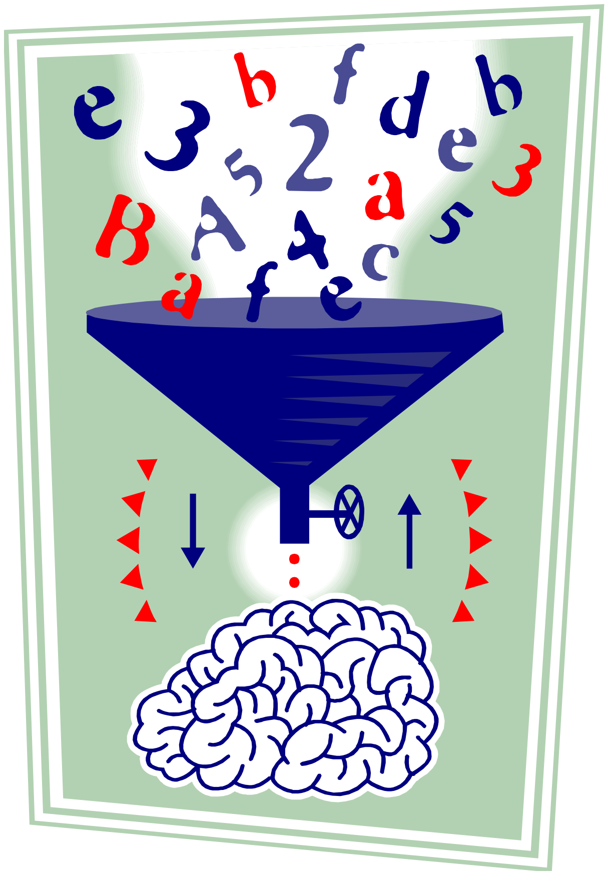
L'AFC pour les nuls

Mise à jour du 8 août 2010

Rémi Bachelet

La version à jour de ce cours d'analyse factorielle des composantes est disponible http://rb.ec-lille.fr

Cette formation est également enregistrée en vidéo



Source des images indiquées au-dessous

Comment utiliser ce cours :

1. Mettre les diapos en format plein écran en cliquant sur
2. Faire défiler l'animation en cliquant **sur** les diapositives

(attention : cliquer sur une image ou un lien ouvre la page web correspondante)

Objectifs du cours d'Analyse Factorielle des Correspondances

Méthode développée notamment par Jean-Paul Benzécri (1970+)

1. Comprendre les fondements de l'Analyse Factorielle des Correspondances
2. Savoir quel est le processus de calcul et ses logiques
3. Pouvoir expliquer le mapping produit par une AFC
4. Également :
   - Connaître quelques logiciels d'administration d'enquêtes et de traitement de données
   - Avoir des éléments de comparaison AFC – ACP (ACP = Analyse en Composantes Principales).

Principes de l'AFC et données d'entrées

1. **Principe général de l'AFC**
2. **Exemples :**
   - **Les limites des représentations graphiques intuitives**
   - **Comment donner du sens aux informations**

Principe général de l'analyse factorielle des correspondances (AFC)

« L'analyse factorielle traite des tableaux de nombres.

*Elle **remplace un tableau de nombres** difficile à analyser par **une série de tableaux plus simples** qui sont une bonne approximation de celui-ci »*

*Ces tableaux sont « simples », car **ils sont exprimables sous forme de graphiques***

Pourquoi « des correspondances » ?

variables numériques $\Rightarrow$ Corrélation

variables nominales $\Rightarrow$ Correspondance

Pourquoi « factorielle » ?

Il s'agit de décomposer le tableau original en une somme de tableaux/matrices qui sont chacun le **produit** de facteurs simples.

Autrement dit, on les « met en facteurs »

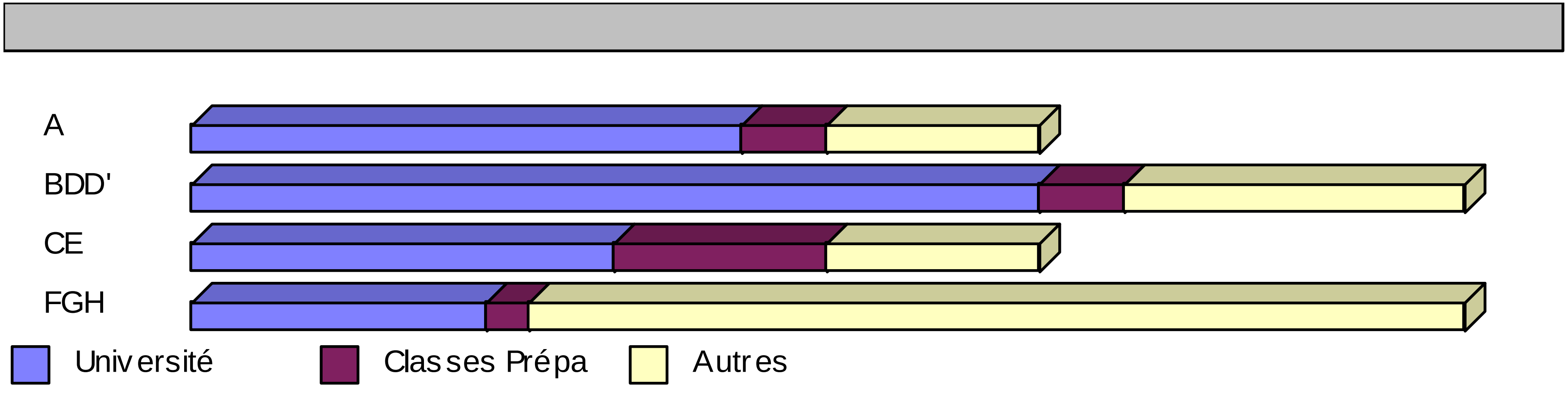
Principale source d'informations, et de l'exemple utilisé pour ce cours : *Que sais-je ? « L' analyse factorielle » - N°2095, Philippe CIBOIS, ed. PUF, épuisé,* extrait disponible en ligne

Exemple : que deviennent les bacheliers ?

	destination			
	université	***classes prépa***	***autres***	***total***
A	13	2	5	**20**
BDD'	20	2	8	**30**
CE	10	5	5	**20**
FGH	7	1	22	**30**
total	**50**	**10**	**40**	**100**

