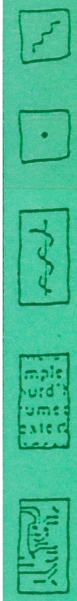
ITUT DE RECHERCHE SUR L'ENSEIGNEMENT DES MATHÉMATIQUES



Analyse factorielle des correspondances

Michel VENDRELY
IREM de BESANÇON

INSTITUT DE RECHERCHE SUR L'ENSEIGNEMENT DES MATHÉ

Analyse factorielle des correspondances

Michel VENDRELY
IREM de BESANÇON

I PRESENTATION

A) En quoi consiste l'analyse factorielle des correspondances.

L'étude des correspondances en statistique, est le traitement des tableaux de contingence (ou tableaux croisés). Il s'agit de tableaux d'effectifs obtenus en croisant les modalités de deux caractères définis sur une même population.

Par exemple, la répartition d'élèves de terminales selon le type de baccalauréat et selon les orientations après le baccalauréat donne le tableau suivant:

Orientation / Types de Baccalauréat	Université	Classes préparatoires	Autres orientations
L	1300	200	500
E	2000	200	800
S	1000	500	500
T	700	100	2200

L'analyse factorielle remplace un tableau A des contingences *de deux caractères qualitatifs*, difficile à lire, par une somme de tableaux plus simples: A_0, A_1,...A_n (les éléments de A sont les sommes des éléments correspondants des tableaux A_0, A_1,...A_n).

La décomposition du tableau A possède une propriété analogue au développement d'une fonction en série entière. Le premier tableau A_0 est la meilleure approximation possible de A, le deuxième est une correction apportée à la première approximation, le troisième n'est qu'une correction de correction, c'est à dire une correction de deuxième espèce et les corrections suivantes sont de ce fait négligeables.

Le tableau A_0 a ses marges M et P qui en constituent un résumé parfait, c'est à dire qu'il s'obtient par simple « produit de facteurs » $M \times P$. Il en est de même des autres tableaux A_1,...A_n, nous pouvons donc écrire: $A = A_0 + A_1 + ... + A_n = M \times P + X' \times X + Y' \times Y + ... + Z' \times Z$,expression qui rappelle celle du produit scalaire.

Cet article a pour but de faire comprendre l'analyse factorielle sans trop faire appel à la théorie mathématique qui la valide. Il propose d'acquérir une expérience personnelle des méthodes statistiques en faisant exécuter des calculs de vérifications de ces méthodes sur un exemple simple.

L'utilisation du produit scalaire éclaire la lecture des graphiques factoriels, comme on le verra par la suite.

B) Contenu et utilisation du document.

Cadre et connaissances mises en jeu : Les exemples qui suivent peuvent être proposés en activité d'option, en première ES sur les tableaux statistiques, les pourcentages, le produit scalaire et les systèmes linéaires.

Intentions pour les élèves: Le but est d'utiliser les propriétés graphiques du produit scalaire de deux vecteurs pour apprécier les informations données par un tableau croisé de deux caractères.

Les notions de pourcentages et les définitions du produit scalaire sont **supposées connues.**

Limite : L'analyse décrite dans ce document reste grossière. Pour l'affiner il faut utiliser le critère du Khi-deux qui déborde le cadre de ce travail mais qui est exploité dans les références indiquées.

Elaboration d'exemples : Le livre d'Economie des élèves possède de nombreux tableaux que l'on peut exploiter directement avec un tableur, en utilisant la méthode des puissances itérées décrite par Philippe Cibois. Cette méthode est décrite dans le paragraphe V

On peut ***ce servir de ce document*** pour ***approcher*** et faire comprendre les méthodes de l'analyse factorielle des correspondances (paragraphe II), pour ***utiliser*** et faire utiliser ces méthodes (paragraphe III et IV) enfin pour ***exploiter d'autres exemples***, en construisant les graphiques factoriels grâce à l'algorithme décrite dans le paragraphe V.

Ce travail a été ***réalisé grâce*** au soutien de la Direction des Lycée et Collèges ***(D.L.C)*** et de la Mission Académique pour la Formation des Personnel de l'Education Nationale ***(M.A.F.P.E.N),*** avec le support logistique de la Direction Générale de l'Enseignement Supérieur ***(D.G.S.E),*** qu'ils en soient remerciés.

Je remercie pour leurs relectures approfondies ***Brigitte CHAPUT*** de la commission Inter I.R.E.M ainsi que les membres du Groupe « Probabilité » de l'***I.R.E.M de Besançon*** : ***Jean-Pierre GRANGE***, le directeur de l'I.R.E.M ***Yves DUCEL*** et ***Michel HENRY*** dont les encouragements et les conseils me furent précieux.

Table des Matières

II Mise en place des objets et des méthodes de l'analyse.

A) Exemple utilisé

Exemple tiré du fascicule « L'enseignement des mathématiques en première ES. Groupe Lycée. I.R.E.M Besançon ».

1.) Présentation d'une statistique.

Le tableau suivant donne la répartition par âges et sections d'élèves de terminale en 1993. Les effectifs étant ramenés à la base de 10 000 élèves pour faciliter la lecture des pourcentages arrondis à 10^{-2} prés.

Tableau A.

Terminales	TA	TB	TC	TD	TE	TF	Total M
Elèves ayant plus de 19 ans	484	404	191	487	188	1686	3440
Elèves ayant 19 ans	425	559	106	442	150	1135	2817
Elèves ayant moins de 19 ans	638	560	903	505	540	597	3743
Total **P**	1547	1523	1200	1434	878	3418	10000

Les effectifs marginaux (total colonne M comme « main » et ligne P comme « pied ») donnent les répartitions de la population suivant chacune de ces différentes modalités. Ils permettent la reconstitution d'une répartition de référence théorique basée sur une absence de liaison entre les lignes et les colonnes comme le présente l'exemple suivant (partie grisée du tableau ci-dessus).

En moyenne, 34,18% des élèves de lycée sont en TF et, 37,43% des élèves du lycée sont sans retard de scolarité. On pourrait donc s'attendre, si l'orientation et le retard scolaire n'étaient pas liés, à ce que les élèves de TF n'ayant pas de retard représentent:

0,3418× 0,3743 × 100, soit 12,79 % des élèves de lycée. On constate qu'en réalité ces élèves représentent 5,97% de la population.

2.) Tableau théorique de référence.

Il donne le niveau zéro des liaisons des caractères. Il est construit sur la base de l'hypothèse théorique d'indépendance entre classes d'âge et sections.

Par exemple : (partie grisée du tableau ci-dessous) 0,3418×0,3743 = 0,1279, soit 12,79%.

On reconstruit ainsi un tableau A_0, à partir des marges M et P du tableau A, sous l'hypothèse que ces deux caractères sont indépendants.

Tableau A_0.

Terminales.	TA	TB	TC	TD	TE	TF	Main M
Elèves ayant plus de 19 ans	532	524	413	493	302	1176	3440
Elèves ayant 19 ans	436	429	338	404	247	963	2817
Elèves ayant moins de 19 ans	579	570	449	537	329	1279	3743
Pied **P**	1547	1523	1200	1434	878	3418	10000

On peut dire que A_0 = M×P représente le mieux possible la répartition « moyenne » de la population.

Comparons maintenant le tableau donné A avec le tableau théorique A_0 en calculant les *écarts*.

3.) Tableau des *écarts* à la référence.

Il est calculé par la différence des deux premiers tableaux.($E_1 = A - A_0$) Il donne *les sous et sur-représentations* du premier tableau.

Tableau E_1..

