



## **COURS DE FOUILLE DE DONNÉES**

M1, P-20- IFI 2016

**Enseignante:** NGUYỄN Thị Minh Huyền

**Rapport TP3 – TP4 – TP5**

**Données Utilisées:** Census Income Data Set

**Étudiant:** Ginel Dorleon

**Gervais Fotsing Sikadie S.**

# Table Des Matières

Rapport TP3 – TP4 – TP5

Données Utilisées: Census Income Data Set

I.	INTRODUCTION GÉNÉRALE.....	3
II.	CONTEXTE.....	3
III.	DESCRIPTION DES VARIABLES.....	4
IV.	STATISTIQUES SUR LES VARIABLES.....	6
	1- Étude d'une variable qualitative- Education.....	6
	2 - Étude d'une variable quantitative – L'Age.....	7
	3 - Corrélation entre chaque paire de variable quantitative.....	8
	4 - Contingence des variables qualitatives.....	10
	5 -Test du Khi 2.....	11
V.	ANALYSE EN COMPOSANTES PRINCIPALE (ACP).....	13
	1- Valeurs Propres.....	13
	2 - Corrélation Entre Les Variables Et Les Axes Principaux.....	15
	3 - Cercle Des Corrélations.....	16
VI.	ANALYSE FACTORIELLE DES CORRESPONDANCES.....	18
	1 - Paramétrage de l'analyse des correspondances.....	18
	2 - KHI-2 (global) de l'écart à l'indépendance.....	18
	3 - Valeurs Propres.....	19
	4 - Choix du nombre d'axes – Règle de Kaiser.....	19
	5 - Représentation des lignes.....	20
	6 - Représentation des colonnes.....	21
VII.	CLUSTERING.....	23
	1 - Introduction.....	23
	2 - Objectif.....	23
	3 - Algorithme de classification utilisé: K-means et Hac.....	23
	4 - Coordonnées des centres des clusters.....	24
	5 - Positionnement des classes dans le plan factoriel.....	26
	6 - Caractérisation des classes – Variables actives et illustratives.....	27
	7 - Interprétation.....	28
VIII.	CONCLUSION.....	29
IX.	RÉFÉRENCES.....	29

## I.INTRODUCTION GÉNÉRALE

De nos jours toutes les entreprises collectent et stockent de grandes quantités de données. Ces mégabases de données, qui ne cessent d'augmenter jour après jour, sont peu exploitées, alors qu'elles cachent de connaissances décisives face au marché et à la concurrence. Pour combler ce besoin, une nouvelle industrie est née : la Fouille de Données ou la Science des Données.

La Fouille de Données a ainsi pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances. [1]

Pour appliquer ainsi les connaissances acquises durant l'introduction du cours de fouille de données, nous allons ainsi mener une étude sur un jeu de données. Le présent rapport est divisé en 3 parties:

Analyse Statistique des Variables - Analyse Factorielle de Correspondance – Clustering.

Pour faire ce travail, nous allons utilisé le logiciel spécialisé, tanagra

## II.CONTEXTE

Jeu de Données: **Census Income Data Set** [2]

L'étude porte sur les rémunérations des salaires aux USA en 2001. Les données que nous utilisons ont été extraites de la base de données d'un bureau de recensement aux USA. 32561 instances avec 15 attributs . L'objectif est d'étudier un ensemble de paramètre et d'établir une tache de prévision afin de déterminer le profil des personnes qui gagnent plus ou moins de 50K par année.

### III.DESCRPTION DES VARIABLES

*Age*: Variable continue représentant l'âge des personnes.

*Workclass*: Variable discrète prenant les valeurs suivantes : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

*Fnlwgt*: Final weight, une variable continue calculée à partir des données d'origine et du sexe de la personne.

*Education*: Variable discrète déterminant le niveau d'éducation de la personne, elle prend les valeurs suivantes : Licence, Certains collège, 11e, HS-grad, Prof-école, Assoc-ACDM, Assoc-voc, 9e, septième-huitième, 12e, Masters, 1ere-4ème, 10e, Doctorat, 5ème-6ème, Maternelle.

*education-num*: continu.

*statut matrimonial*: Marié-civ-conjoint, Divorcé, jamais marié, Séparé, Veuf, Marié-conjoint absent, Marié-AF-conjoint.

*Profession*: Tech-support, Craft-réparation, Autre service, ventes, Exec-gestion, Prof spécialité, Handlers-nettoyants, machine-op-inspct, Adm-clérical, agriculture-pêche, Transport-mobile, Priv-ménage serv, protection-serv, armées-Forces.

*Relationship* : Variable discrète représentant les types de relations des individus. Elle prend les valeurs suivantes : Femme, propre enfant, Mari, Non-in-famille, Autre-parent, Unmarried.

*Race*: Variable discrète représentant la race des individus. Elle prend les valeurs Noir Blanc, Asie-Pac-Islander, Amer-Indian-Eskimo, Autre.

*Sexe*: Variable discrète représentant le sexe de l'individu, 2 valeurs : Femme, Homme.

*Capital-gain*: Variable continue représentant le capital gagné par l'individu.

*Capital-loss*: Variable continue représentant le capital perdu par l'individu.

*Hours-per-week* : Variable continue représentant le nombre d'heur que l'individu travail

*Pays*: Variable discrète représentant le pays d'origine de l'individu.

*Salary*: Notre variable cible  $\leq 50K$ ,  $> 50 K$

# STATISTIQUE SUR LES VARIABLES

## IV. STATISTIQUES SUR LES VARIABLES

Dans la plupart des études sur les données, le nombre de sujets est souvent trop important pour que l'on puisse présenter les données réelles de chaque individu. C'est pourquoi, il est nécessaire de trouver un moyen qui donne le maximum d'informations possible sous le format le plus utile. Une manière courante de présentation de données est la représentation graphique ou le tableau. Les tableaux sont commodes pour présenter l'information relative aux données individuelles et les graphiques pour donner un profil général des observations. Toutefois, il est également utile de donner un résumé chiffré. Pour les variables qualitative ou en catégorie (niveau d'étude, sexe, absence-présence d'une maladie, niveau d'éducation, etc.), la mesure la plus instructive est la proportion d'individus entrant dans chaque catégorie. Les variables quantitatives (poids, taille, âge, salaire, etc.) nécessitent quant à elle deux types de mesure pour avoir une idée complète de la distribution des observations : la mesure de la *position centrale* des observations et la mesure de leur *dispersion*, c'est-à-dire la mesure de la répartition des observations autour de cette position centrale. Ainsi, nous allons faire procéder à l'étude qualitative et quantitative de deux des variables de notre jeu de données.

### 1- Étude d'une variable qualitative- Education

Généralement, en statistique, une variable qualitative, on dit aussi catégorielle, est une variable pour laquelle la valeur mesurée sur chaque individu ne représente pas une quantité. Les différentes valeurs que peut prendre cette variable sont appelées les *catégories*, *modalités* ou *niveaux*.

Ainsi, parmi l'ensemble des variables qualitatives de notre jeu de données, nous avons choisi d'étudier la variable *Education*. En effectuant la statistique primaire de cette variable, on peut constater que la plus grande valeur valeur Hs-grad signifiant le degré en High School. On constate que le mode est cette valeur, Hs-grad. Le nombre d'individu ayant ce niveau est de 10501 et le pourcentage d'individus ayant ce niveau est de 32.25 %. Voir capture ci-dessous

Univariate discrete stat 1				
Parameters				
Attributes : 1				
Examples : 32561				
Results				
Attribute	Gini	Distribution		
		Values	Count	Percent
EDUCATION	0,8096	Bachelors	5355	16,45 %
		HS-grad	10501	32,25 %
		11th	1175	3,61 %
		Masters	1723	5,29 %
		9th	514	1,58 %
		Some-college	7291	22,39 %
		Assoc-acdm	1067	3,28 %
		Assoc-voc	1382	4,24 %
		7th-8th	646	1,98 %
		Doctorate	413	1,27 %
		Prof-school	576	1,77 %
		5th-6th	333	1,02 %
		10th	933	2,87 %
		1st-4th	168	0,52 %

## 2 - Étude d'une variable quantitative – L'Age

Une variable est *quantitative* si elle reflète une notion de grandeur, c'est-à-dire si les valeurs qu'elle peut prendre sont des nombres. Une grandeur quantitative est souvent exprimée avec une unité de mesure qui sert de référence

Parmi l'ensemble des variables quantitatives de notre jeu de données, nous avons choisi d'étudier la variable quantitative Age. En effectuant l'analyse statistique primaire de cette variable, on constate que l'âge minimum est de 17 ans, le max est de 90 ans et l'âge moyenne est 38,5816. Voir capture ci-dessous.

