Charges_réa	4,35	778 680 763	132 605 506	535 172 192	263 383 051	1,37E-05
Actifs	4,03	45 666 209 829	10 373 102 037	8 494 385 935	15 528 829 103	5,60E-05
Actifs.	4,00	45 950 109 884	10 589 005 119	8 721 881 879	15 677 863 451	6,37E-05
S_réa	3,99	363 541 638	64 693 051	133 543 125	132 767 966	6,60E-05
T_bilan	3,96	48 159 545 813	12 069 367 791	10 740 734 673	16 151 047 664	7,45E-05
Primes	3,88	10 864 605 325	3 397 430 373	2 326 470 949	3 407 911 942	1,03E-04
RE	2,67	1 018 254 054	228 437 908	235 767 994	524 060 483	7,54E-03
FG	-2,10	- 1 271 346 786	- 681 381 343	526 555 787	498 831 392	3,60E-02
P_Réass	-2,75	- 645 629 604	- 189 178 751	693 234 950	294 136 377	5,94E-03
Com	-3,65	- 619 110 372	- 216 545 912	185 024 817	195 372 158	2,59E-04
PM	-4,10	- 40 356 335 967	- 9 425 296 831	6 545 735 407	13 374 123 532	4,13E-05
ER	-4,20	- 44 212 388 871	- 10 242 669 996	9 841 731 493	14 346 746 312	2,70E-05
PT	-4,24	- 43 819 987 830	- 10 067 134 499	9 657 388 520	14 108 339 558	2,22E-05
Sinistres	-4,48	- 10 491 886 565	- 2 580 280 465	1 839 501 064	3 130 478 168	7,45E-06
S_payés	-4,82	- 7 683 626 997	- 1 611 517 959	1 850 391 035	2 235 590 167	1,47E-06

**Tableau 75: Données sociétés vie - description de la classe 5 par les variables initiales**

Sur ces 16 variables, les individus de la classe 5 prennent des valeurs significativement différentes de celles de la population.

Les trois compagnies d'assurance de la classe 5 sont caractérisées par des produits financiers (PF), des charges de réassurance (Charges\_réa), des actifs, des provisions techniques (PT), des sinistres, des sinistres payés (S\_payés) ...plus élevés par rapport à la moyenne du marché. La situation financière de ces entreprises est plus satisfaisante que celle des autres entreprises du marché.

Le tableau complet permettant de caractériser la classe 6 par rapport à toutes les variables initiales pertinentes est le suivant :

Classe6_variable_quantitative						
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
RN	5,43	4 052 079 753	247 737 575	NA	700 871 701	5,70E-08
RE	4,37	2 516 237 247	228 437 908	NA	524 060 483	1,27E-05
Actifs.	4,25	77 175 449 375	10 589 005 119	NA	15 677 863 451	2,16E-05
Actifs	4,21	75 779 527 339	10 373 102 037	NA	15 528 829 103	2,53E-05
T_bilan	4,11	78 515 129 058	12 069 367 791	NA	16 151 047 664	3,89E-05
Primes	3,54	15 464 712 711	3 397 430 373	NA	3 407 911 942	3,99E-04
PF	3,04	2 458 369 125	365 788 501	NA	687 411 080	2,33E-03
FP	2,99	5 745 734 640	695 866 984	NA	1 689 036 418	2,79E-03
Ktal	2,69	2 000 000 000	781 222 222	NA	453 030 727	7,14E-03
Com	-2,71	- 745 645 507	- 216 545 912	NA	195 372 158	6,77E-03
FG	-2,94	- 2 150 008 472	- 681 381 343	NA	498 831 392	3,24E-03
Sinistres	-3,12	- 12 340 926 580	- 2 580 280 465	NA	3 130 478 168	1,82E-03
S_payés	-3,13	- 8 616 431 137	- 1 611 517 959	NA	2 235 590 167	1,73E-03
PT	-3,93	- 65 517 939 758	- 10 067 134 499	NA	14 108 339 558	8,48E-05
ER	-3,98	- 67 363 425 565	- 10 242 669 996	NA	14 346 746 312	6,85E-05
PM	-4,19	- 65 403 682 452	- 9 425 296 831	NA	13 374 123 532	2,84E-05

**Tableau 76: Données sociétés vie - description de la classe 6 par les variables initiales**

Sur ces 16 variables, la seule compagnie composant la classe 6 prend des valeurs significativement différentes de celles de la population.

La classe 6 présente des similitudes avec la classe 5. La différence fondamentale entre ces deux classes réside dans le fait que la compagnie de la classe 6 n'effectue quasiment pas d'opérations en réassurance.

La caractérisation de chaque classe par les composantes principales suit la même logique que la description par les variables initiales.

A la suite de la description des classes par les variables initiales et les composantes principales, il est également possible d'effectuer une description à partir des parangons et des individus spécifiques.

### 4.3) Caractérisation des classes par rapport aux composantes principales

Le tableau ci-dessous permet de caractériser les 6 classes examinées ci-dessus par rapport aux composantes principales :

Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Classe1_composantes_principales						
Dim.3	-2,37	-0,41	0,00	0,49	1,47	0,02
Dim.1	-4,43	-1,90	0,00	0,81	3,65	0,00
Classe2_composantes_principales						
Dim.5	-2,35	-0,76	0,00	1,77	0,92	0,02
Dim.2	-2,42	-1,40	0,00	0,89	1,65	0,02
Dim.4	-2,93	-1,01	0,00	1,09	0,98	0,00
Classe3_composantes_principales						
Dim.2	3,84	3,05	0,00	1,02	1,65	0,00
Dim.5	-2,04	-0,91	0,00	0,32	0,92	0,04
Classe4_composantes_principales						
Dim.3	5,59	5,74	0,00	1,15	1,47	0,00
Classe5_composantes_principales						
Dim.1	4,23	8,71	0,00	2,87	3,65	0,00
Dim.4	-2,16	-1,19	0,00	0,11	0,98	0,03
Classe6_composantes_principales						
Dim.1	3,76	13,72	0,00	NA	3,65	0,00
Dim.4	3,32	3,24	0,00	NA	0,98	0,00
Dim.2	-3,02	-4,97	0,00	NA	1,65	0,00

**Tableau 77: Données sociétés vie - description des classes 1 à 6 par les composantes principales**

Les compagnies d'assurance de la classe 1 ont des coordonnées significativement plus faibles que la moyenne des individus sur le 1<sup>er</sup> (faible niveau d'activité) et le 3<sup>ème</sup> axe principal.

Les compagnies d'assurance de la classe 2 ont des coordonnées significativement plus faible que la moyenne des individus sur le 2<sup>ème</sup> (peu de recours à la réassurance), le 4<sup>ème</sup> et le 5<sup>ème</sup> axe principal.

Les compagnies d'assurance de la classe 3 ont des coordonnées significativement plus élevées que la moyenne des individus sur le 2<sup>ème</sup> axe principal (recours intensif à la réassurance). Elles ont des coordonnées significativement plus faibles que la moyenne des individus sur le 5<sup>ème</sup> axe principal.

Les compagnies d'assurance de la classe 4 ont des coordonnées significativement plus élevées que la moyenne des individus sur le 3<sup>ème</sup> axe principal.

Les compagnies d'assurance de la classe 5 ont des coordonnées significativement plus élevées que la moyenne des individus sur le 1<sup>er</sup> axe principal. Elles sont des coordonnées significativement plus faible que la moyenne des individus sur le 4<sup>ème</sup> axe principal.

Les compagnies d'assurance de la classe 6 ont des coordonnées significativement plus élevées que la moyenne des individus sur le 1<sup>er</sup> et le 4<sup>ème</sup> axe principal. Elles sont des coordonnées significativement plus faible que la moyenne des individus sur le 2<sup>ème</sup> axe principal (peu de recours à la réassurance).

A la suite de la description des classes par les variables initiales et les composantes principales, il est également possible d'effectuer une description à partir des parangons et des individus spécifiques.

#### 4.4) Description des classes par les parangons et les individus spécifiques

Les parangons des différentes classes sont listés dans le tableau suivant :

Parangons de la classe 1					
Individus	A31	A30	A28	A33	A37
Distance	0,41	0,57	0,60	0,69	0,72
Parangons de la classe 2					
Individus	A10	A6	A7	A9	A19
Distance	0,78	2,29	2,46	2,57	2,99
Parangons de la classe 3					
Individus	A8	A14	A12	A13	
Distance	1,29	1,44	2,06	2,81	
Parangons de la classe 4					
Individus	A18	A15			
Distance	1,32	1,32			
Parangons de la classe 5					
Individus	A3	A4	A2		
Distance	1,83	3,64	5,36		

**Tableau 78: Données sociétés vie - description des classes 1 à 6 par les parangons**

Le tableau présente pour chaque parangon la distance le séparant du centre de la classe à laquelle il appartient. L'assureur A31 représente au mieux les assureurs de la classe 1. Les assureurs A10, A8, A18 et A3 sont respectivement les meilleurs parangons des classes 2, 3, 4, et 5. L'assureur A1, le seul individu de la classe 6, est le parangon de cette classe.

Les individus spécifiques des différentes classes sont listés dans le tableau suivant :

Individus spécifiques de la classe 1					
Individus	A45	A43	A44	A42	A40
Distance	4,94	4,83	4,78	4,76	4,54
Individus spécifiques de la classe 2					
Individus	A21	A5	A6	A10	A7
Distance	6,69	6,08	4,48	4,24	4,01
Individus spécifiques de la classe 3					
Individus	A13	A14	A8	A12	
Distance	6,66	5,08	4,65	2,67	
Individus spécifiques de la classe 4					
Individus	A18	A15			
Distance	7,49	5,16			
Individus spécifiques de la classe 5					
Individus	A2	A3	A4		
Distance	10,78	6,69	5,49		
Individus spécifiques de la classe 6					
Individus	A1				
Distance	9,22				

**Tableau 79: Données sociétés vie - description des classes 1 à 6 par les individus spécifiques**

Les tableaux détaillés relatifs à la description des classes par les parangons et les individus spécifiques n'ont pas été fournis dans le corps du mémoire. Cependant une analyse y a été effectuée au point suivant : **3.3) Description des classes par les parangons et les individus spécifiques page 65**

### **5) ACP et CAH des sociétés non vie et vie du groupe G**

Dans le cadre de l'ACP et la CAH réalisées sur les données des sociétés vie et non vie du groupe G, plusieurs sorties du logiciels n'ont pas été présentées dans le corps du mémoire du fait des contraintes de volumes imposées. Ces différentes sorties sont cependant présentées et analysées dans les paragraphes ci-dessous.

Cette annexe constitue donc une documentation complémentaire des analyses effectuées aux point suivants : **1) Etude des compagnies d'assurance non vie du groupe G page 67 et 2) Etude des compagnies d'assurance vie du groupe G page 71.**

### 5.1) Analyse en composantes principales des sociétés non vie du groupe G

Les valeurs propres, le pourcentage d'inertie associé à chaque axe et le pourcentage cumulé sont présentés dans le tableau suivant :

Libellés	Eigenvalue	Percentage of variance	Cumulative percentage of variance
comp 1	14,74	77,57	77,57
comp 2	2,73	14,38	91,95
comp 3	1,03	5,41	97,36
comp 4	0,33	1,74	99,09
comp 5	0,17	0,91	100

Tableau 80: Sociétés non vie du groupe G - valeurs propres et pourcentage d'inertie

Le plan principal représente 91,95% de l'inertie totale. Ce taux est excellent dans l'analyse en composantes principales.

L'éboulis des valeurs propres permet de constater une chute brutale d'inertie entre les deux premières composantes principales. Ce constat conforté par le pourcentage d'inertie de 91,95% du plan principal suggère de limiter dans un premier temps l'analyse aux deux premiers axes.

L'éboulis des valeurs propres est présenté comme suit :

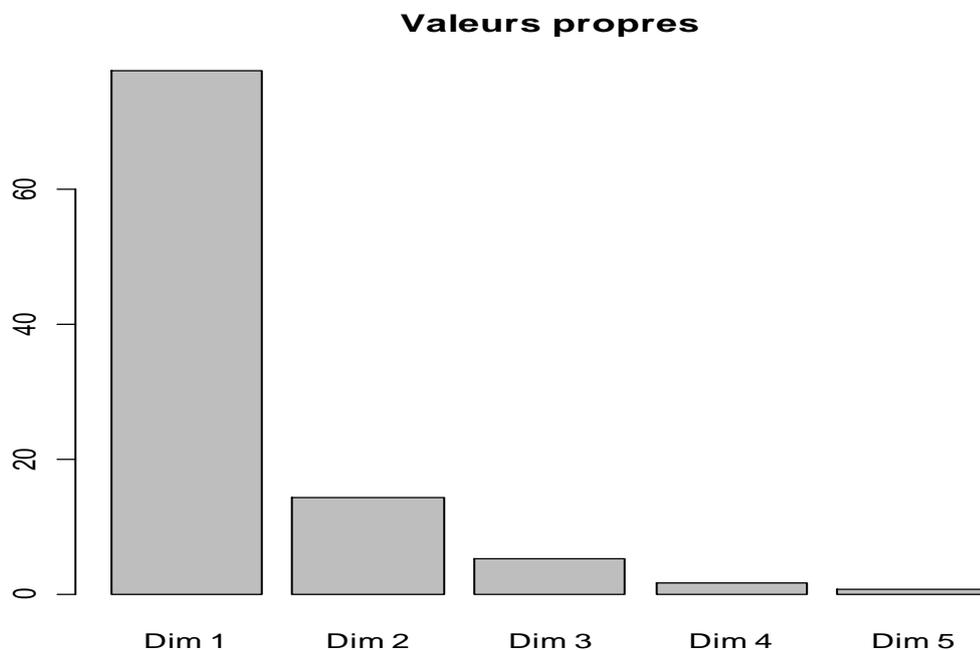


Figure 25: Données sociétés non vie groupe G - éboulis des valeurs propres

Le tableau suivant permet de chiffrer les oppositions entre les différentes entités du groupe notamment à partir des coordonnées de chaque entité sur les 2 premiers axes :

Libellés	Coordonnées				Cos2				Contribution			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
A4_Gabon	<b>4,13</b>	<b>-2,9</b>	-0,72	-0,22	0,65	0,32	0,02	0	19,31	51,46	8,47	2,39
A5_CI	<b>6,23</b>	<b>2,3</b>	0,06	0,37	0,88	0,12	0	0	43,95	32,4	0,07	6,9
A22_Bénin	-1,14	0,81	0,92	-1,11	0,32	0,16	0,21	0,3	1,48	3,98	13,85	62,52
A23_Sénégal	-2,27	-0,22	-0,21	0,08	0,85	0,01	0,01	0	5,84	0,28	0,7	0,36
A29_Congo	-2,51	-0,98	1,51	0,73	0,62	0,1	0,22	0,05	7,13	5,89	37	26,71
A31_Togo	-4,44	0,99	-1,57	0,15	0,85	0,04	0,11	0	22,28	5,99	39,91	1,12

**Tableau 81: Données sociétés non vie groupe G - coordonnées et qualité de représentation des sociétés**

L'axe 1 oppose les compagnies A4 et A5 aux autres entités du groupe. Les sociétés A4 et A5 ont une activité significativement plus importante que celle des autres entités du groupe.