Stats MEN 1975 - 1975 204 489 lycéens

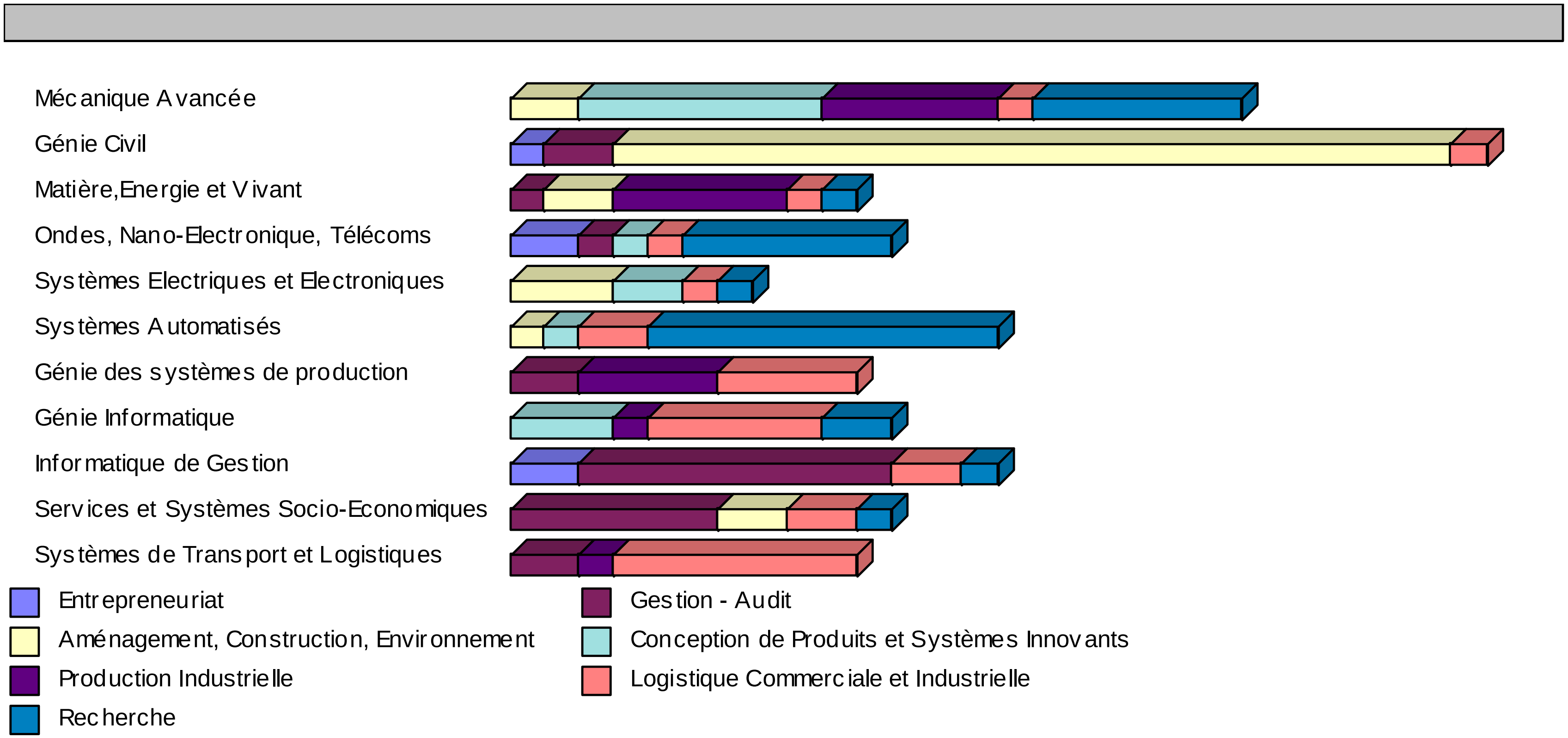
Une représentation graphique intuitive



Exemple : quels souhaits d'orientation ?

Premiers vœux 2003 de Génie / filière.	Entrepreneuriat	Gestion - Audit	Aménagement, Construction, Environnement	Conception de Produits et Systèmes Innovants	Production Industrielle	Logistique Commerciale et Industrielle	Recherche
Mécanique Avancée	0	0	2	7	5	1	6
Génie Civil	1	2	24	0	0	1	0
Matière,Energie et Vivant	0	1	2	0	5	1	1
Ondes, Nano-Electronique, Télécoms	2	1	0	1	0	1	6
Systèmes Electriques et Electroniques	0	0	3	2	0	1	1
Systèmes Automatisés	0	0	1	1	0	2	10
Génie des systèmes de production	0	5	0	0	4	4	0
Génie Informatique	0	0	0	3	1	5	2
Informatique de Gestion	2	11	0	0	0	2	1
Services et Systèmes Socio-Economiques	1	6	3	0	0	2	1
Systèmes de Transport et Logistiques	0	2	0	0	1	8	0

.. Pas toujours suffisante :



Comment donner du sens à ces données

Idée : ce qui est intéressant, c'est de mettre en évidence ce qui est **inattendu** dans ces répartitions

Inattendu = en quoi on dévie d'une répartition uniforme

On va donc :

1. Évaluer ce que serait une situation d'uniformité, d'indépendance
2. Calculer en quoi la situation constatée en diffère
3. Exprimer cette différence graphiquement pour pouvoir l'analyser
4. Interpréter le mapping obtenu …
5. et en optimiser la lisibilité

Première opérations sur les matrices

1. **Matrice « T » des données d'entrée**
   - **Matrice R des écarts à l'indépendance**
2. **Mise en facteur d'une matrice**
   - **Exprimer « simplement » R**

Matrice « T » des données d'entrée

	destination			
	université	**classes prépa**	**autres**	**total**
A	13	2	5	**20**
BDD'	20	2	8	**30**
CE	10	5	5	**20**
FGH	7	1	22	**30**
total	**50**	**10**	**40**	**100**

Ce tableau est aussi une matrice, appellons-la « T »

Quelle matrice aurait-on si la répartition dans les filières post-Bac ne dépendait pas du type de Bac ?

1/ S'il y avait situation d'indépendance…

10 = 50 * 20%

(produit matriciel /100

puisqu'on raisonne en %)

$$
\begin{array}{ccc|c}
\multicolumn{3}{c|}{\left[\begin{array}{ccc} 10 & & \\ & & \\ & & \\ & & \end{array}\right]} & \begin{array}{c} 20 \\ 30 \\ 20 \\ 30 \end{array} \\
50 & 10 & 40 &
\end{array}
$$

On reconstitue la matrice à partir de ses marges

Appellons cette matrice « T_0 »

2/ La matrice des écarts à l'indépendance est

$T - T_0 = R$

$$\begin{bmatrix} 13 & 2 & 5 \\ 20 & 2 & 8 \\ 10 & 5 & 5 \\ 7 & 1 & 22 \end{bmatrix} - \begin{bmatrix} 10 & 2 & 8 \\ 15 & 3 & 12 \\ 10 & 2 & 8 \\ 15 & 3 & 12 \end{bmatrix} = \begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix}$$

Quelle est la particularité de R ?

3/ Comment exprimer simplement R ?

On décompose la matrice des écarts à l'indépendance en une somme de matrices..

$$R = T_{1} + T_{2}$$

.. Chacune de ces matrices étant mise en facteur (le produit d'un vecteur ligne et d'un vecteur colonne).

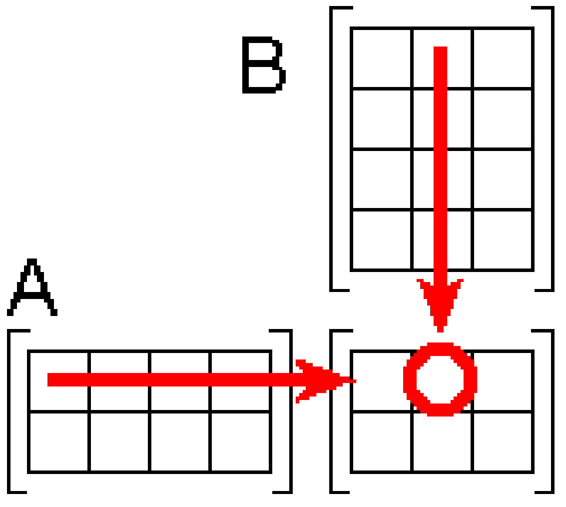
$$T_1 = C_1 L_1$$

(une matrice dont la plus petite dimension est N « rang N » est décomposable au maximum en N matrices pouvant se mettre en facteurs …

ici T = $T_0 + T_1 + T_2$*).*

T est de rang 3, mais R est de rang 2….