Univariate continuous stat 1					
Parameters					
Attributes : 1					
Examples : 32561					
Results					
Attribute	Min	Max	Average	Std-dev	Std-dev/avg
AGE	17	90	38,5816	13,6404	0,3535
Computation time : 0 ms.					
Created at 21/12/2016 15:14:43					

More Univariate cont stat 1					
Parameters					
Attributes : 1					
Examples : 32561					
Results					
Attribute	Stats		Histogram		
AGE	Statistics		Values	Count	Percent
	Average	38,5816	x_<_24,3000	5570	17,11%
	Median	37,0000	24,3000_=<_x_<_31,6000	5890	18,09%
	Std dev. [Coef of variation]	13,6404 [0,3535]	31,6000_=<_x_<_38,9000	6048	18,57%
	MAD [MAD/STDDEV]	11,1892 [0,8203]	38,9000_=<_x_<_46,2000	6163	18,93%
	Min * Max [Full range]	17,00 * 90,00 [73,00]	46,2000_=<_x_<_53,5000	3967	12,18%
	1st * 3rd quartile [Range]	28,00 * 48,00 [20,00]	53,5000_=<_x_<_60,8000	2591	7,96%
	Skewness (std-dev)	0,5587 (0,0136)	60,8000_=<_x_<_68,1000	1595	4,90%
	Kurtosis (std-dev)	-0,1661 (0,0271)	68,1000_=<_x_<_75,4000	496	1,52%
			75,4000_=<_x_<_82,7000	174	0,53%
			x>= 82,7000	67	0,21%

On peut ensuite observer les informations suivantes :

- La médiane est de 37
- La classe modale est [38,9000 ; 46,2000[, le mode est 42.5500
- L'écart-type est de 13,6404

On peut ensuite constater la dispersion de la variable Age.

L'étendue (différence entre la valeur max et la valeur min) est de 73,00

la dispersion interquartile (différence entre le troisième et le premier quartile) est de 20,00;

La représentation graphique ici est sous forme d'histogramme.

La forme de la distribution de l'Age est, suite aux statistiques primaires, étalée à droite car la moyenne est > médiane > mode (on peut aussi le constater en observant que le coefficient d'asymétrie de Fisher (Skewness = 0,5587) est positif.

La forme de la distribution de l'Age est moins aplatie qu'une distribution normale car coefficient d'aplatissement observé ici (Kurtosis = -01661) est négatif.

### **3 - Corrélation entre chaque paire de variable quantitative**

En probabilité et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques *numériques*, c'est étudier l'intensité de la liaison qui peut exister entre ces deux variables X et Y. La mesure de la corrélation linéaire entre les deux se fait alors par le calcul du coefficient de corrélation linéaire, noté r. Cette corrélation est dite:

- Positive, c'est-à-dire à toute augmentation au niveau de X correspond une augmentation au niveau de Y. Les deux variables varient dans le même sens et avec une intensité similaire. Exemple: EDUCATION-NUM et HOURS-PER-WEEK.

- Négative, c'est-à-dire à toute augmentation au niveau de X correspond une diminution au niveau de Y. Les deux variables varient dans deux sens opposés et avec une intensité similaire. Exemple: EDUCATION-NUM et FNLWGT.

Le coefficient:

Le coefficient de corrélation est un indice statistique qui exprime l'intensité et le sens (positif ou négatif) de la relation linéaire entre deux variables quantitatives. C'est une mesure de la liaison linéaire, c'est à dire de la capacité de prédire une variable X par une autre Y à l'aide d'un modèle linéaire. Il permet de mesurer l'intensité de la liaison entre deux caractères quantitatifs. C'est donc un paramètre important dans l'analyse des régressions linéaires (simples ou multiples). En revanche, ce coefficient est nul ( $r = 0$ ) lorsqu'il n'y a pas de relation linéaire entre les variables (ce qui n'exclut pas l'existence d'une relation autre que linéaire). Par ailleurs, le coefficient est de signe positif si la relation est positive (directe, croissante) et de signe négatif si la relation est négative (inverse, décroissante). Ce coefficient varie entre -1 et +1 ; l'intensité de la relation linéaire sera donc d'autant plus forte que la valeur du coefficient est proche de +1 ou de -1, et d'autant plus faible qu'elle est proche de 0. •



Une valeur proche de +1 montre une forte liaison entre les deux caractères, la relation linéaire est ici croissante (c'est-à-dire que les variables varient dans le même sens).•

Une valeur proche de -1 montre également une forte liaison mais la relation linéaire entre les deux caractères est décroissante (les variables varient dans le sens contraire).

- Une valeur proche de 0 montre une absence de relation linéaire entre les deux caractères.

Cependant, il faut noter que l'existence d'une corrélation élevée entre deux variables X et Y ne conduit pas à l'existence d'une relation de cause à effet. On utilise la connaissance de la variable X pour prédire les valeurs de Y ; cela n'implique pas qu'un changement de X cause un changement de Y.

Dans notre cas, le coefficient de corrélation de nos variables se situe dans la majorité des cas entre 0 et 0.2 donc c'est très faible. On en conclut qu'il n'existe pratiquement aucune comparaison entre les paires de variables présentés.

Dans le tableau suivant, nous présentons la corrélation entre chaque paire de nos variables quantitatives.

Dans le tableau ci-dessus, r représente le coefficient de corrélation pour chaque paire de variable,  $r^2$  est l'écart-type

Linear correlation 2					
Parameters					
Cross-tab parameters					
Sort results	non				
Input list	Target (Y) and input (X)				
Results					
Y	X	r	r <sup>2</sup>	t	Pr(> t )
AGE	FNLWGT	-0,0766	0,0059	-13,8709	0,0000
AGE	EDUCATION-NUM	0,0365	0,0013	6,5954	0,0000
AGE	CAPITAL-GAIN	0,0777	0,0060	14,0581	0,0000
AGE	CAPITAL-LOSS	0,0578	0,0033	10,4423	0,0000
AGE	HOURS-PER-WEEK	0,0688	0,0047	12,4358	0,0000
FNLWGT	AGE	-0,0766	0,0059	-13,8709	0,0000
FNLWGT	EDUCATION-NUM	-0,0432	0,0019	-7,8014	0,0000
FNLWGT	CAPITAL-GAIN	0,0004	0,0000	0,0779	0,9379
FNLWGT	CAPITAL-LOSS	-0,0103	0,0001	-1,8499	0,0643
FNLWGT	HOURS-PER-WEEK	-0,0188	0,0004	-3,3872	0,0007
EDUCATION-NUM	AGE	0,0365	0,0013	6,5954	0,0000
EDUCATION-NUM	FNLWGT	-0,0432	0,0019	-7,8014	0,0000
EDUCATION-NUM	CAPITAL-GAIN	0,1226	0,0150	22,2958	0,0000
EDUCATION-NUM	CAPITAL-LOSS	0,0799	0,0064	14,4677	0,0000
EDUCATION-NUM	HOURS-PER-WEEK	0,1481	0,0219	27,0256	0,0000

#### 4 - Contingence des variables qualitatives

Pour évaluer le lien existant entre deux de nos variables qualitatives, on va présenter un tableau de contingence. Le tableau de contingence est une méthode de représentation de données issues d'un comptage permettant d'estimer la dépendance entre deux caractères. Elle consiste à croiser deux caractères d'une population, par exemple dans notre cas, *workclass* et *education*, en dénombrant l'effectif correspondant à la conjonction «caractère 1» et «caractère 2». Les effectifs partiels sont rassemblés dans un tableau à double entrée, par ligne pour le premier caractère, et par colonne en fonction du second caractère: c'est le «tableau de contingence».

Cet outil simple répond à un problème crucial en statistique: la détection d'éventuelles dépendances entre les qualités relevées sur les individus d'une population.

Ainsi, nous présentons le tableau suivant avec la variable *Workclass* d'une part et le niveau d'éducation (variable *education*) de l'individu.