Terminales.	TA	TB	TC	TD	TE	TF	Total
Elèves ayant plus de 19 ans	-48	-120	-222	-6	-114	510	0
Elèves ayant 19 ans	-11	130	-232	38	-97	172	0
Elèves ayant moins de 19 ans	59	-10	454	-32	211	-682	0
Total	0	0	0	0	0	0	0

Par exemple : La valeur **-682** grisée du tableau E_1 ci-dessus est calculée à partir de **597** du tableau A et de **1279** du tableau A_0: (597 - 1279 = -682). Cette valeur montre que les élèves âgés de moins de 19 ans de la classe TF, sont en *sous représentation* de 682 unités par rapport à la référence théorique fabriquée dans le tableau A_0. On peut dire que les modalités « être âgés de moins de 19 ans et appartenir à la classe TF » s'opposent car le déficit de 682 individus sur l'effectif théorique attendu est important.
La valeur **510** montre que les élèves âgés de plus de 19 ans, de la classe TF sont en *sur représentation* de 510 unités par rapport à la référence fabriquée dans le tableau A_0. Nous dirons que ces deux modalités s'attirent.

Plus généralement :
Il y a *attraction*, entre deux modalités (ou *sur-représentation*) quand l'écart correspondant dans le tableau E_1 est positif.
Il y a *conformité avec la référence statistique théorique donnée dans le tableau* A_0 quand l'écart correspondant dans le tableau E_1 est nul ou presque nul.
Il y a *opposition* des modalités (ou *sous-représentation*) quand l'écart correspondant dans le tableau E_1 est négatif.

*Remarque :*La construction de A_0 à partir des marges de A fait que dans le tableau E_1 la somme des éléments d'une ligne ou d'une colonne est nulle.

B) Construction des objets du graphique.

Pour chercher les vecteurs servant à la représentation graphique il faut décomposer A sous la forme $A = A_0 + E_1 = A_0 + A_1 + ... + A_n = M \times P + X' \times X + Y' \times Y + ... + Z' \times Z$.

1.) Principe de la méthode :

Pour calculer le tableau A_1 à partir de E_1 , on recherche une matrice ligne X et une matrice colonne X' en utilisant une méthode itérative décrite au paragraphe V (Cf. page 18), et on construit A_1 par produit matriciel $A_1 = X' \times X$.

Pour calculer le tableau A_2 à partir de $E_2 = E_1 - A_1$. On recherche deux matrices Y et Y' par la même méthode, et on construit A_2 par produit $A_2 = Y' \times Y$.

2.) Résultats numériques obtenus : les tableaux.

Chaque écart du tableau E_1 est décomposé sous la forme x'x + y'y comme dans l'expression analytique du produit scalaire de deux vecteurs du plan dans une base orthonormale.

si $\vec{u} = x.\vec{i} + y.\vec{j}$ et $\vec{u}' = x'.\vec{i} + y'.\vec{j}$ alors $\vec{u}.\vec{u}' = xx' + yy'$

Tableau E_1 (rappel)

Ecarts x'x+ y'y	$\vec{A}$	$\vec{B}$	$\vec{C}$	$\vec{D}$	$\vec{E}$	$\vec{F}$
$\vec{P}$	-48	-120	-222	-6	-114	510
$\vec{N}$	-11	130	-232	38	-97	172
$\vec{M}$	59	-10	454	-32	211	-682

Tableau A_1

Produits x'x	$\vec{A}$	$\vec{B}$	$\vec{C}$	$\vec{D}$	$\vec{E}$	$\vec{F}$	x'
$\vec{P}$	-36	0	-270	18	126	414	**18**
$\vec{N}$	-24	0	-180	12	-84	276	**12**
$\vec{M}$	60	0	450	-30	210	-690	**-30**
x	**-2**	**0**	**-15**	**1**	**-7**	**23**	

Tableau A_2

Produits y'y	$\vec{A}$	$\vec{B}$	$\vec{C}$	$\vec{D}$	$\vec{E}$	$\vec{F}$	y'
$\vec{P}$	-12	-120	48	-24	12	96	**12**
$\vec{N}$	13	130	-52	26	-13	-104	**-13**
$\vec{M}$	-1	-10	4	-2	1	8	**1**
y	**-1**	**-10**	**4**	**-2**	**1**	**8**	

La somme des deux tableaux A_1 et A_2 reconstitue le tableau E_1 des *écarts*.

3.) Extraction des vecteurs servant à la représentation graphique.

Les marges en x'x et y'y des tableaux A_1 et A_2 donnent les coordonnées de ces vecteurs.

Caractères	Classes						Ages			
modalités	TA	TB	TC	TD	TE	TF	plus de 19 ans	19 ans	moins de 19 ans	
Vecteurs	$\vec{A}$	$\vec{B}$	$\vec{C}$	$\vec{D}$	$\vec{E}$	$\vec{F}$	$\vec{P}$	$\vec{N}$	$\vec{M}$	
Abs x	-2	0	-15	1	-7	23	18	12	-30	Abs x'
Ord y	-1	-10	4	-2	1	8	12	-13	1	Ord y'

Nous avons vu en page 4 que la somme des lignes et des colonnes dans E_1 est nulle. Cela ce traduit par : $\vec{A}+\vec{B}+\vec{C}+\vec{D}+\vec{E}+\vec{F}=\vec{P}+\vec{N}+\vec{M}=\vec{0}$

Exemple : décomposition de l'*écart* grisé du tableau E_1.

Le calcul du produit scalaire donne $\vec{F}.\vec{M} = 23 \times (-30) + 8 \times 1 = -690 + 8 = -682$. On retrouve *exactement* l'écart observé dans le tableau E_1, la valeur -690 se trouvant aussi dans le tableau A_1 et la valeur 8 se trouvant dans le tableau A_2.

Nous pourrions vérifier tous les produits scalaires dans les tableaux A_1 et A_2 qui précèdent.

C) Description des méthodes utilisées dans l'analyse du graphique

1.) Aspect graphique du produit scalaire:

$\vec{u}.\vec{v} = \|\vec{u}\|.\|\vec{v}\|.\cos(\vec{u},\vec{v})$.

Le produit scalaire de deux vecteurs est le produit des normes des deux vecteurs par le cosinus de l'angle de ces deux vecteurs. Donc plus les vecteurs sont « grands et rapprochés » plus le produit scalaire et l'écart qu'il représente, est élevé.

2.) Applications.

a) On peut apprécier graphiquement les prédominances.

Si le produit scalaire représente un écart positif, l'angle des deux vecteurs est aigu. Il y a *attraction* (ou *sur-représentation*) des modalités représentées par ces vecteurs. Nous dirons que les vecteurs sont « proches ».

L'intensité de l'attraction dépend à la fois des normes des vecteurs et de leur angle.

Si l'angle des deux vecteurs est droit, le produit scalaire représente un écart nul. Il y a *indépendance* statistique ou *quadrature* des modalités représentées.

Si l'un des vecteurs est presque nul, la modalité du caractère qu'il représente est conforme à la référence statistique théorique induite par les marges du tableau initial.

Si le produit scalaire représente un écart négatif, l'angle des deux vecteurs est obtus. Il y a *opposition* (ou *sous-représentation*) des modalités. Celle-ci est d'autant plus grande que l'angle des vecteurs se rapproche de l'angle plat et que leurs normes sont grandes.

b) Représentativité des axes du repère.

Nous avons vu en page 1 que A_1 représente le mieux possible E_1. A_2 n'est qu'une correction de cette approximation.

Chaque tableau étant représenté par un axe, le premier axe représente le mieux possible des écarts, le second n'est qu'une correction apportée à cette première approximation.