L'axe 2 traduit essentiellement le résultat hors exploitation (RHE). Il oppose principalement la compagnie A4 à la compagnie A5.

L'analyse permet de constater une hétérogénéité des sociétés du groupe et une relative faiblesse du niveau d'activité des filiales A22, A23, A29 et A31 par rapport à A4 et A5.

Les axes 3 et 4 traduisent essentiellement des opérations de réassurance. Même s'ils renferment peu d'inertie, ils constituent dans le cadre d'un audit, des axes de diligences additionnelles du fait des particularités et des oppositions mises en évidence.

Les coordonnées des variables, leur qualité de représentation et leur contribution à la construction des 4 premiers axes sont examinées à travers les paramètres suivants :

- Le coefficient de corrélation de chaque variable initiale avec les nouveaux axes construits,
- Le cosinus carré de l'angle formé par chaque variable avec sa projection sur chacun des axes principaux,
- La contribution de chaque variable à la construction de chacun des axes.

Ces éléments sont consignés dans le tableau suivant :

Libellés	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
PRIMES	0,96	-0,2	-0,19	-0,07	0,92	0,04	0,04	0	6,24	1,45	3,5	1,36
SIN_REG	-0,96	-0,12	0,23	0,07	0,93	0,01	0,05	0	6,3	0,52	5,17	1,43
AR_BRUTS	0,84	-0,49	0,17	0,06	0,71	0,24	0,03	0	4,82	8,79	2,89	1,06
AR_NETS	0,84	-0,49	0,17	0,06	0,71	0,24	0,03	0	4,82	8,79	2,89	1,06
PF_NETS	0,96	0,17	-0,01	-0,19	0,93	0,03	0	0,04	6,3	1,07	0,02	10,87
COM	-0,91	0,3	0,26	0,08	0,84	0,09	0,07	0,01	5,67	3,25	6,69	2,06
FG	-0,96	-0,2	0,19	-0,07	0,92	0,04	0,03	0	6,24	1,48	3,36	1,49
PRI_REA	-0,89	0,38	-0,21	-0,13	0,79	0,15	0,04	0,02	5,35	5,35	4,34	5,34
SIN_REA	0,94	-0,09	-0,15	0,24	0,89	0,01	0,02	0,06	6,06	0,27	2,25	18,05
SAP_REA	0,83	0,51	-0,09	0,23	0,68	0,26	0,01	0,05	4,63	9,37	0,83	15,75
<b>COM_REA</b>	<b>0,59</b>	<b>-0,1</b>	<b>0,78</b>	<b>-0,07</b>	<b>0,35</b>	<b>0,01</b>	<b>0,61</b>	<b>0,01</b>	<b>2,37</b>	<b>0,34</b>	<b>58,92</b>	<b>1,62</b>
PSAP	-0,95	-0,3	-0,01	0,11	0,9	0,09	0	0,01	6,1	3,23	0,02	3,9
PAP	-0,89	0,43	-0,03	0,02	0,79	0,19	0	0	5,36	6,86	0,09	0,12
CAP	0,88	0,47	0,02	-0,03	0,78	0,22	0	0	5,29	8	0,03	0,36
T_BILAN	0,98	0,18	0,04	0,03	0,97	0,03	0	0	6,55	1,13	0,17	0,27
R_EXP	0,93	-0,26	-0,07	0,16	0,86	0,07	0,01	0,03	5,86	2,41	0,52	7,91
<b>RHE</b>	<b>0,09</b>	<b>0,93</b>	<b>0,29</b>	<b>0,18</b>	<b>0,01</b>	<b>0,86</b>	<b>0,08</b>	<b>0,03</b>	<b>0,06</b>	<b>31,47</b>	<b>7,95</b>	<b>10,26</b>
E_R	-0,95	-0,28	-0,01	0,15	0,9	0,08	0	0,02	6,08	2,9	0	6,5
ACTIFS_A	0,93	0,3	0,06	-0,19	0,87	0,09	0	0,03	5,91	3,3	0,38	10,6

**Tableau 82: Données sociétés non vie groupe G - coordonnées et qualité de représentation des variables**

Au cours de l'analyse du nuage des variables des sociétés non vie du groupe G, ce tableau a fait l'objet d'analyse dans le corps du mémoire **(1.2) Etude du nuage des variables à travers une ACP page 69)** même s'il n'y a pas été partiellement reproduit. Cette analyse est rappelée dans les deux paragraphes suivants.

Toutes les variables ont un coefficient de corrélation élevé avec la 1<sup>ère</sup> composante principale à l'exception des variables COM\_REA (Commissions de Réassurance) et RHE. La variable COM\_REA est plutôt bien représentée sur l'axe 3. Elle contribue à 58,92 % à la construction de cet axe. La variable RHE est plutôt bien représentée sur l'axe 2 avec un coefficient de corrélation de 0,93. Les quelques variables contribuant le plus à la construction de l'axe 2 sont les suivantes avec leur taux de contribution : RHE (31,47%), SAP\_REA (9,37%), AR\_BRUTS (8,79%) et AR\_NETS (8,79%).

Les axes 3 et 4 peuvent être intéressants à examiner même s'ils représentent peu d'inertie. Ils mettent en évidence des spécificités, sources de risques dans un groupe où les procédures et traitements devraient être harmonisés.

## 5.2) CAH des sociétés non vie du groupe G

### Représentations de la partition

La forme du dendrogramme et le niveau de coupure optimal calculé par la fonction HCPC suggèrent une partition des sociétés d'assurance vie en 3 classes. Le dendrogramme est produit dans le corps du mémoire (Figure 9: Données sociétés non vie du groupe G - arbre hiérarchique)

Les 3 classes peuvent être mises en évidence et représentées sur le plan principal. Le dendrogramme peut également faire l'objet de représentation en 3 dimensions sur le plan principal.

A la suite de la mise en évidence des différentes classes à partir du dendrogramme, une visualisation peut également être effectuée sur le plan principal.

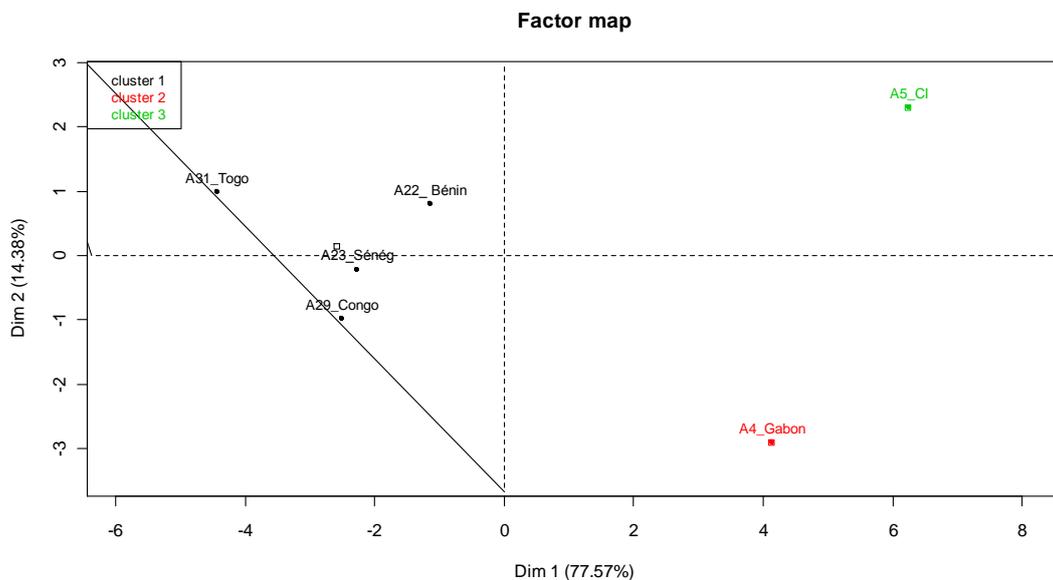
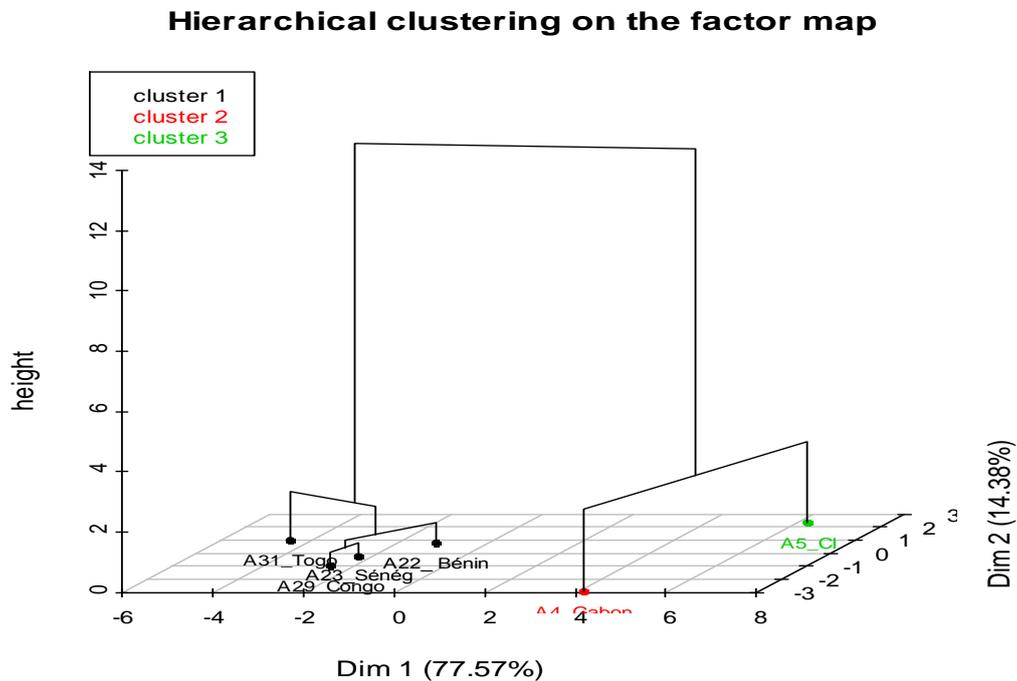


Figure 26: Données sociétés non vie groupe G - représentation de la partition sur le plan principal

La fonction HCPC permet également de représenter le dendrogramme en trois dimensions sur le plan principal. Cette représentation est la suivante :



**Figure 27: Données sociétés non vie du groupe G - représentation du dendrogramme en trois dimensions sur le plan principal**

Ces deux représentations ont été annoncées et analysées dans le corps du mémoire à la section suivante : **1.3) Confirmation des résultats par une CAH page 70.**

### *Caractérisation des classes*

La caractérisation des classes résultant de la CAH sociétés non vie du groupe G n'a pas été systématisée et formalisée dans le corps du mémoire. Cette systématisation est effectuée dans les paragraphes suivants.

Pour 9 variables, les sociétés de la classe 1 prennent des valeurs significativement différentes de celles de la moyenne de la population. Le tableau suivant permet de matérialiser ce constat.

Classe 1_ variable quantitative							
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value	
COM	2,15	-607 123	-1 055 854	147 009	660 276	0,03	
SIN_REG	2,12	-1 620 393	-2 717 183	499 386	1 633 601	0,03	
FG	2,1	-1 182 002	-1 780 338	153 112	900 767	0,04	
PAP	2,04	-209 810	-432 594	134 059	346 124	0,04	
PF_NETS	-1,99	39 404	109 355	52 152	110 976	0,05	
T_BILAN	-2,04	6 882 868	12 523 958	2 611 318	8 748 184	0,04	
R_EXP	-2,08	210 466	465 618	170 709	387 066	0,04	
SIN_REA	-2,14	107 476	296 032	89 485	278 577	0,03	
PRIMES	-2,19	4 904 556	7 586 868	909 910	3 881 541	0,03	

Sur toutes les variables, la société A4 de la classe 2 prend des valeurs proches de celles de la moyenne de la population.

Sur 2 variables présentées dans le tableau ci-dessous, l'assureur A5 de la classe 3 prend des valeurs significativement différentes de la moyenne de la population.

