Enquête 1996 : Utilisation des logiciels statistiques
Groupe logiciels ASU

1. Introduction

Danielle Grangé

CNRS-Montpellier

1.1 Contexte de l'enquête

En 1992, le groupe « Logiciels » de l'ASU décidait de faire une enquête sur les logiciels de Statistique, disponibles sur le marché français (1), afin de connaître, d'une part les logiciels utilisés et d'autre part, l'opinion des utilisateurs sur ces logiciels et l'usage qui en était fait. Quatre ans plus tard, la situation des logiciels a évolué, certains logiciels sont apparus, d'autres ont connu de grands bouleversements, de nouvelles interfaces graphiques ont été développées, de nouvelles fonctionnalités sont disponibles. Il semblait urgent de refaire le point.

1.2 Le questionnaire

Le groupe « Logiciels » a donc décidé de refaire une enquête proche de celle de 1992 pour que les données puissent être comparées. Mais le questionnaire de 1992 devait être allégé pour obtenir le plus de réponses possible et ne pas décourager l'enquêté par un questionnaire trop long. Le questionnaire à donc été revu[1] dans sa forme et dans sa présentation. Il a été découpé en quatre parties. La première partie contenant des questions sur l'utilisateur et son environnement : dans son entreprise, dans son environnement informatique (matériel et logiciel). La deuxième partie comportait des questions sur sa culture statistique : sa formation, son environnement statistique et les méthodes utilisées. La troisième partie devait permettre d'obtenir l'opinion de l'utilisateur sur son environnement logiciel statistique. Partie dans laquelle le répondant peut donner son avis sur trois logiciels dont il a la maîtrise : indiquer les méthodes qu'il utilise avec chaque logiciel, donner son avis sur la qualité des fonctionnalités disponibles et la facilité d'utilisation de chaque produit logiciel. Enfin, une quatrième partie était constituée d'une seule question ouverte dans laquelle l'utilisateur devait décrire librement les qualités d'un bon logiciel de statistique. Cette question ouverte reprend exactement la formulation de 1992 qui était apparue particulièrement riche d'information. Le questionnaire est présenté en annexe.

Dans les parties 1, 2 et 4, chaque répondant est représenté par un seul questionnaire. Alors que dans la partie 3 un répondant peut donner son avis sur 1,2 ou 3 logiciels. Cette partie sera donc traitée de façon différente des autres parties

[1] Le groupe « Constitution du questionnaire » était composé de A. Bringé, D. Grangé, Y. Lechevallier, F. Sermier, N. Valette

1.3 L'enquête

Ne disposant pas de base de sondage, pour constituer un échantillon représentatif, il a été décidé de diffuser le questionnaire très largement en essayant de contacter le maximum d'utilisateurs, recensés soit dans des associations, soit par les sociétés de logiciels. Ce questionnaire a été diffusé dans les congrès de statistique, au printemps 1996 et en particulier au congrès de l'ASU qui se tenait à Québec en Mai 1996. La plupart des sociétés de logiciels contactées ont accepté de diffuser ce questionnaire auprès de leurs clients et nous tenons ici à les en remercier. Afin d'avoir, connaissance de la provenance des questionnaires un code était attribué à chaque source de diffusion. Bien entendu il n'est pas impossible que certaines personnes aient répondues plusieurs fois au questionnaire, obtenu par des sources différentes, mais nous avons essayé de les en dissuader par un courrier d'accompagnement.

542 questionnaires ont été obtenus en septembre 1996. Tout membre de l'ASU pouvait participer au traitement des données à condition de s'engager par écrit à n'utiliser ces données que dans le but exclusif de collaborer au travail du groupe et de publier les résultats avec le groupe « traitement du questionnaire ». Chaque membre du groupe était libre de travailler sur la partie des données qui l'intéressait, d'utiliser les méthodes qu'il souhaitait, le matériel informatique dont il disposait et les logiciels qu'il jugeait les plus appropriés pour traiter les données.

2. Constitution d'une base de données relationnelle concernant l'enquête

Véronique Stephan

INRIA-Rocquencourt

Cette partie concerne la constitution d'une base de données relationnelle à partir des résultats de l'enquête ASU. Nous allons détailler les deux points suivants :

- l'élaboration d'une base pour stocker les résultats d'enquêtes,
- le type de requêtes d'interrogation permettant de construire des tableaux statistiques individus x variables ou des tableaux de comptages.

Les analyses portant sur de tels tableaux seront données et commentées dans les chapitres « Quels logiciels pour quelles techniques » et « Qualité et facilité d'utilisation des logiciels ».

L'intérêt d'utiliser une base comme moyen de stockage des enquêtes est de permettre non seulement d'effectuer les analyses classiques mais de plus d'effectuer des analyses en terme de correspondances entre différents thèmes de l'enquête. Nous montrons dans les chapitres « Quels logiciels pour quelles techniques » et « Qualité et facilité d'utilisation des logiciels », comment effectuer une correspondance entre le profil des utilisateurs et chaque logiciel, ou encore, la description de l'environnement informatique pour chacun d'entre eux.

Cette étude comporte donc trois parties principales : la mise en place d'une base de données relationnelle pour le stockage des résultats du questionnaire ASU, l'élaboration de tableaux de comptage à partir des résultats de requêtes à la base et les analyses proprement dites ainsi que l'interprétation des résultats obtenus. Ces analyses sont dans les chapitres « Quels logiciels pour quelles techniques » et « Qualité et facilité d'utilisation des logiciels ».

2.1 La base de données concernant l'enquête ASU

Nous détaillons l'élaboration d'une base de données relationnelle sous MsAccess. A partir de la base ASU, nous définissons les principales requêtes SQL que nous avons construites. Le fait de passer par une approche base de données pour stocker les résultats d'enquêtes assure une simplification des interrogations.

2.1.1 Elaboration du schéma de relations

Le schéma de relations a été défini conformément aux principes de modélisation et de normalisation propres au domaine des bases de données. De plus, certains schémas de relations ont été redécomposés de manière à respecter les thèmes statistiques abordés dans l'enquête ASU. Il s'agit donc de combiner :

- l'approche de modélisation "base de données" : on s'attache à organiser les données de manière à simplifier les requêtes d'interrogation pour la recherche d'informations. Les décompositions proposées doivent cependant être sans perte d'information. Cela

signifie que le schéma proposé doit permettre de retrouver exactement l'ensemble des informations stockées initialement dans un tableau Excel ;
- l'approche sémantique : on s'attache à décomposer l'information en respectant la structure du questionnaire en différents thèmes comme : l'environnement informatique, la culture statistique ou encore les méthodes utilisées.

Initialement, les données sont stockées sous Excel, sous la forme d'un tableau unique où une ligne correspond à une réponse d'enquête et où une colonne correspond à une question de l'enquête. Dans la mesure où il y a des questions avec réponses multiples, il est alors nécessaire de dupliquer plusieurs fois certaines colonnes. Il en résulte :
- des valeurs manquantes : si un enquêté n'a donné que trois réponses sur les huit possibles, on observe nécessairement des valeurs manquantes pour les cinq autres colonnes ;
- une interrogation difficile pour des recherches du type : « donner le secteur et le domaine d'activité des répondants ayant utilisé le logiciel ADDAD ».

Si nous prenons comme exemple les deux thèmes suivants de l'enquête, la culture statistique de l'utilisateur et l'environnement logiciels statistiques, nous constatons que :
- les questions du premier thème porte sur le répondant. Elles sont pour la plupart des questions à réponse unique. Si l'on restreint l'étude à ce thème, chaque question correspond à une et une seule colonne ;
- les questions du second thème porte sur le croisement entre le répondant et les trois premiers logiciels statistiques qu'il a cité. Les réponses des répondants sont éclatées sur trois colonnes correspondant aux réponses obtenues pour chaque logiciel cité dans les trois premiers rangs.

2.1.2 Modélisation à partir du tableau à plat

L'intérêt du stockage de l'enquête ASU par une base de données est donc de simplifier l'interrogation en décomposant le tableau initial en plusieurs tableaux (que nous appelons tables dans le contexte d'une base de données relationnelle).
En appliquant ce principe, nous pouvons décomposer la table initiale en trois tables. Une première table concerne les caractéristiques des utilisateurs. Elle regroupe les informations sur son environnement et sa culture statistique (partie I,II du questionnaire). La seconde table regroupe les informations sur les couples (utilisateur x logiciel). Elle concerne les réponses données par l'utilisateur pour chaque logiciel cité en rang 1, 2 ou 3. Enfin, une troisième table gère la correspondance entre les utilisateurs et le rang qu'ils donnent aux logiciels statistiques (figure 1). Afin de mieux organiser l'information, une table supplémentaire a été créée. Elle permet de référencer par un numéro, chaque logiciel statistique. Les liens entre les schémas de relations permettent de contrôler la cohérence des informations. Par exemple le lien entre utilisateur et utilisateur_logiciel permet de contrôler que tout numéro apparaissant dans la colonne numero de la table utilisateur_logiciel apparaît aussi dans la table utilisateur. La cardinalité du lien indique qu'à une ligne de la table utilisateur (c'est-à-dire un numéro d'enquête particulier) correspond à plusieurs lignes de la table utilisateur_logiciel.

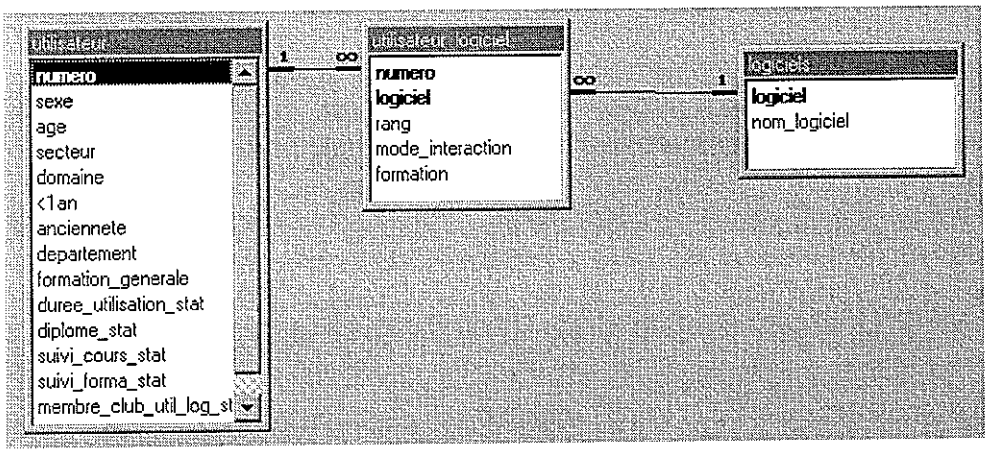


Figure 1 : 1ère décomposition des logiciels en plusieurs tables.

Nous avons donc :

- le numéro d'enquête, noté numéro, qui permet d'identifier chaque utilisateur,
- le numéro de logiciel, noté logiciel, qui permet d'identifier chaque logiciel.

Dans le schéma de relations définis sous MsAccess, si nous considérons les trois principales tables présentées ci-dessus, nous avons :

- numéro qui est la clé primaire du schéma de la table Utilisateur ce qui correspond à la notion d'identifiant. Toute clé primaire est notée en gras.
- logiciel qui est la clé primaire du schéma Logiciel.
- (numéro,logiciel) qui est la clé primaire de Utilisateur_Logiciel (Figure 2). Elle permet d'indiquer pour chaque enquêté, le rang des huit premiers logiciels qu'il a cité (question Q19). Utilisateur_Logiciel indique les caractéristiques de l'utilisation d'un logiciel cité parmi les trois premiers (questions L3,L4,L5). Une telle représentation évite d'avoir à stocker un ensemble de non réponses. Par exemple, l'enquêté numéro 2, n'a cité qu'un logiciel : il n'apparaît donc que sur une seule ligne.

utilisateur_logiciel : Table

numero	logiciel	rang	mode_interaction	formation
1	002	2	programme	aucune
1	136	1	programme	aucune
1	152	3	programme	aucune
2	136	1	programme	externe
3	002	3		
3	067	2	menu	aucune
3	136	1	programme	externe
4	136	1		interne
5	020	2	menu	aucune
5	046	4		
5	113	1	commande	aucune
5	127	3	menu	interne
6	134	3	programme	aucune

Enr : 1 sur 1370

Figure 2 : Table des Utilisateurs x Logiciels

Il reste à définir le moyen de gérer les questions à réponses multiples. Si nous considérons la table Utilisateur et la question Q6, un même enquêté peut donner une ou plusieurs valeurs comme réponse. Si nous ne modifions pas le schéma, nous ne pouvons plus garantir que l'attribut numéro identifie chacune des réponses. Nous effectuons donc la décomposition suivante. Chaque question à réponse multiple donne lieu à la création d'une nouvelle table où

la clé est composée de l'attribut clé initial et de l'attribut correspondant à la question posée. Par exemple pour la question Q6, nous définissons la table Machine (Figure 3) dont le schéma est : (numéro, machine). Cette règle s'applique à la question Q6, Q10, Q11 et Q15 (Figure 4).

sgbd : Table

numero	sgbd
5	Paradox
7	Ingres
8	Db2
9	ISPF
10	Access
10	Oracle
11	SAS
18	Excel
23	SAS
25	Access
25	Ingres
27	SAS
28	Access
28	Oracle

Enr 1

Figure 3 : Définition d'une table pour une question à réponses multiples

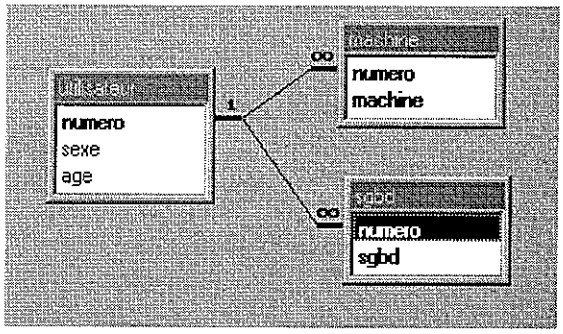


Figure 4 : décomposition liées aux questions à réponses multiples

Considérons à présent la question Q18. Nous définissons de même une nouvelle table correspondant aux réponses données par les répondants sur l'utilisation des méthodes. Cette table est donc composée entre autre de l'attribut numéro et de l'attribut méthode qui constituent la clé de la table. Comparativement à la décomposition précédente, nous ajoutons un attribut supplémentaire au schéma de la table : il permet d'indiquer la fréquence donnée pour chaque couple (utilisateur, méthode) (figure 5). Nous effectuons de même pour la troisième partie de l'enquête.

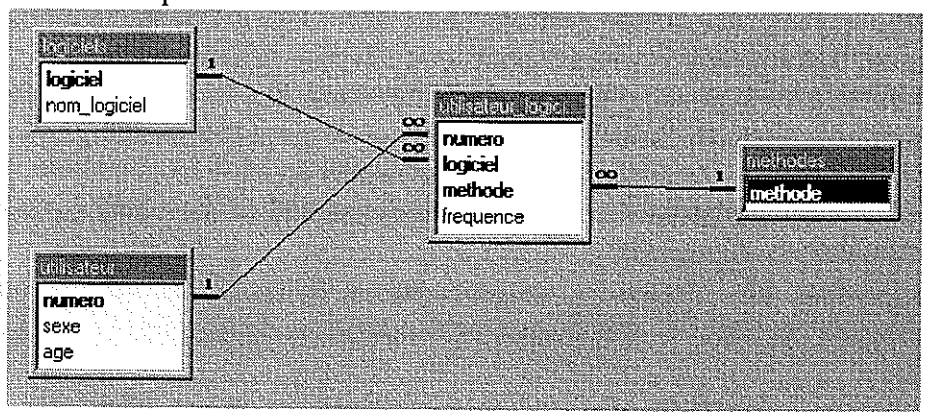


figure 5 : décomposition du schéma pour la question Q18

2.1.3 Décomposition prenant en compte l'organisation du questionnaire

La dernière décomposition consiste à prendre en compte l'organisation de l'information donnée a priori par le questionnaire. Ainsi les tables obtenues sont décomposées, cette fois-ci de manière à être structurée de la "même" façon que dans le questionnaire. Si nous reprenons Utilisateur, nous décomposons le schéma de la table en faisant apparaître :

- la source de l'enquête (table enquête),
- les informations concernant la culture statistique de l'utilisateur (table culture_statistique),
- les informations concernant la profession de l'utilisateur (table entreprise), ...