Vérifions cette propriété par quelques exemples.

Exemple de représentativité de l'axe des abscisses.

Le produit scalaire $\vec{M} \times \vec{C} = (-30) \times (-15) + 4 \times 1 = 454$ est décomposé en deux nombres. L'un xx' = $(-30)\times(-15) = 450$ du tableau A_1 représente $450/454\times100 = 99$ % du produit scalaire, donc du lien entre les modalités TC et « moins de 19 ans » repéré par l'écart 454 du tableau E_1.

Graphiquement les vecteurs $\vec{M}$ et $\vec{C}$ sont proches de l'axe des abscisses, ils transmettent à cet axe une grande part de leur attraction.

Plus généralement,

les valeurs xx' du tableau A_1 sont de bonnes approximations du tableau E_1 des écarts (sauf pour ceux formés avec $\vec{B}$).

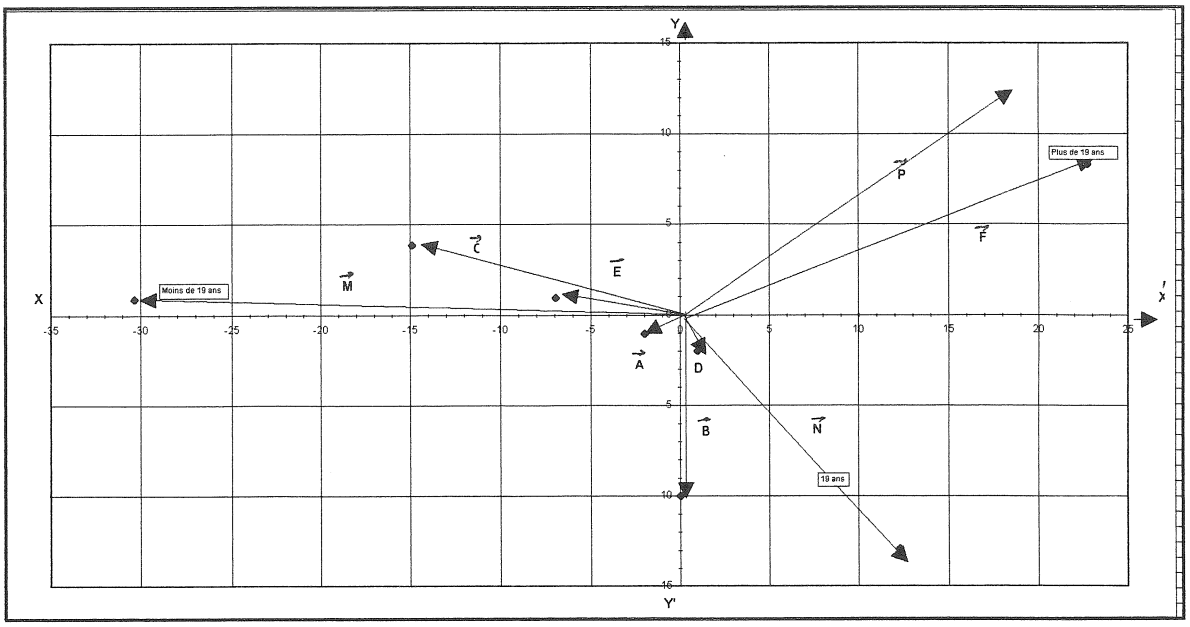
La représentativité d'un tableau peut être donnée par le rapport entre la somme des carrés des éléments qui le constituent et la somme des carrés des écarts du tableau E_1. On calculerait qu'ici l'axe des abscisses représente 92 % des écarts et l'axe des ordonnées les 8 % restant.

Le poids de cet axe dans la représentativité des écarts est visualisé par le fait que les vecteurs lui sont « proches ».

Conclusion :

La part des « x'x » (tableau A_1) des écarts du tableau E_1 est plus importante que celle des « y'y » (tableau A_2). L'axe des abscisses représente 92 % de l'information donnée par le tableau A

c) Identification des axes du repère.



Pour l'axe des abscisses, x'Ox.

Le vecteur $\vec{M}$ est « proche » du demi axe Ox', qui de ce fait, représente bien les élèves âgés de moins de 19 ans.

Les vecteurs $\vec{P}$ et $\vec{N}$, sont « proches » du demi axe Ox qui de ce fait, représente bien les élèves les plus âgés.

L'axe des abscisses qui représente la tendance dominante oppose les élèves les plus jeunes, à gauche, aux plus âgés, à droite (voir le graphique).

Il oppose également les sections scientifiques TC et TE, à gauche, à la section TF à droite. Les vecteurs « proches » de $\vec{0}$ ($\vec{A}$ et $\vec{D}$) ou perpendiculaires ($\vec{B}$) n'apportent rien à cet axe.

Pour l'axe des ordonnées y'Oy.

Cet axe représente dans une proportion de 8 % seulement la deuxième tendance. $\vec{P}$ est proche de la partie positive de cet axe et représente donc les élèves les plus âgés. La partie négative grâce à $\vec{N}$, représente les élèves de 19 ans.

L'axe des ordonnées représente donc les individus les plus âgés : il oppose (en haut) les élèves les plus âgés à ceux de 19 ans (en bas). Il oppose aussi les modalités TF (en haut) et TB (en bas).

D) Utilisation des objets et des méthodes.

Résultat de l'analyse factorielle des âges des bacheliers. (utiliser le graphique)

1.) Les proximités.

a) *Proximité des modalités de caractères différents.*

Les modalités TC et TE sont 'attirées' par la modalité « moins de 19 ans » : on dira que les classes scientifiques sont plutôt composées d'élèves sans retards de scolarité.

La modalité TF est en opposition avec la modalité « moins de 19 ans » (l'angle formé par le vecteur $\vec{F}$ et le vecteur $\vec{M}$ est presque plat).

La modalité TF est attirée par la modalité « plus de 19 ans ».

Comme $\vec{A}$ et $\vec{D}$ sont presque nuls, ils ne visualisent aucune tendance (les classes de TA et TD sont en situation conforme à la référence statistique théorique induite par les marges du tableau initial).

$\vec{B}$ est proche de $\vec{N}$ mais il est en situation d'indépendance (angle droit) avec $\vec{M}$. On peut dire que les élèves de TB sont âgés en moyenne de 19 ans.

b) *Proximité des modalités d'un même caractère.*

Si deux modalités d'un même caractère sont proches, leurs analyses factorielles se ressemblent et on a probablement intérêt à les réunir en une seule modalité.

C'est le cas des modalités TC et TE, qui ont des comportements proches car attirés par la modalité « moins de 19 ans ». C'est aussi le cas des modalités TA et TD, qui sont proches de O.

Les modalités « 19 ans » et « plus de 19 ans » sont indépendantes mais elles s'opposent globalement à la modalité « moins de 19 ans ».

c) *Proximité des modalités avec les axes du repère.*

Nous avons vu que la « proximité » des vecteurs permet *l'identification* des axes. Celui des abscisses, qui représente la tendance dominante, oppose à gauche les élèves sans retard de scolarité aux autres. Il oppose les plus scientifiques à TF.

L'axe des ordonnées représente les individus les plus âgés. Il oppose (en haut) les élèves les plus âgés à ceux de 19 ans (en bas). Il oppose aussi les modalités TF (en haut) et TB (en bas).

2.) Le classement.

L'ordre naturel « moins de 19 ans , 19 ans , plus de 19 ans » des modalités du caractère âge, permet de classer les terminales:

TC < TE < TB < TF . Les classes TA et TD, représentées par des vecteurs presque nuls, sont exclues de ce classement.