Contingency Chi-Square 2												
Parameters												
Cross-tab parameters												
Sort results	non											
Input list	Target (Row) and input (Column)											
Additional information	0											
Contribution threshold	2,0											
Results												
Row (Y)	Column (X)	Statistical indicator										
WORKCLASS	EDUCATION	Stat	Value		Bachelors	HS-grad	11th	Masters	9th	Some-college	Assoc-acdm	Ass
		d.f.	120	State-gov	270	268	14	169	6	325	41	
		Tschuprow's t	0,084918	Self-emp-not-inc	399	866	60	124	34	486	71	
		Cramer's v	0,099369	Private	3551	7780	923	894	387	5094	729	
		Phi²	0,078993	Federal-gov	212	263	9	67	3	254	55	
		Chi² (p-value)	2572,10 (0,0000)	Local-gov	477	503	36	342	23	387	88	
		Lambda	0,000000	?	173	532	118	48	51	514	47	
		Tau (p-value)	0,0177 (0,0000)	Self-emp-inc	273	279	14	79	10	226	35	
		U(R/C) (p-value)	0,0286 (0,0000)	Without-pay	0	9	0	0	0	3	1	
				Never-worked	0	1	1	0	0	2	0	
				Sum	5355	10501	1175	1723	514	7291	1067	

On peut constater que, le secteur prive est le secteur qui offre le plus grand nombre d'emploi.

Les individus avec le niveau de Hs-grad représentent le plus grande valeur des employés. Il n'y a pas de Bacheliers, ni de niveau Masters, ni de niveau de 9th sans emplois ou qui ne sont jamais été employés .

## 5 -Test du Khi 2

Le *Test du Khi 2* est un test statistique permettant de tester l'adéquation d'une série de données à une famille de lois de probabilités ou de tester l'indépendance entre deux variables aléatoires.

Results											
Row (Y)	Column (X)	Statistical indicator									
WORKCLASS	EDUCATION	Stat	Value		Bachelors	HS-grad	11th	Masters	9th	Some-college	Assoc-acdm
		d.f.	120	State-gov	270 (+ 1 %)	268 (- 2 %)	14 (- 1 %)	169 (+ 6 %)	6 (- 0 %)	325 (+ 0 %)	41 (- 0 %)
		Tschuprow's t	0,084918	Self-emp-not-inc	399 (- 0 %)	866 (+ 0 %)	60 (- 0 %)	124 (- 0 %)	34 (- 0 %)	486 (- 0 %)	71 (- 0 %)
		Cramer's v	0,099369	Private	3551 (- 0 %)	7780 (+ 1 %)	923 (+ 1 %)	894 (- 3 %)	387 (+ 0 %)	5094 (+ 0 %)	729 (- 0 %)
		Phi²	0,078993	Federal-gov	212 (+ 1 %)	263 (- 0 %)	9 (- 1 %)	67 (+ 0 %)	3 (- 0 %)	254 (+ 0 %)	55 (+ 1 %)
		Chi² (p-value)	2572,10 (0,0000)	Local-gov	477 (+ 2 %)	503 (- 2 %)	36 (- 1 %)	342 (+ 19 %)	23 (- 0 %)	387 (- 1 %)	88 (+ 0 %)
		Lambda	0,000000	?	173 (- 2 %)	532 (- 0 %)	118 (+ 2 %)	48 (- 1 %)	51 (+ 1 %)	514 (+ 1 %)	47 (- 0 %)
		Tau (p-value)	0,0177 (0,0000)	Self-emp-inc	273 (+ 2 %)	279 (- 1 %)	14 (- 1 %)	79 (+ 0 %)	10 (- 0 %)	226 (- 0 %)	35 (- 0 %)
		U(R/C) (p-value)	0,0286 (0,0000)	Without-pay	0 (- 0 %)	9 (+ 0 %)	0 (- 0 %)	0 (- 0 %)	0 (- 0 %)	3 (- 0 %)	1 (+ 0 %)
				Never-worked	0 (- 0 %)	1 (- 0 %)	1 (+ 0 %)	0 (- 0 %)	0 (- 0 %)	2 (+ 0 %)	0 (- 0 %)
				Sum	5355	10501	1175	1723	514	7291	1067

Il caractérise le degré de liaison entre les variables *Workclass* et *Education* . Il quantifie l'écart entre le tableau construit sur les données et le tableau que l'on aurait obtenu si l'hypothèse d'indépendance entre les variables était vraie. Pour nos deux variables, le *Test du Khi 2* est égal déjà 2572,10.

Observons la p-value de la statistique du KHI-2 de la figure ci-dessus, elle permet de déterminer s'il y a lieu de rejeter ou non l'hypothèse d'indépendance (H0), si elle est inférieure au niveau de signification (par défaut 5% sur Tanagra). Dans notre cas elle est de 0 % donc il ne faut pas rejeter H0. Le CHI-2 étant un critère additif, il serait intéressant pour nous de savoir quelles sont les cases qui y ont le plus contribué.

Les contributions sont indiquées en pourcentage du total. Le signe apparu dans les cases permet de déterminer s'il s'agit d'une attraction ou une répulsion entre les caractéristiques des variables étudiées. Lorsque la contribution d'une case est 2 fois plus élevée que la contribution moyenne, elle est surlignée en rouge, d'où la présence de la couleur rouge dans notre tableau. On peut déduire que la liaison entre la variable secteur de travail(workclass) et Education repose avant tout sur une forte attraction évaluée à 19 % entre la valeur Federal-gov et le niveau d'éducation Masters.

# **APPLICATION DES MÉTHODES FACTORIELLES AUX JEUX DE DONNÉES**

On rappelle que l'idée de notre jeu de données c'est de baser sur des indicateurs afin de déterminer la catégorie d'individus qui gagnent plus ou moins de 50K par année aux USA.

## V. ANALYSE EN COMPOSANTES PRINCIPALE (ACP)

L'analyse en composante principale est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

Il s'agit d'une approche à la fois géométrique (les variables étant représentées dans un nouvel espace, selon des directions d'inertie maximale) et statistique (la recherche portant sur des axes indépendants expliquant au mieux la variabilité — la variance — des données).

Ici, notre objectif de cette analyse en composante principale est de partir d'un ensemble de données contenant 32561 observations et 15 variables continues et discrètes, pour chercher à résumer l'information disponible à l'aide de variables synthétiques appelées composantes principales. Grâce à la fonction «Principal Component Analysis» située sous l'onglet «Factorial Analysis», nous avons réalisé l'ACP. Les résultats obtenus sont présentés dans la suite.

### 1- Valeurs Propres

Le tableau contenant les valeurs propres de la matrice des corrélations nous est présenté dans le tableau ci-dessous. Le tableau contient des valeurs propres « Eigenvalues », avec le pourcentage d'inertie expliquée (individuelle et cumulée) par les axes. Un histogramme permet de situer leur décroissance.

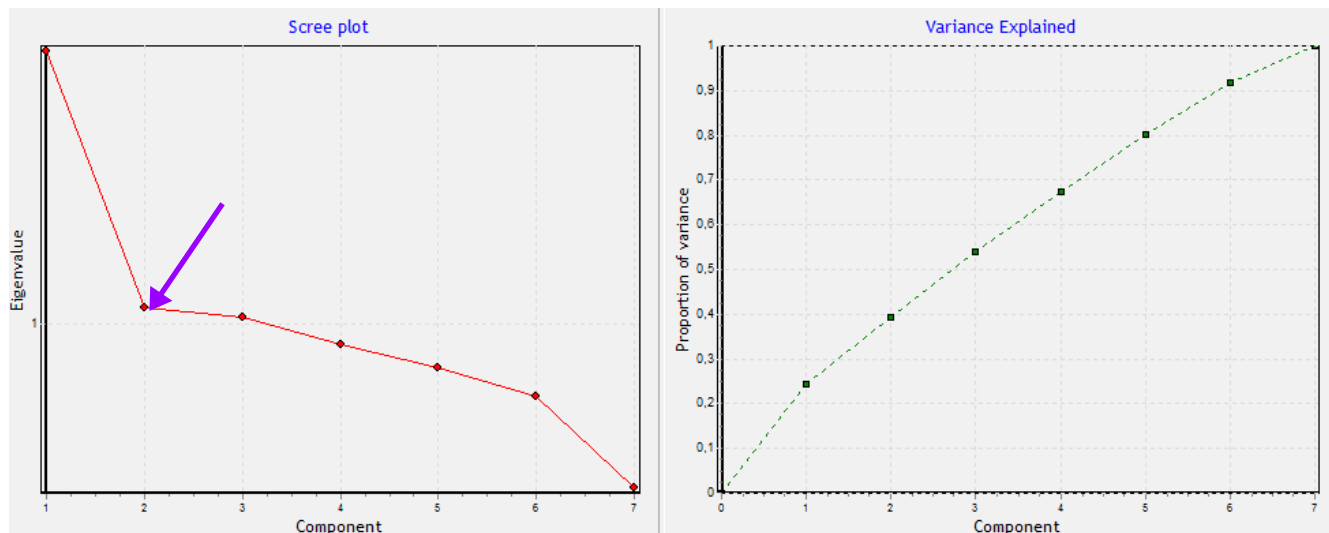
## Eigen values

Matrix trace	7,000000
Average	1,000000

Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	1,711420	0,668368	24,45 %		24,45 %
2	1,043051	0,022248	14,90 %		39,35 %
3	1,020804	0,071918	14,58 %		53,93 %
4	0,948885	0,060728	13,56 %		67,49 %
5	0,888158	0,075683	12,69 %		80,18 %
6	0,812475	0,237267	11,61 %		91,78 %
7	0,575207	-	8,22 %		100,00 %
Tot.	7,000000	-	-	-	-

En étudiant le tableau précédent, on peut constater que l'inertie expliquée par le premier axe principal est évalué à  $\Delta=1.71$ . Un autre constat nous permet de constater que la première composante principale occupe à elle seule 24,45 %.