Classe 3_ variable quantitative							
Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value	
SAP_REA	2,19	4 766 006	1 317 794	NA	1 576 785	0,03	
CAP	2,07	4 984 632	1 818 346	NA	1 533 225	0,04	

La description des différentes classes peut être effectuée également à partir des individus spécifiques et des parangons présentés dans le tableau suivant :

	Individus spécifiques de la classe 1				Parangons de la classe 1			
Individus	A31	A29	A23	A22	A23	A29	A22	A31
Distance	9,46	7,33	7,05	6,72	1,06	1,94	2,07	2,71
	Individus spécifiques de la classe 2				Parangons de la classe 2			
Individus	A4				A4			
Distance	5,70				0			
	Individus spécifiques de la classe 3				Parangons de la classe 3			
Individus	A5				A5			
Distance	5,70				0			

**Tableau 83: Données sociétés non vie groupe G - description des classes 1 à 3 par les parangons et les individus spécifiques**

Les assureurs A5 et A4, singletons dans leurs classes respectives, sont des individus spécifiques et des parangons triviaux. Ils sont à équidistance du centre de la classe 1. Cette classe a pour parangon la société A23 et pour individu spécifique la société A31.

### 5.3) Analyse en composantes principales des sociétés vie du groupe G

Les valeurs propres et le pourcentage d'inertie associé sont présentés dans le tableau suivant :

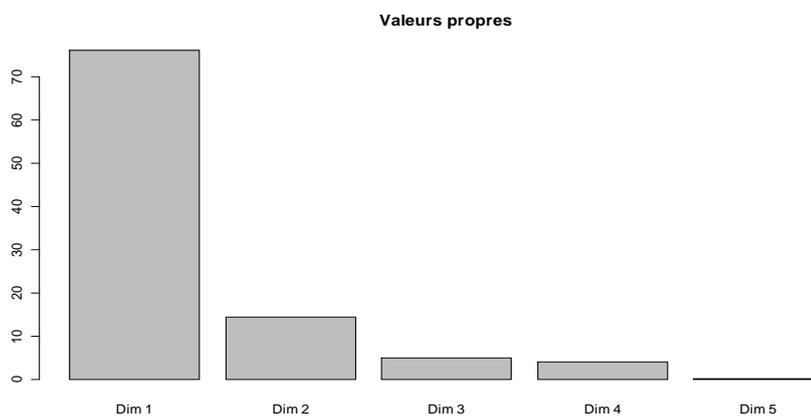
Libellés	Eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	16,73	76,06	76,06
comp 2	3,19	14,51	90,57
comp 3	1,10	5,02	95,59
comp 4	0,90	4,10	99,68
comp 5	0,07	0,32	100,00

**Tableau 84: Données sociétés vie groupe G - valeurs propres et pourcentage d'inertie**

Le plan principal représente 90,57% de l'inertie totale. Ce taux est excellent dans l'analyse en composantes principales.

L'éboulis des valeurs propres permet de constater une chute brutale d'inertie entre les deux premières composantes principales. Ce constat conforté par le pourcentage d'inertie de 90,57% du plan principal suggère de limiter dans un premier temps l'analyse aux deux premiers axes.

L'éboulis des valeurs propres est présenté comme suit :



**Figure 28: Données sociétés vie du groupe G - éboulis des valeurs propres**

Le tableau suivant permet de chiffrer les oppositions entre les différentes entités du groupe notamment à partir des coordonnées de chaque entité sur les 2 premiers axes :

Libellés	Coordonnées		Cos <sup>2</sup>		Contribution	
	Dim.1	Dim.2	Dim.1	Dim.2	Dim.1	Dim.2
A30	1,45	1,24	0,34	0,25	2,1	7,99
A37	1,85	-2,8	0,27	0,62	3,42	40,82
A38	1,62	2,21	0,29	0,54	2,6	25,44
A39	1,59	-1,87	0,29	0,4	2,53	18,32
A40	2,6	1,19	0,54	0,11	6,71	7,43
<b>A5</b>	<b>-9,11</b>	0,03	1	0	82,64	0,01

**Tableau 85: Données sociétés vie groupe G - coordonnées et qualité de représentation des sociétés**

L'axe 1 oppose la compagnie A5 à toutes les autres compagnies du groupe. L'assureur A5 est très bien représenté sur l'axe 1 (coordonnées : -9,11, projection : 1, contribution : 82,64%) par rapport aux autres compagnies.

L'assureur A5 est cependant mal représenté sur l'axe 2 (coordonnées : 0,03, projection : 0, contribution : 0,01%). Cet axe oppose les compagnies A37 et A39 aux compagnies A30, A38 et A40.

Les coordonnées des variables, leur qualité de représentation et leur contribution à la construction des 4 premiers axes sont examinées à travers les paramètres suivants :

- Le coefficient de corrélation de chaque variable initiale avec les nouveaux axes construits,
- Le cosinus carré de l'angle formé par chaque variable avec sa projection sur chacun des axes principaux,
- La contribution de chaque variable à la construction de chacun des axes.

Ces éléments sont consignés dans le tableau suivant :

Libellés	Corrélation des variables p/r axes				Corrélation des variables p/r aux axes (cos <sup>2</sup> )				Contribution des variables à la construction des axes			
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4	Dim.1	Dim.2	Dim.3	Dim.4
Primes	-1,00	0,02	0,00	0,07	1,00	0,00	0,00	0,00	5,95	0,02	0,00	0,47
PF	-0,99	-0,05	-0,05	0,08	0,99	0,00	0,00	0,01	5,90	0,08	0,20	0,75
Sinistres	1,00	0,01	0,00	-0,08	0,99	0,00	0,00	0,01	5,94	0,00	0,00	0,65
Com	0,98	-0,16	-0,07	0,06	0,96	0,02	0,01	0,00	5,74	0,76	0,47	0,33
FG	0,99	-0,08	0,05	0,08	0,98	0,01	0,00	0,01	5,85	0,21	0,23	0,75
P_Réass	0,63	0,69	-0,31	-0,18	0,39	0,47	0,10	0,03	2,35	14,73	8,82	3,60
Charges_réa	-0,23	-0,94	0,22	0,04	0,05	0,87	0,05	0,00	0,31	27,37	4,25	0,19
RE	0,97	0,02	0,02	0,24	0,94	0,00	0,00	0,06	5,63	0,02	0,05	6,16
RN	0,50	0,07	-0,12	0,85	0,25	0,00	0,01	0,73	1,52	0,13	1,31	80,69
S_payés	1,00	0,01	0,07	-0,04	0,99	0,00	0,01	0,00	5,94	0,00	0,41	0,16
S_réa	-0,98	-0,08	-0,13	-0,08	0,97	0,01	0,02	0,01	5,77	0,21	1,56	0,73
Com_réa	0,07	-0,87	0,49	-0,04	0,00	0,75	0,24	0,00	0,03	23,55	21,43	0,17
PM	1,00	0,00	0,03	-0,08	0,99	0,00	0,00	0,01	5,93	0,00	0,10	0,69
A_bruts	-0,39	0,72	0,57	0,05	0,15	0,52	0,32	0,00	0,92	16,43	28,93	0,31
A_nets	-0,39	0,72	0,57	0,05	0,15	0,52	0,32	0,00	0,92	16,43	28,93	0,31
Ktal	-1,00	0,01	-0,07	0,05	0,99	0,00	0,01	0,00	5,93	0,00	0,46	0,30
FP	-0,98	-0,04	-0,15	-0,07	0,96	0,00	0,02	0,00	5,76	0,06	1,97	0,48
PT	1,00	0,00	0,04	-0,08	0,99	0,00	0,00	0,01	5,93	0,00	0,11	0,63
T_bilan	-1,00	0,00	-0,05	0,06	0,99	0,00	0,00	0,00	5,94	0,00	0,22	0,46
Actifs	-1,00	0,00	-0,05	0,08	0,99	0,00	0,00	0,01	5,92	0,00	0,21	0,77
ER	1,00	0,00	0,04	-0,08	0,99	0,00	0,00	0,01	5,93	0,00	0,11	0,64
Actifs.	-1,00	0,00	-0,05	0,08	0,99	0,00	0,00	0,01	5,92	0,00	0,22	0,76

**Tableau 86: Données sociétés vie groupe G - corrélation et qualité de représentation des variables**

Toutes les variables sont bien représentées dans le plan principal à l'exception de la variable résultat net (RN). Cette variable contribue à 80,69% à la construction du 4<sup>ème</sup> axe, non représenté ici. Le 4<sup>ème</sup> axe représente 4,10% de l'inertie totale. La représentativité est donc faible. Il permet cependant d'attirer l'attention sur l'absence de corrélation entre le résultat net par rapport au résultat d'exploitation.

Le 1<sup>er</sup> axe représente le niveau d'activité des compagnies. Les opérations de réassurance et les arriérés contribuent pour 98,5% à la construction de l'axe 2.

A la suite de l'ACP réalisée sur les données des sociétés vie du groupe, une CAH peut être mise en œuvre pour avoir un autre angle d'analyse sur les mêmes données.

## 5.4) CAH des sociétés vie du groupe G

### Représentations de la partition

La forme du dendrogramme et le niveau de coupure optimal calculé par la fonction HCPC suggèrent une partition des sociétés d'assurance vie en 3 classes. Le dendrogramme est produit dans le corps du mémoire (Figure 12: Données sociétés vie du groupe G - arbre hiérarchique)

Les 3 classes peuvent être mises en évidence et représentées sur le plan principal. Le dendrogramme peut également faire l'objet de représentation en 3 dimensions sur le plan principal.

A la suite de la mise en évidence des différentes classes à partir du dendrogramme, une visualisation peut également être effectuée sur le plan principal.

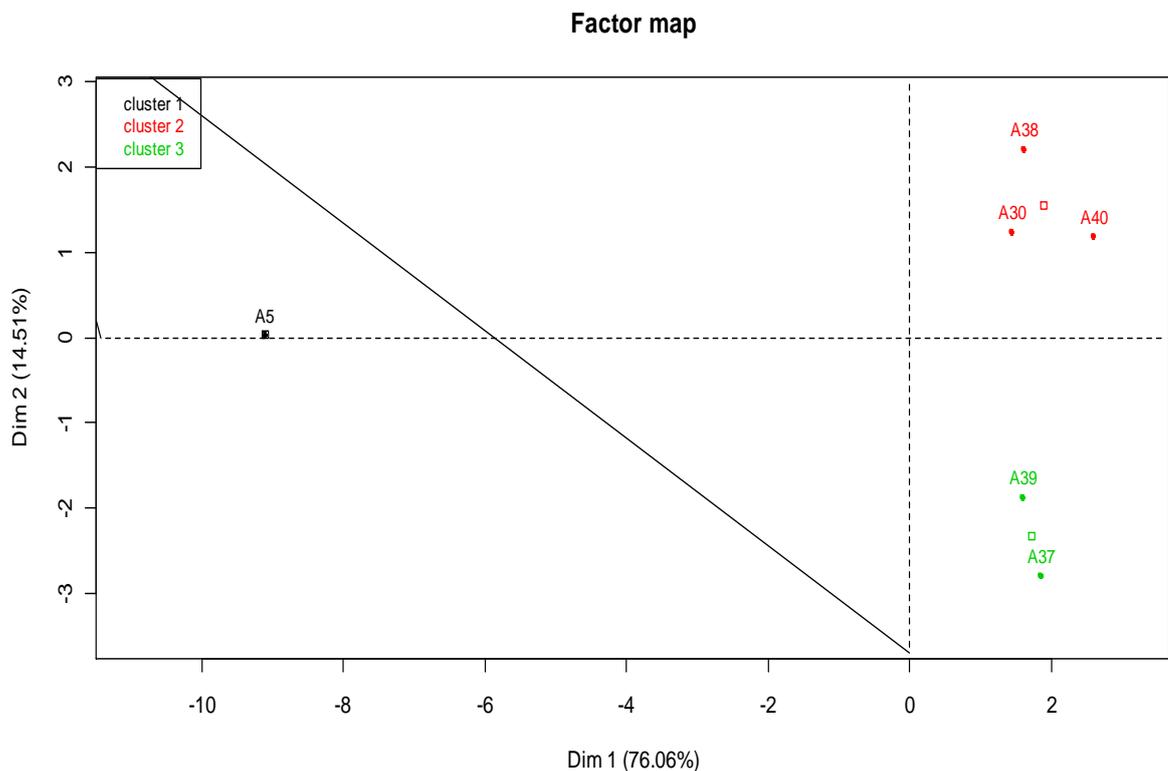


Figure 29: Données sociétés vie du groupe G - représentation de la partition sur le plan principal

La fonction HCPC permet également de représenter le dendrogramme en trois dimensions sur le plan principal. Cette représentation est la suivante :