Le schéma général obtenu est donné dans la figure 6.

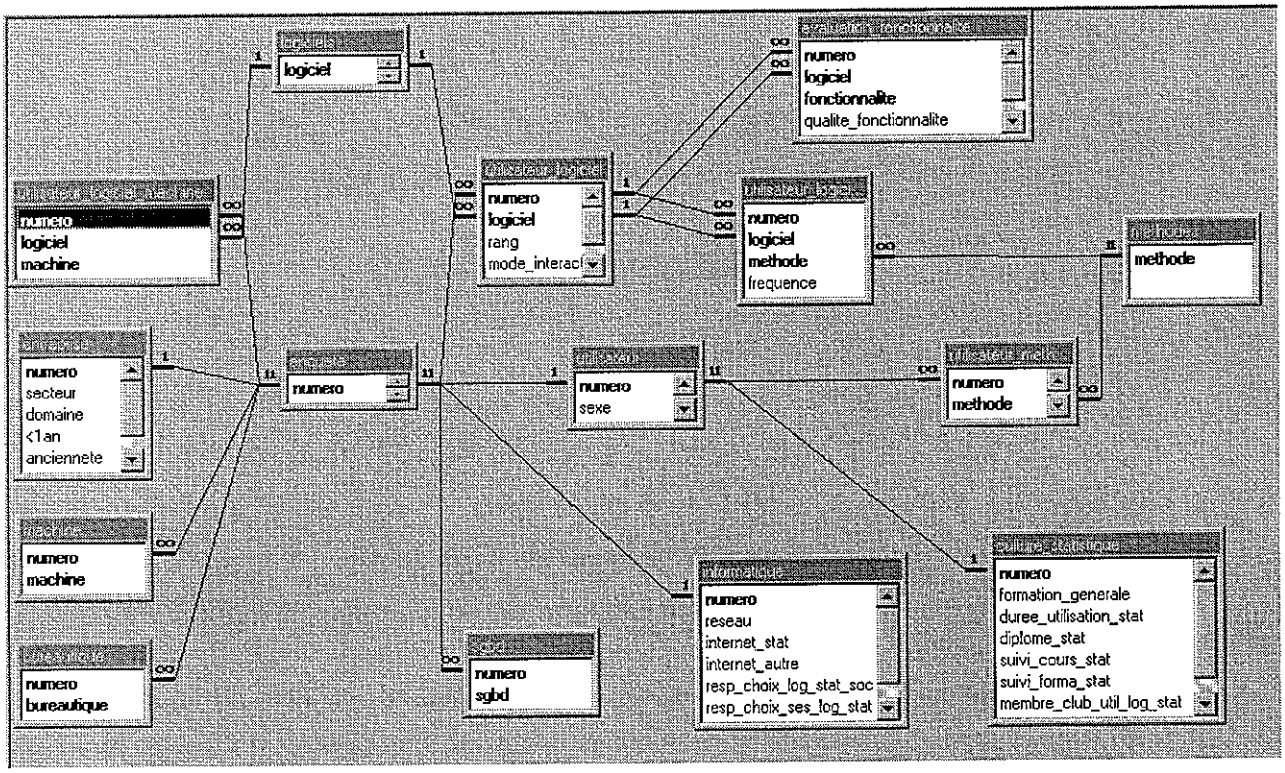


Figure 6 : Schéma de relations concernant la totalité de l'enquête ASU

2.2 Sélection des informations et construction des tableaux de comptages

La première analyse est celle des logiciels en fonction du profil des utilisateurs. On regroupe pour un même logiciel, l'ensemble des utilisateurs qui l'ont cités en rang 1,2 ou 3. On construit alors un nouveau tableau où chaque ligne est un logiciel et où chaque colonne est un attribut descriptif des utilisateurs. Dans une case, on indique alors soit :

- la distribution des fréquences observées sur l'attribut descriptif au sein du groupe d'utilisateurs du logiciel,
- le vecteur des comptages correspondant à l'effectif des valeurs observées sur l'attribut descriptif au sein du groupe d'utilisateurs du logiciel.

2.2.1 Les logiciels retenus

La construction des tableaux de comptages s'est faite par une application à l'extérieur de MsAccess. L'accès aux données se fait par une requête SQL transmise à la base via l'interface ODBC. L'élaboration d'un tableau de comptages se fait donc à partir des résultats d'une requête lancée sur la base. Les résultats sont traités de manière à obtenir un tableau de comptages qui sert en entrée aux méthodes d'analyse des données (voir la partie 3). Nous détaillons le passage entre les données dans la base ASU et le tableau de comptages sur les logiciels Nous illustrons ce passage sur le premier type d'analyse.

Les requêtes portent sur deux groupes de logiciels :

- le premier groupe est celui des logiciels ayant été cités par plus de 10 répondants. Ces logiciels ont été stockés dans une nouvelle table *selog10*, où l'on a comme attribut, le numéro du logiciel, le nom, et le nombre d'enquêtes où il apparaît. La table *selog10* regroupe 20 logiciels ;
- le second groupe est celui des logiciels ayant été cités par au moins un enquêté. Le résultat est stocké dans la table *selog1*. Cette table a la même structure que *selog10*. Certains logiciels apparaissant dans cette table mais n'ayant pas été décrits ont été supprimés. La table *selog1* regroupe 77 logiciels

2.2.2 Construction d'un tableau de comptages à partir d'une requête SQL de sélection

Nous décrivons à présent les principales opérations pour obtenir un tableau de comptage exploitable par des méthodes d'analyse des correspondances binaires ou de partitionnement croisé :

Nous définissons tout d'abord une requête à la base ASU permettant d'extraire le tableau des caractéristiques des utilisateurs en indiquant pour chaque utilisateur un des trois logiciels qu'il a cité dans les trois premiers rangs. L'interrogation de la base MsAccess se fait via ODBC.

- Pour toutes les requêtes, nous faisons une correspondance entre les unités statistiques (par exemple les utilisateurs) et les logiciels qu'ils ont cités. Cette correspondance est donnée de la manière suivante. La première colonne du résultat de la requête est l'identifiant de l'unité statistique alors que la deuxième colonne est l'un des logiciels cité.

La requête correspondant à la description des utilisateurs suivant les informations entreprises est la suivante (figure 7).

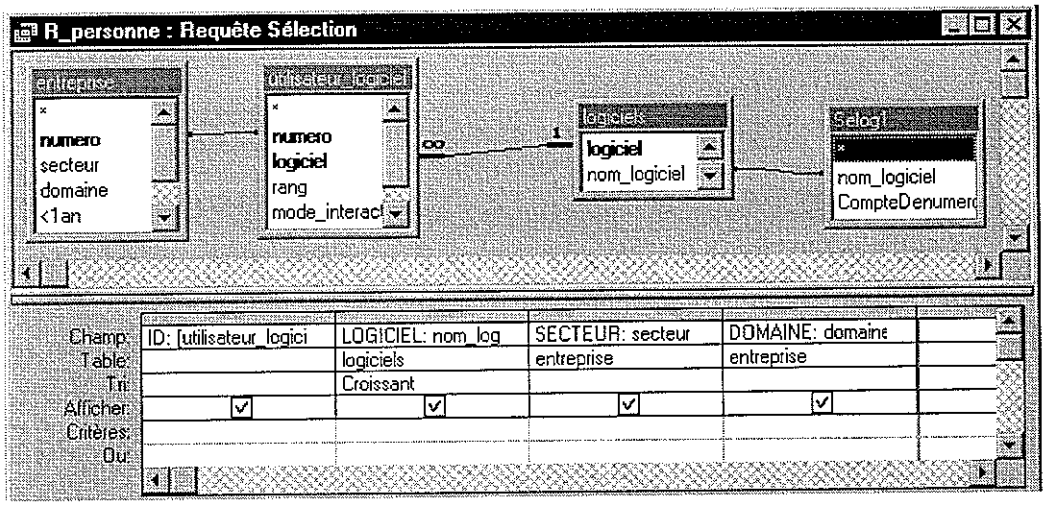


Figure 7 : Requête sous MsAccess pour obtenir les informations entreprises des enquêtés avec un filtre sur les logiciels

La requête complète que nous avons exécutée concerne l'ensemble des informations (de la table entreprise, utilisateur, culture_statistique) des utilisateurs avec un filtre sur les logiciels cités (on ne travaille que sur les logiciels de la table selog1). L'identifiant de l'unité statistique « enquêté » est fourni en première colonne. En seconde colonne, on indique le nom du logiciel qu'il a cité. Ainsi, un même enquêté peut apparaître sur différentes lignes du résultat de la requête (figure 8).

ID	LOGICIEL	SEXE	AGE	SECTEUR	DOMAINE	FORMATION_(	DPSTAT
349001	Access	Homme	24	Prive	Autre	3emecycle	non
295001	Access	Homme	27	Public	Banque/Assura	3emecycle	oui
438001	Access	Homme	29	Prive	Pharmacie	licence/maitrise	non
469001	Access	Femme	27	Prive	Banque/Assura	3emecycle	oui
170002	Addad		34	Administration	Enseign /Rech.	3emecycle	oui
52002	Addad	Femme	36	Administration	Enseign./Rech.	Grd Ecole	non
40002	Addad		25	Prive	Agro-alimentair	3emecycle	
181002	Addad	Homme	26	Public	Tranport/Energi	3emecycle	oui
330002	Addad	Femme	34	Collectivite	Etudes&service	3emecycle	non
31002	Addad	Homme	37	Public	Enseign./Rech.	3emecycle	non
267002	Addad	Homme	49	Administration	Enseign./Rech	3emecycle	non
169002	Addad	Homme	35	Administration	Enseign./Rech	3emecycle	non
57002	Addad	Homme	58	Administration	Enseign./Rech	Grd Ecole	non

Enr: 1 sur 1252

Figure 8 : Résultat de la requête sur les caractéristiques des employés

- Pour chaque logiciel (2ème attribut de la requête de sélection), nous définissons les caractéristiques de comptage sur chacun des attributs sélectionnés par l'utilisateur dans la requête.

Par exemple, nous obtenons pour le logiciel SAS, le vecteur de description suivant :

Attribut	Vecteur de comptage ou Intervalle associé
SEXE	{"Homme"(170), "Femme"(97), "NULL"(19)}
AGE	[22.00: 63.00]
SECTEUR	{"Prive"(141), "Public"(20), "Administration"(119), "Collectivite"(5), "NULL"(1)}
DOMAINE	{"Autre"(14), "Banque/Assurances"(29), "Pharmacie"(38), "Enseign./Rech."(102), "Agro-alimentaire"(9), "Tranport/Energie"(15), "Etudes&services"(42), "Administration"(12), "Economie/Finance"(5), "industrie"(12), "Sondages"(4), "NULL"(4)}
FORMATION_GENERALE	{"3emecycle"(167), "licence/maitrise"(24), "Grd.Ecole"(74), "Bac"(3), "IUT"(16), "DEUG"(1), "NULL"(1)}
DPSTAT	{"non"(90), "oui"(190), "NULL"(6)}
COURS	{"oui"(261), "non"(9), "NULL"(16)}
FORMATION	{"non"(112), "oui"(147), "NULL"(27)}
CLUB_UTI	{"non"(210), "oui"(74), "NULL"(2)}
CLUB_STAT	{"non"(180), "oui"(106)}

De manière à faciliter la lecture des informations de ce tableau, nous pouvons introduire les fréquences d'observations. Par exemple, la description du logiciel SAS est la suivante, où l'on indique que 59 % des répondants ayant cités SAS sont des hommes, 34% des femmes et que leur âge varie entre 22 et 63 ans ...

SEXE	{"Homme"(0.59), "Femme"(0.34), "NULL"(0.07)}
AGE	[22.00: 63.00]
SECTEUR	{"Prive"(0.49), "Public"(0.07), "Administration"(0.42), "Collectivite"(0.02), "NULL"(0.01)}
DOMAINE	{"Autre"(0.05),"Banque/Assurances"(0.10), "Pharmacie"(0.13), "Enseign./Rech."(0.36), "Agro-alimentaire"(0.03), "industrie"(0.04), "Tranport/Energie"(0.05), "Etudes&services"(0.15), "Sondages"(0.01), "Administration"(0.04), "Economie/Finance"(0.02), "NULL"(0.02)}
FORMATION_GENERALE	{"3emecycle"(0.58), "licence/maitrise"(0.08), "Grd.Ecole"(0.26), "Bac"(0.01), "IUT"(0.06), "DEUG"(0.01}
FORMATION	{"non"(0.40), "oui"(0.51), "NULL"(0.09)}
CLUB_UTI	{"non"(0.73), "oui"(0.26), "NULL"(0.01)}
CLUB_STAT	{"non"(0.63), "oui"(0.37)}

Les résultats sont fournis sous la forme d'un tableau de données complexes où chaque ligne correspond à un vecteur de description où un élément du vecteur peut être :

- une distribution de fréquences sur les modalités d'un attribut descriptif,
- un vecteur de comptages sur les modalités d'un attribut descriptif,

- un intervalle de valeurs correspondant à la variation des observations sur un attribut numérique.

On indique les non réponses par la valeur NULL. L'effectif associé à cette non réponse n'est cependant pas transmis dans le tableau de comptages. En effet, les modalités NULL n'ajoutent rien à l'analyse du tableau et, au contraire, rendent plus difficile l'interprétation des résultats.

- A partir du tableau de description des logiciels, on effectue un codage permettant de se ramener à un tableau classique. Chaque attribut descriptif est décomposé en autant de colonnes que l'on observe de modalités. Les attributs numériques ne sont pas pris en compte dans le tableau résultat. Ce tableau sert alors en entrée de l'analyse. Nous donnons un exemple du tableau de comptage trouvé :

Logiciel	Sex-Hom	Sex-Fem	Sec-Pri	Sec-Pub	...
SAS	170	97	141	20	
S-PLUS	29	13	7	4	
SPADN	55	30	38	9	
...					

Tableau final de comptage sur les enquêtés

2.2.3 Construction d'un tableau de comptages à partir d'une requête d'analyse croisée

Un deuxième type d'analyse est celui du traitement de la fréquence d'utilisation des méthodes par logiciel. Dans ce cas, la requête SQL n'est plus triviale. Elle consiste en une analyse croisée où l'on a :

- en ligne, la description du couple (utilisateur, logiciel),
- en colonne, les différentes méthodes cités dans le questionnaire.

Au croisement d'une ligne et d'une colonne, on trouve la fréquence donnée pour l'utilisateur, le logiciel concerné et la méthode citée. Par rapport au schéma de la base, nous sommes donc obligé de passer par une requête d'analyse croisée pour transposer les valeurs de l'attribut méthode en colonne. Nous donnons la formulation correspondante de la requête en QBE.