Produit matriciel : exemple



$$c_{12} = \sum_{r=1}^{4} a_{1r} b_{r2} = a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} + a_{14}b_{42}$$

Mise en facteur d'une matrice: exemple

T = CL

On met en facteur **T** comme le produit d'une matrice colonne **C** par une matrice ligne **L**

- T (2X2)
- C (1X2)
- L (2X1)

Attention les règles de présentation du produit matriciel ne sont pas bien respectées dans nos diapos

De plus, la multiplication des matrices n'est pas commutative ($LC \neq CL$)

$$R = T_1 + T_2 = C_1 L_1 + C_2 L_2$$

$$\begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ 2 & 2 & -4 \\ -4 & -4 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \\ -4 \end{bmatrix} + \begin{bmatrix} 2 & -1 & -1 \\ 4 & -2 & -2 \\ -2 & 1 & 1 \\ -4 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & -2 \end{bmatrix} \qquad \begin{bmatrix} 2 & -1 & -1 \end{bmatrix}$$

Attention le sens de multiplication écrit ici
est LC au lieu de CL

D’une matrice à une présentation graphique

Production et interprétation du mapping

- **Vecteurs colonne et vecteurs ligne**
- **Produit scalaire**

3/ bis Comment représenter graphiquement la décomposition ?

Un vecteur colonne (resp. ligne) correspond à une modalité des données en colonnes (resp. lignes)

Un axe unidimensionnel + un axe unidimensionnel = un repère

Un vecteur colonne correspond à une modalité des données en colonnes

$$
\begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix}
=
\begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ 2 & 2 & -4 \\ -4 & -4 & 8 \end{bmatrix}
\begin{bmatrix} 1 \\ 1 \\ 2 \\ -4 \end{bmatrix}
+
\begin{bmatrix} 2 & -1 & -1 \\ 4 & -2 & -2 \\ -2 & 1 & 1 \\ -4 & 2 & 2 \end{bmatrix}
\begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix}
$$

$$
\begin{bmatrix} 1 & 1 & -2 \end{bmatrix} \qquad \begin{bmatrix} 2 & -1 & -1 \end{bmatrix}
$$

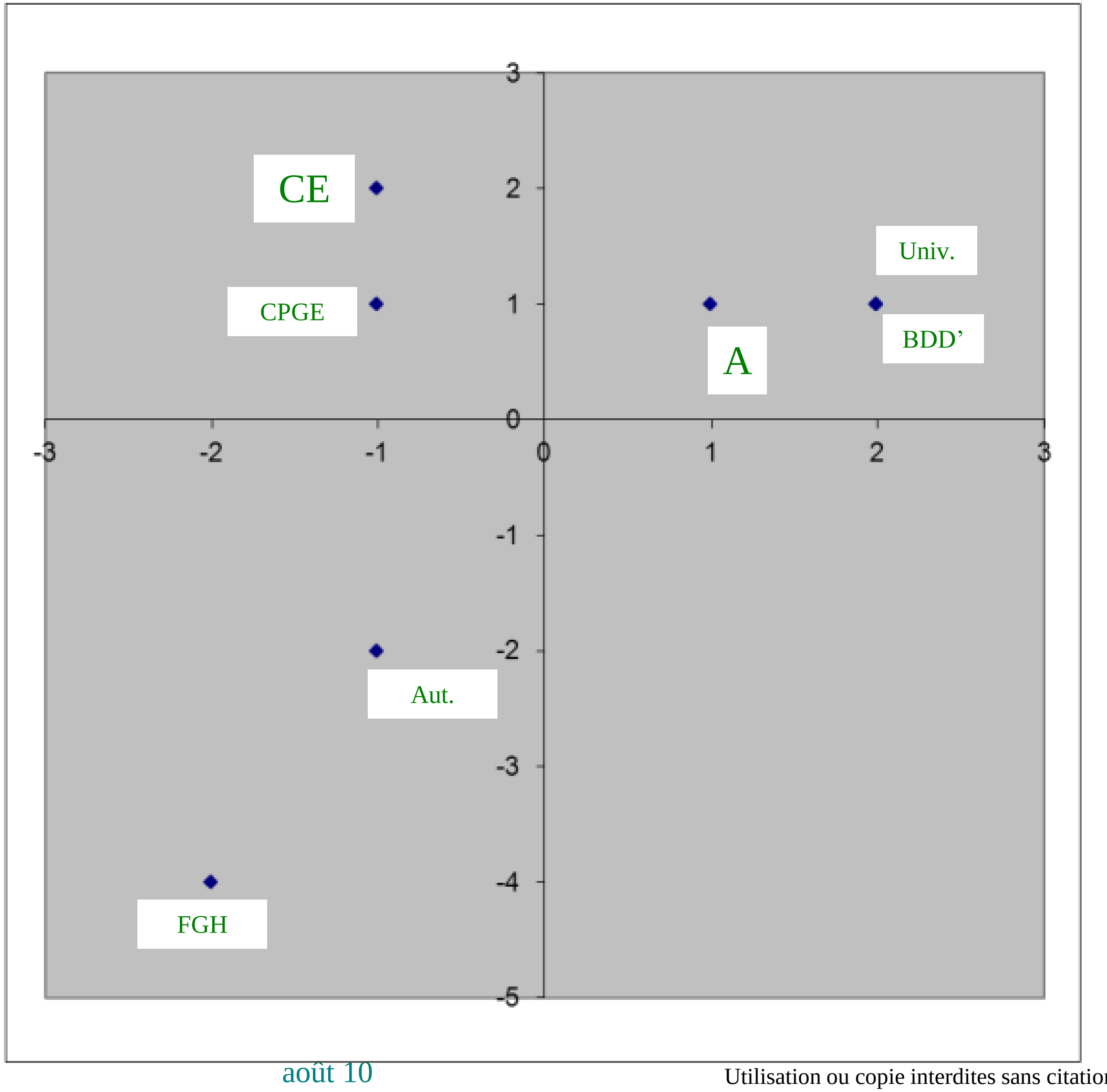
Un vecteur colonne correspond à une modalité des données en colonnes

$$\begin{array}{cc} \text{A} & \\ \text{BDD'} & \\ \text{CE} & \\ \text{FGH} & \end{array} \begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix}$$

$$\begin{array}{ccc} \text{Univ} & \text{CPGE} & \text{Autres} \end{array}$$
$$\begin{bmatrix} 2 & -1 & -1 \end{bmatrix}$$