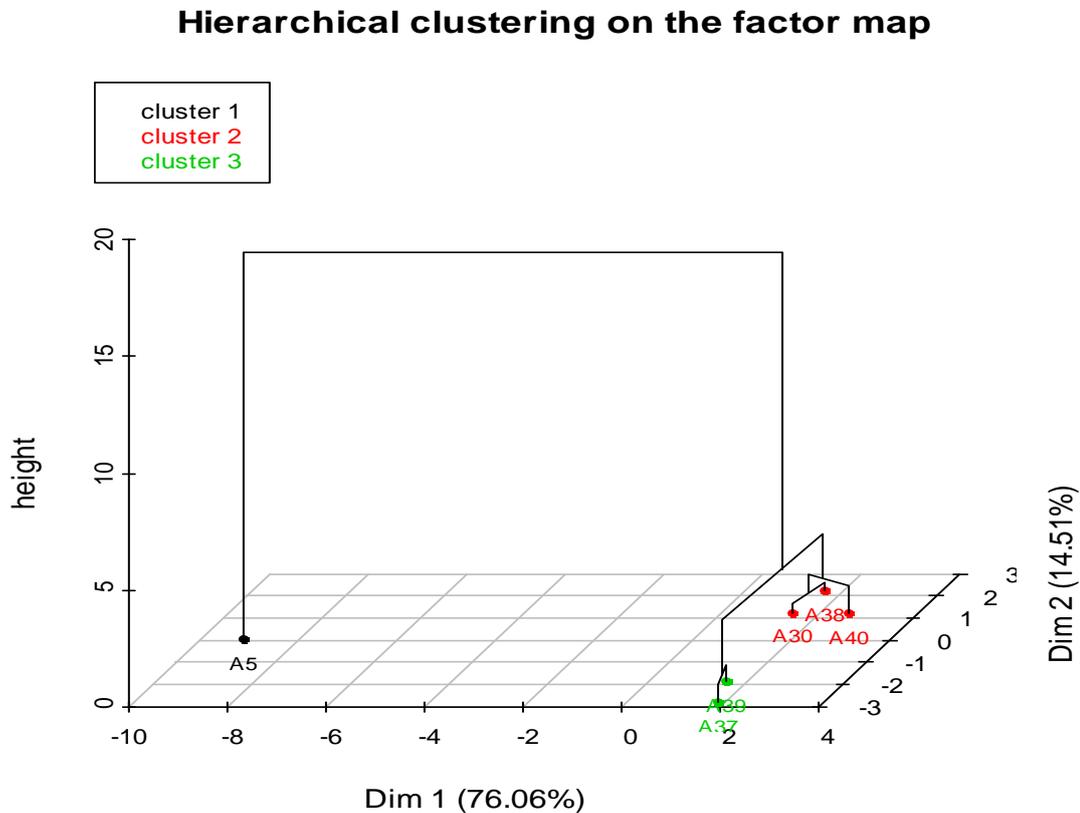


Figure 30: Données sociétés vie groupe G - représentation du dendrogramme en trois dimensions dans le plan principal

Ces deux représentations ont été annoncées et analysées dans le corps du mémoire à la section suivante: **2.3) Confirmation des résultats par une CAH page 74.**

### *Caractérisation des classes*

La caractérisation des classes résultant de la CAH sociétés vie du groupe G n'a pas été systématisée et formalisée dans le corps du mémoire. Cette systématisation est effectuée dans les paragraphes suivants.

Pour la classe 1 composée seulement de l'assureur A5, la valeur test associée à 17 variables sur 19 est en valeur absolue supérieure à 2. Sur ces 17 variables, les valeurs prises par l'assureur A5 sont significativement différentes de celles prises par les autres sociétés du groupe.

Le tableau de ces 17 variables est le suivant :

Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Ktal	2,24	2 250 000 000	791 666 667	NA	652 186 493	0,03
T_bilan	2,24	37 191 912 108	7 350 373 211	NA	13 350 544 941	0,03
Actifs.	2,23	32 441 101 889	6 137 945 181	NA	11 772 835 792	0,03
Actifs	2,23	32 318 166 830	6 091 356 894	NA	11 738 797 651	0,03
PF	2,23	608 371 204	113 297 096	NA	221 995 649	0,03
Primes	2,23	8 341 051 061	2 108 507 622	NA	2 795 568 020	0,03
FP	2,20	1 598 503 536	523 458 408	NA	489 396 193	0,03
S_réa	2,20	35 289 886	6 984 478	NA	12 887 092	0,03
RE	- 2,14	- 673 562 635	- 182 666 749	NA	229 478 824	0,03
Com	- 2,16	- 457 083 650	- 146 146 199	NA	143 781 343	0,03
FG	- 2,21	- 1 848 691 767	- 613 154 211	NA	560 139 086	0,03
Sinistres	- 2,23	- 7 233 171 469	- 1 610 240 858	NA	2 522 491 532	0,03
PM	- 2,23	- 26 552 241 254	- 5 050 329 653	NA	9 626 830 562	0,03
ER	- 2,23	- 27 911 057 875	- 5 347 942 919	NA	10 100 891 120	0,03
PT	- 2,23	- 27 811 849 585	- 5 302 274 616	NA	10 076 883 884	0,03
S_payés	- 2,24	- 3 785 537 580	- 687 180 346	NA	1 386 085 643	0,03

**Tableau 87: Données sociétés vie du groupe G - description de la classe 1 par les variables initiales**

Pour la classe 2, une seule valeur test associée à la variable charges de réassurance (Charges\_réa) est supérieure en valeur absolue à 2. Sur cette variable, la moyenne de la classe est significativement inférieure à la moyenne de la population. Cette variable testée est présentée dans le tableau suivant :

Libellés	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Charges_réa	- 2,02	5 983 616	31 678 952	5 267 993	28 426 067	0,04

Pour la classe 3, aucune valeur test calculée n'est supérieure en valeur absolue à 2. Sur les 19 variables, la moyenne de la classe est proche de la moyenne de la population.

Les individus spécifiques et les parangons des différentes classes sont présentés dans le tableau suivant :

	Individus spécifiques de la classe 1			Parangons de la classe 1		
Individus	A5			A5		
Distance	11,10			0		
	Individus spécifiques de la classe 2			Parangons de la classe 2		
Individus	A38	A40	A30	A38	A30	A40
Distance	4,64	4,28	3,96	1,53	1,57	2,19
	Individus spécifiques de la classe 3			Parangons de la classe 3		
Individus	A37	A39		A37	A39	
Distance	4,46	3,89		1,48	1,48	

**Tableau 88: Données société vie groupe G - description des classes 1 à 3 par les parangons et les individus spécifiques**

Les sociétés A5, A38 et A37 représentent le mieux leurs classes respectives. Elles sont les parangons et les individus spécifiques de leurs classes.

### **III. ANNEXES RELATIVES A LA MISE EN ŒUVRE DES MODELES PREDICTIFS SUR LES DONNEES COMPTABLES ET FINANCIERES DES COMPAGNIES D'ASSURANCE**

Ces annexes relatives aux méthodes prédictives sont composées des différents documents suivants :

- Annexes relatives à la régression linéaire multiple sur les données des sociétés non vie et vie ;
- Annexes relatives à l'analyse discriminante sur les données des sociétés non vie et vie.

#### **1) Régression linéaire multiple sur les données des sociétés non vie et vie**

Cette annexe relative à la régression linéaire multiple sur les données des sociétés non vie et vie est composée des 4 points suivants :

- Généralités sur les provisions techniques des sociétés d'assurance ;
- Modèle avec PSAP comme variable à expliquer ;
- Modèle avec PAP comme variable à expliquer ;
- Modèle avec PM comme variable à expliquer.

##### **1.1) Généralités sur les provisions techniques des sociétés d'assurance**

Le code des assurances des Etats membres de la CIMA définit en ses articles 334-2 et suivants les provisions techniques des opérations d'assurance.

Ainsi l'article 334-2 définit les provisions mathématiques comme le différentiel entre la valeur actuelle probable des engagements de l'assureur et la valeur actuelle probable des engagements des assurés.

L'article 334-8 définit la provision pour sinistres à payer comme la valeur estimative des dépenses en principal et en frais nécessaires au règlement de tous les sinistres survenus et non payés. Selon les dispositions de l'article 334-12, les provisions pour sinistres à payer doivent être évaluées dossier par dossier. La CRCA peut cependant accorder des dérogations pour l'estimation d'une partie de la provision pour sinistres à payer. Les compagnies peuvent mettre en place d'autres méthodes d'évaluation notamment statistiques pour s'assurer que les montants résultant de l'évaluation dossier par dossier sont prudents. Les méthodes statistiques généralement utilisées

dans ce cadre sont du type chain ladder. Ces méthodes permettent d'évaluer les provisions pour sinistres à payer à partir de la cadence de paiement des exercices antérieurs.

La provision pour annulations de primes est définie comme une provision destinée à faire face aux annulations probables à intervenir sur les primes émises et non encaissées. La CRCA a mis en place une circulaire permettant de déterminer statistiquement la provision pour annulations de primes à partir de la cadence d'annulations des primes des exercices précédents.

## 1.2) Modèle avec PSAP comme variable à expliquer

Le logiciel  offre de belles possibilités graphiques pour visualiser un phénomène. A partir de la représentation ci dessous, on peut voir comment la variable à expliquer PSAP se comporte par rapport aux deux variables explicatives ACTIFS\_A et SIN\_REA. La représentation ci-dessous est obtenue à partir des commandes suivantes :

```
> donnees_non_vie<-read.table("donnees_non_vie.csv", header=TRUE, sep=';',
dec='.', row.name=1) # importation des données
> library("scatterplot3d") # On charge le package
> scatterplot3d(donnees_non_vie [,"PSAP"], donnees_non_vie [,"SIN_REA"],
donnees_non_vie [,"ACTIFS_A"], type="h", pch=16, box = FALSE, xlab ="ACTIFS_A",
ylab="SIN_REA", zlab="PSAP") # Représentation en 3 dimensions
```

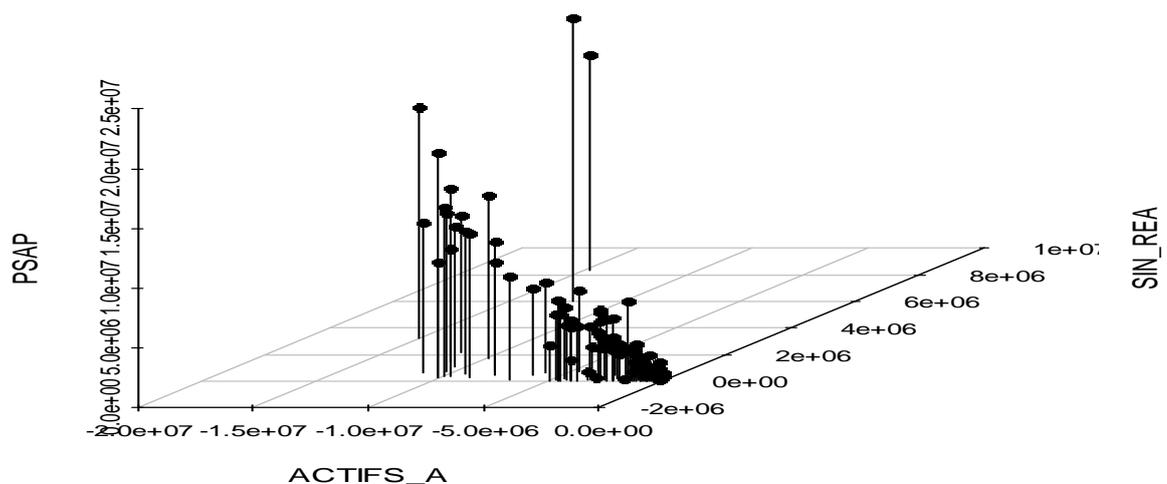


Figure 31: Représentation des PSAP en fonction des ACTIFS\_A et des SIN\_REA

Deux individus extrêmes pourraient perturber le modèle. Ils peuvent éventuellement être éliminés avant d'effectuer la modélisation. Lorsqu'ils sont éliminés, on aboutit à la représentation suivante qui paraît plus homogène :

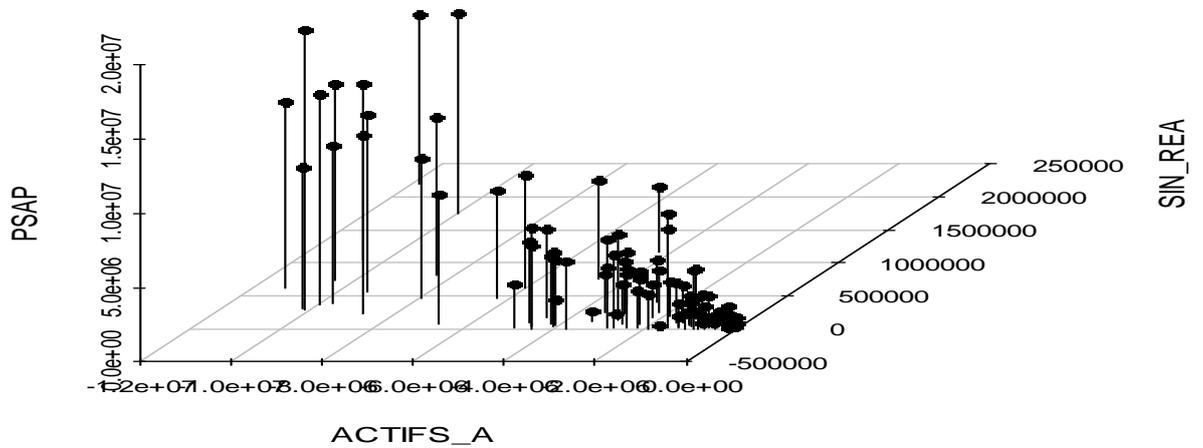


Figure 32: Représentation des PSAP en fonction des ACTIFS\_A et des SIN\_REA sans deux individus extrêmes

Le tableau suivant donne pour chaque compagnie non vie ayant un numéro pair, l'estimation effectuée avec le modèle pour obtenir le montant des PSAP avec un intervalle de confiance à 95%.