Ce classement est visualisé dans le graphique page 21 par un trait qui passe par les modalités « Plus de 19 ans », « 19 ans » et « moins de 19 ans » dans cet ordre.

Il permet au caractère quantitatif, de classer le caractère qualitatif.

3.) Introduction d'une nouvelle classe.

Imaginons que nous voulions placer dans cette analyse les élèves d'une classe de TG qui ne figurait pas dans le tableau initial.

Supposons qu'une statistique établie sur 1000 élèves de cette classe donne:

Ages	Répartition statistique
Plus de 19 ans	590
19 ans	299
Moins de 19 ans	111
Total	1000

Comparons cette classe à la répartition marginale donnée dans le tableau A de la page 4.

Ages	Répartition statistique	Marge M de A	Répartition de référence	Ecarts à la référence
Plus de 19 ans	590	34,4%	344	246
19 ans	299	28,17%	281,7	17
Moins de 19 ans	111	37,43%	374,3	-263
Total	1000	100 %	1000	0

Pour trouver les coordonnées (x, y) du vecteur $\vec{G}$ représentant la classe de TG, on écrit les conditions nécessaires portant sur les produits scalaires: $\begin{cases} \overrightarrow{G} \times \overrightarrow{P} = 246 \\ \overrightarrow{G} \times \overrightarrow{N} = 17 \\ \overrightarrow{G} \times \overrightarrow{M} = -263 \end{cases}$ il faut résoudre le système $\begin{cases} 18x + 12y = 246 \\ 12x - 13y = 17 \\ -30x + y = -263 \end{cases}$ chaque écart devant être représenté par un produit scalaire des vecteurs correspondants.

Ces équations sont compatibles car la somme des colonnes du tableau E_1 est nulle. Ce qui se traduit par : $\vec{0} = \vec{P} + \vec{N} + \vec{M}$, qui entraîne que le produit scalaire $\vec{G}.(\vec{P} + \vec{N} + \vec{M}) = \vec{G}.\vec{P} + \vec{G}\,\vec{N} + \vec{G}\,\vec{M} = 0$.

Il ne reste plus qu'à placer le vecteur, $\vec{G}\begin{pmatrix} 9 \\ 7 \end{pmatrix}$ solution du système, dans le graphique page 21, pour comparer la classe de TG aux autres classes.

On trouve TC < TE < TB < TG < TF.

Ce classement est légèrement différent de celui que nous pourrions faire avec les moyennes arithmétiques des âges.

Terminales.	TC	TE	TB	TG	TF
Ages moyens	18,4	18,5	18,8	19,4	19,3

Ce classement, donné principalement par l'axe des abscisses, est construit à partir des écarts à la référence A_0 des élèves les plus âgés et les moins âgés.

Terminales.	TC	TE	TB	TG	TF
Elèves de plus de 19 ans	-222	-114	-120	246	510
Elèves de moins de 19 ans	454	210	-10	-263	-682

Une correction étant apportée par l'axe des ordonnées pour TE et TB avec les élèves de 19 ans.

III Travaux pratiques : fiche élève.

1.) Enoncé .

Orientation des Bacheliers en 1975 (Statistiques du 15 octobre 1976). Ce tableau donne la répartition observée, à l'époque, des différentes orientations des bacheliers Littéraires (L), Economiques (E), Scientifiques (S) et Techniques (T) vers l'Université (U), vers les classes Préparatoires aux grandes écoles (P) et vers les Autres orientations (A).

(Exemple d'analyse factorielle extrait du livre de Philippe Cibois cité en référence 2)

2.) Tableaux numériques utilisés.

a) Tableau A des données statistiques.

Orientation des bacheliers.

Tableau A	U	P	A	Total M
L	1300	200	500	
E	2000	200	800	
S	1000	500	500	
T	700	100	2200	
Total P	5000	1000		10000

Questions 1:

Complétez le tableau A ci dessus.

Quel est le pourcentage des bacheliers allant à l'université ?

Quel est le pourcentage des bacheliers de série L. ?

b) Tableau A_0 des références théoriques.

Tableau A_0	U	P	A	Total
L			800	2000
E			1200	3000
S	1000	200	800	2000
T	1500	300	1200	3000
Total	5000	1000	4000	10000

Questions 2:

Si indépendamment de leur série, 50 % des bacheliers allaient à l'université, combien de bacheliers de série L devraient aller à l'université?

En utilisant le même raisonnement, compléter le tableau A_0.

Quel est le déficit (en pourcentage) des élèves du technique pour l'université.

c) *Tableau E des écarts au tableau de référence.*

Il est obtenu par différence entre les tableaux A et A_0.

Tableau E	U	P	A	Total
L			-300	
E			-400	
S	0	300	-300	
T	-800	-200	1000	
Total				

Les écarts positifs représentent des choix privilégiés pour les modalités correspondantes (on dit qu'il y a sur représentation).

Exemple : les modalités S et P sont en sur représentation de 3 %.

Les écarts négatifs représentent des déficits (il y a sous-représentation).

Les écarts nuls représentent des choix conformes à la référence donnée dans le tableau A_0.

Questions 3 :

Compléter le tableau E.

Citer les sur-représentations, les sous-représentations, et les orientations conformes à la répartition du tableau A_0.

3.) Utilisation des produits scalaires.

a) Rappel

Le produit scalaire de deux vecteurs dépend des composantes de ces vecteurs dans une base orthonormale :.

si $\vec{u} = x.\vec{i} + y.\vec{j}$ et $\vec{u}' = x'.\vec{i} + y'.\vec{j}$ alors $\vec{u}.\vec{u}' = xx' + yy'$

b) Résultats numériques de la décomposition factorielle.

Pour visualiser graphiquement les sur et sous représentations données dans le tableau E des écarts, il existe un procédé (qui dépasse le cadre de cette étude) qui permet de représenter chaque ligne et chaque colonne du tableau E des écarts par un vecteur. Les écarts indiqués dans ce tableau sont alors les produits scalaires de ces vecteurs.

Le tableau suivant donne les coordonnées (arrondies) des vecteurs du plan.

Vecteurs	Abscisses	Ordonnées	
L	-9	5	Vecteurs lignes
E	-11	14	
S	-10	-17	
T	31	-1	
U	-25	15	Vecteurs colonnes
P	-7	-13	
A	33	-2	

Question 4 :

Vérifier que le produit scalaire des vecteurs L et U correspond approximativement à l'écart donné dans le tableau E

c) Décomposition factorielle des écarts (ou des produits scalaires).

La vérification de tous les écarts par tous les produits scalaires des vecteurs lignes et colonnes correspondants est fastidieuse, elle est donnée ci-dessous :

Tableau des produits xx' et tableau des produits yy' pour la vérification des produits scalaires x x' + y y'.

Tableau A_1	U	P	A	x'
L	226,44	63,64	-290,08	-8,89
E	290,16	81,55	-371,71	-11,39
S	261,72	73,56	-335,28	-10,28
T	-778,32	-218,76	997,08	30,56
x	-25,47	-7,16	32,62	

+

Tableau A_2	U	P	A	y'
L	73,56	-63,64	-9,92	4,82
E	209,84	-181,55	-28,29	13,74
S	-261,72	226,44	35,28	-17,14
T	-21,68	18,76	2,92	-1,42
y	15,27	-13,21	-2,06	

Le tableau E des écarts est la somme des tableaux A_1 et A_2.

d) Les représentations graphiques.

Le but de cette représentation est de faire apparaître l'information pertinente en associant à chaque tableau de décomposition, un axe.