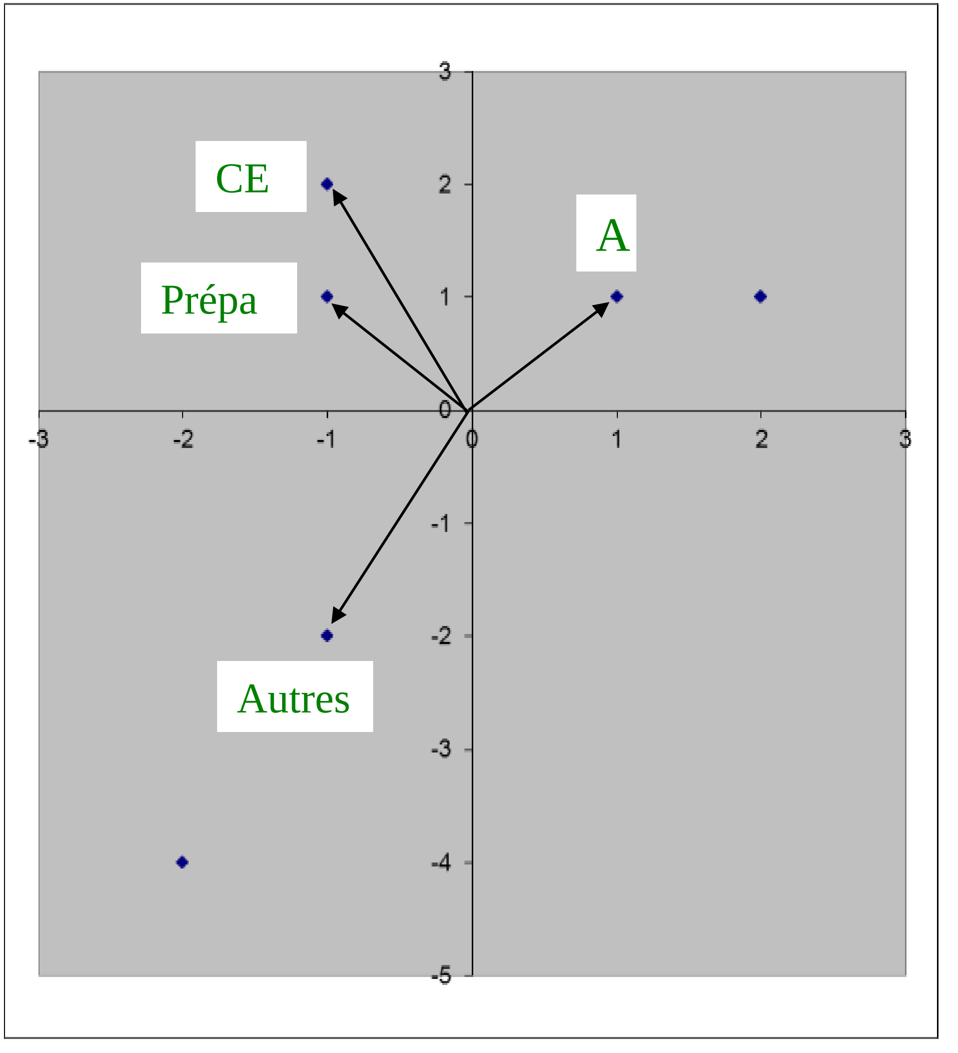
Un axe unidimensionnel + un axe unidimensionnel = un repère

A	1	1
BDD’	2	1
CE	-1	2
FGH	-2	-4
Univ	2	1
CPGE	-1	1
Autres	-1	-2



4/ Que veut dire ce mapping ?

1. Conjonction :
   Produit scalaire positif
   Les Bac CE ont une affinité pour la prépa
2. Opposition
   Produit scalaire négatif
   Les Bacs A ne vont pas vers les « autres » (IUT, BTS)
3. Quadrature
   Produit scalaire nul
   Les bacs A ne vont ni plus ni moins vers les prépas que la moyenne des bacheliers



Optimisation de la factorisation

1. **Le Chi-2 (χ^2) comme métrique**
   - **Degrés de liberté**
2. **Retour aux applications**
   - **Analyse de mappings**

5/ Mais ….. Quelle est la meilleure décomposition possible pour R ?

En effet $R = T_1 + T_2$… mais il existe aussi

$$R = T'_1 + T'_2 = T''_1 + T''_2 \ldots$$

Quel est le critère (la métrique) qui permet de définir les meilleurs T_1 et T_2?

Pour une matrice de rang n, on cherche d'abord à trouver la meilleure T_1, puis la meilleure T_2 de telle manière à ce que le premier axe soit celui qui exprime le plus de sens..

La métrique que nous cherchons, c'est le Chi-2 (χ^2)

Le χ^2 représente l'écart à l'indépendance

- or cette indépendance, est exprimée par T_0
- … l'écart à l'indépendance peut donc se mesurer comme l'écart à T_0

À partir de la matrice des données pour chaque cellule de T_1 et T_2, on calcule

1. L'écart avec la cellule correspondante de T_0 au carré (d'où le « ^{2} » du χ^2)
2. On divise par l'effectif théorique de cette cellule (on parle de χ^2 pondéré)
3. Le χ^2 de la matrice est la somme de toutes les « contributions au χ^2 » de ses cellules
4. Le pourcentage des contributions de T_1 et T_2 par rapport au χ^2 de R donne les contributions relatives de T_1 et T_2 au χ^2 de T

Note sur le χ^2 : ses degrés de liberté

$$\chi^2(R) = \chi^2(T_1) + \chi^2(T_2)$$

$$2491 = 1998 + 493$$

Attention à considérer le χ^2 en proportion de la richesse en information le la matrice = de son nombre de ddl.

À partir des distributions marginales on peut obtenir plusieurs matrices T_n, mais pour chaque ligne et chaque colonne, la dernière "case" est imposée par la contrainte du total marginal

Définition :

- On appelle degré de liberté par ligne (ddll) le nombre de colonnes (de modalités) diminué de 1.
- On appelle degré de liberté par colonne (ddlc) le nombre de lignes (de modalités) diminué de 1.
- Le degré de liberté du khi-deux de la matrice est le produit ddll x ddlc = ddl.
- Pour une matrice donnée, le χ^2 à prendre en compte est en fait χ^2 / ddl

http://brassens.upmf-grenoble.fr/IMSS/MathSHS/SHS1/Stat1/Diapo/COURS9_fichiers/frame.htm