Libellés	Fit	lwr	upr	Test1	Test2	PSAP	Ecart	En %
A2	- 22 350 430	- 25 961 018	- 18 739 841	NON	INFERIEUR	-12 267 624	-10 082 806	82%
A6	- 25 040 795	- 30 475 402	- 19 606 187	NON	INFERIEUR	-14 716 191	-10 324 604	70%
A10	- 10 754 096	- 12 738 294	- 8 769 898	OUI	INFERIEUR	-9 981 689	-772 407	8%
A14	- 10 601 836	- 12 736 223	- 8 467 449	OUI	INFERIEUR	-9 735 175	-866 661	9%
A18	- 2 519 024	- 4 313 619	- 724 428	OUI	INFERIEUR	-2 362 919	-156 105	7%
A22	- 4 680 204	- 6 504 077	- 2 856 332	OUI	INFERIEUR	-4 548 846	-131 358	3%
A26	- 6 106 300	- 8 033 970	- 4 178 629	OUI	INFERIEUR	-4 686 841	-1 419 459	30%
A28	- 3 729 246	- 5 534 576	- 1 923 917	OUI	INFERIEUR	-3 194 085	-535 161	17%
A36	- 884 076	- 2 749 711	981 559	OUI	INFERIEUR	-660 348	-223 728	34%
A40	- 2 292 550	- 4 096 633	- 488 467	OUI	INFERIEUR	-1 543 251	-749 299	49%
A42	- 1 066 153	- 2 857 485	725 178	OUI	INFERIEUR	-875 876	-190 277	22%
A44	- 3 908 434	- 5 728 031	- 2 088 836	OUI	INFERIEUR	-2 890 000	-1 018 434	35%
A46	- 2 212 239	- 4 001 766	- 422 712	OUI	INFERIEUR	-2 211 896	-343	0%
A48	- 2 426 934	- 4 216 818	- 637 051	OUI	INFERIEUR	-2 310 425	-116 509	5%
A50	- 1 645 246	- 3 439 120	148 627	OUI	INFERIEUR	-1 198 764	-446 482	37%
A56	- 3 045 893	- 4 957 088	- 1 134 698	OUI	INFERIEUR	-1 939 503	-1 106 390	57%
A58	- 650 533	- 2 456 242	1 155 177	OUI	INFERIEUR	-425 118	-225 415	53%
A60	- 1 952 990	- 3 745 002	- 160 978	OUI	INFERIEUR	-1 408 974	-544 016	39%
A62	- 1 168 486	- 2 965 667	628 695	OUI	INFERIEUR	-475 621	-692 865	146%
A70	- 1 719 933	- 3 521 391	81 526	OUI	INFERIEUR	-1 330 222	-389 711	29%
A72	- 2 813 035	- 4 631 820	- 994 250	OUI	INFERIEUR	-2 717 813	-95 222	4%
A74	- 762 054	- 2 556 838	1 032 730	OUI	INFERIEUR	-309 992	-452 062	146%
A78	- 931 613	- 2 732 597	869 372	OUI	INFERIEUR	-760 047	-171 566	23%
A80	- 676 341	- 2 472 825	1 120 143	OUI	INFERIEUR	-568 300	-108 041	19%
A82	- 1 212 840	- 3 015 921	590 241	OUI	INFERIEUR	-188 461	-1 024 379	544%
A84	- 675 163	- 2 471 636	1 121 311	OUI	INFERIEUR	-181 797	-493 366	271%
A86	- 762 613	- 2 567 711	1 042 485	OUI	INFERIEUR	-40 278	-722 335	1793%
A88	- 649 129	- 2 452 319	1 154 061	NON	INFERIEUR	-	#VALEUR!	#VALEUR!
A90	- 265 173	- 2 074 507	1 544 160	OUI	INFERIEUR	-119 856	-145 317	121%
A4	- 5 867 961	- 7 742 882	- 3 993 039	OUI	SUPERIEUR	-5 995 145	127 184	-2%
A8	- 3 943 728	- 6 960 217	- 927 239	NON	SUPERIEUR	-7 933 254	3 989 526	-50%
A12	- 8 425 224	- 10 278 754	- 6 571 694	OUI	SUPERIEUR	-9 365 585	940 361	-10%
A16	- 9 967 616	- 11 931 594	- 8 003 639	OUI	SUPERIEUR	-9 994 856	27 240	0%
A20	- 6 256 251	- 8 058 663	- 4 453 840	OUI	SUPERIEUR	-6 264 136	7 885	0%
A24	- 1 692 689	- 3 621 789	236 410	OUI	SUPERIEUR	-3 201 612	1 508 923	-47%
A30	- 2 773 814	- 4 561 982	- 985 646	OUI	SUPERIEUR	-3 000 005	226 191	-8%
A32	- 4 155 517	- 5 974 213	- 2 336 821	OUI	SUPERIEUR	-4 537 375	381 858	-8%
A34	- 2 731 568	- 4 560 574	- 902 562	OUI	SUPERIEUR	-2 848 770	117 202	-4%
A38	- 1 783 167	- 3 604 716	38 382	NON	SUPERIEUR	-4 899 929	3 116 762	-64%
A52	- 1 061 808	- 2 949 258	825 643	OUI	SUPERIEUR	-1 127 721	65 913	-6%
A54	- 2 932 226	- 5 051 578	- 812 874	OUI	SUPERIEUR	-3 236 573	304 347	-9%
A64	- 2 683 492	- 4 472 395	- 894 590	OUI	SUPERIEUR	-2 732 516	49 024	-2%
A66	- 1 970 555	- 3 758 304	- 182 806	OUI	SUPERIEUR	-2 232 799	262 244	-12%
A68	- 3 550 960	- 7 103 413	1 494	OUI	SUPERIEUR	-6 716 140	3 165 180	-47%
A76	- 1 605 936	- 3 492 315	280 443	OUI	SUPERIEUR	-1 710 892	104 956	-6%

Tableau 89: Données sociétés non vie - prédiction des PSAP

Un extrait de ce tableau a été produit et a fait l'objet d'analyse dans le corps du mémoire (Tableau 22: Données sociétés non vie - prédiction des PSAP - extrait )

### 1.3) Modèle avec PAP comme variable à expliquer

Le tableau suivant donne pour chaque compagnie non vie ayant un numéro pair, l'estimation effectuée avec le modèle pour obtenir le montant des PAP avec un intervalle de confiance à 95%.

Libellés	fit	lwr	upr	Test1	Test2	PAP	Ecart	En %	
A2	-	51 108	- 1 002 583	900 367	NON	SUPERIEUR	- 2 141 000	2 089 892	-98%
A4	-	1 674 094	- 2 358 890	- 989 297	OUI	INFERIEUR	- 1 038 363	-635 731	61%
A6		2 633 636	1 498 592	3 768 681	NON	SUPERIEUR	- 433 174	3 066 810	-708%
A8	-	2 279 195	- 2 935 013	- 1 623 377	NON	INFERIEUR	- 702 216	-1 576 979	225%
A10		116 849	- 467 771	701 468	OUI	SUPERIEUR	- 286 457	403 306	-141%
A12	-	653 099	- 1 225 122	- 81 076	OUI	INFERIEUR	- 258 078	-395 021	153%
A14	-	970 289	- 1 539 382	- 401 196	NON	INFERIEUR	- 98 669	-871 620	883%
A16	-	522 274	- 1 107 173	62 625	OUI	INFERIEUR	- 301 495	-220 779	73%
A18	-	129 684	- 691 371	432 004	OUI	INFERIEUR	-	-129 684	
A20	-	232 140	- 781 711	317 431	OUI	SUPERIEUR	- 761 638	529 498	-70%
A22		103 131	- 483 893	690 155	OUI	SUPERIEUR	- 305 716	408 847	-134%
A24	-	766 996	- 1 325 877	- 208 115	NON	INFERIEUR	- 146 697	-620 299	423%
A26	-	729 718	- 1 338 784	- 120 652	OUI	SUPERIEUR	- 762 618	32 900	-4%
A28		99 186	- 450 826	649 197	OUI	SUPERIEUR	- 323 309	422 495	-131%
A30	-	77 223	- 623 364	468 918	OUI	SUPERIEUR	- 104 500	27 277	-26%
A32	-	332 435	- 881 468	216 599	OUI	INFERIEUR	- 232 260	-100 175	43%
A34	-	824 954	- 1 386 594	- 263 314	OUI	SUPERIEUR	- 904 895	79 941	-9%
A36	-	11 844	- 572 980	549 291	OUI	INFERIEUR	-	-11 844	
A38	-	348 958	- 902 014	204 098	OUI	INFERIEUR	- 112 882	-236 076	209%
A40	-	140 484	- 696 675	415 708	NON	SUPERIEUR	- 1 100 000	959 516	-87%
A42	-	451 505	- 1 008 807	105 797	OUI	INFERIEUR	- 231 375	-220 130	95%
A44		112 477	- 437 888	662 841	OUI	SUPERIEUR	- 285 000	397 477	-139%
A46	-	42 172	- 609 112	524 767	OUI	SUPERIEUR	- 90 000	47 828	-53%
A48	-	29 360	- 520 383	579 102	OUI	SUPERIEUR	- 5 600	34 960	-624%
A50	-	240 006	- 786 885	306 872	OUI	SUPERIEUR	- 324 470	84 464	-26%
A52	-	999 463	- 1 572 259	- 426 668	NON	INFERIEUR	-	-999 463	
A54	-	1 998 962	- 2 635 288	- 1 362 637	OUI	INFERIEUR	- 1 917 594	-81 368	4%
A56	-	1 294 236	- 1 880 339	- 708 133	NON	SUPERIEUR	- 2 863 107	1 568 871	-55%
A58	-	431 624	- 995 342	132 093	OUI	INFERIEUR	-	-431 624	
A60		130 978	- 422 373	684 329	OUI	SUPERIEUR	- 117 179	248 157	-212%
A62	-	211 961	- 768 085	344 162	OUI	INFERIEUR	-	-211 961	
A64	-	232 857	- 789 416	323 703	OUI	INFERIEUR	- 207 045	-25 812	12%
A66	-	62 356	- 612 201	487 490	OUI	INFERIEUR	- 2 118	-60 238	2844%
A68	-	1 802 284	- 2 428 434	- 1 176 134	NON	INFERIEUR	- 663 912	-1 138 372	171%
A70	-	162 818	- 715 920	390 284	OUI	INFERIEUR	- 52 677	-110 141	209%
A72		83 070	- 466 977	633 117	OUI	SUPERIEUR	-	83 070	
A74	-	141 739	- 694 984	411 507	OUI	INFERIEUR	-	-141 739	
A76	-	292 801	- 846 001	260 398	OUI	INFERIEUR	-	-292 801	
A78		17 023	- 535 086	569 132	OUI	SUPERIEUR	-	17 023	
A80	-	77 223	- 629 194	474 749	OUI	INFERIEUR	- 50 655	-26 568	52%
A82		177 358	- 377 044	731 759	OUI	SUPERIEUR	-	177 358	
A84	-	9 497	- 561 747	542 753	OUI	INFERIEUR	-	-9 497	
A86		37 612	- 515 766	590 989	OUI	SUPERIEUR	-	37 612	
A88		192 775	- 361 704	747 254	OUI	SUPERIEUR	-	192 775	
A90		158 603	- 395 497	712 704	OUI	SUPERIEUR	- 92	158 695	-172495%

Tableau 90: Données sociétés non vie - prédiction des PAP

Un extrait de ce tableau a été produit et a fait l'objet d'analyse dans le corps du mémoire (Tableau 23: Données sociétés non vie - prédiction des PAP - extrait)

#### 1.4) Modèle avec PM comme variable à expliquer

Le tableau suivant donne pour chaque compagnie vie ayant un numéro pair, l'estimation effectuée avec le modèle pour obtenir le montant des PM avec un intervalle de confiance à 95%.

Libellés	Fit	Lwr	Upr	Test1	Test2	PM	Ecart	En %
A2	-48 369 900	-52 530 828	-44 208 971	OUI	INFERIEUR	-48 092 231	-277 669	1%
A4	-37 581 496	-39 733 523	-35 429 469	NON	INFERIEUR	-32 085 330	-5 496 166	17%
A6	-19 980 823	-21 043 874	-18 917 772	OUI	INFERIEUR	-19 885 955	-94 868	0%
A8	-16 032 881	-17 783 196	-14 282 566	NON	INFERIEUR	-11 877 989	-4 154 892	35%
A10	-13 840 566	-15 702 088	-11 979 044	OUI	SUPERIEUR	-14 190 010	349 444	-2%
A12	-8 148 277	-9 092 545	-7 204 010	OUI	SUPERIEUR	-8 317 261	168 984	-2%
A14	-15 213 212	-17 581 568	-12 844 856	NON	INFERIEUR	-11 527 536	-3 685 676	32%
A16	-8 730 869	-10 154 471	-7 307 268	NON	INFERIEUR	-5 495 963	-3 234 907	59%
A18	-6 323 453	-7 274 338	-5 372 568	OUI	INFERIEUR	-6 200 872	-122 581	2%
A20	-7 624 333	-8 417 329	-6 831 337	OUI	SUPERIEUR	-8 137 911	513 578	-6%
A22	-7 049 539	-7 996 808	-6 102 270	NON	INFERIEUR	-5 562 564	-1 486 975	27%
A24	-3 673 785	-4 611 588	-2 735 982	OUI	SUPERIEUR	-4 162 999	489 214	-12%
A26	-7 155 891	-8 283 421	-6 028 360	NON	INFERIEUR	-4 404 668	-2 751 222	62%
A28	-4 990 257	-5 846 434	-4 134 081	OUI	INFERIEUR	-4 980 470	-9 787	0%
A30	-1 154 379	-1 946 970	-361 787	OUI	SUPERIEUR	-1 420 974	266 595	-19%
A32	-1 901 899	-2 695 948	-1 107 849	OUI	SUPERIEUR	-2 346 992	445 094	-19%
A34	-1 878 669	-2 727 419	-1 029 920	OUI	INFERIEUR	-1 128 982	-749 687	66%
A36	-1 036 280	-1 889 154	-183 405	OUI	INFERIEUR	-730 148	-306 132	42%
A38	-942 257	-1 784 056	-100 459	OUI	INFERIEUR	-439 241	-503 016	115%
A40	-653 731	-1 458 075	150 613	OUI	INFERIEUR	-352 903	-300 828	85%
A42	-140 113	-950 807	670 581	OUI	SUPERIEUR	-147 616	7 503	-5%
A44	16 173	-792 969	825 316	OUI	SUPERIEUR	-101 958	118 131	-116%

**Tableau 91: Données sociétés vie - prédiction des PM**

Un extrait de ce tableau a été produit et a fait l'objet d'analyse dans le corps du mémoire (Tableau 26: Données sociétés non vie - prédiction de PM - extrait )

#### 2) Analyse discriminante sur les données des sociétés non vie et vie

Cette annexe relative à l'analyse discriminante des sociétés non vie est composée des 2 points suivants :

- Modèles pour les sociétés d'assurance non vie,
- Modèles pour les sociétés d'assurance vie.