Le tableau A_1 des abscisses x et x' peut être représenté sur l'axe des abscisses.

x'						O							x
-30	-25	-20	-15	-10	-5	0	5	10	15	20	25	30	35

Travail 5 :

Placer sur cet axe, les abscisses des modalités des caractères étudiés et repérer graphiquement les sur et sous représentations.

Ce graphique ne contient qu'une partie de l'information car les situations d'indépendance y sont mal représentées.

Le tableau A_2 des ordonnées y et y' peut être représenté sur l'axe des ordonnées.

y'						O							y
-30	-25	-20	-15	-10	-5	0	5	10	15	20	25	30	35

Placer sur cet axe, les ordonnées des modalités des caractères étudiés.

e) Utilisation du produit scalaire des vecteurs du plan.

Travail 6 :

Dans un repère orthonormal tracer les vecteurs donnés par leurs coordonnées..

(Le graphique est donné en page 22)

Le produit scalaire $\vec{u}.\vec{v} = \|\vec{u}\|.\|\vec{v}\|.\cos(\vec{u},\vec{v})$ de deux vecteurs dépend des normes des deux vecteurs et de l'angle de ces deux vecteurs : plus les vecteurs sont « grands et rapprochés » plus le produit scalaire, donc l'écart qu'il représente, est important.
On peut donc apprécier graphiquement les prédominances des caractères.

Si le produit scalaire représente un écart positif, l'angle des deux vecteurs est aigu. Il y a *attraction* (ou *sur représentation*) des modalités représentées par les vecteurs.

L'intensité de la conjonction dépend à la fois des normes des vecteurs et de leur angle.

Si l'angle des deux vecteurs est droit, le produit scalaire représente un écart nul. Il y a *indépendance* statistique ou *quadrature* des modalités représentées.

Si le produit scalaire représente un écart négatif, l'angle des deux vecteurs est obtus. Il y a *opposition* (ou *sous représentation*) des modalités d'autant plus grande que l'angle des vecteurs se rapproche de l'angle plat et que leurs normes sont grandes.

Trouver graphiquement les conjonctions, les oppositions et enfin les quadratures.

f) Introduction d'une classe supplémentaire : la classe G.

Pour avoir une idée de l'évolution des orientations des bacheliers des sections économiques, on désire placer dans le plan ci-dessus, les bacheliers des sections économiques de 1985. Une statistique de cette classe, à cette époque, donne la répartition suivante

Orientations	Répartition statistique: classe G	Pourcentage moyen donné par la marge du premier tableau)	Réparation référence attendue	Ecarts à la référence
Université	1650	50 %	1000	650
Classes Prép.	210			
Autres	140			
Total	2000	100 %	2000	0

Travail 6 :

Compléter le tableau ci dessus.

Le vecteur $\vec{G}$ de coordonnées (x,y) représentant la classe appelée G des sections économiques de 1985, vérifie les égalités suivantes :

$$\begin{cases} \overrightarrow{G} \times \overrightarrow{U} = 650 \\ \overrightarrow{G} \times \overrightarrow{P} = 10 \\ \overrightarrow{G} \times \overrightarrow{A} = -660 \end{cases}$$

car chaque écart est représenté par un produit scalaire des vecteurs correspondants..

Remarque : $\vec{U} + \vec{P} + \vec{A} = \vec{0}$ car la somme des colonnes du tableau E est nulle,

on obtient le système $\begin{cases} -25x + 15y = 650 \\ -7x - 13y = 10 \\ -33x - 2y = -660 \end{cases}$

Résoudre ce système et placer le vecteur, $\overrightarrow{G}\begin{pmatrix} x \\ y \end{pmatrix}$ **solution du système, dans le graphique précédent.**

Comparer la classe G aux autres classes.

IV Exemple de représentation dans l'espace

1.) Enoncé.

D'après Sciences humaines N°45, dec 1994. Enquête FQP 1993 INSEE.

Dans un échantillon de 10000 familles françaises, en 1993, on a repéré la profession des parents (en colonne) et celle des enfants (en ligne).

Le tableau ci dessous donne la répartition en catégories socio-professionnelles des enfants et de leur parents.

Emploi des enfants / Des Parents	agriculteur	artisan	cadre	P indép	employé	ouvriers	Total
Agriculteur	428	135	179	254	139	606	1742
Artisan	23	409	299	279	93	277	1380
Cadre	5	90	442	173	70	56	835
P Indép	8	93	374	316	101	161	1053
Employé	2	81	246	357	123	299	1108
Ouvriers	32	339	379	943	417	1773	3882
Total	498	1146	1918	2321	943	3173	10000

2.) Résultats de l'analyse factorielle : les coordonnées des vecteurs

Ce tableau de 6 lignes et 6 colonnes nécessite la décomposition des écarts en produits scalaires dans un espace de dimension 5. Le plan des deux premiers axes en constitue une bonne projection horizontale (graphique page 23). Le plan des deuxième et troisième axes en constitue une projection verticale (graphique page 24).

Vecteurs	Abs X	Ord Y	Cote Z		Abs X	Ord Y	Cote Z
Parents				Enfants			
Agriculteur	22	-10	1	Agriculteur	12	-7	0
Artisan	-6	-4	-15	Artisan	-3	-2	-15
Cadre	-8	-9	4	Cadre	-14	-15	6
P Indép	-7	-4	4	P Indép	-5	4	4
Employé	-4	1	4	Employé	0	2	2
Ouvriers	2	25	2	Ouvriers	11	19	2

3.) Exemple de vérification approximative.

Vérifions les valeurs obtenues dans les cases grisées (des agriculteurs)

$\overrightarrow{Ag}.\overrightarrow{ag}$ = 22× 12 + (-10) ×(-7) + 1× 0 = 334. Or si l'on calcule l'écart 428 - 0,17 × 0,0498 × 10000 entre l'élément du tableau A et le produit des marges correspondantes, on voit que la représentation est bonne.

Question posée : Quels groupes socioprofessionnels offrent la moins grande et la plus grande perspective de mobilité sociale ascendante ?

V Recherche pratique des vecteurs du graphique.

1.) Méthode pratique de recherche des tableaux de la page 6

a) Référence :

En utilisant la méthode des puissances itérées décrite par Philippe Cibois (Référence 2), l'analyse factorielle décompose le tableau E_1 des écarts, en une somme de deux tableaux A_1 et A_2. Les éléments de ces tableaux sont respectivement les produits de leurs marges.

b) Description de la recherche pratique du tableau A_1. (Voir la feuille de calcul en page 20)

Le tableau A_1 est obtenu à la suite d'un calcul dont la justification théorique n'est pas l'objectif de cet article. Les étapes essentielles de la procédure sont décrites ici pour permettre l'exploitation d'autres exemples.

On choisit un vecteur colonne initial V_1 quelconque.