Pour compléter ce résultat, un graphique Scree plot est proposé dans le second onglet de la fenêtre de visualisation.



Il y a manifestement un « coude » au niveau de la 2ème composante, mais cependant la valeur propre associée n'est relativement pas faible. Le graphique des inerties expliquées cumulées « Variance explained » confirme cette idée. Le gain d'inertie en passant du 1er facteur au 3ème semble décisif.

Se basant sur le critère de Kaiser-Guttman utilisé par tanagra pour la définition des composantes principales les plus significatives, on observe que seules les trois premières valeurs propres sont importantes. Nous ne retiendrons que ces derniers dans la suite de notre étude. L'intensité de la couleur de fond rouge des cellules du tableau dépend du nombre de règles de détection du nombre d'axes déclenchée. S'il n'y en a aucune, elle est simplement grisée. Ici, nous constatons qu'une ACP en 3 axes semble être la bonne solution. A partir du 3ème, aucune règle n'est activée.

## Significance of Principal Components

Global critical values	
Kaiser-Guttman	1
Karlis-Saporta-Spinaki	1,02715

Eigenvalue table - Test for significance

Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	1,711420	2,592857
2	1,043051	1,592857
3	1,020804	1,092857
4	0,948885	0,759524
5	0,888158	0,509524
6	0,812475	0,309524
7	0,575207	0,142857

( Errata: *Salary-per-week* est une erreur de renommage, elle représente *Salary-per-year*)

## 2 - Corrélation Entre Les Variables Et Les Axes Principaux

Dans la seconde partie des résultats, on peut constater suivant une indication claire, la corrélation qui existe entre les variables avec les axes factoriels, voir schéma suivant. Le tableau « Factor Loadings [Communalities] » indique les corrélations et les  $\cos^2$  des variables avec les facteurs. Dans la dernière ligne, nous observons les inerties exprimées. Les deux facteurs cumulés restituent 80% de l'information disponible.

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)
d2c_SALARY-PER-WEEK_1	-0,78860	62 % (62 %)	0,04826	0 % (62 %)
EDUCATION-NUM	-0,61179	37 % (37 %)	0,14711	2 % (40 %)
FNLWGT	0,11014	1 % (1 %)	0,63644	41 % (42 %)
CAPITAL-LOSS	-0,28974	8 % (8 %)	-0,47708	23 % (31 %)
AGE	-0,42235	18 % (18 %)	-0,40206	16 % (34 %)
HOURS-PER-WEEK	-0,49640	25 % (25 %)	0,09042	1 % (25 %)
CAPITAL-GAIN	-0,44089	19 % (19 %)	0,46540	22 % (41 %)
Var. Expl.	1,71142	24 % (24 %)	1,04305	15 % (39 %)

Que peut-on en déduire du tableau précédent?

Nous notons que le premier axe est fortement corrélé négativement avec les variables SALARY-PER-YEAR, AGE, HOURS-PER-WEEK, excepté la variable FNLWGT c'est-à-dire qu'il prend en compte le niveau d'éducation des gens et aussi le nombre d'heures de travail par semaine. On peut cependant constater que de la variable opposée associée à cet axe, est corrélée positivement pour ces mêmes attributs. Le premier axe ainsi défini est donc crucial dans la détermination des indices pouvant influencer le revenu annuel des gens. Le deuxième axe par contre est corrélé négativement pour les variables AGE, CAPITAL-LOSS, ce qui signifie qu'il fait référence aux données liées aux origines des gens (FNLWGT), de pertes ou dépenses qui peuvent affecter les indices.

La prochaine étape de notre analyse en composante principale concerne le cercle des corrélations.

### 3 - Cercle Des Corrélations

Le cercle de corrélation défini par deux composantes principales est la représentation graphique des variables en fonction de leurs coefficients de corrélation avec les composantes principales. Ainsi:

Un point proche du cercle caractérise bien la variable correspondante. Un point proche du centre indique une variable dont les propriétés ne sont pas mises en évidence par le cercle de corrélation.

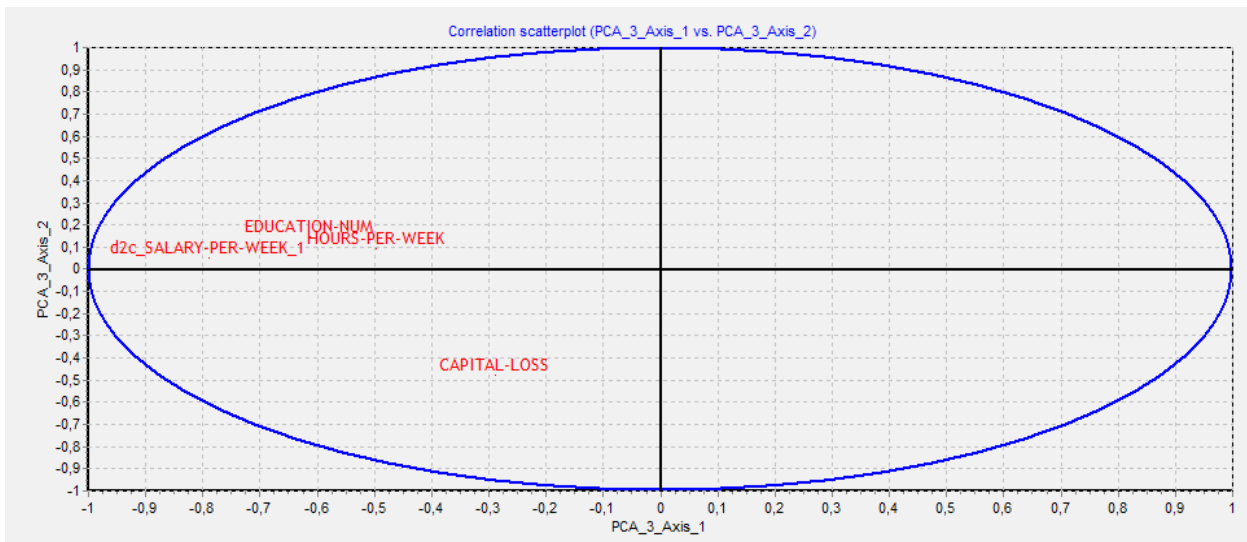
Deux points proches du cercle et l'un de l'autre indiquent une forte corrélation positive entre les variables qu'ils caractérisent.

Deux points proches du cercle et opposés indiquent une forte corrélation négative. Deux points proches du centre du cercle ne donnent aucune indication sur la corrélation des variables qu'ils représentent.