Du fait des sorties relativement longues du logiciel, plusieurs éléments ci-dessous ont été présentés brièvement dans la partie suivante du mémoire : 0

**Analyse discriminante page 94 et suivantes.** Dans cette annexe, ces éléments sont reproduits exhaustivement.

## 2.1) Modèles pour les sociétés d'assurance non vie

### **Modèle avec toutes les variables explicatives**

Le modèle prenant en compte toutes les variables explicatives donne les résultats suivants :

```

library(MASS)

donnees_analyse_discriminante_non_vie<-
read.table("donnees_analyse_discriminante_non_vie.csv", header=TRUE, sep=';',
dec='.', row.name=1)

modele_non_vie1 <-lda(STATUT ~ ., data=donnees_analyse_discriminante_non_vie)

modele_non_vie1
Call:
lda(STATUT ~ ., data = donnees_analyse_discriminante_non_vie)

Prior probabilities of groups:
  Faillite      Sain      Suivi
0.3043478 0.3043478 0.3913043

Group means:
          PRIMES      SIN_REG AR_BRUTS AR_NETS      PF_NETS      COM      FG
Faillite 1915710 -522528.6 1743008 1574887 -40.14286 -190044.0 -945064.9
Sain     8652195 -3704561.1 2978797 2831870 361094.14286 -1222979.3 -1973125.3
Suivi    4266061 -1600991.4 2225718 2159492 57592.44444 -622778.2 -1145794.0
          PRI_REA      SIN_REA      SAP_REA      COM_REA      PSAP      PAP      CAP
Faillite -101420.9 3216.429 159927.4 14646.0 -1598618 -442068.9 660793.1
Sain     -3186212.1 1213813.286 1747225.6 497615.1 -7921324 -590320.0 3456210.9
Suivi    -736416.6 128561.889 312487.6 156123.6 -2376681 -264669.1 936425.9
          T_BILAN      R_EXP      RHE      E_R      ACTIFS_A
Faillite 4292860 -6223.571 42858.29 -2028024 2162746
Sain     19354525 642187.857 -30993.86 -9553997 11354555
Suivi    6665309 351327.222 -113706.89 -3165475 2834822

Coefficients of linear discriminants:
          LD1      LD2
PRIMES 7.559773e-06 -4.380770e-06
SIN_REG -2.399760e-05 -9.639854e-07
AR_BRUTS 2.037457e-05 2.723350e-06
AR_NETS -1.401986e-05 -1.035819e-06
PF_NETS 6.911027e-05 -4.993773e-05
COM 9.231007e-05 -1.251428e-05
FG -2.706630e-05 -1.136737e-05
PRI_REA -4.370264e-05 -1.520821e-05
SIN_REA -3.178401e-05 -4.227863e-06
SAP_REA 3.411298e-05 3.641870e-06
COM_REA -1.778787e-04 -4.465498e-05
PSAP 1.134420e-06 1.654884e-06
PAP -2.026720e-05 -2.564823e-06
CAP 1.505347e-05 1.294635e-06
T_BILAN -2.196527e-05 -3.502334e-06
R_EXP 3.697927e-05 7.260948e-06
RHE -1.001813e-04 7.056691e-06
E_R -1.037900e-05 -1.784480e-06
ACTIFS_A 2.210181e-06 3.058657e-06

Proportion of trace:
          LD1      LD2
0.8806 0.1194

```

**Tableau 92: Données sociétés non vie - analyse discriminante - modèle complet**

Les sorties du logiciel fournissent les probabilités a priori du modèle, le centre de gravité de chacun des deux groupes et les coefficients de la variable canonique.

Le logiciel renvoie par convention les coefficients des variables canoniques de sorte que la variance intraclasse soit égale à 1.

### **Modèle avec 6 variables explicatives**

Plusieurs modèles peuvent être conçus et comparés en fonction notamment du taux de **mauvais classement**. Sur la base de ce critère, le modèle prenant en compte les 6 variables explicatives PF\_NETS+ COM+ FG+ PRI\_REA+COM\_REA+RHE, est meilleur que celui comprenant toutes les variables explicatives. Le choix de ces 6 variables vient du fait qu'elles présentent les coefficients discriminants les moins faibles. L'expérience montre également que les 6 variables font l'objet de moins d'erreurs de mesure au niveau des compagnies d'assurance. Du fait de l'ordre de grandeur des variables, les coefficients discriminants sont tous proches de 0.

Le modèle construit à partir de 6 variables explicatives sélectionnées se présente comme suit :

```
> modele_non_vie2 <- lda(STATUT ~ PF_NETS+ COM+ FG+ PRI_REA+COM_REA+RHE,
data=donnees_analyse_discriminante_non_vie)
> modele_non_vie2

Call:
lda(STATUT ~ PF_NETS + COM + FG + PRI_REA + COM_REA + RHE, data =
donnees_analyse_discriminante_non_vie)

Prior probabilities of groups:
  Faillite      Sain      Suivi
0.3043478 0.3043478 0.3913043

Group means:
      PF_NETS      COM      FG      PRI_REA      COM_REA      RHE
Faillite  -40.14286 -190044.0 -945064.9 -101420.9  14646.0  42858.29
Sain      361094.14286 -1222979.3 -1973125.3 -3186212.1 497615.1 -30993.86
Suivi     57592.44444  -622778.2 -1145794.0  -736416.6 156123.6 -113706.89

Coefficients of linear discriminants:
      LD1      LD2
PF_NETS  5.393783e-06  1.458827e-06
COM      -1.260151e-06 -2.423329e-06
FG        1.482703e-06  1.294582e-06
PRI_REA   1.728263e-06  1.867329e-06
COM_REA   1.494415e-05  8.633694e-06
RHE       2.628078e-06 -2.377660e-06

Proportion of trace:
      LD1      LD2
0.8519 0.1481
```

**Tableau 93: Données sociétés non vie - analyse discriminante - modèle à 6 variables**

Les sorties du logiciel fournissent les probabilités a priori du modèle, le centre de gravité de chacun des deux groupes et les coefficients des variables canoniques.

Le logiciel renvoie par convention les coefficients des variables canoniques de sorte que la variance intraclasse soit égale à 1.

### **Modèle avec toutes les variables explicatives**

La fonction `lda` permet d'estimer le taux de mauvais classement par validation croisée. Il suffit pour cela d'ajouter l'argument `CV=TRUE` lors de l'appel de la fonction. On obtient les labels prédits par le modèle avec la commande :

```
> prevision1 <- lda(STATUT ~ ., data=donnees_analyse_discriminante_non_vie,
CV=TRUE)$class
> table(prevision1, donnees_analyse_discriminante_non_vie$STATUT)

prevision1 Faillite Sain Suivi
Faillite   1     1     3
Sain       3     4     3
Suivi      3     2     3
```

**Tableau 94: Données sociétés non vie - analyse discriminante - classement modèle complet**

La commande suivante permet une estimation du taux de mauvais classement:

```
> sum(prevision1!= donnees_analyse_discriminante_non_vie$STATUT)/nrow
(donnees_analyse_discriminante_non_vie)
[1] 0.6521739
```

Du point de vue des erreurs de classement, le modèle avec toutes les variables explicatives se comporte comme un modèle aléatoire. Un tel modèle présenterait deux chances sur trois de faire un mauvais classement. Le modèle avec toutes les variables explicatives ne fait pas mieux qu'un modèle aléatoire. Il est donc de peu d'intérêt.

### **Modèle avec 6 variables explicatives**

Pour les 6 variables explicatives retenues, les labels prédits par le modèle et les taux de mauvais classement sont obtenus avec la commande suivante :

```
> prevision2 <- lda(STATUT ~ PF_NETS+ COM+ FG+ PRI_REA+COM_REA+RHE,
data=donnees_analyse_discriminante_non_vie, CV=TRUE)$class
> table(prevision2, donnees_analyse_discriminante_non_vie$STATUT)

prevision2 Faillite Sain Suivi
Faillite   5     0     2
Sain       0     4     1
Suivi      2     3     6
```

**Tableau 95: Données sociétés non vie - analyse discriminante - classement 6 variables**

On obtient une estimation du taux global de mauvais classement par :

```
> sum(prevision2!= donnees_analyse_discriminante_non_vie$STATUT)/nrow
(donnees_analyse_discriminante_non_vie)
[1] 0.3478261
```

Du point de vue des erreurs de classement, le modèle avec les 6 variables explicatives PF\_NETS+ COM+ FG+ PRI\_REA+COM\_REA+RHE donne de meilleurs résultats que le modèle avec toutes les variables explicatives.

La fonction `greedy.wilks()` du package `klAR` permet cependant d'identifier une combinaison de variables discriminant le mieux les individus. Cette fonction identifie la variable `PF_NETS` comme la variable la plus discriminante.

### ***Prédiction avec le modèle à une variable explicative***

Pour prédire le statut des autres sociétés n'ayant pas fait l'objet d'un contrôle, les données sont consignées dans un tableau ayant une structure identique à celui ayant permis de construire le modèle. La fonction `predict` permet d'affecter chaque individu dans un des groupes déterminés. Les résultats de la fonction sont les suivants :

```
> nouvelles_donnees<-read.table("nouvelles_donnees_analyse_discriminante.csv",
header=TRUE, sep=';', dec='.', row.name=1)
> predict(modele_non_vie3, newdata=nouvelles_donnees)
$class
 [1] Sain      Sain      Sain      Sain      Faillite suivi      Sain      Sain
 [9] Sain      Suivi     Suivi     Sain      Suivi      Sain      Suivi     Suivi
[17] Sain      Suivi     Suivi     Sain      Suivi      Suivi     Suivi     Sain
[25] Suivi     Suivi     Suivi     Suivi     Faillite suivi     Suivi     Suivi
[33] Suivi     Suivi     Suivi     Suivi     Suivi      Suivi     Suivi     Sain
[41] Sain      Suivi     Suivi     Suivi     Suivi      Suivi     Suivi     Suivi
[49] Suivi     Suivi     Sain      Suivi     Suivi      Suivi     Suivi     Suivi
[57] Faillite suivi     Suivi     Suivi     Suivi      Suivi     Suivi     Suivi
[65] Suivi     Suivi     Suivi
Levels: Faillite sain suivi

$posterior
      Faillite      Sain      Suivi
A1  2.138101e-03 0.9869893660 1.087253e-02
A3  2.046862e-03 0.9874710606 1.048208e-02
A6  1.426206e-08 0.9999994989 4.868483e-07
A7  2.031410e-02 0.9085124932 7.117341e-02
A8  5.016600e-01 0.0075712228 4.907688e-01
A9  3.693132e-01 0.0790414519 5.516453e-01
A10 2.255475e-11 0.9999999978 2.154992e-09
A13 1.028572e-04 0.9990466272 8.505156e-04
A14 5.349791e-02 0.7894624151 1.570397e-01
A15 3.352606e-01 0.1206544752 5.440849e-01
A17 4.183684e-01 0.0375764575 5.440552e-01
A19 4.238516e-03 0.9764708516 1.929063e-02
A20 3.067334e-01 0.1633506429 5.299159e-01
A21 1.657655e-02 0.9232751132 6.014834e-02
A22 3.227581e-01 0.1385191136 5.387228e-01
A24 4.170594e-01 0.0384109512 5.445297e-01
A25 1.517895e-03 0.9903252523 8.156852e-03
A26 3.431190e-01 0.1101244870 5.467565e-01
A27 3.195970e-01 0.1432480107 5.371550e-01
A28 9.240478e-02 0.6656776370 2.419176e-01
A29 4.415791e-01 0.0249896342 5.334312e-01
A30 3.010569e-01 0.1726509977 5.262921e-01
A31 4.437268e-01 0.0240231656 5.322500e-01
A32 1.471200e-01 0.5101381985 3.427418e-01
A33 3.984193e-01 0.0518611617 5.497196e-01
A34 2.779406e-01 0.2131230179 5.089364e-01
A36 3.603520e-01 0.0889714112 5.506766e-01
```

A38 4.698896e-01 0.0145513010 5.155591e-01  
A39 6.011725e-01 0.0007859683 3.980416e-01  
A40 4.321338e-01 0.0296240218 5.382422e-01  
A41 3.752056e-01 0.0729141265 5.518803e-01  
A43 4.526015e-01 0.0203518917 5.270466e-01  
A44 3.787034e-01 0.0694277824 5.518688e-01  
A46 3.739306e-01 0.0742129240 5.518565e-01  
A47 3.752234e-01 0.0728961235 5.518805e-01  
A48 4.932279e-01 0.0090419996 4.977301e-01  
A50 4.205428e-01 0.0362209680 5.432362e-01  
A51 2.008144e-01 0.3752975548 4.238881e-01  
A54 1.865765e-01 0.4093745776 4.040489e-01  
A55 4.377164e-01 0.0268079578 5.354756e-01  
A57 6.551722e-02 0.7498071780 1.846756e-01  
A58 4.452850e-01 0.0233414585 5.313735e-01  
A59 3.338160e-01 0.1226493361 5.435346e-01  
A60 3.944859e-01 0.0550857974 5.504283e-01  
A61 4.790985e-01 0.0120957041 5.088058e-01  
A62 4.471772e-01 0.0225352822 5.302876e-01  
A63 4.945451e-01 0.0087962985 4.966586e-01  
A64 4.263851e-01 0.0327658734 5.408491e-01  
A65 3.119546e-01 0.1550266733 5.330187e-01  
A66 3.838764e-01 0.0644772278 5.516463e-01  
A67 1.851213e-03 0.9885137943 9.634993e-03  
A68 2.496454e-01 0.2680089615 4.823457e-01  
A70 4.230425e-01 0.0347096848 5.422478e-01  
A73 3.774620e-01 0.0706522119 5.518858e-01  
A74 4.161780e-01 0.0389807666 5.448413e-01  
A75 4.324887e-01 0.0294382209 5.380731e-01  
A76 5.701001e-01 0.0016439408 4.282560e-01  
A77 3.075953e-01 0.1619613013 5.304434e-01  
A80 4.540407e-01 0.0198034164 5.261559e-01  
A81 4.302538e-01 0.0306239823 5.391222e-01  
A83 4.468464e-01 0.0226745290 5.304791e-01  
A84 4.436454e-01 0.0240592417 5.322954e-01  
A85 4.470761e-01 0.0225777600 5.303462e-01  
A86 4.657236e-01 0.0157990099 5.184774e-01  
A87 4.360066e-01 0.0276464616 5.363470e-01  
A88 4.467413e-01 0.0227189189 5.305398e-01  
A90 2.263876e-01 0.3172108877 4.564015e-01