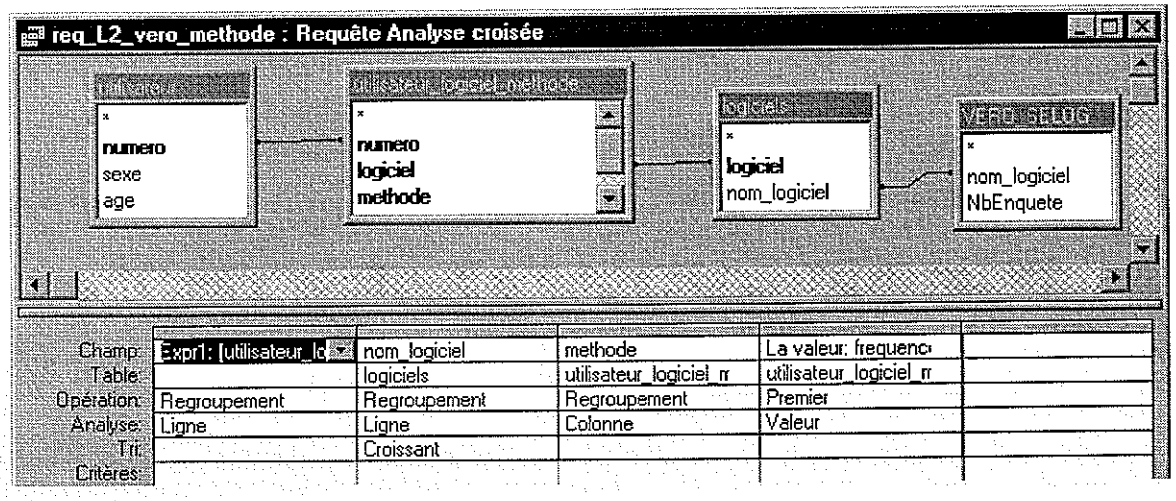


Figure 8 : Requête correspondant à l'analyse des méthodes

La 4ème colonne de la requête QBE indique le fait que pour chaque utilisateur et pour chaque méthode, on indique la première valeur observée dans l'attribut fréquence (c'est-à-dire « jamais », « ponctuellement », « souvent », « très souvent »). Cette valeur est bien unique et donc l'opération valide.

Le principe de construction du tableau de comptage est alors le même que précédemment. Nous pourrions faire de même pour obtenir un tableau sur les utilisateurs et les méthodes indépendamment du logiciel utilisé (question Q18). Il est alors possible d'effectuer la jointure de ces deux tableaux. Cette opération consiste à concaténer les variables des deux tableaux et d'effectuer la jointure de chaque ligne en fonction du logiciel. On obtient alors pour chaque ligne les caractéristiques du logiciel suivant les méthodes utilisées par les enquêtés qui ont cités ce logiciel (1ère sous colonne) ainsi que suivant les méthodes utilisées spécifiquement pour ce logiciel (2ème sous colonne de Table 2).

Logiciel	**Stat-Desc**		**A-Fact**		**Classif**	
	(J,P,S,TS)	**(J,P,S,TS)**	**(J,P,S,TS)**	**(J,P,S,TS)**	**(J,P,S,TS)**	**(J,P,S,TS)**
SAS	(1,20,79,184)	(11,34,65,137)	(41,121,104,53	(71,103,40,27)	(52,108,69,50)	(71,95,45,26)
SPADN	(1,8,20,58)	(24,18,13,14)	(1,21,24,41)	(2,18,17,31)	(2,23,28,34)	(4,18,17,31)
...						

Table 2 : Description des logiciels suivant les méthodes utilisées

3. Les répondants

Danielle Grangé

CNRS-Montpellier

3.1 Leur origine

Les 542 personnes qui ont répondu au questionnaire représentent 61% d'hommes et 32% de femmes, 6% n'ayant pas répondu à la question. Ils proviennent de l'enseignement ou de la recherche (35%), des « études et services » (14%), de l'industrie pharmaceutique (9%). Ces chiffres correspondent sensiblement aux chiffres de 1992. La distribution de l'âge des répondants montre une population assez jeune qui présente un mode vers 30 ans. 60% sont depuis moins de 5 ans dans l'entreprise. Le niveau de formation est élevé : 90% ont un niveau supérieur à bac + 2 et 80 % ont le niveau grandes écoles ou thèse. 50% seulement ont un diplôme en statistique mais 88% ont eu au cours de leur cursus un enseignement de statistique. Ils sont pour la moitié en relation avec des statisticiens dans leur entreprise. C'est une population ouverte sur l'extérieur puisque 68% répondent avoir des contacts avec des statisticiens à l'extérieur de leur entreprise. Par contre 20% seulement participent à un club d'utilisateurs de logiciels et 26% sont membres d'une association de statisticiens. Ces deux derniers chiffres montrent que l'échantillon observé n'a pas une sur représentation des membres de l'ASU ni une sur représentation des membres du club SAS/statistique.

3.2 Leur environnement informatique

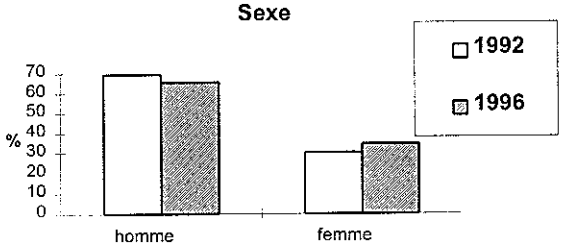
On constate une évolution depuis 1992 de l'environnement matériel. En effet, 90% disent travailler sur PC mais ils ne sont plus que 16% à travailler sur site central alors qu'ils étaient 45% en 1992. L'informatique se décentralisant, ils sont beaucoup plus nombreux à être responsables du choix de leur logiciel de statistique (79%) alors qu'en 1992 ils étaient 50%.
La procédure DEMOD du logiciel SPADN nous permet de donner un profil des utilisateurs suivant la machine qu'ils utilisent.
Ceux qui travaillent sur PC (90%) sont des utilisateurs de tableurs à 81% et plutôt des hommes. Le logiciel Statgraphics est plus souvent cité sur ce matériel qu'en moyenne. Ils sont responsables du choix de leur logiciel, utilisent le SGBD Access et font très souvent des comptages.
Parmi les utilisateurs du Mac (15%), 63% travaillent dans l'enseignement ou la recherche alors que ce domaine d'activité ne représente que 35% dans l'ensemble de la population observée. Les utilisateurs Mac ne travaillent pas sur PC mais 34% utilisent également une station de travail. Ils disent faire très souvent des classifications et des analyses factorielles, des statistiques non paramétriques, du modèle log-linéaire et ponctuellement des séries chronologiques. Ils utilisent un grapheur, du traitement de texte. Ils ont une formation supérieure, et utilisent pour 60% d'entre eux l'Internet pour leur travail statistique. Ils citent dans 16% des cas le logiciel SAS en deuxième position alors que ce logiciel ne représente que 8% des citations en 2ème positions. Ils travaillent sur plus de 2 machines, utilisent plusieurs logiciels de bureautique sont significativement plus âgés (39 ans) que la moyenne générale (36 ans).

Les utilisateurs de station (21%) ne sont pas des utilisateurs du PC. 13% d'entre eux citent S-Plus comme 1er logiciel utilisé alors que ce logiciel n'est cité que dans 4% des cas en premier logiciel Ils citent SAS en 1er ou 2ème logiciel. Ils ont un niveau d'éducation élevé (supérieur au 3ème cycle), sont membres d'association de statistique. Ils ont beaucoup de relations avec leurs collègues statisticiens dans leur environnement de travail et à l'extérieur Ils proviennent essentiellement du secteur public : « enseignement ou recherche ». Ils travaillent en réseau et utilisent Internet intensivement aussi bien pour leur travail statistique que non statistique. Ils utilisent parfois les SGBD Oracle et Ingres, font très souvent des régressions et des classifications, occasionnellement des méthodes neuronales. S'ils travaillent sur une deuxième machine, s'est essentiellement sur un Mac.

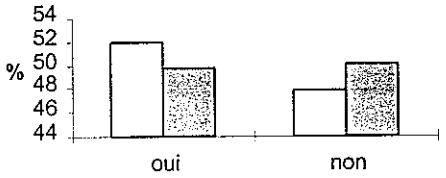
Ils ne sont que 6% a travailler sur mini. Sur ce type de machine, on trouvera des utilisateurs d'Oracle provenant des établissements publics. Ils ont des relations avec leurs collègues statisticiens, travaillent sur l'Internet, en réseau, et éventuellement sur site central, font très souvent des statistiques descriptives, du modèle log-linéaire, des comptages, et ponctuellement des plans d'expériences. Ce sont des gens qui travaillent en moyenne sur 3 machines différentes Ils sont plus âgés que la moyenne (40 ans) et ont plus d'ancienneté dans leur entreprise que la moyenne (12 ans).

16% travaillent sur Site Central citent en 1er le logiciel SAS travaillent en réseau, ont des relations avec leurs collègues statisticien dans leur entreprise. Ils proviennent du milieu banque et assurance. Ils sont diplômés en statistique. Ils utilisent les SGBD Oracle ou Ingres, font parfois de l'analyse discriminante, très souvent des comptages et des séries chronologiques et souvent des plans d'expériences. Ils sont membres d'une association de statisticiens. Ils travaillent également sur mini. Ils ont une moyenne de 10 ans d'ancienneté dans leur entreprise et sont âgés de 38 ans en moyenne.

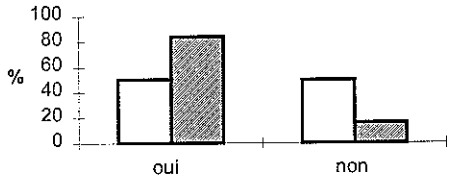
Sexe



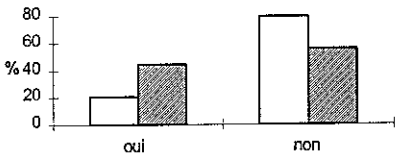
Diplôme en Statistique



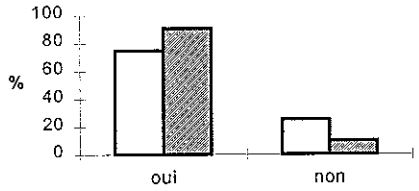
Responsable du choix de ses logiciels



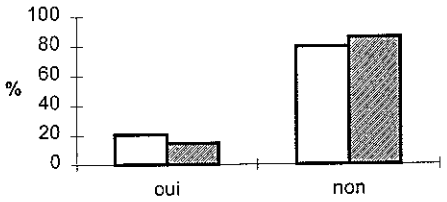
Responsable du choix des logiciels pour la société



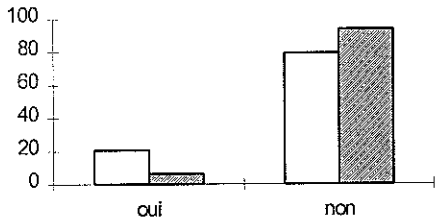
Utilise un PC



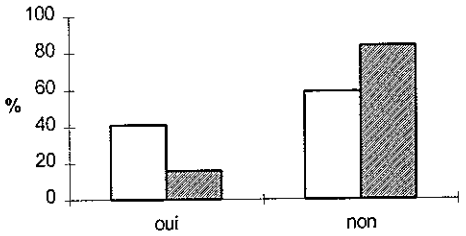
Utilise un Mac



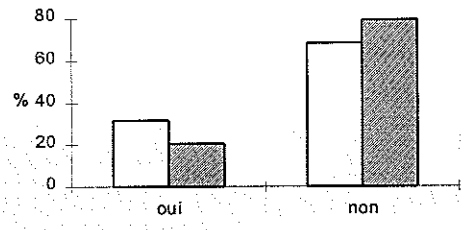
Utilise un Mini



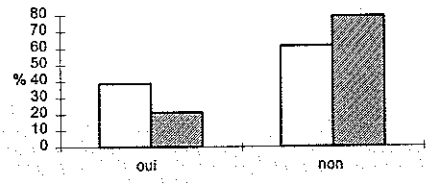
Travaille sur Site Central



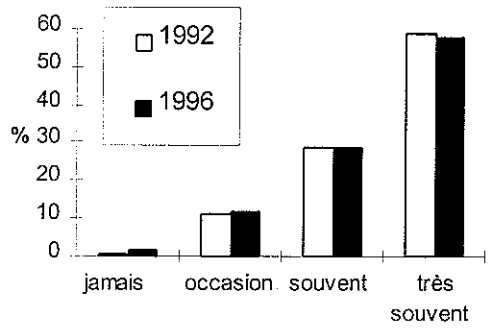
Utilise une Station de Travail



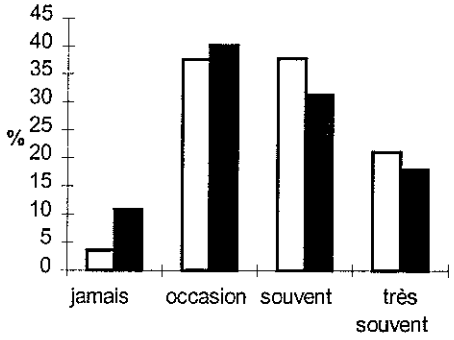
Membre d'un club d'utilisateurs de statistiques