Matrice T_1 maximisant le χ^2 dans notre cas

$$\chi^2(R) = \chi^2(T_1) + \chi^2(T_2)$$

2491 = 1998 + 493

100% = 80.2% + 19.8%

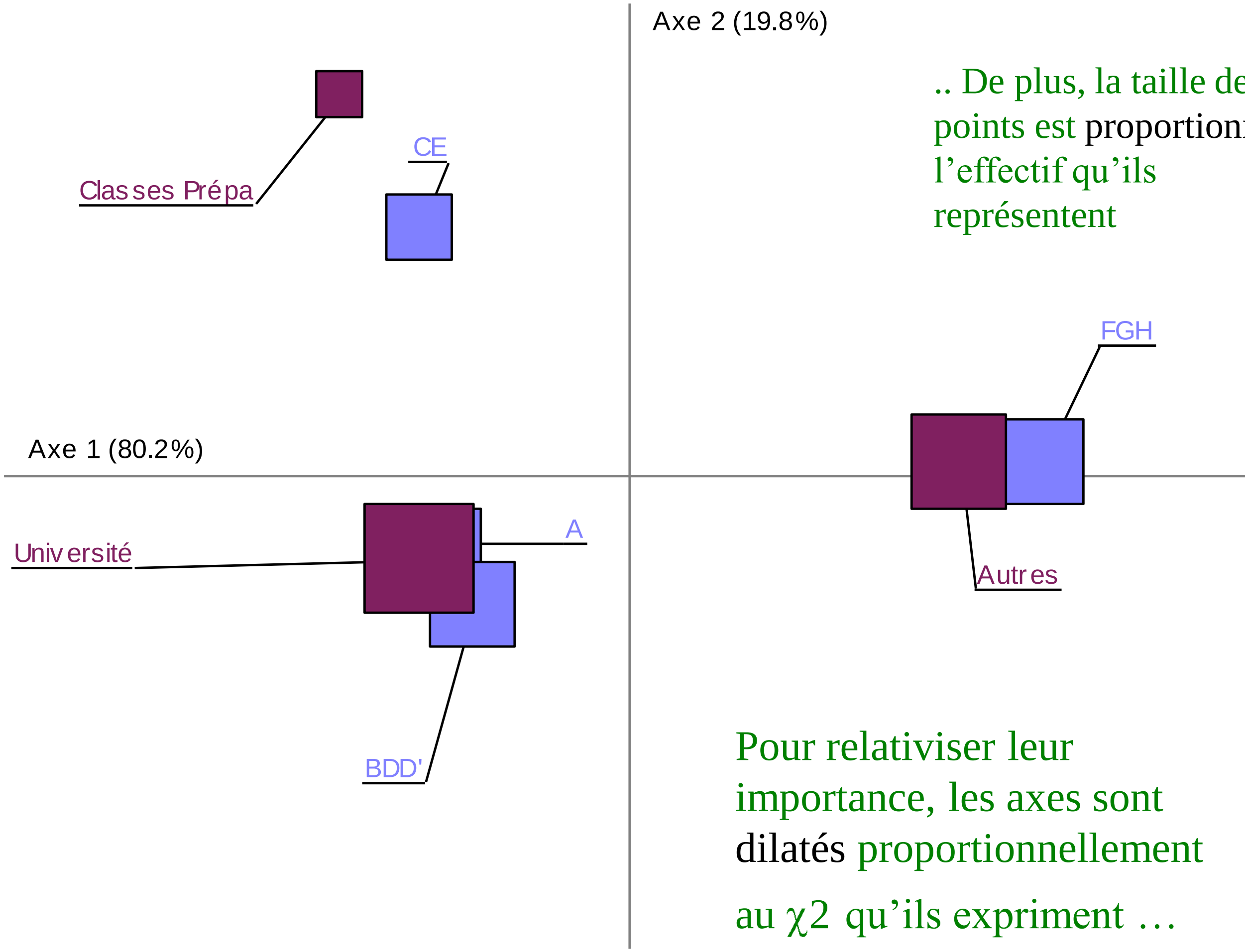
Cette ‘concentration’ de ce que l’on appelle le pourcentage de la variance expliquée par un axe est particulièrement intéressante lorsque la taille du tableau de données augmente…

$$\chi^2(R) = \chi^2(T_1) + \chi^2(T_2) + \chi^2(T_3) + \chi^2(T_4) \;..$$

Pourquoi ?

➔ On ne peut que représenter que deux axes à la fois sur un mapping … autant représenter les plus significatifs.

On obtient alors ce nouveau mapping



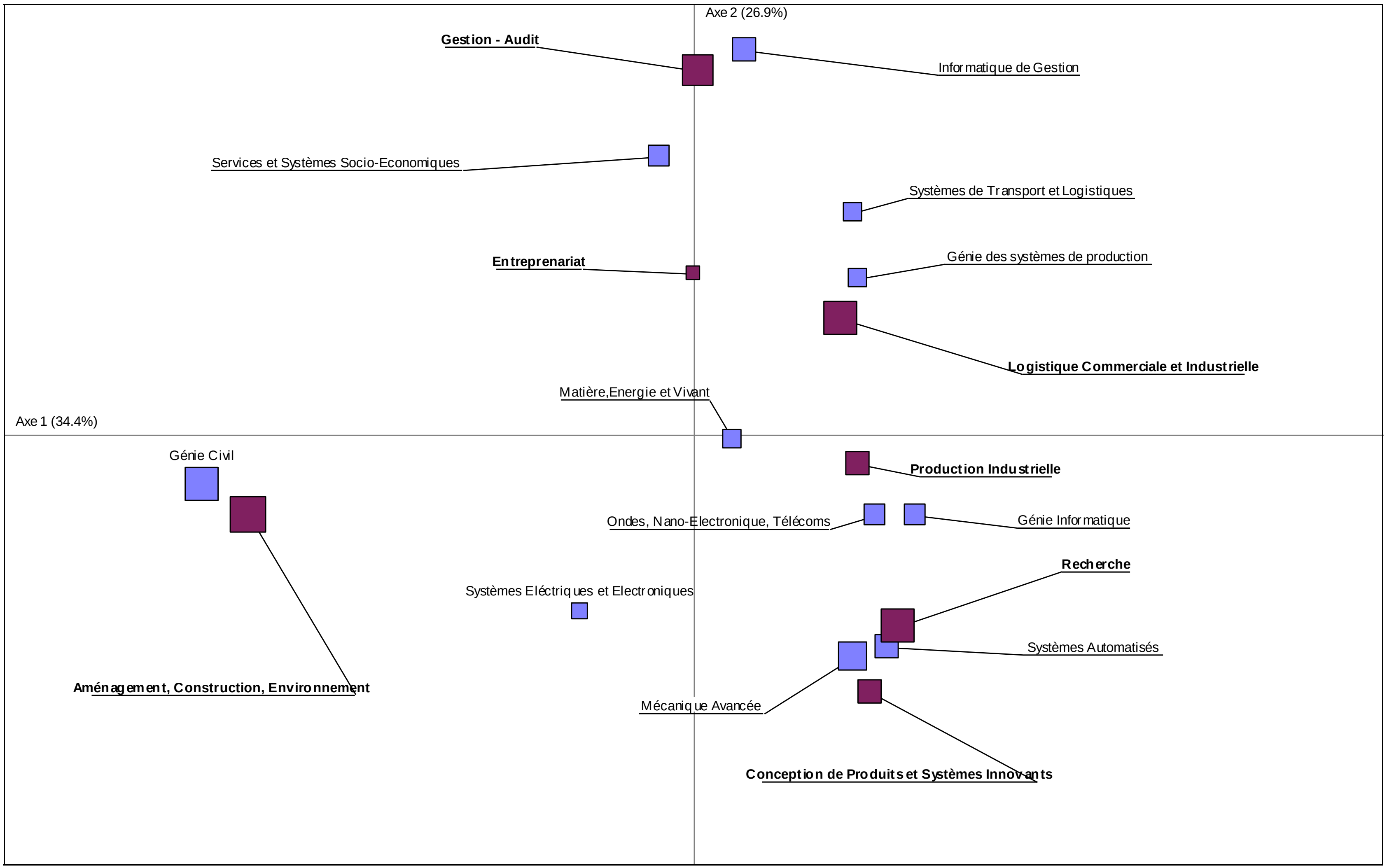
.. De plus, la taille des points est proportionnelle à l'effectif qu'ils représentent

Pour relativiser leur importance, les axes sont dilatés proportionnellement au $\chi 2$ qu'ils expriment …

Application : quels souhaits d'options?