\$x

LD1

A1 2.848281296  
A3 2.866428794  
A6 7.758145119  
A7 1.887681803  
A8 -1.402068667  
A9 -0.310749320  
A10 10.412235225  
A13 4.101979369  
A14 1.431390315  
A15 -0.096882030  
A17 -0.668080758  
A19 2.562268113  
A20 0.064393178  
A21 1.977988652  
A22 -0.024419884  
A24 -0.657752092  
A25 2.990655383  
A26 -0.143996891  
A27 -0.006554994  
A28 1.136293289  
A29 -0.858168567  
A30 0.094867786  
A31 -0.876396810  
A32 0.835389333  
A33 -0.515384734  
A34 0.214411150  
A36 -0.251939852  
A38 -1.106285306  
A39 -2.408719876

A40 -0.779260258  
 A41 -0.350469261  
 A43 -0.952795288  
 A44 -0.374450604  
 A46 -0.341802602  
 A47 -0.350590379  
 A48 -1.322036652  
 A50 -0.685333329  
 A51 0.581028293  
 A54 0.647057733  
 A55 -0.825648412  
 A57 1.326778096  
 A58 -0.889686135  
 A59 -0.088356676  
 A60 -0.486477930  
 A61 -1.190334398  
 A62 -0.905895747  
 A63 -1.334471423  
 A64 -0.732266514  
 A65 0.035923744  
 A66 -0.410476450  
 A67 2.908207740  
 A68 0.352896007  
 A70 -0.705311053  
 A73 -0.365905064  
 A74 -0.650821457  
 A75 -0.782187273  
 A76 -2.083202072  
 A77 0.059723410  
 A80 -0.965344448  
 A81 -0.763804268  
 A83 -0.903056205  
 A84 -0.875703747  
 A85 -0.905027735  
 A86 -1.068765674  
 A87 -0.811363229  
 A88 -0.902154550  
 A90 0.462500974

**Tableau 96: Données sociétés non vie - analyse discriminante - prédiction**

La fonction predict renvoie une liste de longueur 3 :

- Le 1<sup>er</sup> élément (class) contient les groupes prédits ;
- Le 2<sup>ème</sup> élément (posterior) est une matrice à 3 colonnes donnant les probabilités a posteriori d'appartenance aux groupes « faillite », « sain » et « suivi » pour chaque individu,
- Le troisième élément (x) désigne le score de chaque individu.

Les résultats obtenus ne présentent pas beaucoup d'intérêt du fait de la mauvaise qualité du modèle. L'intérêt principal réside dans la démarche mise en œuvre et la possibilité de la reproduire dans d'autres situations similaires. Un autre aspect porte sur la nécessité de mieux fiabiliser les informations avant la construction d'un tel modèle.

## 2.2) Modèles pour les sociétés d'assurance vie

### *Modèle avec toutes les variables explicatives*

Le modèle prenant en compte toutes les variables explicatives donne les résultats suivants :

```
> modele_non_vie1 <-lda(STATUT ~ ., data=donnees_analyse_discriminante_non_vie)
Message d'avis :
In lda.default(x, grouping, ...) : les variables sont collinéaires
> modele_non_vie1
Call:
lda(STATUT ~ ., data = donnees_analyse_discriminante_non_vie)

Prior probabilities of groups:
Faillite      Sain      Suivi
      0.2      0.4      0.4

Group means:
          Primes      PF      Sinistres      Com      FG      P_Réass
Faillite 2598950750  99652033 -1270577073  -68619142 -960450355 -106144083
Sain     6214592598 1017514690 -5378443585 -380957104 -916394364 -597062903
Suivi    4178375608 208035445  -3325182194 -332257065 -669499703 -75797874
          Charges_réa      RE      RN      S_.payés      S_réa      Com_réa
Faillite  70543876 363356007 162685268  -942691804 66279847 18012491
Sain     494991712 454241046 203359293 -3402572191 92899269 105956747
Suivi    42253980 25928197 43918860 -1273653735 4442867 25460597
          PM      A_bruts      A_nets      Kta1      FP      PT
Faillite -4934972415 1267214661 1267214661 1125000000 1042221534 -5256710345
Sain     -18272150242 464685057 405123161 581250000 1359778980 -21016041103
Suivi    -9281634378 541544607 532775084 500000000 312790528 -9824520347
          T_bilan      Actifs      ER      Actifs.
Faillite 7818699067 4184032951 -5317063029 4520487585
Sain     23675559410 21193401526 -21259341671 21448209085
Suivi    11212735228 10092031518 -10360736225 10126832768

Coefficients of linear discriminants:
          LD1      LD2
Primes    1.189591e-10 1.708156e-10
PF        -3.353892e-10 -1.671197e-10
Sinistres -1.243910e-10 -1.959543e-10
Com       -1.017679e-09 2.027474e-09
FG        3.902416e-10 6.463907e-10
P_Réass   7.820323e-10 8.471704e-10
Charges_réa -8.340445e-10 -3.044141e-10
RE        -1.062266e-09 4.656110e-10
RN        -1.374697e-09 1.924556e-09
S_.payés  6.332409e-11 8.654254e-11
S_réa     -2.638363e-09 3.408513e-09
Com_réa   3.609437e-09 -1.443515e-08
PM        -8.624154e-12 1.283721e-12
A_bruts   1.563382e-10 1.001815e-09
A_nets    1.314610e-10 6.338144e-11
Kta1     -3.933352e-09 -2.256314e-09
FP        5.288746e-10 -2.837401e-10
PT        -3.908591e-12 2.758208e-12
T_bilan   8.556237e-12 -1.981038e-12
Actifs    1.000637e-11 -3.681723e-12
ER        -7.405086e-12 7.327757e-13
Actifs.   9.311687e-12 -2.713398e-12

Proportion of trace:
          LD1      LD2
0.8014 0.1986
```

**Tableau 97: Données sociétés vie - analyse discriminante - modèle complet**

Pour le modèle comprenant toutes les variables explicatives, les labels prédits et les taux de mauvais classement sont obtenus avec les commandes suivantes :

```
> prevision1 <-lda(STATUT ~ ., data=donnees_analyse_discriminante_non_vie,
CV=TRUE)$class
Message d'avis :
In lda.default(x, grouping, ...) : les variables sont collinéaires
> table(prevision1, donnees_analyse_discriminante_non_vie$STATUT)

prevision1 Faillite Sain Suivi
  Faillite      1  2   2
    Sain       1  1   0
    Suivi      0  1   2
> sum(prevision1!= donnees_analyse_discriminante_non_vie$STATUT)/nrow
(donnees_analyse_discriminante_non_vie)
[1] 0.6
```

Tableau 98: Données sociétés vie - analyse discriminante - classement

Le taux de mauvais classement de 60% n'est pas satisfaisant. Du point de vue des erreurs de classement, le modèle avec les 2 variables explicative  $Kta1 + A\_nets$  donne de meilleurs résultats que le modèle avec toutes les variables explicatives.

### ***Prédiction avec le modèle sélectionné $Kta1 + A\_nets$***

Pour prédire le statut des autres sociétés n'ayant pas fait l'objet d'un contrôle, les données sont consignées dans un tableau ayant une structure identique à celui ayant permis de construire le modèle. La fonction `predict` permet d'affecter chaque individu dans un des groupes déterminés. Les résultats de la fonction sont les suivants :

```
nouvelles_donnees<-
read.table("nouvelles_donnees_analyse_discriminante_vie.csv", header=TRUE,
sep=';', dec='.', row.name=1)
> predict(modele_vie2, newdata=nouvelles_donnees)
$class
 [1] Faillite Faillite Faillite Sain      Suivi   Sain    Suivi   Faillite
Suivi   Sain    Sain    Sain    Sain    Sain    Sain    Faillite Sain
[17] Faillite Suivi   Sain    Sain    Faillite Sain    Suivi   Faillite Sain
Suivi   Sain    Sain    Sain    Sain    Sain    Sain    Sain    Sain
[33] Sain      Sain      Sain
Levels: Faillite Sain Suivi

$posterior
      Faillite      Sain      Suivi
A1  1.000000e+00 5.381765e-17 8.954865e-24
A3  9.956947e-01 4.255189e-03 5.013354e-05
A5  1.000000e+00 7.866305e-20 1.550538e-27
A6  1.194482e-03 6.760569e-01 3.227486e-01
A9  4.803091e-07 1.206573e-01 8.793422e-01
A11 2.780305e-04 5.455823e-01 4.541397e-01
A12 2.408208e-05 3.338503e-01 6.661257e-01
A13 1.000000e+00 2.970376e-08 8.322713e-12
A14 3.343185e-06 2.049620e-01 7.950347e-01
A16 1.551629e-03 6.978426e-01 3.006058e-01
A17 9.415280e-04 6.556491e-01 3.434094e-01
A18 3.657226e-01 4.401594e-01 1.941180e-01
A19 7.532243e-03 7.862172e-01 2.062505e-01
A20 1.384159e-03 6.884213e-01 3.101946e-01
A21 1.000000e+00 1.588292e-16 4.287758e-23
A22 9.699786e-04 6.582298e-01 3.408002e-01
A24 9.926893e-01 7.203215e-03 1.075341e-04
```

A25	2.758483e-04	4.932439e-01	5.064803e-01
A28	6.016991e-04	6.160425e-01	3.833558e-01
A29	1.000000e+00	9.965214e-09	1.069163e-12
A30	2.430076e-04	5.331987e-01	4.665583e-01
A31	1.121827e-04	4.628119e-01	5.370759e-01
A33	9.099507e-01	8.782534e-02	2.223961e-03
A34	1.551629e-03	6.978426e-01	3.006058e-01
A35	1.395114e-06	1.622225e-01	8.377761e-01
A36	5.559373e-04	5.445227e-01	4.549214e-01
A37	7.278971e-04	6.330509e-01	3.662212e-01
A38	2.652738e-04	5.412618e-01	4.584729e-01
A39	9.840312e-04	6.594740e-01	3.395420e-01
A40	6.629247e-04	6.247259e-01	3.746111e-01
A41	1.240332e-02	7.923164e-01	1.952803e-01
A42	4.097218e-04	5.811740e-01	4.184162e-01
A43	1.389215e-03	6.887242e-01	3.098866e-01
A44	4.293547e-04	5.854503e-01	4.141204e-01
A45	1.322196e-03	6.846032e-01	3.140746e-01

\$x

	LD1	LD2
A1	-13.3812269	-3.7713230
A3	-3.3417452	-0.8996461
A5	-15.3997488	-4.0632569
A6	0.2742304	-0.6808325
A9	2.3071872	1.0816835
A11	0.6888994	-0.3213262
A12	1.3409635	0.2439949
A13	-6.9853443	-1.1424086
A14	1.8367986	0.6738699
A16	0.1975622	-0.7473017
A17	0.3433731	-0.6208878
A18	-1.2382545	2.9653155
A19	-0.2578414	-0.8505103
A20	0.2311208	-0.7182073
A21	-13.0202831	-3.4583948
A22	0.3347544	-0.6283599
A24	-3.1651763	-0.7465659
A25	0.7086930	0.1624159
A28	0.4719236	-0.5094381
A29	-7.4448569	-2.4156335
A30	0.7261463	-0.2890342
A31	0.9367274	-0.1064664
A33	-2.4290809	-1.2748460
A34	0.1975622	-0.7473017
A35	2.0502817	0.8589538
A36	0.5206485	0.1160320
A37	0.4175043	-0.5566182
A38	0.7019126	-0.3100442
A39	0.3305870	-0.6319729
A40	0.4442727	-0.5334107
A41	-0.3898632	-0.6733559
A42	0.5806554	-0.4151707
A43	0.2300517	-0.7191342
A44	0.5674901	-0.4265847
A45	0.2445408	-0.7065726

**Tableau 99: Données sociétés vie - analyse discriminante - prédiction**

Les résultats obtenus ne présentent pas beaucoup d'intérêt du fait de la mauvaise qualité du modèle. L'intérêt principal réside dans la démarche mise en œuvre et la possibilité de la reproduire dans d'autres situations similaires. Un autre aspect porte sur la nécessité de mieux fiabiliser les informations avant la construction d'un tel modèle.