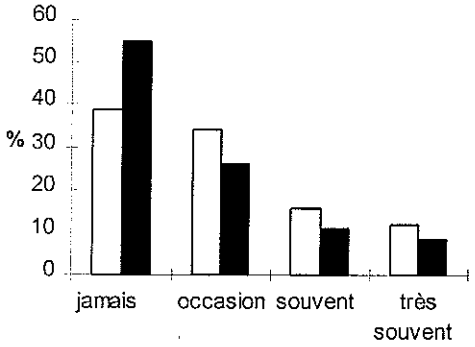
Statistiques Descriptives



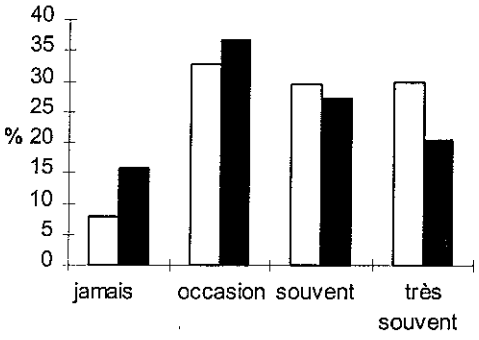
Régressions



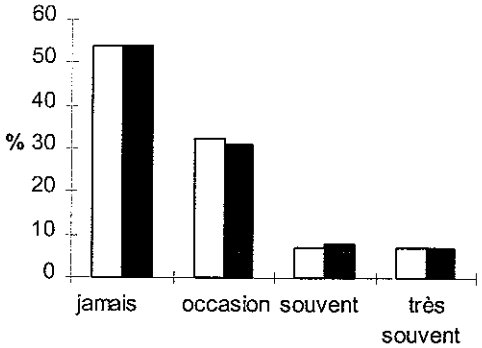
Plans d'Expériences



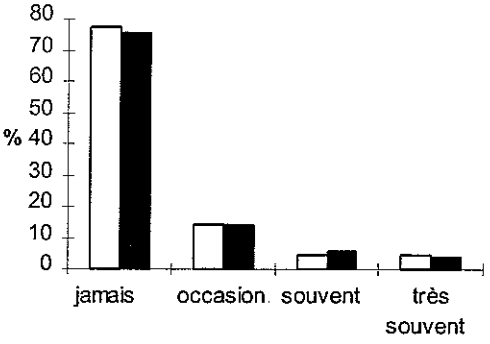
Analyse de la Variance



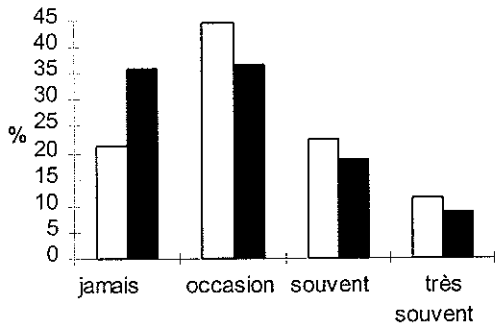
Séries Chronologiques



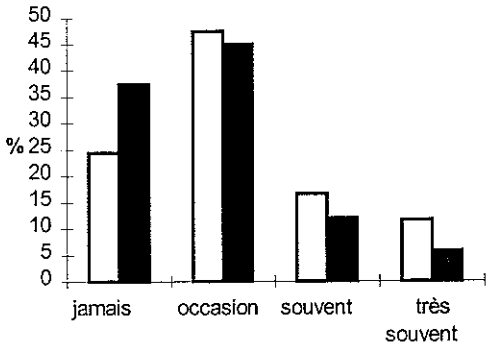
Contrôle de Qualité



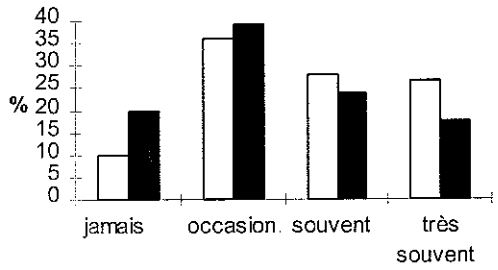
Tests non-paramétriques



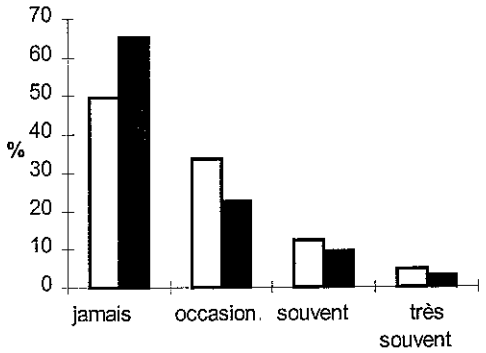
Discriminante



Analyses Factorielles ou Classifications



Modèle Log-Linéaire ou Survie



4. Caractérisation des répondants selon les techniques statistiques et les logiciels utilisés.

Dominique Drouet

IUT Vannes

Parmi les logiciels statistiques, certains se prétendent universels, d'autres sont dévolus à des utilisations plus spécifiques (tests exacts, Plans d'expérience, enquêtes...). Les répondants, eux, appartiennent à des domaine d'activité très différents et n'ont pas tous la même culture statistique. Cette diversité, se reflètent-elles dans les techniques et les logiciels utilisés? Peut-on identifier des groupes spécifiques pour les techniques ou les logiciels utilisés ? A l'inverse, la pratique d'un logiciel particulier induit-elle un usage particulier des techniques ?

4.1 Le domaine d'activité

DOMAINE D'ACTIVITE	Nombre	Nombre de Techniques*			Nombre de logiciels		
		Q1	Med	Q3	Q1	Med	Q3
Ensemble	542	2	4	6	1	2	3
Administration	26	1	2	5	1	1	2
Industrie Pharmaceutique	48	3	4	6	1,5	3	3
Enseignement/Recherche	190	3	4	6	2	3	4
Industrie agro-alimentaire	18	3	4,5	6	1	2	3
Transport/Energie	29	1	2	5	1	2	2
Industrie (autre)	43	3	3	5	1	2	3
Banques/Assurance	35	1	4	5	1	2	3
Instituts de Sondage	17	2	4	5	1	2	2
Economie et Finance	11	2	4	6	1	2	4
Autres Etudes et services	75	2	3	6	1	2	3

(La catégorie « comptages » n'est pas comptée comme une technique statistique. les techniques statistiques sont celles qui ont été déclarées être utilisées souvent ou très souvent)

Tableau 1 : Nombre de Techniques et de Logiciels selon le domaine d'activité

C'est dans l'administration que les nombres de techniques et de logiciels sont les moins élevés, et dans l'industrie agro-alimentaire, l'enseignement et la recherche, et l'industrie pharmaceutique qu'ils sont les plus élevés.

4.2 La responsabilité du choix des logiciels statistiques dans l'entreprise

Responsable des choix de l'entreprise	Nombre	Nombre de	Q1	Med	Q3
Oui	220	techniques	3	5	6
		logiciels	2	3	4
Non	275	techniques	2	3	5
		logiciels	1	2	3

Tableau 2 : Nombre de techniques et nombre de logiciels selon la responsabilité des choix de l'entreprise

Responsable de ses choix	Nombre	Nombre de	Q1	Med	Q3
Oui	429	techniques	3	4	6
		logiciels	2	2	3
Non	82	techniques	1	2,5	4
		logiciels	1	2	2

Tableau 3 : Nombre de techniques et nombre de logiciels selon la responsabilité de ses choix

Le nombre de techniques et de logiciels utilisés augmentent selon le degré de responsabilité dans l'entreprise.

4.3 Lien entre le nombre d'années d'utilisation des méthodes statistiques et le nombre de techniques et de logiciels utilisés

Nombre de techniques:	Nombre	Moyenne	Ecart-type
0*	30	6,93	6,18
1	65	5,72	5,09
2	55	8,27	7,00
3	89	8,59	6,16
4	88	9,12	7,54
5	73	9,55	7,69
6	63	9,02	8,20
≥7	79	11,37	8,01

Tableau 4 : Nombre d'années d'utilisation des méthodes selon le nombre de techniques

* Les trente personnes dans la catégorie « 0 » techniques n'utilisent les techniques statistiques que de façon ponctuelle, et/ou n'utilisent que les comptages

Les personnes qui utilisent le plus de techniques statistiques sont aussi celles qui ont le plus d'ancienneté en statistique.

Nombre de Logiciels :	Nombre	Moyenne	Ecart-type
0	3	20,00	28,28
1	146	6,66	5,44
2	160	7,70	6,41
3	131	10,48	7,68
≥4	88	11,14	7,57

Tableau 5 : Nombre d'années d'utilisation des méthodes selon le nombre de logiciels

Le nombre de logiciels utilisés est aussi lié positivement à l'ancienneté en statistique.

4.4 Lien entre nombre de techniques et nombre de logiciels

Nombre de Techniques	Nombre	Q1	Médiane	Q3
≤ 3	334	1	2	3
4 - 5	161	2	2	3
≥ 6	142	2	3	4

Tableau 6 : Nombre de Logiciels selon le nombre de Techniques

Le besoin en logiciels augmentent avec le nombre de techniques pratiquées.

4.5 Nombre de techniques (utilisées « souvent » ou « très souvent ») et nombre de logiciels

On cherche maintenant si le nombre de techniques et de logiciels utilisés est lié à une pratique spécifique de la statistique. Certaines techniques statistiques (hormis les comptages) sont utilisées «souvent» ou «très souvent» par (presque) la majorité des répondants :

- statistique descriptive : 86,7 %
- régression linéaire : 49,2 %
- analyse de variance : 47,6%
- analyse factorielle : 40,9 %
- classification : 41,3%

D'autres le sont par des minorités :

- contrôle de qualité : 10%
- techniques marché : 15,1%
- survie : 10,9%
- séries chronologiques : 15%

Les méthodes neuronales ne sont utilisées que par 16 (3,2%) répondants.

4.5.1 Description des groupes minoritaires

Ces quatre sous-groupes, sauf les pratiquants de «séries chronologiques», sont relativement exclusifs :

	Contrôle Qualité	survie	marketing	séries chronologiques
Contrôle Qualité	51	4	6	13
survie		55	7	5
marketing			75	15
séries chronologiques				77

Tableau 7 : Effectifs dans les groupes «minoritaires»

Le sous-groupe «marketing» se distingue par une variété de logiciels moins grande que dans les autres sous-groupes .

Techniques utilisées souvent / très souvent	Nombre	Nombre de	Q1	Médiane	Q3
marketing	75	techniques	4	5	7
		logiciels	1	2	3
contrôle de qualité	51	techniques	4	5	6
		logiciels	2	2	3
séries chronologiques	77	techniques	3	5	7
		logiciels	2	2	3
survie	55	techniques	4	6	7
		logiciels	2	3	4

Tableau 8 : Nombre de techniques et de logiciels dans les groupes « minoritaires »

Dans le groupe «techniques de marketing», ce qui prédomine (hormis les statistiques descriptives) est le trio «segmentation», «classification», «analyse factorielle» (de 50 à 70 % de répondants) Le groupe «contrôle de qualité » est caractérisé par une utilisation plus importante que la moyenne de « plans d'expérience» (38% des répondants), et des techniques « régression» (66%), «analyse de variance» (62%).
Dans le groupe «survie», les techniques les plus associées sont «modèles log-linéaires» et «tests non-paramétriques» (environ 50% des répondants)
Dans le groupe «séries chronologiques», les techniques les plus associées sont «classification» et «régression linéaire» (50% et 66% des répondants).

4.5.2 Y a-t-il des logiciels spécifiques dans les groupes «minoritaires » ?:

Ce qui est remarquable :

- dans le groupe «contrôle qualité », la position en premier de STATG (43,1% des répondants) (vs 39,2% pour SAS)
- dans le groupe «techniques de marketing» , la deuxième position de EXCEL (17,3%) ex-aequo avec SPSS, derrière SAS (45,3%), et la présence de logiciels non cités dans les autres groupes de techniques tels DESTIN (11 réponses), QUESTION (6).
- dans le groupe «survie», la présence massive de SAS (80%) des répondants.

4.6 Nombre de Techniques et nombre de logiciels selon le logiciel utilisé

Etudions maintenant le problème de façon duale : en quoi l'utilisation d'un logiciel particulier est liée à des pratiques spécifiques en statistique ?
Etant donnée la place particulière occupé par SAS (52,7 % d'utilisateurs), il est intéressant d'étudier les quatre sous-groupes suivants : les mono-utilisateurs de SAS, les mono-utilisateurs d'autres logiciels (non-SAS), et les parmi les pluri-utilisateurs de logiciels , ceux qui utilisent SAS et ceux qui ne l'utilisent pas (pour ces deux derniers groupes, le rang du logiciel n'est pas pris en compte).

Techniques «souvent » ou « très souvent »	mono non-SAS	mono SAS	pluri non-SAS	pluri SAS
Nombre	78	68	175	218
Médiane	2	4	4	4
Quartiles	[1 - 3]	[2 - 6]	[2 - 6]	[3 - 6]
Maximum	10	10	15	12

Tableau 9 : Nombre de Techniques selon la catégorie « Logiciels »

Les mono-utilisateurs non-SAS constituent un groupe très particulier qui utilisent beaucoup moins de techniques statistiques que les autres sous-groupes : le pourcentage de non-utilisateurs de techniques dans ce sous-groupe est toujours supérieur aux trois autres groupes. Parmi les techniques les plus utilisées :

jamais (%)	statistiques descriptives	régression linéaire	analyse de la variance	analyse factorielles	classifica-tion
mono non-SAS	4	21	28	37	36
mono SAS	0	12	15	19	19
pluri non-SAS	2	8	13	15	21
pluri SAS	0,5	9	14	13	18

Tableau 10 A : Pourcentage de non-utilisateurs de techniques selon la catégorie « logiciels »

Parmi les techniques peu utilisées :

jamais (%)	séries chronologiq ues.	Contrôle Qualité	tests. non paramétri ques	survie	marketing
mono-non SAS	67	76	60	91	74
mono-SAS	58	78	41	68	73
pluri non SAS	54	71	35	83	69
pluri SAS	48	79	27	64	76

Tableau 10 B : Pourcentage de non-utilisateurs de Techniques selon la catégorie « Logiciels »

On retrouve dans le dernier tableau la place particulière des techniques : «contrôle de qualité »et « marketing », qui sont les plus utilisées dans le groupe pluri non-SAS.
En revanche les tests non-paramétriques, les séries chronologiques et la survie sont les plus utilisées dans le groupe pluri SAS.

4.7 Conclusion

Les statisticiens, dont c'est le métier d'analyser la diversité du réel, n'échappent pas à cette diversité : plus de la moitié des répondants utilisent (souvent ou très souvent) plus de quatre techniques et plus de deux logiciels. Parmi les captifs d'un seul logiciel , ceux qui n'utilisent pas SAS n'ont l'usage que de deux techniques pour la moitié d'entre eux. Au point de vue des techniques il est plus difficile de caractériser, parmi les pluri-utilisateurs, ceux qui utilisent SAS. Cependant, les groupes particuliers émergent : «Contrôle Qualité» et «Marketing» qui utilisent des logiciels spécifiques.

5. Les associations logicielles.

Arnaud BRINGE

INED

5.1 Un ou plusieurs logiciels utilisés ?

5.1.1 Logiciels les plus utilisés et les rang d'utilisation :

D'après les résultats de l'enquête, le tableau suivant donne les logiciels les plus utilisés en nombre et le rang moyen d'utilisation de chaque logiciel :

Logiciel	Effectif	Rang moyen d'utilisation
SAS	289	1.39
Statgraphics	106	1.70
SPSS	89	1.77
SPAD-N	88	2.32
Excel	63	2.38
StatLab	62	2.34
S-Plus	47	2.34
Addad	44	2.61
StatITCF	39	2.92
Destin	25	1.16
StatView	25	2.00
Systat	23	2.43
Le Sphinx	21	1.86
BMDP	18	2.50
Question	16	2.12
Epi-info	14	2.36
SPAD-T	14	3.29
StatXact	13	2.92
Statistica	12	2.75
Unistat	11	2.54

Certains logiciels semblent plus utilisés comme logiciel principal (SAS, Destin) voire unique, tandis que d'autres, faisant appel à des techniques spécialisées apparaissent comme des logiciels de complément (SPAD-T, Stat-ITCF, StatXact).

5.1.2 Utilisation exclusive d'un logiciel :

Pour chaque logiciel a été calculée la proportion d'utilisation exclusive, qui donne le pourcentage d'utilisateurs n'utilisant *que* ce logiciel :

Logiciel	Nombre d'utilisateurs	Nombre d'utilisateurs exclusifs	Ratio
Destin	25	9	36%
Le Sphinx	21	7	33%
SAS	289	65	22%
StatLab	62	11	18%
Statgraphics	106	16	15%
SPSS	89	12	13%
S-Plus	47	6	13%

On retrouve des logiciels transversaux (SAS, StatGraphics, SPSS), mais aussi des logiciels plus spécialisés destinés à un public précis (Sphinx). Inversement d'autres logiciels sont nettement moins utilisés de manière exclusive :

Logiciel	Nombre d'utilisateurs	Nombre d'utilisateurs exclusifs	Ratio
Epi-info	14	1	7%
Addad	44	2	4%
Excel	63	1	1%
SPAD-N	88	1	1%

Ces logiciels sont plutôt des logiciels de complément, axés sur des techniques peu diffusées dans les logiciels américains (Analyse de données avec SPAD-N ou ADDAD), ou de reporting (Excel).

5.1.3 Utilisation de plusieurs logiciels :

Le tableau suivant indique le nombre moyen de logiciels par domaine d'activité :

Domaine d'activité	Nombre moyen de logiciels utilisés	Ecart Type	Nombre de personnes interrogées
Administration	1.61	1.20	26
Pharmacie	2.54	1.11	48
Enseign./Rech.	3.04	1.60	190
Agro-alimentaire	2.39	1.50	18
Tranport/Energie	2.07	1.38	29
industrie	2.37	1.34	43
Banque/Assurances	2.29	1.07	35
Sondages	2.10	0.99	19
Economie/Finance	3.00	2.24	11
Etudes & services	2.35	1.46	75

Le domaine Enseignement/Recherche est le secteur où les personnes utilisent le plus de logiciels. Ceci est probablement dû à l'utilisation de techniques de pointe, et à la comparaison de différents produits du marché à des fins pédagogiques.
Inversement, le secteur Administration n'utilise en moyenne qu'un nombre de logiciels assez faible.
Le chiffre apparemment élevé dans le secteur Economie Finances ne peut être réellement interprété, compte tenu du faible effectif et de l'écart type important.
Afin de capter les déterminants majeurs de l'utilisation de plusieurs logiciels, une régression logistique basée sur l'utilisation de plusieurs logiciels a mis en évidence les critères les plus discriminants parmi les variables suivantes :

- Domaine d'activité (10 postes),
- Ancienneté,
- Responsable du choix de ses logiciels statistiques (Oui/Non),
- Degré d'utilisation des 16 différentes techniques,
- Formation,
- Appartenance à un club d'utilisateurs,
- Nombre de machines utilisées.