Par exemple $V_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$

On calcule le produit matriciel $V_2 = {}^{T}V_1 \times E_1$

$$V_2 = \begin{pmatrix} 1 & 1 & -1 \end{pmatrix} \times \begin{pmatrix} -48 & -120 & -222 & -6 & -114 & 510 \\ -11 & 130 & -232 & 38 & -97 & 172 \\ 59 & -10 & 454 & -32 & 211 & -682 \end{pmatrix}$$

$V_2 \approx$ (-117 , 20 , -907 , 63 , -422 , 1364)

On calcule la norme du Khi deux de ce vecteur. C'est la racine carrée de la somme des carrés des coordonnées du vecteur divisée par l'élément de la marge P (pied) correspondant:

$$K_2 \approx \sqrt{\frac{(-117)^2}{1547} + \frac{(20)^2}{1523} + \ldots + \frac{(1364)^2}{3418}} \approx 38$$

On calcule le vecteur réduit pondéré V'_2 correspondant à V_2 en divisant V_2 par sa norme K_2 et en divisant chaque coordonnées de ce vecteur par son correspondant pris dans la marge P du tableau initial.

$$V'_2 \approx \frac{1}{K_2}\left(\frac{-117}{1547}; \frac{20}{1523}; \ldots; \frac{1364}{3418}\right)$$

$V'_2 \approx$ ($-2{,}03.10^{-3}$; $7{,}18.10^{-5}$; $-1{,}98.10^{-2}$; $1{,}11.10^{-3}$; $-1{,}26.10^{-2}$; $1{,}06.10^{-2}$)

On calcule le vecteur colonne : $V_3 = E_1 \times {}^{T}V'_2$.

$$V_3 \approx \begin{pmatrix} -48 & -120 & -222 & -6 & -114 & 510 \\ -11 & 130 & -232 & 38 & -97 & 172 \\ 59 & -10 & 454 & -32 & 211 & -682 \end{pmatrix} \times \begin{pmatrix} -0,0020 \\ +0,0003 \\ -0,0199 \\ +0,0011 \\ -0,0127 \\ +0,0105 \end{pmatrix}$$

$$V_3 \approx \begin{pmatrix} 11 \\ 7 \\ -19 \end{pmatrix}$$

On recommence avec V_3 ce qui a été fait pour V_2 .

C'est à dire qu'en utilisant comme pondération, la marge M (main) du tableau initial, on calcule sa norme du Khi deux (ici $K_3 \approx 0,39$) pour pouvoir le réduire et le pondérer et ainsi obtenir le vecteur V'_3 qui servira pour la quatrième étape ...etc...On continue ainsi jusqu'à ce que la différence entre deux normes successivement calculées soit suffisamment petite. Dans l'exemple traité, la dernière étape est celle du calcul de V_5.

Calcul des marges du tableau A_1 :

Les deux derniers vecteurs trouvés par cette méthode dans le tableau de la page 20 sont V_4 et V_5. Ils sont divisés par la racine carrée de leurs normes K_4 et K_5

$$X' = \frac{1}{K_4} V_4 \approx \frac{1}{\sqrt{0,39}} (-1,24;\ 0,04;\ -9,33;\ 0,62;\ -4,36;\ 14,26)$$

$X' \approx (-1,97;\ 0,07;\ -14,87;\ 0,99;\ -6,95;\ 22,73) \approx (-2;\ 0;\ -15;\ 1;\ -7;\ 23)$

et

$$X = \frac{1}{K_5} V_5 \approx \frac{1}{\sqrt{0,39}} \begin{pmatrix} 11,31 \\ 7,71 \\ -19,02 \end{pmatrix} \approx \begin{pmatrix} +18 \\ +12 \\ -30 \end{pmatrix}$$

Chaque élément $a_{i,j}$ de A_1 s'obtient en faisant le produit de x'_i de X' par x_j de X et on vérifie que les marges de A_1 sont bien X' et X. On peut donc dire que A_1 est le produit $X' \times X$ de ces marges.

c) Calcul de A_2

Posons $E_2 = E_1 - A_1$, comme $A = A_0 + E_1$, nous avons : $A = A_0 + A_1 + E_2$

La décomposition de E_2 se fait de la même manière que pour E_1. .

Dans le cas général, on démontre que le nombre de matrices A_i à obtenir est égal à la plus petite dimension de A.

Dans notre exemple A est une matrice (3,6) donc E_2 est la dernière différence calculée, elle est donc égale à A_2 d'où $A = A_0 + A_1 + A_2 = M \times P + X' \times X + Y' \times Y$

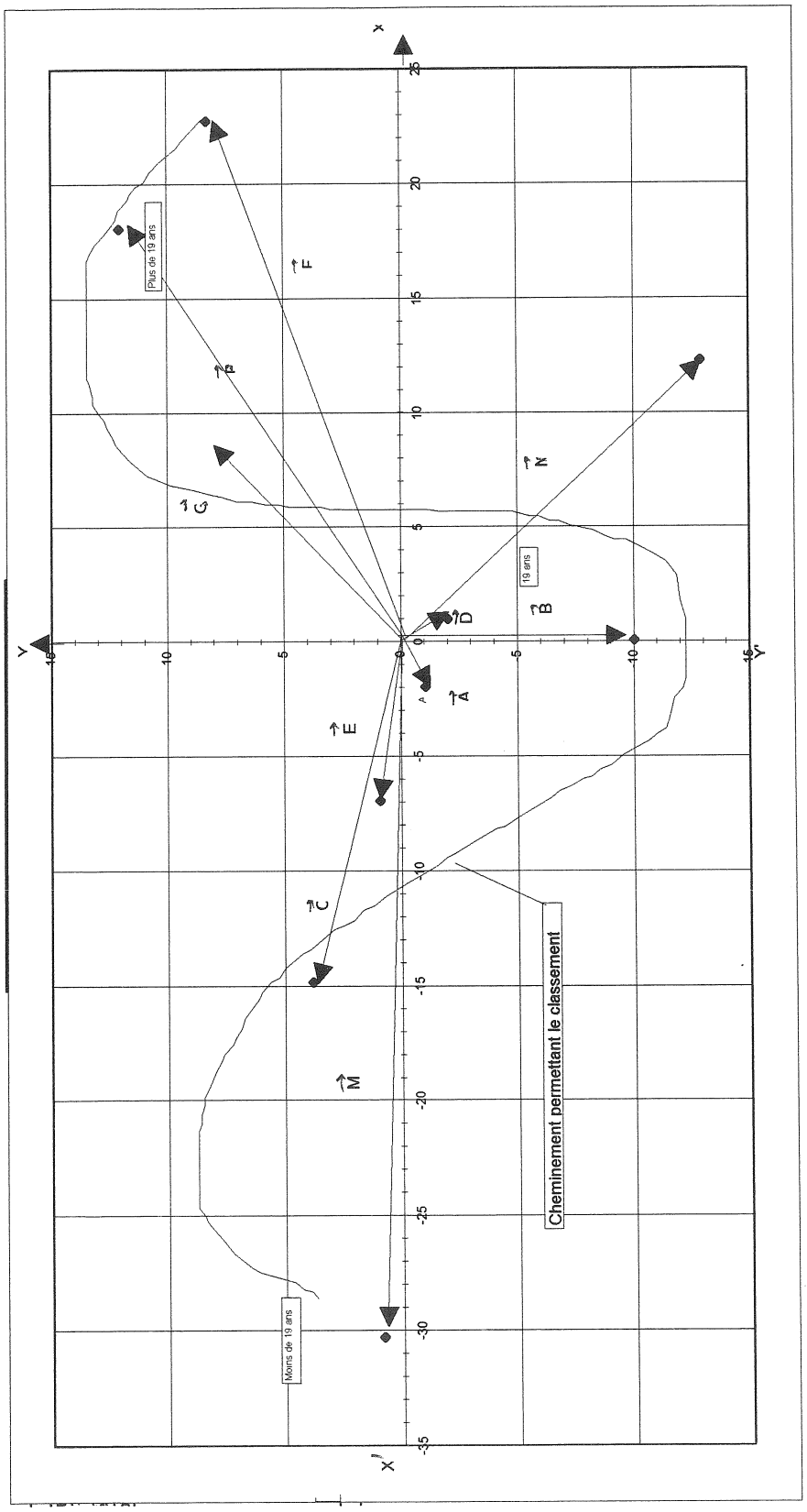
2.) Disposition des calculs de A_1 dans un tableur