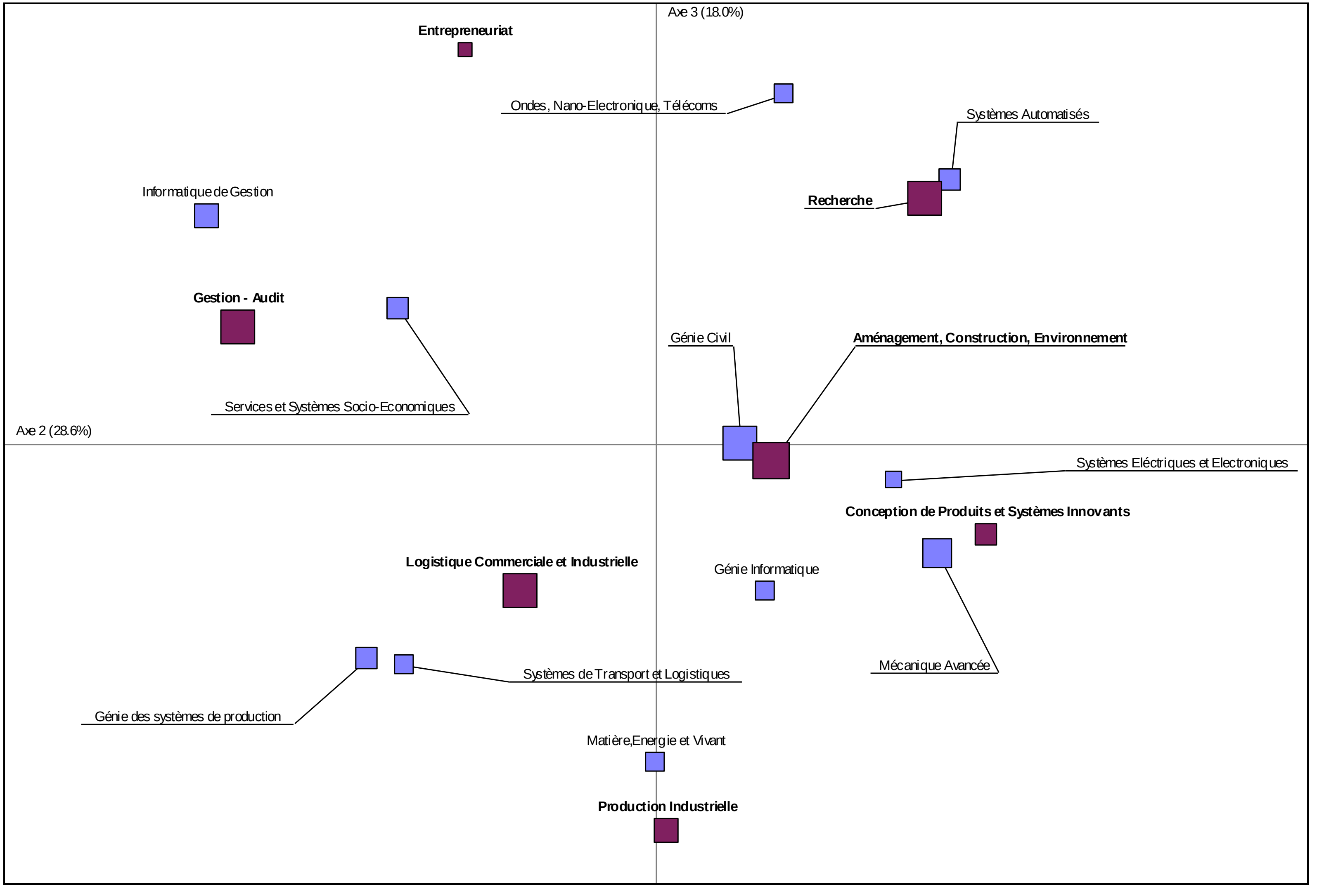
Premiers vœux 2003 de Génie / filière.	Entrepreneuriat	Gestion - Audit	Aménagement, Construction, Environnement	Conception de Produits et Systèmes Innovants	Production Industrielle	Logistique Commerciale et Industrielle	Recherche
Mécanique Avancée	0	0	2	7	5	1	6
Génie Civil	1	2	24	0	0	1	0
Matière,Energie et Vivant	0	1	2	0	5	1	1
Ondes, Nano-Electronique, Télécoms	2	1	0	1	0	1	6
Systèmes Electriques et Electroniques	0	0	3	2	0	1	1
Systèmes Automatisés	0	0	1	1	0	2	10
Génie des systèmes de production	0	5	0	0	4	4	0
Génie Informatique	0	0	0	3	1	5	2
Informatique de Gestion	2	11	0	0	0	2	1
Services et Systèmes Socio-Economiques	1	6	3	0	0	2	1
Systèmes de Transport et Logistiques	0	2	0	0	1	8	0

Mapping des choix de filière / génie



Premiers choix de génie / filière des 147 G2 en 2003

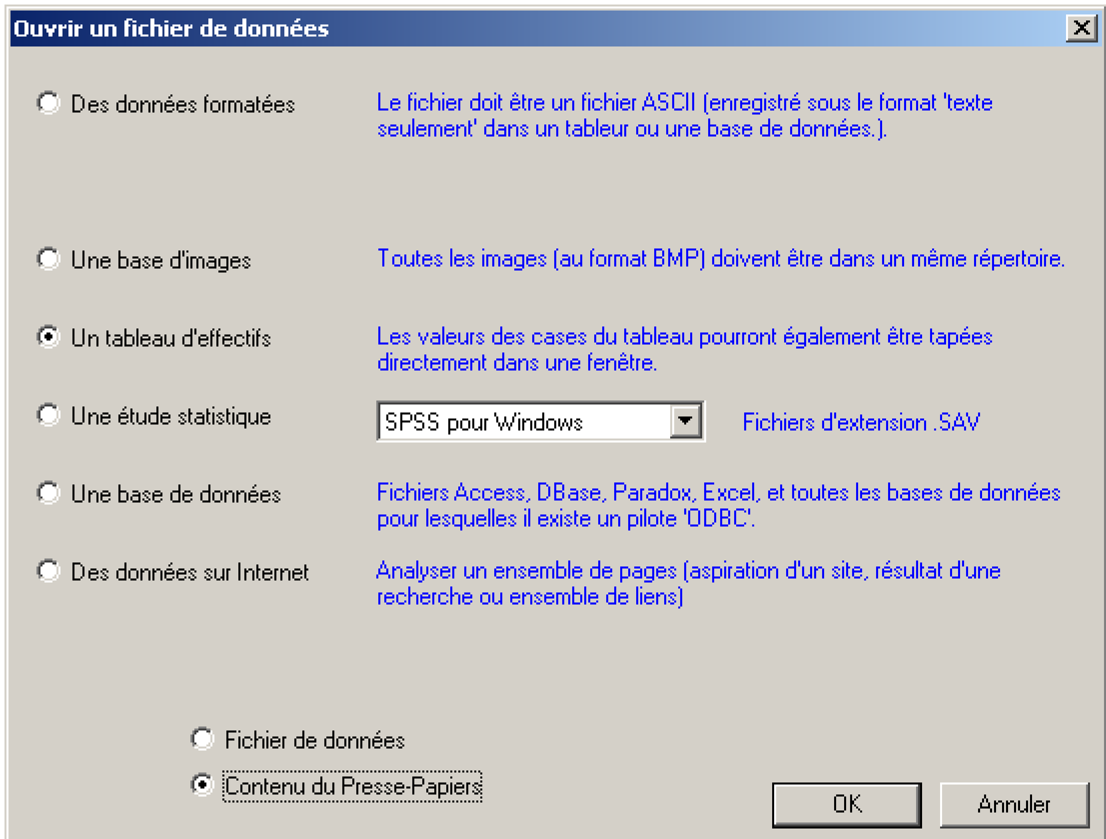
C'était les deux premiers axes = 62% de la variance expliquée On peut aussi regarder l'axe 3.. = 18%