Les éléments les plus influents de manière positive sur le faut d'utiliser plusieurs logiciels sont les suivants :

- Plans d'expérience,
- Enseignement Recherche,
- Responsable du choix de ses logiciels,
- Nombre de machines utilisées

Le domaine d'activité Administration est l'élément le plus influent négativement sur le fait d'utiliser plusieurs logiciels statistiques.

5.2 Les associations les plus fréquentes :

Le tableau ci-dessus indique les associations les plus fréquentes. Les associations SAS*SPAD-N, Excel*SAS, SAS*SPSS, Addad*SAS, SAS*StatGraphics apparaissent comme fortement utilisées, alors que des associations comme SPADN*SPADT, Addad*SPADN, apparaissent plutôt comme des associations de confort, moins utilisées en rangs 1 à 3 (donc prioritairement), par des utilisateurs ayant à leur disposition plus de logiciels.

Logiciel	Logiciel	Effectif	Dont de rang 1,2 et 3
SAS	**SPAD-N**	54	44 (78%)
Excel	**SAS**	34	28 (82%)
SAS	**Statgraphics**	33	26 (79%)
SAS	**SPSS**	32	26 (81%)
S-Plus	**SAS**	28	18 (64%)
Addad	**SAS**	26	22 (85%)
SAS	**StatITCF**	21	13 (62%)
SPAD-N	**SPSS**	21	13 (62%)
SAS	**StatLab**	17	12 (70%)
BMDP	**SAS**	13	10 (77%)
Statgraphics	**SPSS**	13	9 (69%)
SAS	**StatXact**	12	9 (75%)
SPAD-N	**SPAD-T**	12	6 (50%)
Addad	**SPAD-N**	11	3 (27%)
Excel	**Statgraphics**	11	9 (82%)
SAS	**Systat**	11	8 (73%)
SPAD-N	**StatLab**	11	7 (64%)
Statgraphics	**Unistat**	11	10 (91%)

5.3 Les techniques les plus fréquentes dans les principales associations de logiciels

On a sélectionné les 8 couples de logiciels les plus fréquents (les 8 premières lignes du tableau précédent).
On n'a indiqué que pour chaque couple de logiciels les techniques qui les différencient dans leur utilisation. Par exemple parmi les personnes utilisant les logiciels SAS et SPADN (44 personnes), celles qui veulent faire des comptages font appel à SAS prioritairement (25 personnes pour SAS contre 6 personnes pour SPADN).

5.3.1 Couple SAS-SPAD (44 personnes)

Technique	SAS	SPAD	Ensemble des répondants
Comptages	25 personnes	6 personnes	62%
Statistiques descriptives	30 personnes	10 personnes	85%
Segmentation	8 personnes	26 personnes	19%
Analyses factorielles	8 personnes	30 personnes	39%

Lecture : Quand SAS et SPAD sont utilisés simultanément par une personne, SAS est utilisé pour faire des comptages dans 57% (25 personnes sur 44) des cas, par contre SPAD est utilisé dans 17% des cas. La valeur 62% de la colonne « Ensemble des répondants » correspond au pourcentage des répondants ayant utilisé « souvent » ou « très souvent » les comptages. Le cardinal de l'ensemble des répondants est égal à 542.

Conclusion : SAS est utilisé majoritairement pour les comptages et statistiques descriptives, SPAD pour les analyses factorielles et la segmentation.

5.3.2 Couple EXCEL-SAS : 28 personnes

Technique	EXCEL	SAS	Ensemble des répondants
Comptages	9	20	62%
Statistiques descriptives	17	22	85%
Régressions linéaires	7	8	47%
Analyse de la variance	6	13	45%

SAS est utilisé majoritairement pour les comptages. Il est à noter que Excel est utilisé dans un nombre de cas non négligeable pour effectuer des statistiques descriptives et des régressions. La non prise en compte initiale de distributeurs de produits Add-Ins d'Excel (par exemple Statbox) ne permet pas d'évaluer l'importance de l'utilisations de « statistiques avancées » dans un environnement de type tableur.

5.3.3 Couple SAS-StatGraphics : 26 personnes

Technique	SAS	StatGraphics	Ensemble des répondants
Comptages	9	6	62%
Statistiques descriptives	14	16	85%
Régressions linéaires	14	12	47%
Analyse de la variance	16	9	45%
Séries chronologiques	6	8	14%

On ne trouve aucun secteur où un des deux logiciels n'est utilisé de manière spécifique

5.3.4 Couple SAS-SPSS : 26 personnes

Technique	SAS	SPSS	Ensemble des répondants
Comptages	13	11	62%
Statistiques descriptives	17	14	85%
Régressions linéaires	8	8	47%
Analyse de la variance	15	15	45%
Analyse discriminante	10	11	14%
Analyses factorielles	8	7	39%

5.3.5 Couple SAS-Splus : 18 personnes

Technique	Splus	SAS	Ensemble des répondants
Comptages	4	7	62%
Statistiques descriptives	5	10	85%
Régression	8	8	47%
Analyses factorielles	5	4	39%
Segmentation	6	6	19%

Les utilisations sont équivalentes, en termes de techniques utilisées.

5.3.6 Couple SAS-Addad : 22 personnes

Technique	SAS	Addad	Ensemble des répondants
Comptages	15	0	62%
Statistiques descriptives	18	0	85%
Analyse de la variance	11	3	45%
Analyses factorielles	9	18	39%
Segmentation	9	15	19%

SAS est nettement plus utilisé pour les comptages et statistiques descriptives. Addad est clairement utilisé pour les méthodes d'Analyse de Données (Analyses factorielles, segmentation), bien que SAS soit aussi utilisé, de manière non négligeable, pour ces méthodes.

5.3.7 Couple SAS-Stat Itcf : 13 personnes

Technique	SAS	Stat Itcf	Ensemble des répondants
Comptages	5	0	62%
Statistiques descriptives	8	2	85%
Régressions linéaires	8	10	47%
Analyse de la variance	10	10	45%

SAS est nettement plus utilisé pour les comptages et statistiques descriptives. Les régressions linéaires et analyses de la variance sont utilisées dans ce groupe, aussi bien avec SAS qu'avec Stat-Itcf.

5.3.8 Couple SPADN - SPSS : 13 personnes

Technique	SPAD-N	SPSS	Ensemble des répondants
Comptages	1	8	62%
Statistiques descriptives	4	9	85%
Analyse de la variance	1	1	45%
Analyses factorielles	9	4	39%
Tests non param.	4	3	26%
Segmentation	9	5	19%

SPSS est nettement plus utilisé pour les comptages et statistiques descriptives. L'analyse de données est prise en charge prioritairement par SPAD, mais est aussi effectuée sous SPSS. Les tests paramétriques et les techniques de marketing demeurent plus l'apanage de SPSS.

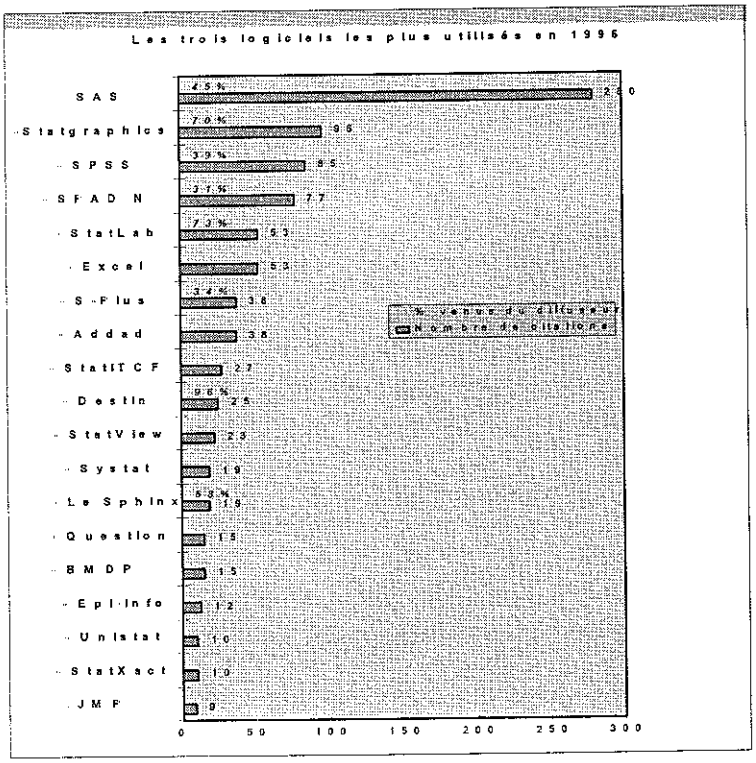
6. Techniques utilisées selon les logiciels

Laurence Hauesler

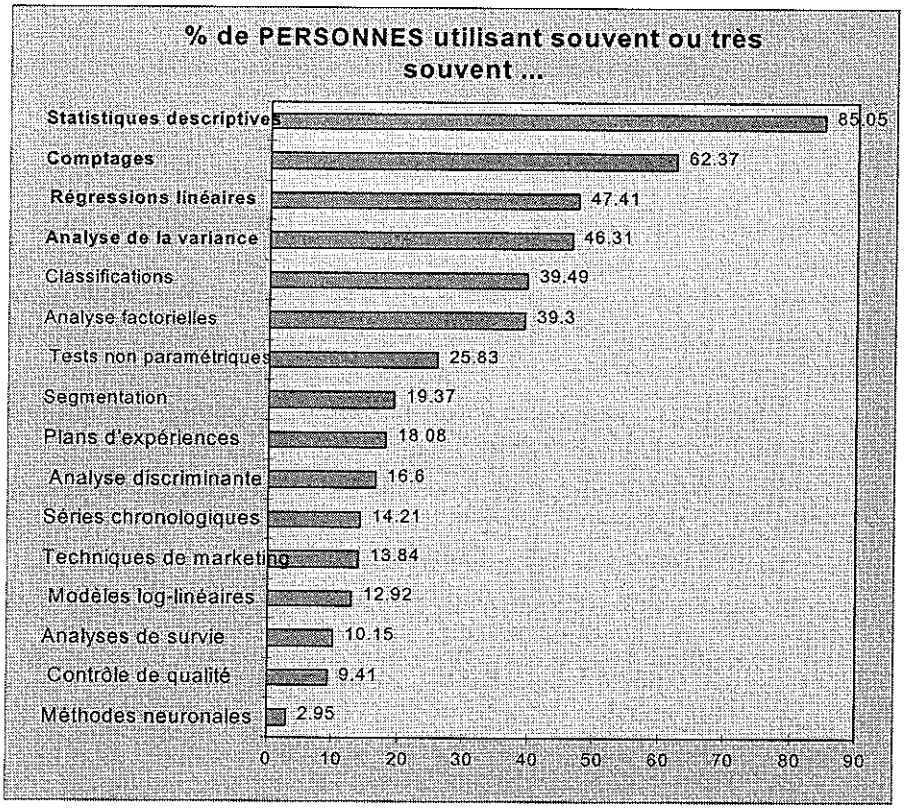
Planistat France

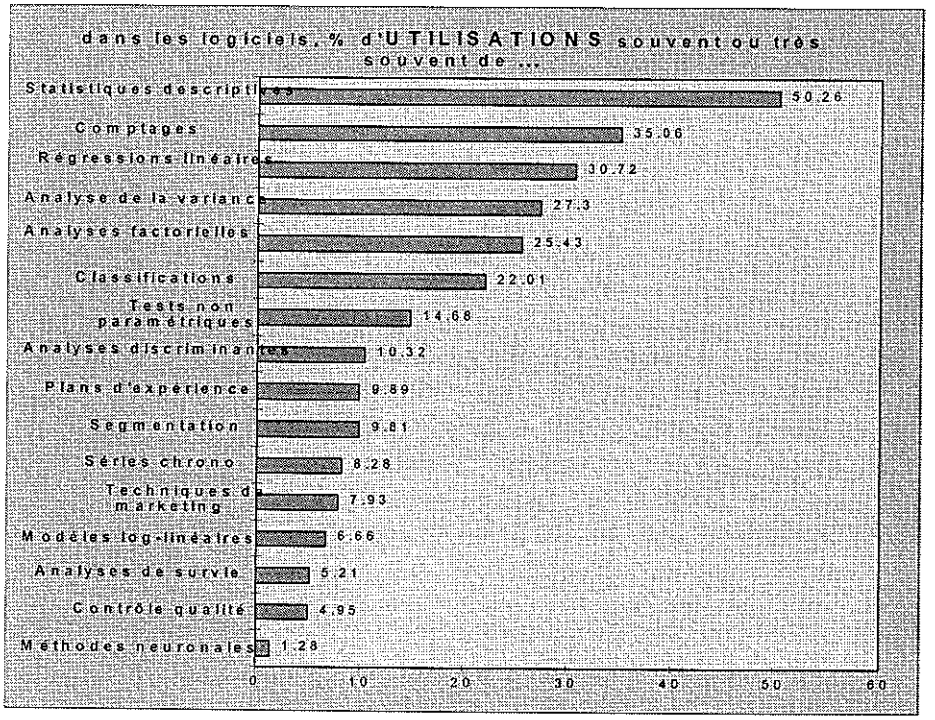
6.1 SAS toujours en tête

En 1996, les logiciels de statistique les plus utilisés dans l'enquête sont SAS (280), Statgraphics (96), SPSS (85), SPAD.N (77), Statlab(53) et Excel (53).
Il est extrêmement difficile d'établir des comparaisons avec l'enquête de 1992, le champ de l'enquête étant intimement lié à la diffusion du questionnaire par les éditeurs de logiciels. Le principal élément de stabilité demeure la prédominance de SAS. L'élément de surprise le plus net (sauf évidemment pour les inconditionnels des tableurs) est l'apparition d'EXCEL dans les logiciels de statistique les plus utilisés. Ce phénomène, qui ne peut être imputé au mode de diffusion du questionnaire, Microsoft ne faisant pas partie des partenaires de l'ASU, est certainement la conséquence des efforts importants d'intégration de fonctions statistiques dans le tableur. On pourrait imaginer également qu'il est la conséquence de l'incapacité des fabricants de logiciels statistiques à fournir des fonctions graphiques simples et puissantes.
Un peu moins d'une dizaine d'éditeurs avaient accepté de diffuser le questionnaire auprès de leurs clients. Pour certains logiciels, les réponses proviennent presqu'exclusivement de questionnaires diffusés par leur éditeur (Destin à 96%, Statgraphics à 70%, Statlab à 73%). Pour d'autres, les réponses proviennent majoritairement d'autres sources que leur propre éditeur (SAS, SPSS, SPAD.N).



Les techniques utilisées dans les logiciels, comme d'ailleurs celles employées par les statisticiens n'ont guère évolué depuis 1992 : les statistiques descriptives en tête, suivies par les régressions et les analyses de variance; les classifications et les analyses factorielles.