Dans le diagramme ci-dessous, nous visualisons immédiatement les attractions et oppositions entre les variables. Nous devinons la superposition entre la variable *Éducation* et le *nombre d'heures de travail par semaine*



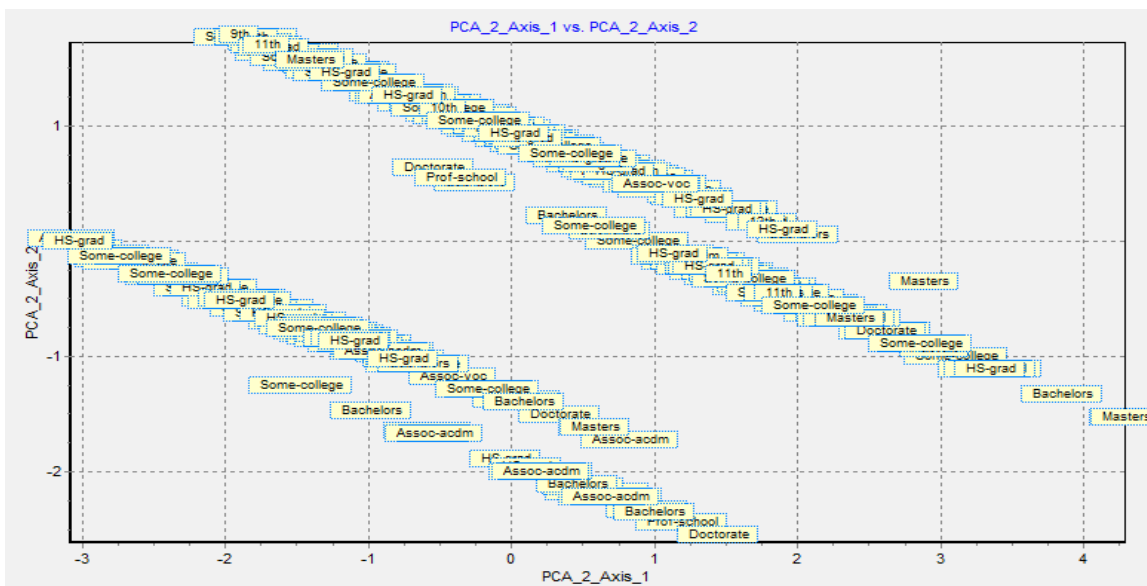
Nous



constatons que les variables liées aux indices SEX, HOURS-PER-WEEK et SALARY-PER-YEAR sont très éloignées du cercle et du deuxième axe principal ce qui confirme leur forte corrélation positive avec la première composante principale.

#### 4 - Plans Factoriels

Considérant les résultats du tableau suivant, on peut constater que la majorité des gens qui travaillent sont de sexe M et ont aussi un niveau d'éducation plus élevé, ces variables ont toutes une valeur positive pour l'axe principal choisi; on peut aussi constater que les femmes avec un niveau d'éducation élevé et qui travaillent beaucoup par semaine gagnent. Ce qui équivaut à des valeurs élevées des indices du de CAPITAL-LOSS de l'enquête. Autrement dit, dans la majorité des cas, les gens de sexe M (Male) travaillent plus en semaine et ont un niveau d'éducation plus élevé.



## VI. ANALYSE FACTORIELLE DES CORRESPONDANCES

L'analyse des correspondances est une méthode statistique de réduction de dimension. Elle propose une vision synthétique de l'information intéressante d'un tableau de contingence. Son pouvoir de séduction repose en grande partie sur les représentations graphiques qu'elle propose.

Elles nous permettent de situer facilement les similarités (dissimilarités) et les attractions (répulsions) entre les modalités. L'AFC est bien une technique factorielle.

Les facteurs – les variables latentes – qui en sont issus sont des combinaisons linéaires des points modalités (lignes ou colonnes) exprimés par des profils (lignes ou colonnes).

Nous avons effectué une analyse factorielle des correspondances entre les variables qualitatives de notre jeu de données. La relation entre ces variables est présentée dans le tableau de contingence à la figure du paragraphe étude d'une variable qualitative. Les sorties sont subdivisées en plusieurs zones, nous allons les énumérer tour à tour dans ce qui suit.

### 1 - Paramétrage de l'analyse des correspondances

Nous devons définir le rôle des variables à l'aide du composant DEFINE STATUS. Nous plaçons Occupation( c'est-à-dire l'emploi de l'individu, étiquette des lignes) en TARGET, les autres variables (les colonnes) en INPUT.

#### Résultats de l'AFC

### 2 - KHI-2 (global) de l'écart à l'indépendance

Le premier tableau indique la statistique du test du Chi2 d'écart à l'indépendance<sup>6</sup>. Ce résultat est fondamental. En effet, si la liaison globale est trop faible, l'étude des relations entre les modalités ne sert à rien. Il faut s'assurer qu'il existe une information exploitable dans le tableau.

#### CHI-SQUARE statistic

Trace	0,0550
Chi <sup>2</sup>	159653,31
d.f	65120
p-value	0,0000

En l'occurrence, nous avons  $\text{Chi}^2 (\text{h}^2) = 159653.31$ , avec un degré de liberté égal à 65120, la liaison est très significative ( $p\text{-value} < 0.0001$ ). De plus, Tanagra fournit la valeur du coefficient  $\phi^2$  (Trace), avec  $\phi^2 = \text{h}^2/n = 0.0550$ . L'AFC va décomposer cette quantité - qui symbolise l'information disponible dans le tableau de contingence - sur les différents

axes factoriels.

Le test du  $\chi^2$  n'est pas strictement applicable ici. En effet chaque individu a pu choisir plusieurs couples de valeurs (Occupation x Niveau scolaire par exemple). De fait, les observations ne sont pas indépendantes. Il faut dès lors voir le  $\chi^2$  plutôt comme un indicateur de la quantité d'information exploitable dans le tableau. A partir de  $\phi^2 > 0.2$ , la valeur «Trace» on peut penser que le tableau recèle des informations intéressantes .

### 3 - Valeurs Propres

Tanagra fournit le tableau des valeurs propres « Eigenvalues », avec le pourcentage d'inertie expliquée (individuelle et cumulée) par les axes. Un histogramme permet de situer leur décroissance.



Les valeurs propres expriment la part d'inertie expliquée par les axes. Ainsi, puisque la décomposition est orthogonale, elles s'additionnent et la somme ( $0.43633 + 0,011397$  est égale à  $\phi^2 = 0.055031$ .

Le tableau suivant permet d'observer la valeur propre calculée, le pourcentage d'inertie associée à chaque axe et le pourcentage cumulé qui permet d'avoir une idée du nombre d'axes à retenir.

#### Eigen values

Matrix trace = 0,0550

SQRT(Matrix trace) = 0,2346

Axis	Eigen value	% explained	Histogram	% cumulated
1	0,043633	79,29%		79,29%
2	0,011397	20,71%		100,00%
Tot.	0,055031	-	-	-

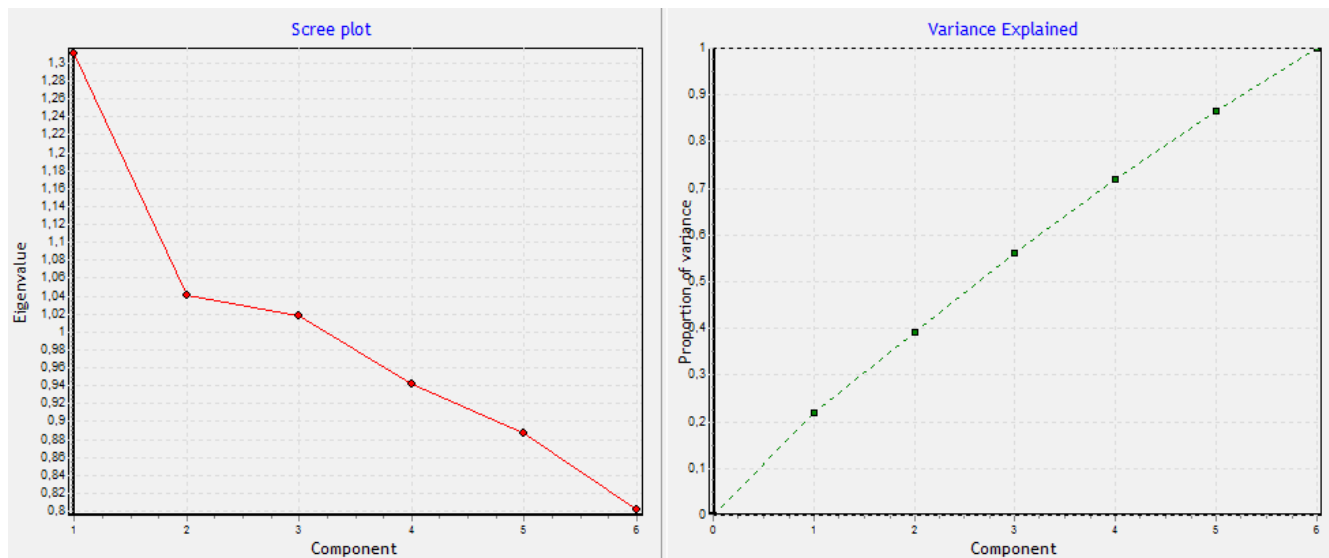
### 4 - Choix du nombre d'axes – Règle de Kaiser

Maintenant, il vient alors une question récurrente : combien d'axes devons-nous retenir dans l'analyse ?

Une règle très simple consiste à choisir les axes portés par une valeur propre supérieure à leur moyenne donc nous avons bien les 2 axes.

La représentation de Scree plot ci-dessous permet de confirmer l'hypothèse du choix de ces deux axes.