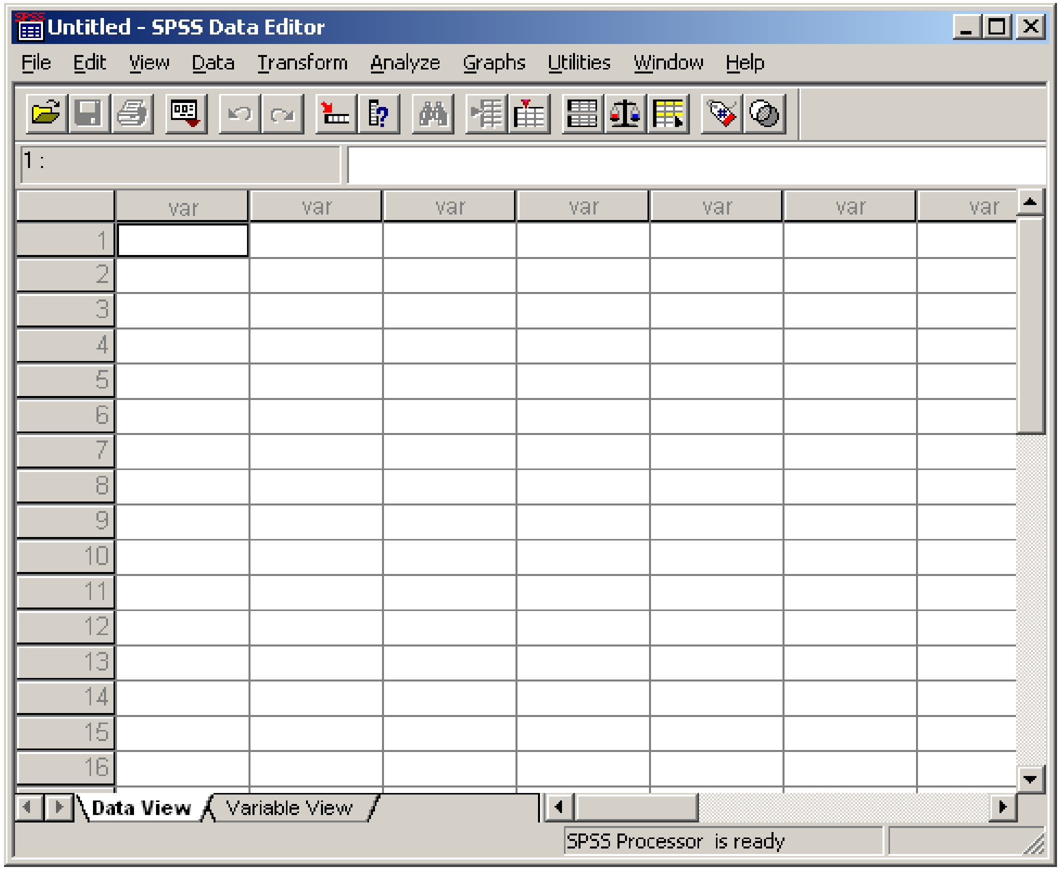
Conclusion

1. **Mise en œuvre logicielle**
   - **Sphinx, SPSS, SAS**
2. **Généralisation de l'AFC**
   - **Comparaison avec l'Analyse en Composantes Principales (ACP)**
   - **Généralisation de l'AFC**
   - **Pour approfondir**

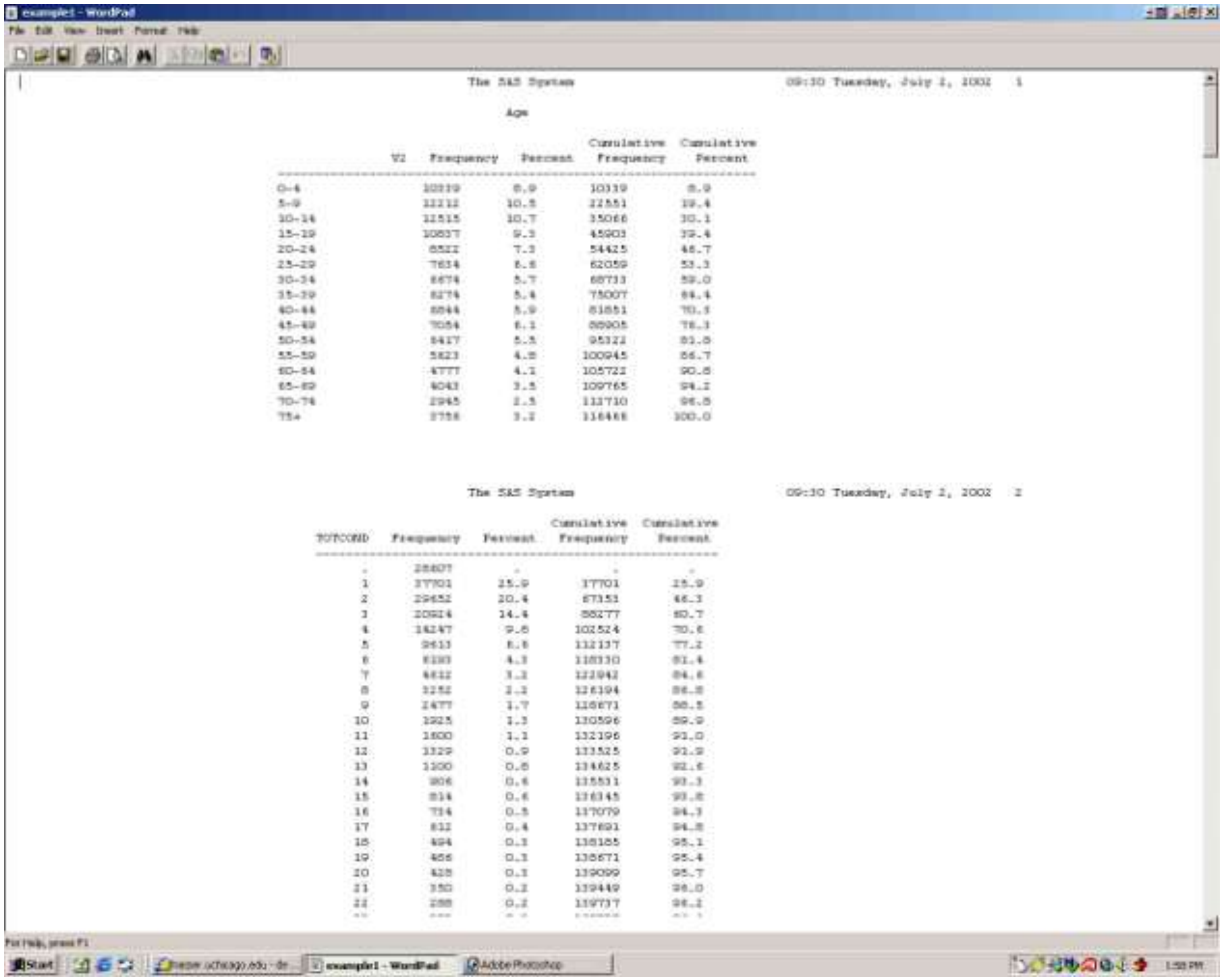
Mise en œuvre logicielle de l'AFC : Sphinx