6.2 Peu de techniques par logiciel

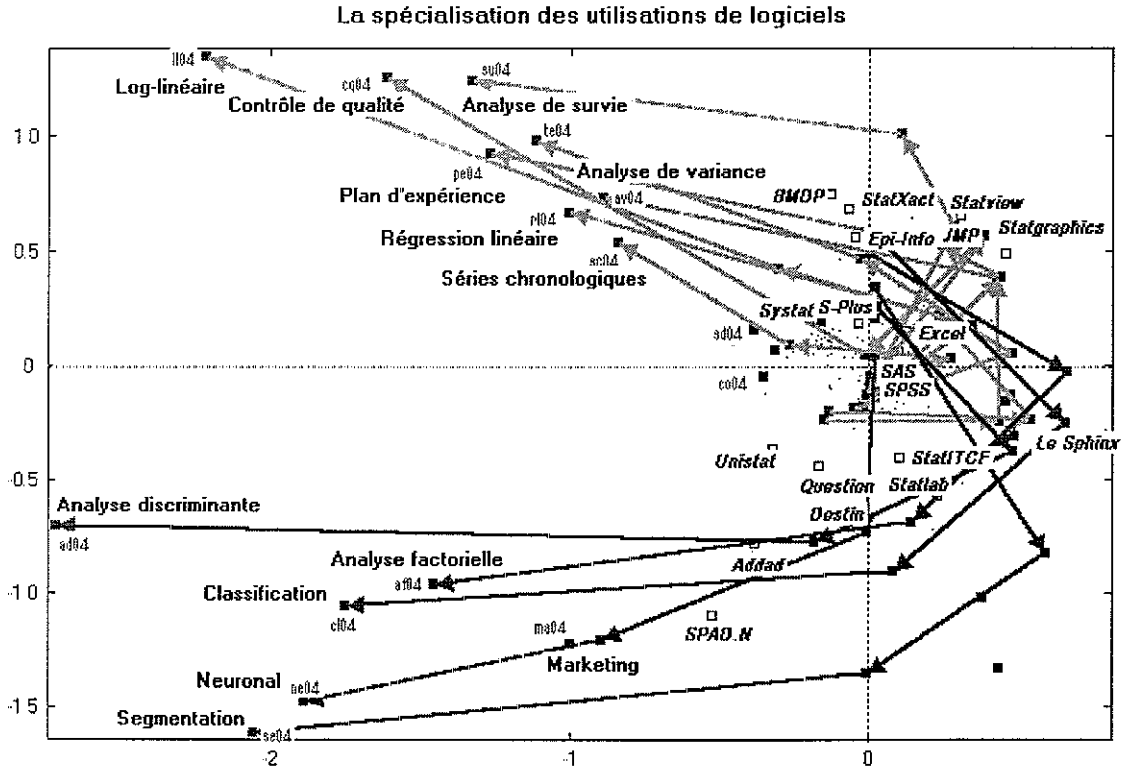
Comme en 1992, la majorité des utilisations de logiciels sont ciblées, et ne concernent que quelques techniques : 11% n'emploient aucune des 17 techniques ou familles de techniques proposées dans le questionnaire, 27% n'en emploient pas plus de 3. Dans le premier cas, il semble bien que ce soient des utilisations très spécifiques (comme le confirme la présence de S+) et pas un abandon complet du logiciel, car les opinions ne sont pas forcément négatives, encore que présentant beaucoup de non-réponses.

6.3 Deux familles d'utilisations : qualitatif et quantitatif ?

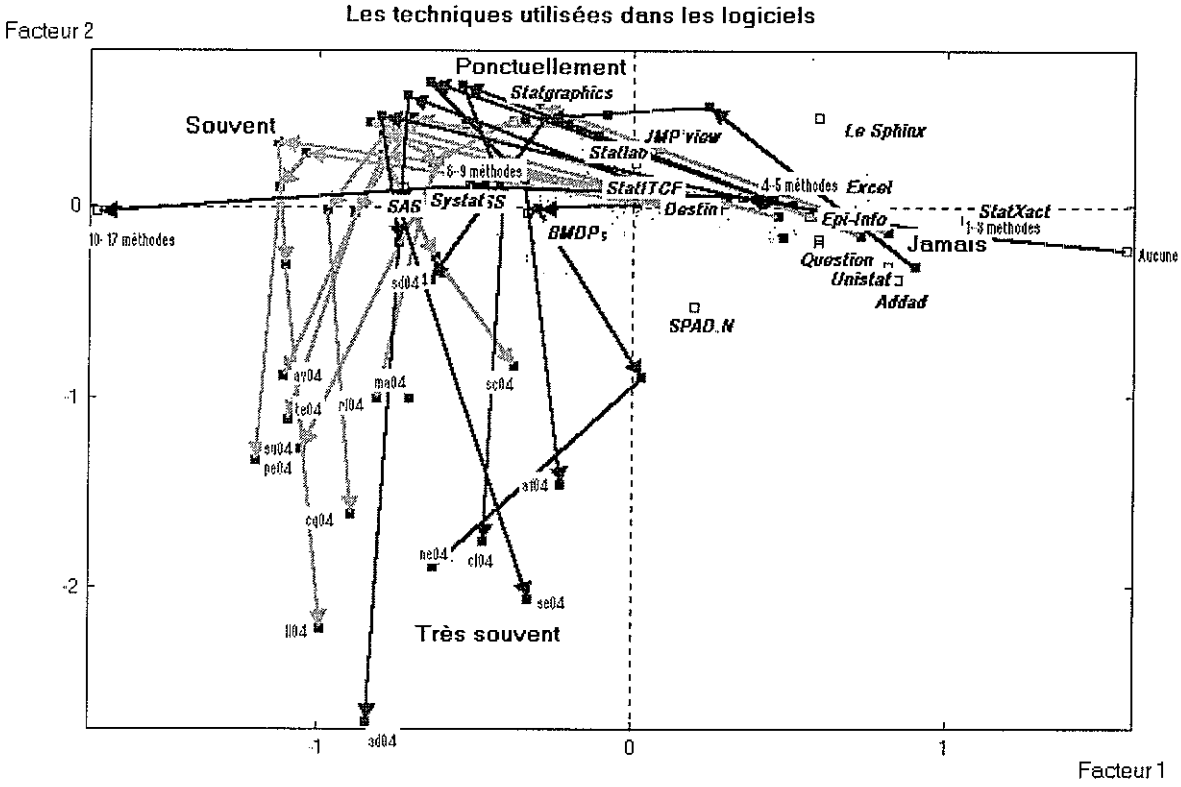
Comme pour l'enquête de 1992, deux grandes familles d'utilisations se dégagent, recoupant plus ou moins une opposition entre modélisations quantitatives et descriptions qualitatives, ou plutôt analyse de variables qualitatives et analyse de variables quantitatives:

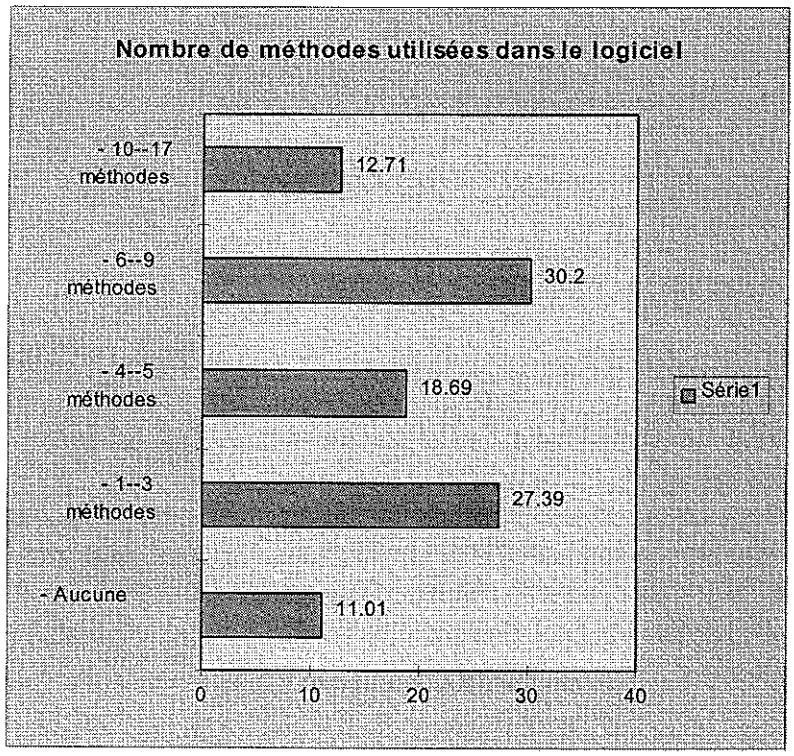
- d'un côté la régression, l'analyse de variance, les séries chronologiques, les plans d'expérience, le contrôle de qualité et les modèles log-linéaires
- de l'autre, l'analyse discriminante, l'analyse factorielle, la classification, la segmentation et les méthodes neuronales.

Cette dichotomie est clairement mise en évidence sur le plan (2,3) de l'ACM ayant comme variables actives les variables de fréquences d'utilisation des techniques dans les logiciels.



Le plan (1,2) représente quant à lui une image de l'évolution des réponses quand le nombre de techniques utilisées augmente





On peut mettre en évidence des utilisations-types de logiciels, qui, malgré les modifications de questionaire, ont peu varié depuis 1992. Les utilisations de logiciels dans lesquelles aucune des techniques proposées n'a été utilisée ont été exclus de l'analyse.

6.4 Huit types d'utilisations de logiciels

Une coupure relativement fine, en 8 classes a été choisie ici.

6.4.1 Les utilisations faibles

Les deux classes les plus importantes sont les deux classes d'utilisation faibles:

- dans l'une (23% des utilisations), quelques statistiques classiques sont utilisés « ponctuellement »: statistiques descriptives, comptages, analyse de la variance et régression. Les méthodes plus spécialisées ne sont jamais employées. 3,8 méthodes sont utilisées en moyenne La moitié des utilisations d'EXCEL sont de ce type.
- dans l'autre (22% des utilisations), il s'agit apparamment de logiciels très spécialisés, que l'on n'emploie que pour un usage très spécifique. 2 méthodes sont employées en moyenne. Presque toutes les utilisations d'ADDAD sont de ce type. Il s'agit toujours d'un logiciel d'appoint, deuxième ou troisième logiciel Les utilisateurs sont assez volontiers critiques quant à la gestion ou l'édition des données.

6.4.2 L'utilisation ponctuelle de nombreuses techniques

Ensuite vient une classe d'utilisation « ponctuelle » d'un nombre assez important de techniques (12% des utilisations). Le nombre moyen de techniques employées est élevé (8 en moyenne). Presque toutes les techniques sont ici utilisées « ponctuellement »: classifications, segmentations, analyses factorielles, analyse discriminante, régressions, marketing, analyse de la variance. Comptages et statistiques descriptives sont utilisés plus fréquemment. Il s'agit souvent du premier logiciel. SAS est le logiciel le plus sur-représenté dans cette classe d'utilisations. Banques, assurances, autres études et services ont fréquemment ce type d'utilisation des logiciels de statistique.

6.4.3 Les utilisations fréquentes plutôt spécialisées

Enfin arrivent les classes d'utilisations fréquentes, relativement plus spécialisées :

- « Souvent » Statistiques descriptives, comtages, analyses de variance; « jamais » de segmentation ou de classification. 12% des utilisations. 7,4 méthodes en moyenne. Logiciel typique : Statview.
- « Souvent » les classifications, segmentations, analyses factorielles, analyses discriminantes. « Jamais » d'analyse de survie. 10% des utilisations. Logiciels typiques : SPAD.N, Statlab
- « Très souvent » les classifications, segmentations, analyses factorielles, analyses discriminantes, techniques de marketing. 6% des utilisations. Nombre moyen de méthodes 7,5. Instituts de sondage.

- « Souvent » : Modèles log-linéaires, analyses de survie, analyse de variance, analyse discriminante, tests non paramétriques, régressions. Beaucoup de méthodes: plus de 10 méthodes en moyenne. 4% des utilisations. Logiciel majoritaire: SAS
- « Très souvent » : analyse de la variance, statistiques descriptives, régressions linéaires, mais aussi plans d'expérience et tests non paramétriques. 11% des utilisations. Logiciel typique: SAS. Nombre moyen d'utilisations 8,8.

6.5 Répartition des utilisations des logiciels les plus fréquents dans les classes

6.5.1 SAS

21% de Classe 1:	utilisation ponctuelle de beaucoup de méthodes
21% de classe 5:	très souvent analyse de variance, régression
16% de classe 7 :	utilisation ponctuelle de quelques méthodes classiques

6.5.2 STATGRAPHICS

39% de Classe 2:	utilisation « souvent » de statistiques descriptives, comptages, analyses de variance. Jamais de classification
17% de classe 7 :	utilisation ponctuelle de quelques méthodes classiques

6.5.3 SPSS

23% de classe 7 :	utilisation ponctuelle de quelques méthodes classiques
22% de Classe 1:	utilisation ponctuelle de beaucoup de méthodes
15% de classe 5:	très souvent analyse de variance, régression

6.5.4 SPAD.N

27% de classe 8 :	utilisation très spécialisée de quelques méthodes
23% de classe 6 :	utilisation très souvent de classification, segmentation, analyse factorielle et discriminante.
22% de classe 3 :	utilisation souvent de classification, segmentation, ana fac, ana. disc

6.5.5 EXCEL

54% de classe 7 :	utilisation ponctuelle de quelques méthodes classiques
18% de classe 8 :	utilisation très spécialisée pour une ou 2 méthodes

6.5.6 STATLAB

23% de classe 7 :	utilisation ponctuelle de quelques méthodes classiques
23% de classe 3 :	utilisation souvent de classification, segmentation, ana fac, ana. disc
19% de Classe 1:	utilisation ponctuelle de beaucoup de méthodes

6.5.7 ADDAD :

75% de classe 8

7. Quels logiciels pour quelles techniques ?

Yves Lechevallier

INRIA_Rocquencourt

Dans la troisième partie du questionnaire chaque utilisateur pouvait décrire, au maximum, trois logiciels. La description de chaque logiciel comprenait :
1. les techniques utilisées par l'utilisateur dans ce logiciel,
2. une évaluation de la qualité des fonctionnalités et de la facilité d'utilisation,
3. le cadre d'utilisation de ce logiciel (environnement informatique, formation, ...).

Pour répondre à cette question nous avons construit un tableau de comptage à partir des réponses à la question L2 qui était « Avec quelle fréquence utilisez-vous les techniques statistiques dans ce logiciels ? ». Les réponses possibles étaient « Jamais », « Ponctuellement », « Souvent », « Très souvent » et la liste des techniques statistiques utilisées était la suivante :

Techniques utilisées	Techniques utilisées
Comptages	**Classification**
Statistiques Descriptives	**Segmentation**
Régressions linéaires	**Test non paramétriques**
Analyse de la variance	**Analyse discriminante**
Plans d'expérience	**Modèles Log-linéaires**
Contrôle de la qualité	**Analyse de Survie**
Analyses factorielles	**Techniques de marketing**
Méthodes neuronales	**Séries chronologiques**

Les colonnes, représentant les variables de ce tableau des données, contiennent le nombre de réponses des utilisateurs et des logiciels pour chaque technique utilisée. Chaque ligne correspond à l'un des logiciels choisis. Nous avons sélectionné les logiciels qui ont été cités dans la question Q19 par, au moins, deux personnes. Cet ensemble représente 90 logiciels mais, parmi eux, il n'y a eu que 77 logiciels décrits par les répondants car ils ne pouvaient, au maximum, décrire que 3 logiciels alors qu'ils pouvaient en citer 7. Chaque case de ce tableau contient le nombre répondants ayant sélectionné cette réponse à cette technique utilisée pour le logiciel décrit.

7.1 Analyses Factorielles

A partir de ce tableau nous avons réalisé l'analyse factorielle binaire de ce tableau de comptage (étape CORBI de SPADN).