Il s'agit de repérer le « coude » dans la décroissance des valeurs propres. Dans notre cas, il survient au second axe, ce qui corrobore la conclusion de la règle de Kaiser.



## 5 - Représentation des lignes

### Rows analysis

Characterization				Coord.		Contributions (%)		COS²	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos² 1	cos² 2
Adm-clerical	0,00003	0,00800	0,00000	-0,00256	0,08943	0,00	0,00	0,00 (0,00)	1,00 (1,00)
Exec-managerial	0,00003	0,32264	0,00001	-0,52102	0,22621	0,02	0,01	0,84 (0,84)	0,16 (1,00)
Handlers-cleaners	0,00003	0,00094	0,00000	-0,00045	-0,03065	0,00	0,00	0,00 (0,00)	1,00 (1,00)
Handlers-cleaners	0,00003	0,04456	0,00000	-0,17187	-0,12253	0,00	0,00	0,66 (0,66)	0,34 (1,00)
Prof-specialty	0,00003	0,04096	0,00000	0,14819	0,13784	0,00	0,00	0,54 (0,54)	0,46 (1,00)
Exec-managerial	0,00003	0,01671	0,00000	0,02438	0,12695	0,00	0,00	0,04 (0,04)	0,96 (1,00)
Other-service	0,00002	0,29181	0,00001	-0,53291	-0,08837	0,02	0,00	0,97 (0,97)	0,03 (1,00)
Exec-managerial	0,00004	0,01659	0,00000	-0,09994	-0,08122	0,00	0,00	0,60 (0,60)	0,40 (1,00)
Prof-specialty	0,00003	0,04822	0,00000	0,19937	0,09207	0,00	0,00	0,82 (0,82)	0,18 (1,00)
Exec-managerial	0,00003	0,00753	0,00000	-0,03762	0,07817	0,00	0,00	0,19 (0,19)	0,81 (1,00)
Exec-managerial	0,00004	0,12513	0,00001	0,32664	-0,13576	0,01	0,01	0,85 (0,85)	0,15 (1,00)
Prof-specialty	0,00003	0,03028	0,00000	0,11781	0,12809	0,00	0,00	0,46 (0,46)	0,54 (1,00)
Adm-clerical	0,00002	0,07862	0,00000	0,11397	0,25619	0,00	0,01	0,17 (0,17)	0,83 (1,00)
Sales	0,00003	0,03509	0,00000	0,18478	0,03083	0,00	0,00	0,97 (0,97)	0,03 (1,00)
Craft-repair	0,00003	0,00108	0,00000	-0,01985	0,02615	0,00	0,00	0,37 (0,37)	0,63 (1,00)
Transport-moving	0,00003	0,05573	0,00000	0,09987	-0,21391	0,00	0,01	0,18 (0,18)	0,82 (1,00)
Farming-fishing	0,00002	0,02047	0,00000	0,13627	0,04361	0,00	0,00	0,91 (0,91)	0,09 (1,00)
Machine-op-inspct	0,00003	0,00689	0,00000	0,08194	-0,01314	0,00	0,00	0,98 (0,98)	0,03 (1,00)

A

B

C

D

La représentation des lignes couvre plusieurs informations: les statistiques sur les points lignes Age, Niveau d'éducation, Nombre d'Heures par Semaine, FNLWGT (A); les coordonnées factorielles (B); les contributions (CTR) aux axes (en %) (C) ; et la qualité de représentation (COS<sup>2</sup>) par axe et cumulée (D).

Les COS<sup>2</sup> indiquent la qualité de représentation individuelle et cumulée des modalités sur les K premiers facteurs. Dans notre exemple, toutes les variables de lignes sont bien résumées par les axes

Dans notre exemple, nous constatons que le premier facteur est défini par l'opposition entre (Exec Managerial et Other-Service). Sur le second, nous avons une opposition entre (Prof-Speciality, Admin-Clerical) et Transport-moving. Cependant, la lecture n'est pas trop aussi évidente qu'on pourrait.

L'inertie ici est fixée à 0.00003 pour ce cluster.

## 6 - Représentation des colonnes

La représentation des colonnes obéit à la même logique.

### Columns analysis

Characterization				Coord.		Contributions (%)		COS <sup>2</sup>	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos <sup>2</sup> 1	cos <sup>2</sup> 2
AGE	0,43302	0,05483	0,02374	-0,23237	-0,02848	53,58	3,08	0,98 (0,98)	0,01 (1,00)
HOURS-PER-WEEK	0,45384	0,04623	0,02098	0,20982	-0,04696	45,79	8,78	0,95 (0,95)	0,05 (1,00)
EDUCATION-NUM	0,11314	0,09101	0,01030	0,04791	0,29787	0,60	88,08	0,03 (0,03)	0,97 (1,00)

D'emblée, on sait que « Age » et, dans une moindre mesure, « Hours-per-Week » vont beaucoup peser dans l'étude de détermination du salaire annuel. En effet, « Age » compte pour 38,1% (0.02098 / 0.0550) de l'inertie totale, « Hours-per-Week » pour 42,24% (0,023237 / 0.0,0550). Et effectivement, ces modalités déterminent en grande partie les deux premiers facteurs.

# CLUSTERING

## VII. CLUSTERING

### 1 - Introduction

Le clustering est une des méthodes d'analyse des données. Elle vise à diviser un ensemble de données en différents «paquets» homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité informatique) que l'on définit en introduisant des mesures et classes de distance entre objets.

Pour obtenir un bon partitionnement, il convient d'à la fois:

- Minimiser l'inertie intra-classe pour obtenir des grappes (cluster en anglais) les plus homogènes possibles;
- Maximiser l'inertie inter-classe afin d'obtenir des sous-ensembles bien différenciés.

### 2 - Objectif

L'objectif de notre classification c'est que, se basant sur les critères représentés par les variables telles que Age, Secteur de Travail (work-class), Education, on doit arriver à des groupes d'individu dont le salaire annuel (Salary-peryear) est soit inférieur soit supérieur déjà 50k.

### 3 - Algorithme de classification utilisé: K-means et Hac

Afin de classer les individus de notre ensemble de données, nous avons appliqué, l'algorithme K-means sur les facteurs de l'ACP. Nous insérons de nouveau DEFINE STATUS dans le diagramme, nous plaçons en INPUT les variables calculées PCA\_1\_Axis\_1 à PCA\_1\_AXIS\_5. Nous plaçons le composant K-MEANS (onglet CLUSTERING).

Nous le paramétrons en actionnant le menu contextuel PARAMETERS.

Parameters	
K-Means parameters	
Clusters	10
Max Iteration	8
Trials	1
Distance normalization	none
Average computation	McQueen
Seed random generator	Standard

Le nombre de clusters spécifié est 10 (Number of Clusters). Nous ne réalisons qu'un seul essai d'optimisation (Number of trials = 1), avec un nombre d'itération maximum à 8 (Max iterations). Attention,

les variables (axes factoriels) ne doivent pas être standardisées, nous les mettons à NONE le paramètre DISTANCE NORMALIZATION. Nous validons et nous actionnons le menu VIEW.

La figure ci-dessous montre la répartition des individus dans les différents clusters

### Global evaluation

Within Sum of Squares	55825,1258
Total Sum of Squares	182742,6841
R-Square	0,6945

### Cluster size and WSS

Clusters	10		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	2596	6524,1719
cluster n°2	c_kmeans_2	5315	5865,8660
cluster n°3	c_kmeans_3	5261	5811,0784
cluster n°4	c_kmeans_4	1463	6429,6366
cluster n°5	c_kmeans_5	2963	4123,0418
cluster n°6	c_kmeans_6	4103	9850,5218
cluster n°7	c_kmeans_7	6170	5870,6555
cluster n°8	c_kmeans_8	1478	4627,7680
cluster n°9	c_kmeans_9	159	443,9524
cluster n°10	c_kmeans_10	3053	6278,4334

### R-Square for each attempt

Number of trials	1
Trial	R-square
1	0,694515

Tanagra énumère nos 10 groupes avec les effectifs associés. Nous retiendrons que la part de variance expliquée par le partitionnement est de 69,4%.

Le cluster n°2 a le plus grand nombre d'individu avec un total de 5315.