Mise en œuvre logicielle : SPSS



Mise en œuvre logicielle : SAS



The SAS System — 09:30 Tuesday, July 2, 2002 — 1

Age

V2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-4	10319	8.9	10319	8.9
5-9	12232	10.5	22551	19.4
10-14	12515	10.7	35066	30.1
15-19	10837	9.3	45903	39.4
20-24	8522	7.3	54425	46.7
25-29	7634	6.6	62059	53.3
30-34	6674	5.7	68733	59.0
35-39	6274	5.4	75007	64.4
40-44	6844	5.9	81851	70.3
45-49	7054	6.1	88905	76.3
50-54	6417	5.5	95322	81.8
55-59	5623	4.8	100945	86.7
60-64	4777	4.1	105722	90.8
65-69	4043	3.5	109765	94.2
70-74	2945	2.5	112710	96.8
75+	3758	3.2	116468	100.0

The SAS System — 09:30 Tuesday, July 2, 2002 — 2

TOTCOND	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	28607	.	.	.
1	37701	25.9	37701	25.9
2	29652	20.4	67353	46.3
3	20924	14.4	88277	60.7
4	14247	9.8	102524	70.6
5	9613	6.6	112137	77.2
6	6193	4.3	118330	81.4
7	4612	3.2	122942	84.6
8	3252	2.2	126194	86.8
9	2477	1.7	128671	88.5
10	1925	1.3	130596	89.9
11	1600	1.1	132196	91.0
12	1329	0.9	133525	91.9
13	1100	0.8	134625	92.6
14	906	0.6	135531	93.3
15	814	0.6	136345	93.8
16	734	0.5	137079	94.3
17	612	0.4	137691	94.8
18	494	0.3	138185	95.1
19	486	0.3	138671	95.4
20	428	0.3	139099	95.7
21	350	0.2	139449	96.0
22	288	0.2	139737	96.2

Généralisations de l'AFC

- Les 'catégories' des questionnaires sont souvent mutuellement exclusives :
  - Sexe : H ou F
  - Politique : gauche, centre, droite
  - ➔ Tableau disjonctif

- Aux croisements de plus de deux caractéristiques : Analyse des Composantes Multiples (ACM)
  - Bac X Orientation X sexe
  - ➔ Tableau de Burt

$$
\left(\begin{array}{ccc|cc|ccc}
1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 2 & 0 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 2 & 1 & 1 & 1 & 0 & 0 \\
\hline
0 & 1 & 1 & 2 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 3 & 1 & 1 & 1 \\
\hline
0 & 1 & 1 & 1 & 1 & 2 & 0 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 2 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1
\end{array}\right)
$$

Autre méthode d'analyse de données proche : l'Analyse en Composantes Principales

	AFC	ACP
Données	Catégorielles	Métriques
Décomposition	T - T0 = T1 + T2	T = T1 + T2 + T3
Métrique	χ^2 pondéré	χ^2

Attention, le poids des cellules à faible effectif est renforcé

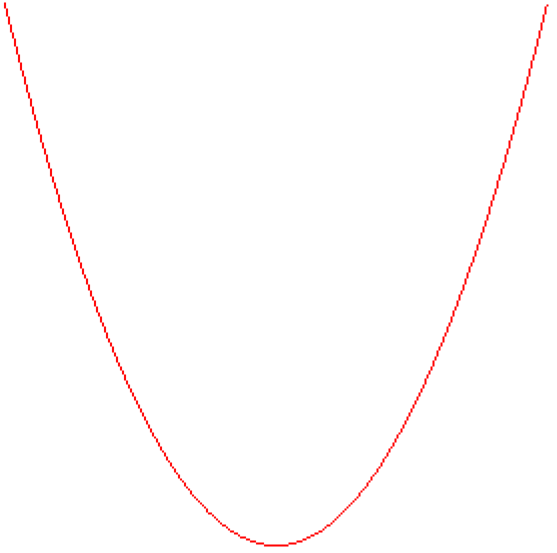
Rapports entre ACP et AFC

- Si on a des données permettant de faire une AFC, peut-on y appliquer une ACP ?
  - Non
- Si on a des données permettant de faire une ACP, peut-on y appliquer un AFC ?
  - Oui !
- .. Mais alors ?
  - .. Alors on traite les données numériques, les nombres comme des catégories
  - Si par exemple on travaille sur des notes, 18/20 n'est plus « supérieur à » 10/20, il n'est pas non plus « plus proche » de 16/20 que de 10/20.

Effet particulier lorsque l’on traite des Likert

- Que voit-on sur une AFC s’il existe une relation linéaire entre deux Likert corrélées, comme par exemple
  - Q1 Aimez-vous les mathématiques (beaucoup/assez/un peu/pas du tout)
  - Q2 Avez-vous de bonnes notes en mathématiques (très bonnes/bonnes/moyennes/mauvaises)

→ Les points du mapping suivent une parabole (c’est l'effet Guttman)



Pour en savoir plus

- Approches simples : rares
  - Site web de Philippe Cibois, professeur émérite de sociologie
    - texte d’où est tiré l’exemple développé dans ce cours
    - Trideux : logiciel libre de dépouillement d'enquête
  - Analyse factorielle des correspondances dans Wikipédia
  - Leçon Analyse factorielle des correspondances du CNAM
- Plus complexe : de nombreuses références
  - "Statistique textuelle" de Lebart et Salem, Chapitre 3
  - …

Autres cours de méthodologie:

1. Explorer ou vérifier ? Deux catégories d'approches
2. Éventails des démarches de recueil de données
3. Conception de questionnaires
4. Techniques d'entretien et reformulation
5. Validité et Fiabilité des données
6. Mesurer, tester des hypothèses

Merci de votre attention !

Rémi BACHELET

Enseignant-chercheur,
Ecole Centrale de Lille

Mon CV est disponible http://rb.ec-lille.fr

Mes principaux cours à Centrale
Gestion de projet, sociologie des organisations, recueil, analyse et traitement de données, prévention du plagiat, module de marchés financiers, cours de qualité et méthodes de résolution de problèmes, établir des cartes conceptuelles, utiliser Wikipédia et CentraleWiki, formation au coaching pédagogique et à l'encadrement