Valeur propre	*Pourcentage*	*Pour. Cumulé*
1	**24.72**	**24.72**
2	**14.11**	**38.83**
3	**9.34**	**48.17**
4	**5.67**	**53.84**

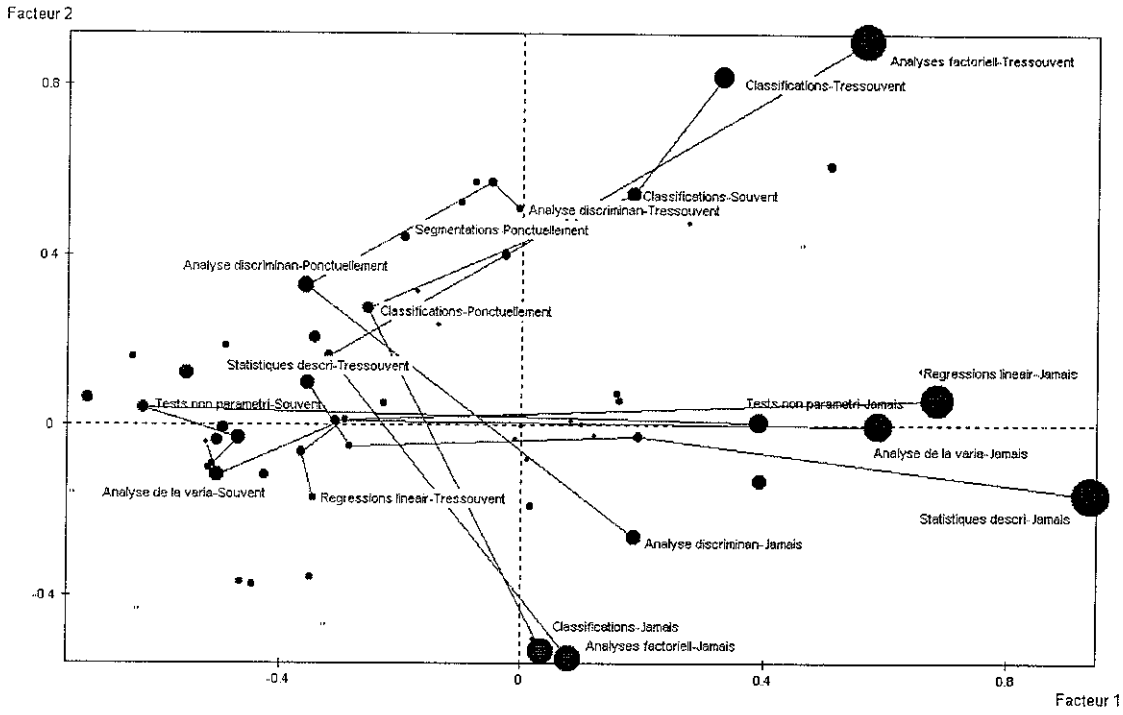


figure 1, le premier plan factoriel

L'interprétation des deux premiers axes est facile. Sur le premier axe s'opposent les logiciels qui sont jamais utilisés pour les statistiques descriptives, la régression linéaire, l'analyse de variance et les tests non paramétriques à ceux qui sont utilisés pour ces techniques. Sur cet axe la graduation d'utilisation de ces méthodes n'est pas significative, on peut regrouper les trois modalités qui caractérisent l'utilisation des logiciels Pour simplifier le discours nous appellerons « méthodes de l'axe1 » ces méthodes.

Sur les deuxième axe s'opposent les logiciels qui sont jamais utilisés pour la classification, l'analyse factorielle et de moindre mesure pour l'analyse discriminante et la segmentation aux autres. Nous appellerons « méthodes de l'axe2 » ces méthodes.

Par contre la graduation de l'utilisation de ces logiciels permet de remarquer que les logiciels utilisant les méthodes de l'axe1 sont aussi « Ponctuellement » utilisés pour les « méthodes de l'axe2 ».

L'axe3 est plus difficile à interpréter. Parmi les logiciels utilisant les « méthodes de l'axe1 », le troisième axe sépare les logiciels utilisés pour les plans d'expériences aux logiciels utilisant « Très souvent » les comptages et les techniques de marketing. Par exemple cet axe oppose Statgraphics et SPSS, Statbox.

Pour les logiciels utilisés « Souvent », « Très souvent » pour les « méthodes de l'axe2 » cet axe sépare les logiciels utilisés « Souvent » pour les techniques de marketing et « Très souvent » pour la segmentation et les méthodes neuronales aux logiciels utilisés « Très souvent » pour l'analyse factorielles et « Jamais » pour les comptages et les statistiques descriptives. Nous trouvons ici une opposition entre Modulad, Addad et SPADS, KnowledgeSeeker.

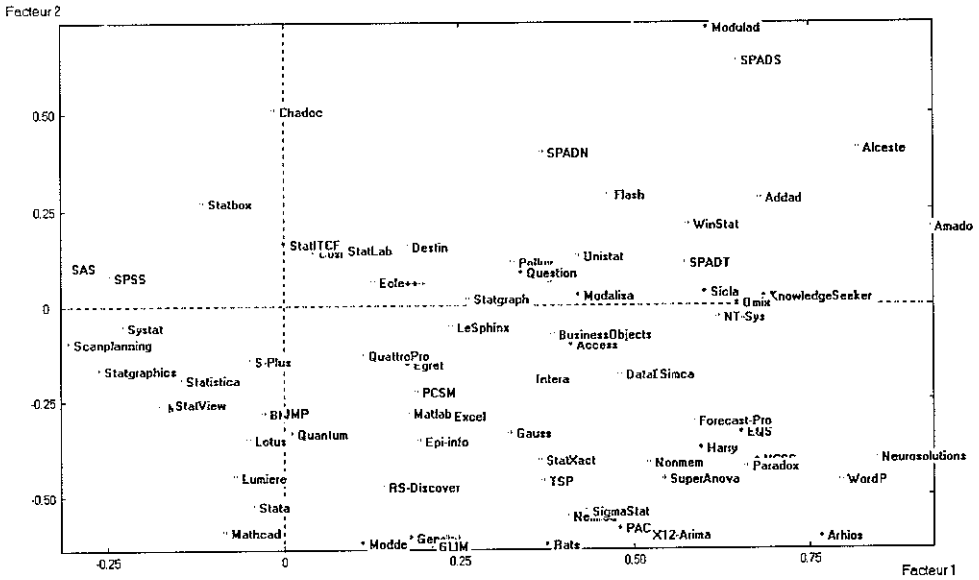


figure 2, les 77 logiciels sur le premier plan factoriel

Nous avons sélectionné les 20 logiciels les plus utilisés. Les deux premiers axes de cette nouvelle analyse factorielles sont identiques aux deux premiers axes de l'analyse factorielle sur les 77 logiciels de la figure 2.

7.2 Classifications des logiciels

Nous avons réalisé une classification en 5 classes de ces logiciels sur les 10 premiers axes factoriels. Nous avons pris comme variables illustratives les variables de la première partie et de la deuxième partie de ce questionnaire. Voici l'interprétation de cette partition en 5 classes.

CLASSE 1 / 5

Voici la liste des 36 logiciels de cette classe. Les logiciels ayant le nom en gras appartiennent aux logiciels les plus cités dans cette enquête.

Arhios	DataDesk	Egret	**Epi-info**	EQS	**Excel**
Forecast-Pro	Gauss	Genstat	GLIM	Harry	Intera
Lotus	Mathcad	Matlab	Modde	NCSS	Nemrod
Neurosolutions	Nonmem	PAC	Paradox	PCSM	Quantum
QuattroPro	Rats	RS-Discover	SigmaStat	Simca	Stata
StatXact	StatXP	SuperAnova	TSP	Word	X12-Arima

Les logiciels de cette classe sont « jamais » utilisés pour les « méthodes de l'axe 1 » et de « l'axe 2 ». Cependant certains sont « très souvent » utilisés pour les séries chronologiques. Les variables illustratives les plus liées à cette classe sont la question Q2 « domaine d'activité » dont les modalités les plus significatives sont l'« industrie pharmaceutique » et la « banque/assurance » et la question Q1 « secteur d'activité » la modalité « privé » est significative. Ces utilisateurs ont un diplôme en statistique mais leur formation est une formation interne et la base de données plutôt utilisée est Oracle.

CLASSE 2 / 5

Voici la liste des 10 logiciels de cette classe.

Access	BusinessObjects	Cosi	**Destin**	Ecle+++	**Flash**
KnowledgeSeeker	**LeSphinx**	Pollux	**Question**		

Les logiciels de cette classe ne sont « jamais » utilisés par les « méthodes de l'axe 1 » mais utilisés de manière intensive pour les techniques de marketing et de segmentation.
Ici le domaine d'activité des utilisateurs est le « sondage » et « études&services ». Ils n'ont pas de diplôme en statistique, ni de formation en statistique et n'utilisent pas le réseau Internet.

CLASSE 3 / 5

Voici la liste des logiciels de cette classe.

Addad	Alceste	Amado	Eyelid	Modalisa	Modulad
NT-Sys	Omix	Sicla	SPADN	SPADS	SPADT
Statgraph	Unistat	WinStat			

Les 15 logiciels de cette classe ne sont « jamais » utilisés pour les « méthodes de l'axe 1 » mais utilisés de manière intensive pour les « méthodes de l'axe 2 » et pour la segmentation. De plus ils ne sont « jamais » utilisés pour les comptages, les plans d'expérience et les séries chronologiques.
Les utilisateurs sont dans le secteur d'activité de l'administration et ont une formation en statistique et appartiennent à un club d'utilisateurs de logiciels et sont sur le réseau Internet.

CLASSE 4 / 5

Voici la liste des logiciels de cette classe.

BMDP	JMP	Lumiere	Minitab	S-Plus	Scanplanning
Statgraphics	Statistica	StatITCF	StatLab	StatView	Systat

Les 12 logiciels de cette classe sont « souvent» et même « très souvent » utilisés pour les « méthodes de l'axe 1 » mais utilisés de manière ponctuelle pour les plans d'expérience et le contrôle de qualité. Les utilisateurs ont comme secteur d'activité le secteur de l'administration, de l'enseignement et de la recherche. Ils ont eu une formation en statistique mais n'ont pas de diplôme en statistique. La base de données utilisée est Access.

CLASSE 5 / 5

Chadoc	SAS	SPSS	Statbox

Les 4 logiciels de cette classe sont « très souvent » utilisés pour les comptages et pour les « méthodes de l'axe 1 » Ils sont utilisés de manière ponctuelle pour de nombreuses méthodes, ce sont les logiciels les plus généralistes. La base de données utilisée est la base SAS et les utilisateurs ont eu une formation interne.
Nous retrouvons cette classification sur l'arbre hiérarchique obtenu à partir des 20 principaux logiciels utilisés.

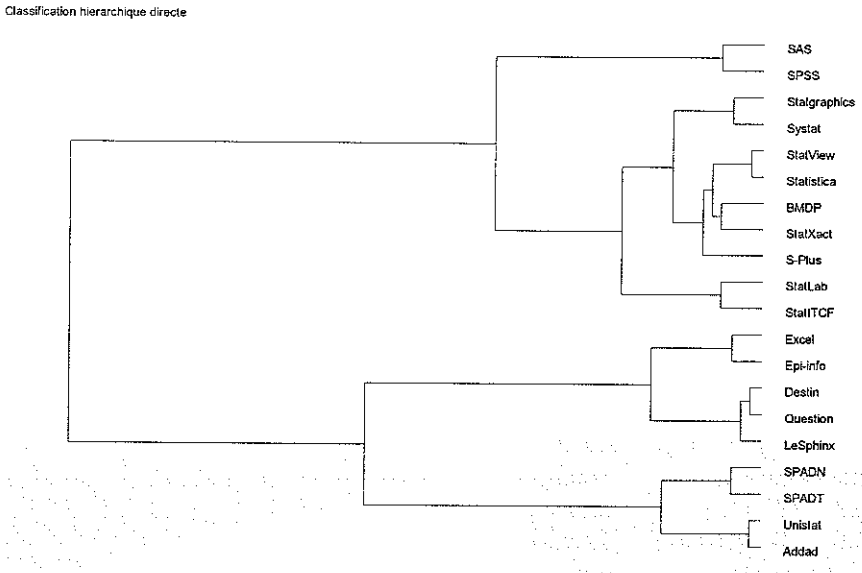


figure 3, classification hiérarchique des 20 logiciels les plus cités

7.3 Classification croisée entre les logiciels et les méthodes

Cette méthode (commande CROKI2 de SICLA) permet de réaliser une classification simultanée des lignes et des colonnes d'un tableau de comptages.
Nous avons sélectionné une classification en 5 classes des logiciels et une classification en 7 classes des réponses aux méthodes utilisées. La partition en 5 classes des logiciels est très semblable à la partition obtenue dans le paragraphe précédent. La classification en 7 classes des réponses aux méthodes statistiques, décrite ci dessous, se retrouve sur le premier plan factoriel.

Classe 1/7 des méthodes statistiques. Cette classe correspond à une utilisation intensive des « méthodes de l'axe 2 ».

Analyse discriminante	Très souvent
Analyses factorielles	Très souvent
Classification	Souvent, Très Souvent
Méthodes neuronales	Souvent
Segmentation	Souvent

Classe 2/7 des méthodes statistiques. Cette classe correspond à un usage très spécifique de la segmentation et à une absence d'utilisation des « méthodes de l'axe 2 ».

Analyse discriminante	Jamais
Analyses factorielles	Jamais
Classification	Jamais
Segmentation	Très souvent

Classe 3/7 des méthodes statistiques. Cette classe regroupe toutes les « méthodes de l'axe 1 ».

Analyse de la variance	Ponctuellement ,Souvent, Très Souvent
Plans d'expérience	Ponctuellement ,Souvent, Très Souvent
Contrôle de la qualité	Ponctuellement ,Souvent, Très Souvent
Analyses factorielles	Ponctuellement
Statistiques Des	Souvent
Régressions linéaires	Ponctuellement ,Souvent, Très Souvent
Séries chronologiques	Ponctuellement ,Souvent
Comptages	Souvent
Test non paramétriques	Ponctuellement ,Souvent, Très Souvent

Classe 4/7 des méthodes statistiques. Cette classe représente une absence d'utilisation des méthodes des statistiques élémentaires.

Statistiques Des.	Jamais
Comptages	Jamais

Classe 5/7 des méthodes statistiques. Cette classe correspond à une absence d'utilisation des « méthodes de l'axe 1 ».

Plans d'expérience	Jamais
Contrôle de la qualité	Jamais
Analyse de Survie	Jamais
Modèles Log-linéaires	Jamais
Méthodes neuronales	Jamais
Segmentation	Jamais
Analyses factorielles	Souvent
Statistiques Des.	Ponctuellement
Séries chronologiques	Jamais
Comptages	Ponctuellement
Tech. de marketing	Jamais

Classe 6/7 des méthodes statistiques correspond à une utilisation ponctuelle des méthodes.

Analyse discriminante	Ponctuellement ,Souvent
Analyse de Survie	Ponctuellement ,Souvent, Très Souvent
Contrôle de la qualité	Ponctuellement ,Souvent, Très Souvent
Analyses factorielles	Ponctuellement
Classification	Ponctuellement
Méthodes neuronales	Ponctuellement
Statistiques Des	Très Souvent
Modèles Log-linéaires	Ponctuellement ,Souvent, Très Souvent
Segmentation	Ponctuellement ,Souvent
Comptages	Très Souvent
Tech. de marketing	Ponctuellement ,Souvent, Très Souvent

Classe 7/7 des méthodes statistiques. Cette classe correspond à un usage très spécifique des méthodes neuronales et à une absence d'utilisation des « méthodes de l'axe 1 ».