## 4 - Coordonnées des centres des clusters

Les coordonnées des centres de chacun des 10 clusters sont présentées à la figure suivante

### Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4	Cluster n°5	Cluster n°6	Cluster n°7	Cluster n°8	Cluster n°9	Cluster n°10
PCA_2_Axis_1	0,648897	0,770610	0,242640	-1,701790	1,480870	-1,775617	0,372420	0,930791	-6,521573	-1,410484
PCA_2_Axis_2	-0,891827	0,620032	-0,567692	-2,195357	0,138143	-0,166590	-0,013923	1,752266	6,309272	0,650256
PCA_2_Axis_3	0,987318	-0,392474	0,602337	-2,998426	0,090526	0,356805	0,098503	-1,301076	5,382105	-0,174257
PCA_2_Axis_4	-1,652528	0,011566	-0,079948	-0,655875	0,260095	0,474584	1,039208	-1,241658	-1,889584	-0,453825
PCA_2_Axis_5	0,177371	0,311764	0,439759	-1,213201	-1,364745	0,379979	-0,078813	0,426277	-4,422835	0,127093

Use GROUP CHARACTERIZATION for detailed comparisons

Nous souhaitons maintenant réaliser la HAC (Hierarchical Agglomerative Clustering) en prenant comme groupes de départ, ceux produits par les K-MEANS. Nous insérons encore une fois le composant DEFINE STATUS. Nous plaçons en TARGET la variable indicatrice des sous-groupes CLUSTER\_KMEANS\_1, produite par le composant KMEANS.

Cette spécification est importante. Si nous l'omettons, Tanagra essaiera de partir des observations individuelles, soit plus de 32500 individus, c'est suicidaire. En INPUT, nous plaçons les axes factoriels.



Nous introduisons ensuite le composant CAH. Nous demandons à utiliser une distance non normalisée afin que les facteurs pèsent selon la part d’inertie rapportée par l’AFDM. Le nombre de clusters est détecté automatiquement (Tanagra recherche l’écart le plus élevé entre deux paliers consécutifs du dendrogramme, en ignorant la partition triviale). Nous lançons les calculs en actionnant le menu contextuel VIEW. Le rapport apparaît. Nous avons les effectifs dans chaque cluster. Plus bas, Tanagra nous indique la proportion d’inertie expliquée par la partition.

Dans les captures ci-dessous, nous présentons les résultats du HAC



Clustering results

Clusters	From the dendrogram	After one-pass relocation
cluster n°1	23783	23783
cluster n°2	1463	1463
cluster n°3	7315	7315

Best cluster selection

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,2022	0,4155
3	0,3304	0,1696
4	0,4283	0,0218
5	0,5224	0,1740
6	0,5854	0,1379
7	0,6239	0,0579
8	0,6520	0,0260
9	0,6755	0,0240
10	0,6948	0,1080

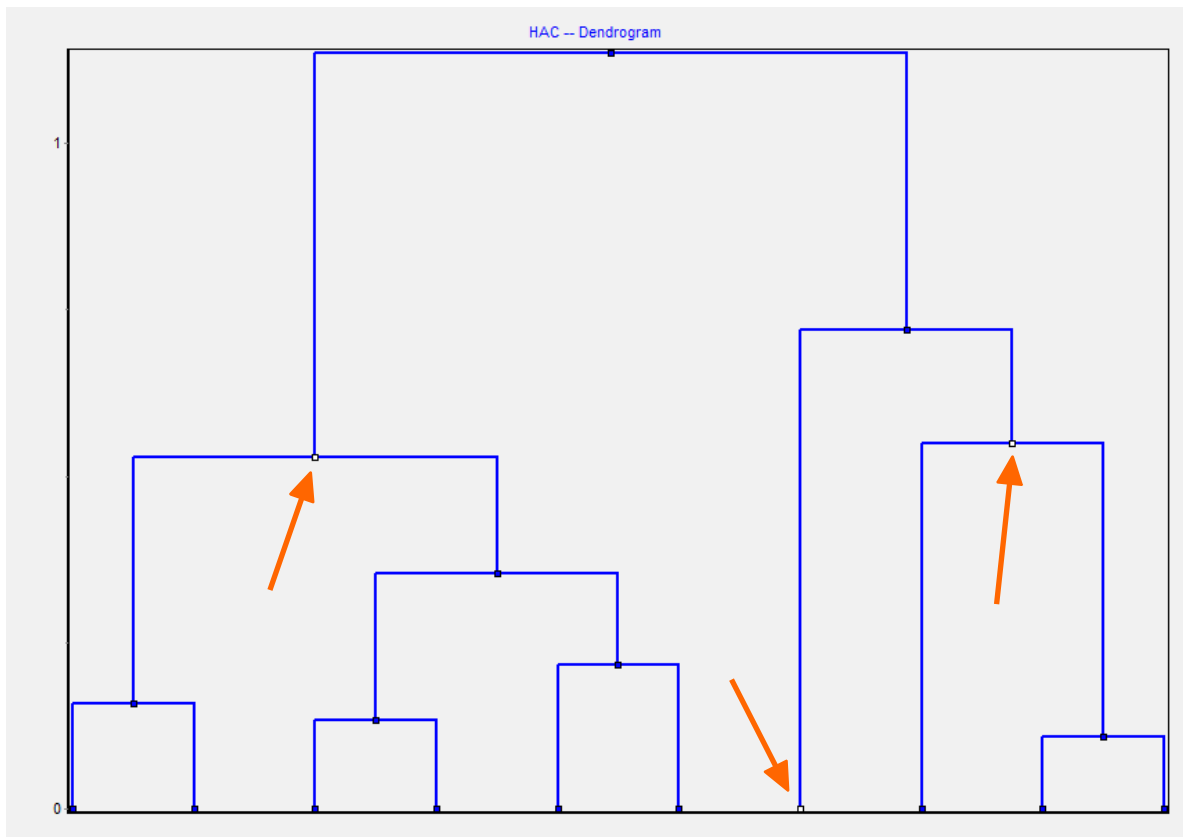
Tanagra produit 3 classes d’effectifs respectifs : 23783, 1463 et 7315 individus. Les effectifs sont différents de ceux observés dans le dendrogramme car Tanagra, depuis la version 1.4.48, effectue une dernière passe sur les données pour affecter les individus aux barycentres de classes qui leur sont le plus proches. L’objectif est d’obtenir des groupes avec une meilleure cohésion, la partition initiale étant contrainte par la structure hiérarchique de la recherche des solutions. Le dendrogramme montre que la partition en 3 classes est la plus évidente .

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
PCA_2_Axis_1	0,635673	-1,701790	-1,726383
PCA_2_Axis_2	0,038133	-2,195357	0,315091
PCA_2_Axis_3	0,109279	-2,998426	0,244391
PCA_2_Axis_4	0,029361	-0,655875	0,035714
PCA_2_Axis_5	0,022330	-1,213201	0,170039

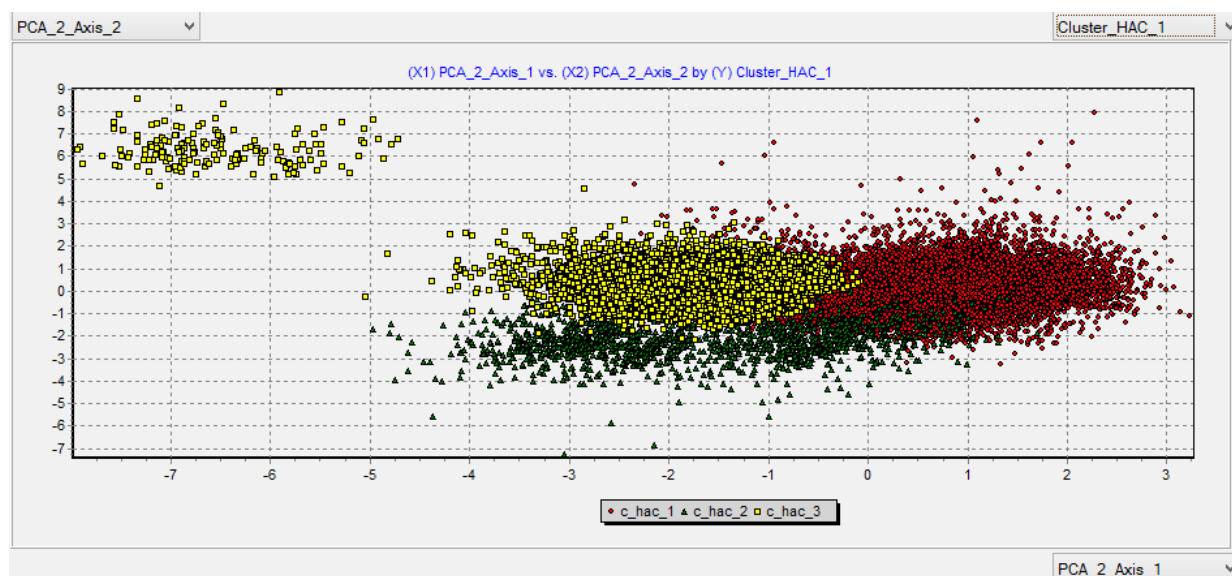
Use GROUP CHARACTERIZATION for detailed comparisons

Dans la partie «Cluster Centroids», Tanagra fournit les coordonnées des centres de classes. Cette information est importante pour le déploiement. En effet, on s’appuiera sur ces éléments pour rattacher les individus supplémentaires aux catégories.