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Etape 1, 2 et 3											
2	Classes	TA	TB	TC	TD	TE	TF	Main M	V1	V3		V '3	Vec Propre
3	Plus de 19	-48	-120	-222	-6	-114	510	3440	1	11,26		8,31E-03	17,94
4	19ans	-11	130	-232	38	-97	172	2817	1	7,76	Norme K3	7,00E-03	12,38
5	Moins de 1	59	-10	454	-32	211	-682	3743	-1	-19,02	0,39	-1,29E-02	-30,32
6	Pied P	1547	1523	1200	1434	878	3418	10000					
7	V2	- 117,92	20,12	-907,68	63,49	-422,73	1364,71						
8						Norme K2=	38,04						
9	V '2	-2,00E-03	3,47E-04	-1,99E-02	1,16E-03	-1,27E-02	1,05E-02						
10	Vprop2	-19,12	3,26	-147,17	10,29	-68,54	221,27						

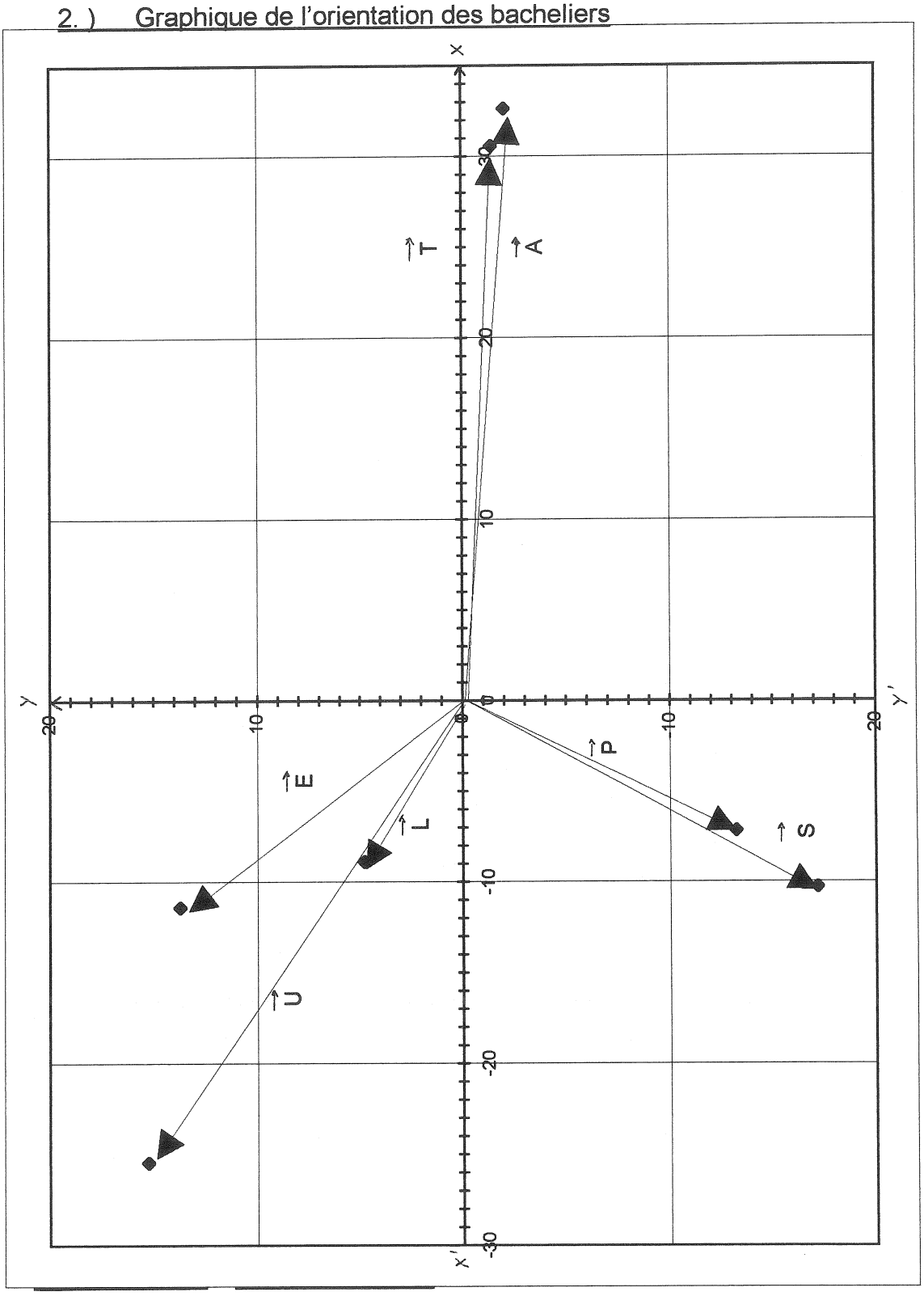
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Dernière étape											
2	Classes	TA	TB	TC	TD	TE	TF	Main M	V1	V5		V'5	X
3	Plus de 19	-48	-120	-222	-6	-114	510	3440	1	11,31		8,35E-03	18,03
4	19ans	-11	130	-232	38	-97	172	2817	1	7,71	Norme K5	6,95E-03	12,29
5	Moins de 1	59	-10	454	-32	211	-682	3743	-1	-19,02	0,39	-1,29E-02	-30,31
6	Pied P	1547	1523	1200	1434	878	3418	10000					
7	V4	- 1,24	0,04	- 9,33	0,62	- 4,36	14,26						
8						Norme K4	0,39						
9	V'4	-2,03E-03	7,18E-05	-1,98E-02	1,11E-03	-1,26E-02	1,06E-02						
10	X'	-1,97	0,07	-14,87	0,99	-6,95	22,73						

VI Les Graphiques

1.) Graphique de l'âge des élèves de terminale.