# INDEX DES TABLEAUX

Tableau 1: Forme générale d'un tableau de contingence	20
Tableau 2: Forme générale d'un tableau disjonctif complet	21
Tableau 3: Tableau d'analyse de la variance	25
Tableau 4: Synthèse des techniques d'exploration et d'analyse de données	38
Tableau 5: Données sociétés non vie - éloignement du centre de gravité	42
Tableau 6: Données sociétés non vie - six premières valeurs propres et pourcentage d'inertie	44
Tableau 7: Données sociétés non vie - indicateurs pour 10 sociétés sur les quatre premiers axes - extrait -	45
Tableau 8: Données sociétés non vie - coordonnées et qualité de représentation des variables - extrait-	49
Tableau 9: Données sociétés non vie - effectifs des classes résultant de la CAH	51
Tableau 10: Données sociétés non vie - description de la classe 1 par les variables initiales - extrait -	52
Tableau 11: Données sociétés vie - éloignement du centre de gravité	57
Tableau 12: Données sociétés vie - six premières valeurs propres et pourcentage d'inertie	58
Tableau 13: Données sociétés vie - coordonnées et qualité de représentation de 10 sociétés - extrait -	60
Tableau 14: Données sociétés vie - corrélation et qualité de représentation des variables - extrait -	62
Tableau 15: Données société vie - effectif des classes résultant de la CAH	64
Tableau 16: Données sociétés vie - description de la classe 1 par les variables initiales - extrait -	65
Tableau 17: Synthèse des travaux effectués dans le cadre de l'ACP et de la CAH	76
Tableau 18: Données sociétés non vie - régression avec PSAP comme variable à expliquer	79
Tableau 19: Données sociétés non vie - régression avec PAP comme variable à expliquer	82
Tableau 20: Données sociétés non vie - régression ajustée avec PSAP comme variable à expliquer	83
Tableau 21: Données sociétés non vie - régression ajustée avec PAP comme variable à expliquer	84
Tableau 22: Données sociétés non vie - prédiction des PSAP - extrait	87
Tableau 23: Données sociétés non vie - prédiction des PAP - extrait	89
Tableau 24: Données sociétés vie - régression avec PM comme variable à expliquer	90
Tableau 25: Données sociétés vie - régression ajustée avec PM comme variable à expliquer	91
Tableau 26: Données sociétés non vie - prédiction de PM - extrait	93
Tableau 27: Données sociétés non vie - analyse discriminante - sélection des variables	95
Tableau 28: Données sociétés non vie - analyse discriminante - modèle sélectionné	96
Tableau 29: Données sociétés non vie - analyse discriminante - classement	96
Tableau 30: Données sociétés vie - analyse discriminante - sélection des variables	98
Tableau 31: Données sociétés vie - analyse discriminante - modèle sélectionné	99
Tableau 32: Données sociétés vie - analyse discriminante - classement	99
Tableau 33: Données sociétés non vie - régression logistique - modèle brut	102
Tableau 34: Données sociétés non vie - régression logistique - modèle sélectionné	103
Tableau 35: Données sociétés non vie - prédiction d'anomalies	105
Tableau 36: Données sociétés vie - régression logistique - modèle brut	107
Tableau 37: Données sociétés vie - régression logistique - modèle sélectionné	107
Tableau 38: Données sociétés vie - prédiction d'anomalies	110
Tableau 39: Synthèse des travaux effectués dans le cadre de la mise en œuvre des méthodes prédictives	111
Tableau 40: Logiciels commerciaux	131
Tableau 41: Logiciels libres	133
Tableau 42: Variables relatives aux données des compagnies d'assurance non vie	135
Tableau 43: Variables relatives aux données des compagnies d'assurance vie	136
Tableau 44: Données sociétés non vie - tableau A	137
Tableau 45 : Données sociétés non vie - tableau B	138
Tableau 46: Données sociétés non vie - tableau C	139
Tableau 47: Données sociétés non vie - tableau D	140
Tableau 48: Données sociétés non vie - données résumées	141

Tableau 49: Données sociétés vie - tableau A	142
Tableau 50: Données sociétés vie - tableau B	143
Tableau 51: Données sociétés vie - données résumées	144
Tableau 52: Données sociétés non vie - ajustements et statuts	145
Tableau 53: Données sociétés vie - ajustements et statuts	146
Tableau 54: Données sociétés non vie - valeurs propres et pourcentage d'inertie	148
Tableau 55: Données sociétés non vie - coordonnées et qualité de représentation des sociétés - tableau A	149
Tableau 56: Données sociétés non vie - coordonnées et qualité de représentation des sociétés - tableau B	150
Tableau 57 : Données sociétés non vie - matrice de corrélation	151
Tableau 58: Données sociétés non vie - coordonnées et qualité de représentation des variables	152
Tableau 59: Données sociétés non vie - description de la classe 1 par les variables initiales	156
Tableau 60: Données sociétés non vie - description de la classe 2 par les variables initiales	157
Tableau 61: Données sociétés non vie - description de la classe 3 par les variables initiales	158
Tableau 62: Données sociétés non vie - description de la classe 1 par les composantes principales	159
Tableau 63: Données sociétés non vie - description de la classe 2 par les composantes principales	159
Tableau 64: Données sociétés non vie - description de la classe 3 par les composantes principales	160
Tableau 65: Données sociétés non vie - description des classes 1, 2 et 3 par les parangons	160
Tableau 66: Données sociétés non vie - description des classes 1, 2, et 3 par les individus spécifiques	161
Tableau 67: Données sociétés vie - valeurs propres et pourcentage d'inertie	162
Tableau 68: Données sociétés vie - coordonnées et qualité de représentation des sociétés	164
Tableau 69: Données sociétés vie - matrice de corrélation	165
Tableau 70: Données sociétés vie - corrélation et qualité de représentation des variables	166
Tableau 71: Données sociétés vie - description de la classe 1 par les variables initiales	170
Tableau 72: Données sociétés vie - description de la classe 2 par les variables initiales	171
Tableau 73: Données sociétés vie - description de la classe 3 par les variables initiales	171
Tableau 74: Données sociétés vie - description de la classe 4 par les variables initiales	172
Tableau 75: Données sociétés vie - description de la classe 5 par les variables initiales	173
Tableau 76: Données sociétés vie - description de la classe 6 par les variables initiales	174
Tableau 77: Données sociétés vie - description des classes 1 à 6 par les composantes principales	175
Tableau 78: Données sociétés vie - description des classes 1 à 6 par les parangons	176
Tableau 79: Données sociétés vie - description des classes 1 à 6 par les individus spécifiques	177
Tableau 80: Sociétés non vie du groupe G - valeurs propres et pourcentage d'inertie	178
Tableau 81: Données sociétés non vie groupe G - coordonnées et qualité de représentation des sociétés	179
Tableau 82: Données sociétés non vie groupe G - coordonnées et qualité de représentation des variables	180
Tableau 83: Données sociétés non vie groupe G - description des classes 1 à 3 par les parangons et les individus spécifiques	183
Tableau 84: Données sociétés vie groupe G - valeurs propres et pourcentage d'inertie	184
Tableau 85: Données sociétés vie groupe G - coordonnées et qualité de représentation des sociétés	185
Tableau 86: Données sociétés vie groupe G - corrélation et qualité de représentation des variables	186
Tableau 87: Données sociétés vie du groupe G - description de la classe 1 par les variables initiales	189
Tableau 88: Données société vie groupe G - description des classes 1 à 3 par les parangons et les individus spécifiques	190
Tableau 89: Données sociétés non vie - prédiction des PSAP	194
Tableau 90: Données sociétés non vie - prédiction des PAP	195
Tableau 91: Données sociétés vie - prédiction des PM	196
Tableau 92: Données sociétés non vie - analyse discriminante - modèle complet	197

<b>Tableau 93: Données sociétés non vie - analyse discriminante - modèle à 6 variables</b>	<b>198</b>
<b>Tableau 94: Données sociétés non vie - analyse discriminante - classement modèle complet</b>	<b>199</b>
<b>Tableau 95: Données sociétés non vie - analyse discriminante - classement 6 variables</b>	<b>199</b>
<b>Tableau 96: Données sociétés non vie - analyse discriminante - prédiction</b>	<b>202</b>
<b>Tableau 97: Données sociétés vie - analyse discriminante - modèle complet</b>	<b>203</b>
<b>Tableau 98: Données sociétés vie - analyse discriminante - classement</b>	<b>204</b>
<b>Tableau 99: Données sociétés vie - analyse discriminante - prédiction</b>	<b>205</b>

# INDEX DES FIGURES

Figure 1 : Données des sociétés non vie - nuage des individus	41
Figure 2: Données sociétés non vie - éboulis des valeurs propres	44
Figure 3: Données sociétés non vie - graphe des variables	48
Figure 4: Données sociétés vie - nuage des individus	56
Figure 5: Données sociétés vie - éboulis des valeurs propres	59
Figure 6: Données sociétés vie - graphe des variables	61
Figure 7: Données sociétés non vie groupe G - nuage des individus	68
Figure 8: Données sociétés non vie du groupe G - nuage des variables	69
Figure 9: Données sociétés non vie du groupe G - arbre hiérarchique	70
Figure 10: Données sociétés vie groupe G - nuage des individus	71
Figure 11: Données sociétés vie groupe G - nuage des variables	73
Figure 12: Données sociétés vie du groupe G - arbre hiérarchique	74
Figure 13: Données sociétés non vie - modèle PSAP - choix des variables avec BIC	83
Figure 14: Données sociétés non vie - modèle PAP - choix de variables avec BIC	84
Figure 15: Données sociétés non vie - modèle PSAP - représentation des résidus	85
Figure 16: Données sociétés non vie - modèle PAP - représentation des résidus	86
Figure 17 : Données sociétés vie - choix des variables avec BIC	91
Figure 18: Données sociétés vie - représentation des résidus	92
Figure 19: Données sociétés non vie - arbre hiérarchique	153
Figure 20: Données sociétés non vie - représentation de la partition sur le plan principal	154
Figure 21: Données sociétés non vie - représentation du dendrogramme en trois dimensions sur le plan principal	155
Figure 22: Données sociétés vie - arbre hiérarchique	167
Figure 23: Données sociétés vie - représentation de la partition sur le plan principal	168
Figure 24: Données sociétés vie - représentation du dendrogramme en trois dimensions sur le plan principal	169
Figure 25: Données sociétés non vie groupe G - éboulis des valeurs propres	178
Figure 26: Données sociétés non vie groupe G - représentation de la partition sur le plan principal	181
Figure 27: Données sociétés non vie du groupe G - représentation du dendrogramme en trois dimensions sur le plan principal	182
Figure 28: Données sociétés vie du groupe G - éboulis des valeurs propres	184
Figure 29: Données sociétés vie du groupe G - représentation de la partition sur le plan principal	187
Figure 30: Données sociétés vie groupe G - représentation du dendrogramme en trois dimensions dans le plan principal	188
Figure 31: Représentation des PSAP en fonction des ACTIFS_A et des SIN_REA	192
Figure 32: Représentation des PSAP en fonction des ACTIFS_A et des SIN_REA sans deux individus extrêmes	193

# SOMMAIRE

NOTE DE SYNTHÈSE	6
SIGLES ET ABBREVIATIONS	8
INTRODUCTION	9
<b>PREMIÈRE PARTIE : PRÉSENTATION DES MÉTHODES, DES OUTILS STATISTIQUES ET DES DONNÉES UTILISÉES</b>	<b>12</b>
CHAPITRE I. STATISTIQUE EXPLORATOIRE ET ANALYSE DE DONNÉES	13
CHAPITRE II. MODÈLES PRÉDICTIFS	22
CHAPITRE III. PRÉSENTATION DE QUELQUES LOGICIELS DE TRAITEMENTS STATISTIQUES ET DES DONNÉES UTILISÉES	32
<b>DEUXIÈME PARTIE: ANALYSE DES DONNÉES COMPTABLES ET FINANCIÈRES DES COMPAGNIES D'ASSURANCE</b>	<b>39</b>
CHAPITRE I. RÉALISATION ET INTERPRÉTATION D'UNE ANALYSE EN COMPOSANTES PRINCIPALES (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE(CAH) SUR LES DONNÉES DES COMPAGNIES D'ASSURANCE NON VIE DU MARCHÉ	40
CHAPITRE II. RÉALISATION ET INTERPRÉTATION D'UNE ANALYSE EN COMPOSANTES PRINCIPALES (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE(CAH) SUR LES DONNÉES DES COMPAGNIES D'ASSURANCE VIE DU MARCHÉ	55
CHAPITRE III. RÉALISATION ET INTERPRÉTATION D'UNE ANALYSE EN COMPOSANTES PRINCIPALES (ACP) ET D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE(CAH) SUR LES DONNÉES DES COMPAGNIES D'ASSURANCE DU GROUPE G	67
<b>TROISIÈME PARTIE : MISE EN ŒUVRE DES MODÈLES PRÉDICTIFS SUR LES DONNÉES COMPTABLES ET FINANCIÈRES DES COMPAGNIES D'ASSURANCE DE L'ESPACE CIMA</b>	<b>77</b>
CHAPITRE I. RÉGRESSION LINÉAIRE MULTIPLE	78
CHAPITRE II. ANALYSE DISCRIMINANTE	94
CHAPITRE III. RÉGRESSION LOGISTIQUE	101
CONCLUSION	112
BIBLIOGRAPHIE	115
ANNEXES	120
I. ANNEXES RELATIVES A LA PRÉSENTATION DES MÉTHODES, DES OUTILS STATISTIQUES ET DES DONNÉES	121
II. ANNEXES RELATIVES A L'ANALYSE DES DONNÉES COMPTABLES ET FINANCIÈRES DES COMPAGNIES D'ASSURANCE	147
III. ANNEXES RELATIVES A LA MISE EN ŒUVRE DES MODÈLES PRÉDICTIFS SUR LES DONNÉES COMPTABLES ET FINANCIÈRES DES COMPAGNIES D'ASSURANCE	191
INDEX DES TABLEAUX	206
INDEX DES FIGURES	209
SOMMAIRE	210