## 5 - Positionnement des classes dans le plan factoriel

Pour mieux apprécier la qualité de la partition, nous visualisons les groupes dans le premier plan factoriel. Nous utilisons le composant SCATTERPLOT (onglet DATA VISUALIZATION). Nous plaçons PCA\_2\_AXIS\_1 en abscisse et PCA\_2\_AXIS\_2 en ordonnée. Nous illustrons les points à l'aide de la variable CLUSTER\_HAC\_1 générée automatiquement par le composant HAC.



Nous observons dans la figure ci-dessus les 3 groupes d'observations mises en évidence par l'algorithme de classification HAC. Le premier facteur permet d'isoler le premier cluster (C\_HAC\_1) en rouge, le second permet de distinguer le second (C\_HAC\_2) en vert et le troisième (C\_HAC\_3) en jaune. Il n'y a empiètement entre les classes – dans le premier plan factoriel. Ce qui est tout à fait normal dans la mesure où les 3 groupes présentent des traits relativement similaires.

## 6 - Caractérisation des classes – Variables actives et illustratives

Il s'agit justement de comprendre les caractéristiques sous-jacentes aux classes, en utilisant d'une part les variables ayant participé à la construction de la typologie, d'autre part la variable supplémentaire SCORE indiquant l'appréciation du client par le conseiller clientèle. Nous utilisons pour la 3ème fois le composant DEFINE STATUS. Nous plaçons en TARGET la variable CLUSTER\_HAC\_1. Elle associe chaque individu à la classe qui lui a été affectée. Nous mettons en INPUT toutes les variables de l'étude, y compris SCORE que nous utiliserons pour illustrer les groupes.

Nous insérons le composant GROUP CHARACTERIZATION (onglet STATISTICS). Le tableau est scindé en deux : pour les variables quantitatives (CONTINUOUS), les moyennes conditionnellement aux groupes sont comparées aux moyennes globales, calculées sur la totalité de l'échantillon ; pour les variables qualitatives (DISCRETE), les fréquences conditionnelles sont opposées aux fréquences globales.

Results											
Description of "Cluster_HAC_1"											
Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples		[ 73,0 % ] 23783		Examples		[ 4,5 % ] 1463		Examples		[ 22,5 % ] 7315	
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
FNLWGT	6,47	192079,32 (109885,67)	189778,37 (105549,98)	CAPITAL-LOSS	176,84	1908,04 (318,65)	87,30 (402,96)	EDUCATION-NUM	68,23	11,89 (2,19)	10,08 (2,57)
CAPITAL-GAIN	-36,98	158,12 (984,19)	1077,65 (7385,29)	EDUCATION-NUM	14,60	11,04 (2,64)	10,08 (2,57)	HOURS-PER-WEEK	57,08	47,69 (12,14)	40,44 (12,35)
AGE	-44,24	36,55 (14,03)	38,58 (13,64)	HOURS-PER-WEEK	9,67	43,49 (12,14)	40,44 (12,35)	AGE	42,60	44,56 (10,28)	38,58 (13,64)
HOURS-PER-WEEK	-58,21	38,02 (11,48)	40,44 (12,35)	AGE	8,94	41,70 (12,57)	38,58 (13,64)	CAPITAL-GAIN	42,15	4282,81 (15046,70)	1077,65 (7385,29)
CAPITAL-LOSS	-62,81	2,09 (50,00)	87,30 (402,96)	FNLWGT	-2,23	183769,48 (93347,70)	189778,37 (105549,98)	FNLWGT	-5,78	183499,13 (92346,83)	189778,37 (105549,98)
EDUCATION-NUM	-70,99	9,47 (2,39)	10,08 (2,57)	CAPITAL-GAIN	-5,71	0,00 (0,00)	1077,65 (7385,29)	CAPITAL-LOSS	-20,99	0,22 (11,44)	87,30 (402,96)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			
SALARY-PER-WEEK= <=50K	153,14	[ 94,3 % ] 98,0 %	75,9 %	SALARY-PER-WEEK= >50K	26,20	[ 9,8 % ] 52,7 %	24,1 %	SALARY-PER-WEEK= >50K	149,82	[ 84,0 % ] 90,0 %	24,1 %
OCCUPATION= Other-service	28,63	[ 94,0 % ] 13,0 %	10,1 %	OCCUPATION= Exec-managerial	8,28	[ 7,0 % ] 19,5 %	12,5 %	OCCUPATION= Prof-specialty	38,46	[ 45,8 % ] 25,9 %	12,7 %
EDUCATION= HS-grad	26,66	[ 82,5 % ] 36,4 %	32,3 %	EDUCATION= Doctorate	8,23	[ 12,8 % ] 3,6 %	1,3 %	EDUCATION= Bachelors	36,75	[ 41,6 % ] 30,5 %	16,4 %
WORKCLASS= Private	18,41	[ 76,0 % ] 72,6 %	69,7 %	EDUCATION= Masters	8,20	[ 8,5 % ] 10,0 %	5,3 %	EDUCATION= Masters	35,94	[ 57,6 % ] 13,6 %	5,3 %
EDUCATION= 11th	17,66	[ 95,5 % ] 4,7 %	3,6 %	OCCUPATION=		[ 6,8 % ]		OCCUPATION=		[ 44,1 % ]	

## 7 - Interprétation

Détaillons le tableau pour bien préciser les idées en considérant le troisième cluster par exemple.

Variables Quantitatives:

Le Niveau d'éducation moyen dans la population (en prenant en compte la totalité de l'échantillon) est de 10,08, avec un écart type de 2,57. Dans ce cluster, il devient 11,89 (avec un écart type de 2,19). De fait, les personnes du troisième groupe présentent un Niveau d'éducation significativement plus élevé comme l'atteste la valeur test (TEST VALUE ) de 68,23. Le Niveau d'éducation est de 10,08 dans la population, dans ce groupe il est de 11,89. Ces personnes sont mieux payées par les employeurs. Pas de surprise : plus on est qualifié, plus on intéresse les organismes qui recrutent. Le contraire eut été très étonnant. •

Variables Qualitatives:

Dans le Groupe 1: On retrouve les individus dont le salaire est dont l'occupation est dans le secteur des autres services, le niveau d'éducation est «high school», et ils sont pour la plupart embauchés dans le secteur privée des petites affaires. Leur revenu annuel est inférieur ou égal à 50K ( $\leq 50K$ ).

Dans le Groupe 2: On retrouve les individus dont le salaire annuel sont uniquement supérieur à 50K. Leur niveau d'éducation est le doctorat, masters et occupent des postes exécutifs.

Dans le Groupe 3: On retrouve aussi les individus dont le salaire annuel sont uniquement supérieur à 50K. Leur niveau d'éducation est bachelier, masters et occupent des postes de spécialisation.

## VIII. Conclusion

Ce travail pratique effectué dans le cadre du cours de Fouille de données nous a permis d'approfondir nos connaissances sur les aspects vus en cours. Nous avons présenté la statistique sur les variables, l'Analyse Factorielle des Correspondances de notre jeu de données, l'analyse factorielle, le clustering. Dans le clustering nous nous sommes basés sur l'algorithme de Kmeans sur lequel on a effectué aussi le Hac. Considérant les résultats, toutes les méthodes utilisées ne sont pas optimales, un autre jeu de données plus approprié donnerait peut-être des résultats beaucoup plus satisfaisants. Néanmoins on a une connaissance pratique de ces notions.

## IX. RÉFÉRENCES

- [1] <http://www.analyse-donnees.fr/services-analyse-enquetes-traitement/>
- [2] <http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names>
- [3] Cours de Fouilles de Données, Master 1, IFI 2016, Enseignante NGUYỄN Thị Minh Huyền

Tutoriels:

- <http://tutoriels-data-mining.blogspot.com/2009/05/analyse-factorielle-des-correspondances.html>
- <http://tutoriels-data-mining.blogspot.com/search/label/App.%20Supervis%C3%A9%20-%20Scoring>