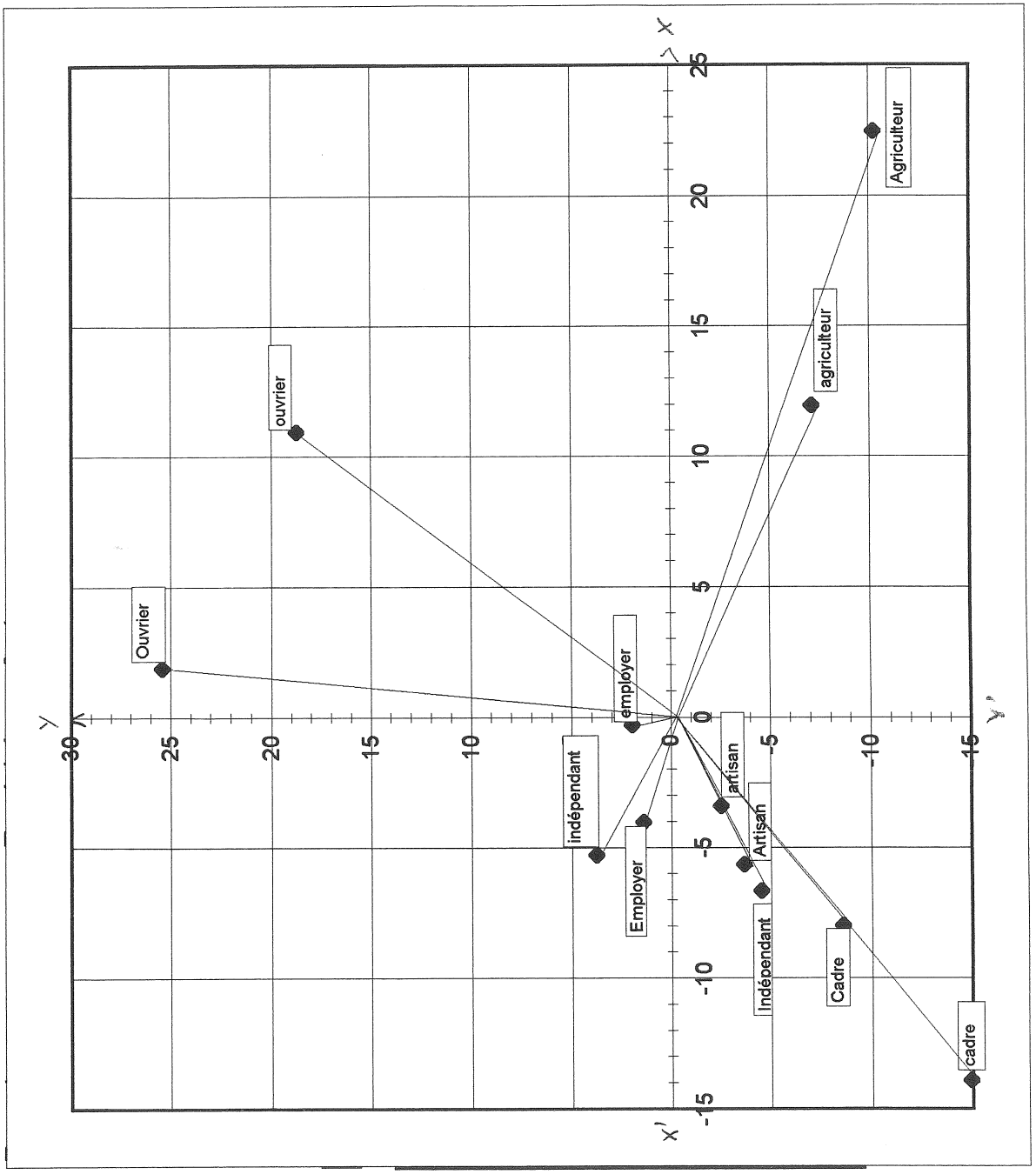


2.) Graphique de l'orientation des bacheliers

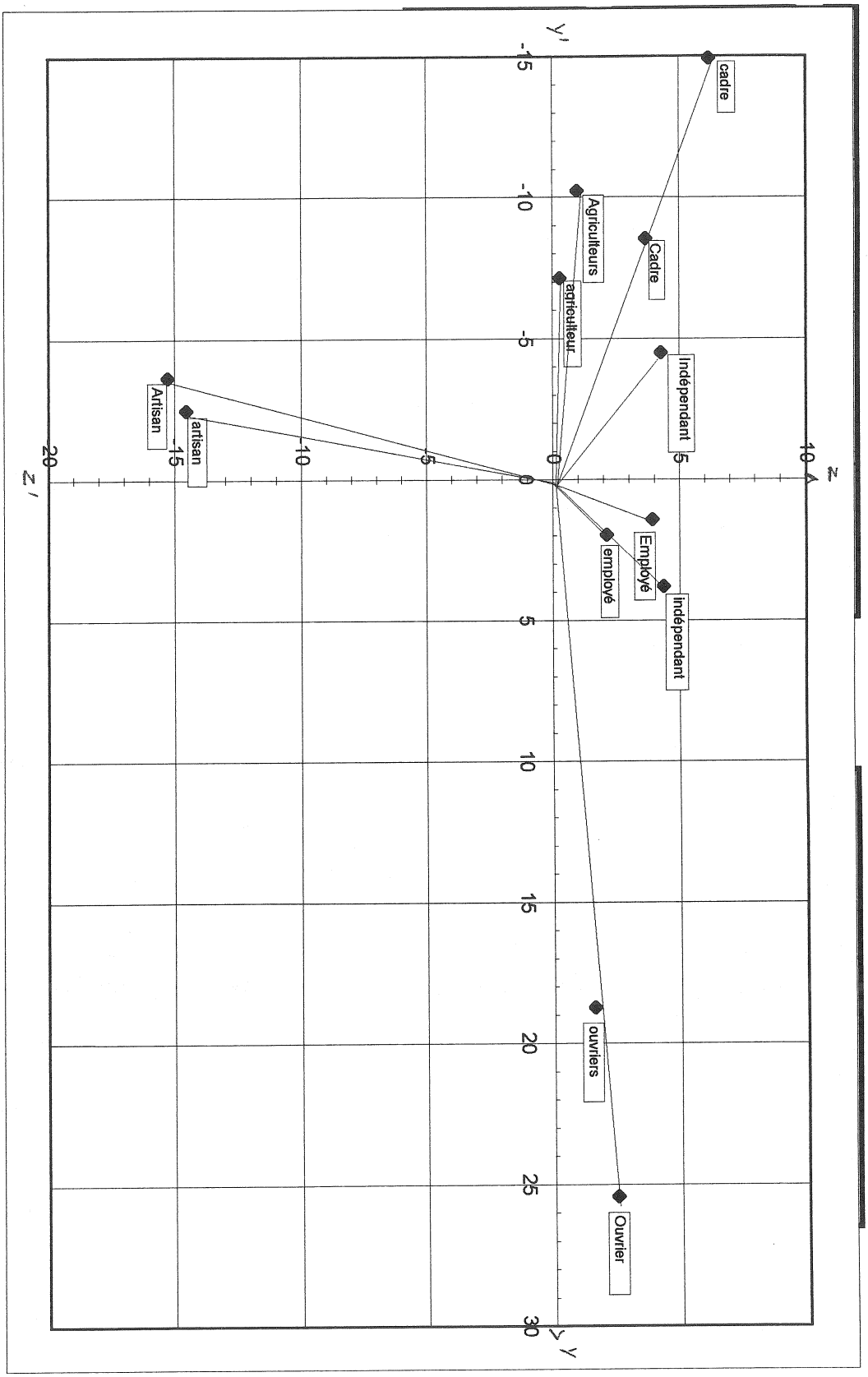


3.) Graphiques des catégories socio-professionnelles.

a) Graphique du plan horizontal.



b) Graphique du plan vertical



VII BIBLIOGRAPHIE

1) **L'enseignement des mathématiques en première ES**.
Groupe Lycée.(I.R.E.M. Besançon).

2) **L'analyse Factorielle**.
Philippe Cibois (Presses Universitaires de France, 1983).

3) **Initiation à l'analyse des données**
de Jean de Lagarde (Dunod, 1986).

I.R.E.M. de Franche-Comté
UFR des Sciences et Techniques
16, route de Gray, La Bouloie
F-25030 BESANÇON cedex
Téléphone : 81.66.61.92 - Télécopie : 81.66.61.99
Cour. électr. : iremfc@math.univ-fcomte.fr

TITRE : Analyse factorielle des correspondances

AUTEUR : Michel VENDRELY

DATE : Octobre 1996

MOTS CLÉS : Statistique, caractère qualitatif, tableau de contingence, indépendance, sous et sur représentation, pourcentage, produit scalaire.

RÉSUMÉ : Cette brochure propose au lecteur une approche de l'analyse factorielle des correspondances à l'aide du produit scalaire. Partant d'un exemple, l'auteur décrit la construction des objets et des méthodes de cette analyse. La présentation de l'algorithme utilisé pour la construction des graphiques factoriels permet, en outre, d'exploiter d'autres exemples.

Format A4- Nombre de pages : 25 - Poids : 70 g

IREM DE BESANÇON

Dépot Légal : 96/101

Numéro ISBN : 2-909963-12-8