Analyse de la variance	Jamais
Régressions linéaires	Jamais
Test non paramétriques	Jamais
Méthodes neuronales	Très souvent

A partir du tableau de contingence entre les classes de méthodes statistiques et classes de logiciels, le nombre mis dans une case de ce tableau de contingence est égal au nombre d'utilisateurs de logiciels appartenant à cette classe de logiciels (classe en ligne) ayant décrits ces logiciels par les modalités de variables appartenant à la classe de méthodes statistiques (classe en colonne). Ainsi nous pouvons quantifier les relations entre méthodes statistiques et logiciels.

méthodes	1/7	2/7	3/7	4/7	5/7	6/7	7/7	total
Classe1/5	*4*	**190**	*112*	**105**	520	*38*	**146**	1115
Classe2/5	61	344	384	49	1322	316	**355**	2831
Classe3/5	**198**	104	*170*	*142*	950	185	**286**	2035
Classe4/5	*63*	335	861	131	1364	346	*153*	3253
Classe5/5	162	351	1171	*68*	1898	**1072**	*239*	4961
total	488	1324	2698	495	6054	1957	1179	14195

figure 4, tableau de contingence entre les classes de logiciels et les classes des méthodes statistiques

Une case en *italique* signale que l'effectif des réponses est inférieur à la moitié de l'effectif des réponses qui aurait été obtenu dans le cas d'indépendance entre ces deux classes. Une case en **gras** signale que l'effectif des réponses est supérieur au double de l'effectif des réponses qui aurait été obtenu dans le cas d'indépendance entre ces deux classes.

Nous pouvons avoir une double lecture de ce tableau, par exemple la classe 3/5 des logiciels est caractérisée par un nombre de réponses élevé aux modalités des méthodes statistiques des classes 1/7 et 7/7 (utilisation de manière fréquente des « méthodes de l'axe 2 » et non utilisation des « méthodes de l'axe1 ») et par un nombre faible de réponses aux modalités des méthodes statistiques des classes 3/7 et 4/7. De même la classe 1/7 des méthodes statistiques est caractérisée par un nombre d'utilisations faible des logiciels des classes 1/5 et 4/5 et par nombre d'utilisations élevé des logiciels de la classe 3/5. Cette interprétation peut être confirmée en réalisant une analyse factorielle de ce tableau de contingence ; le premier plan factoriel est un simplification et une synthèse du premier plan factoriel des figures 1 et 2.

La figure 5, qui est une représentation graphique sous forme d'un histogramme, permet d'avoir une synthèse des relations entre logiciels et méthodes statistiques.

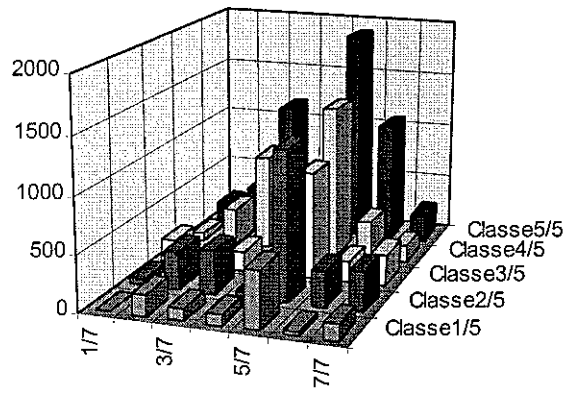


figure 5, Visualisation du tableau des comptages entre les classes de logiciels et les classes des méthodes statistiques

La figure 6 visualise la répartition des réponses aux classes des méthodes statistiques des classes de logiciels. Chaque ligne représente une classe des cinq classes de logiciels et sur cette ligne est donné la proportion des réponses dans chacune des classes des méthodes statistiques des répondants utilisant les logiciels de cette classe. Les réponses appartenant aux modalités de la classe 5/7 des méthodes statistiques sont les plus nombreuses dans chacune des classes de logiciels ; elles correspondent à la réponse « jamais » aux méthodes peu utilisées. Les utilisateurs des logiciels de la classe 3/5 ont proportionnellement beaucoup plus choisi les modalités de la classe 1/7 des méthodes statistiques, ce qui correspond à une utilisation de manière intensive des méthodes d'analyse de données. Par contre les classes 4/5 et 5/5 des logiciels sont associées aux réponses de la classe 3/7, caractérisant une utilisation des méthodes d'analyse de la variance, des plans d'expérience, contrôle de qualité et régression linéaire.

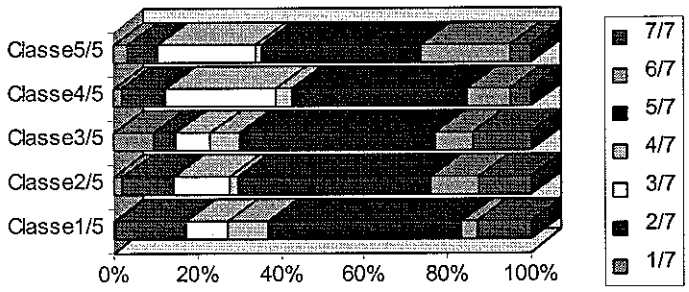


figure 3c, Proportions des classes des méthodes statistiques dans les classes de logiciels

Cette classification croisée permet de synthétiser les informations du tableau de comptages entre les logiciels et le mode d'utilisation des méthodes statistiques. Ce résultat est conforme à l'interprétation du premier plan factoriel mais ici nous pouvons quantifier cette interprétation. Par exemple les logiciels de la classe 4/5 représentent 22,9% (3253/14195) des réponses mais 40,5% (198/488) des réponses correspondant aux modalités de la classe 1/7 des méthodes statistiques.

8. Qualité et facilité d'utilisation des logiciels

Yves Lechevallier

INRIA_Rocquencourt

L'analyse de la qualité et de la facilité d'utilisation des logiciels est faite à de la question L5 « Donner votre avis sur ces logiciels ». Les réponses possibles étaient « Médiocre », « Peu satisfaisant », « Assez satisfaisant » et « Très satisfaisant ». Le répondant devait donner son avis sur « la qualité des fonctionnalités du logiciel » et sur « la facilité d'utilisation du logiciel » d'une manière globale et localement sur les huit fonctionnalités suivantes :

Appréciation d'ensemble	Aide à l'interprétation
Gestion des données	Graphisme
Import/export des données	Documentation technique
Edition des données	Documentation statistique
Import/export des résultats	

8.1 Analyse des deux questions d'appréciation d'ensemble

La première analyse factorielle a été réalisée sur les deux questions « appréciation d'ensemble ».

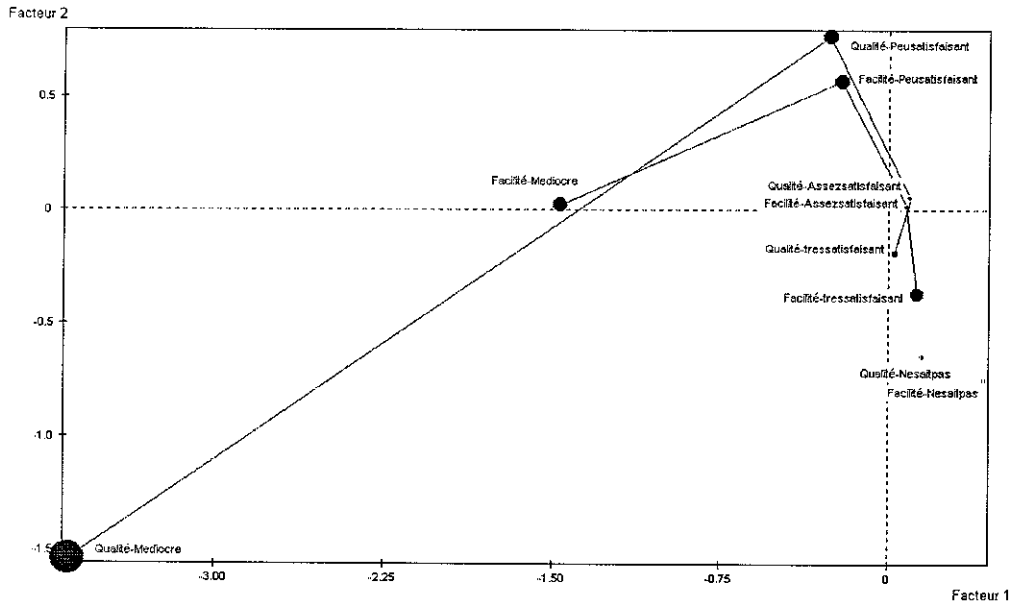


figure 1, Analyse factorielle sur l'appréciation d'ensemble

Le premier plan factoriel oppose les jugements « médiocres » aux autres jugements. Nous pouvons aussi remarquer une liaison forte entre ces deux variables d'appréciation qui indique une corrélation forte entre ces deux jugements de qualité et de facilité, cette corrélation se retrouve pour toutes les fonctionnalités locales.

8.2 Analyse des questions sur la qualité et la facilité

8.2.1 Analyses Factorielles

Nous allons analyser l'ensemble des questions de la rubrique L5 de la troisième partie du questionnaire. Voici les résultats obtenus par l'analyse factorielle.

Valeur propre	*Pourcentage*	*Pour . Cumulé*
1	16.86	16.86
2	11.68	28.51
3	9.85	38.36
4	6.83	45.19

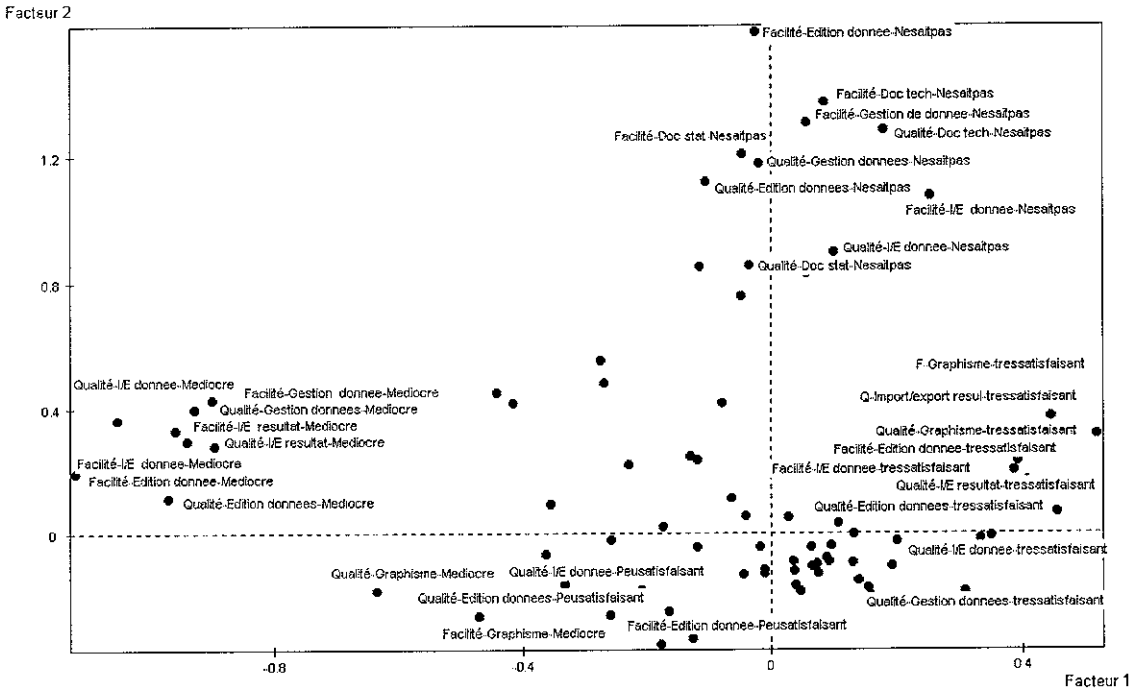


figure 5, la qualité et la facilité sur le premier plan factoriel

Nous pouvons remarquer que le premier axe correspond à une graduation de l'appréciation ; à gauche l'avis est « médiocre » à droite l'avis est « très satisfaisant ». Nous retrouvons ici pour toutes les fonctionnalités une corrélation entre la qualité et la facilité. Les fonctionnalités les plus contributives sont : Gestion des données, Edition des données, Import/Export des données, Import/Export des résultats et Graphisme. Le second axe est caractérisé par la modalité « je ne sais pas » sur la documentation technique et statistique. Sur le troisième axe est caractérisé par un avis médiocre sur la document technique et statistique et sur l'aide à l'interprétation.

8.2.2 Typologie des logiciel

Nous avons choisi une partition des logiciels en 6 classes. Comme la qualité et la facilité sont très liées cette information ne sera donnée que quand il y a une différence entre les deux.
La classe 1/6 représente 11 logiciels et est caractérisée par une documentation technique et statistique et aide à l'interprétation médiocres mais une gestion des données, une édition des données, un Import/Export des données, un Import/Export des résultats et un graphisme assez et même très satisfaisant. On les tableurs et les bases de données dans cette classe.
La classe 2/6 représente 22 logiciels. Dans cette classe on retrouve les fonctionnalités assez et très satisfaisantes de la classe précédente mais avec une bonne documentation technique. Dans cette classe se trouvent l'ensemble les logiciels de la classe 4/5 de l'analyse précédente sur l'utilisation des méthodes statistiques.
La classe 3/6 représente 4 logiciels dont SAS. On retrouve les fonctionnalités de la classe précédente sauf qu'ici le graphisme est jugé « peu satisfaisant » ou « médiocre » par contre la documentation statistique est jugée « très satisfaisante ».
La classe 4/6 représente 10 logiciels. Toutes les fonctionnalités n'ont pas été évaluées par les utilisateurs car cette classe est caractérisée par la modalité « je ne sais pas ».
La classe 5/6 représente 7 logiciels. Cette classe est caractérisée par la modalité « médiocre » pour l'édition, la gestion et Import/Export des données.
La classe 6/6 représente 22 logiciels On retrouve les fonctionnalités de la classe 2 sauf que le jugement est plus sévère (on trouve beaucoup de « peu satisfaisant » et quelques « médiocre ») sauf pour l'aide à l'interprétation qui est ici jugée « très satisfaisante » pour sa qualité. On trouve ici beaucoup des logiciels orientés vers l'analyse des données.

8.2.3 Analyse des 20 logiciels les plus cités

Dans cette analyse nous avons retenu que les 20 logiciels les plus cités. La classification hiérarchique nous donne l'arbre suivant :

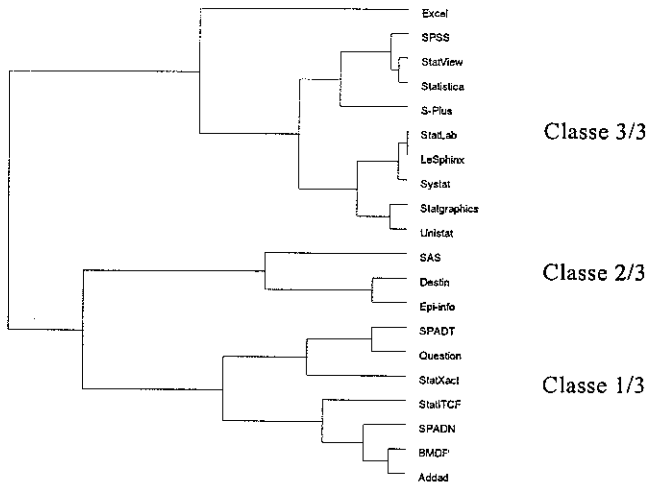


figure 2, la classification des logiciels les plus cités

Puis nous avons réalisé une analyse factorielle.

Valeur propre	*Pourcentage*	*Pour . Cumulé*
1	27.82	27.82
2	16.02	43.85
3	11.50	55.34
4	10.71	66.05

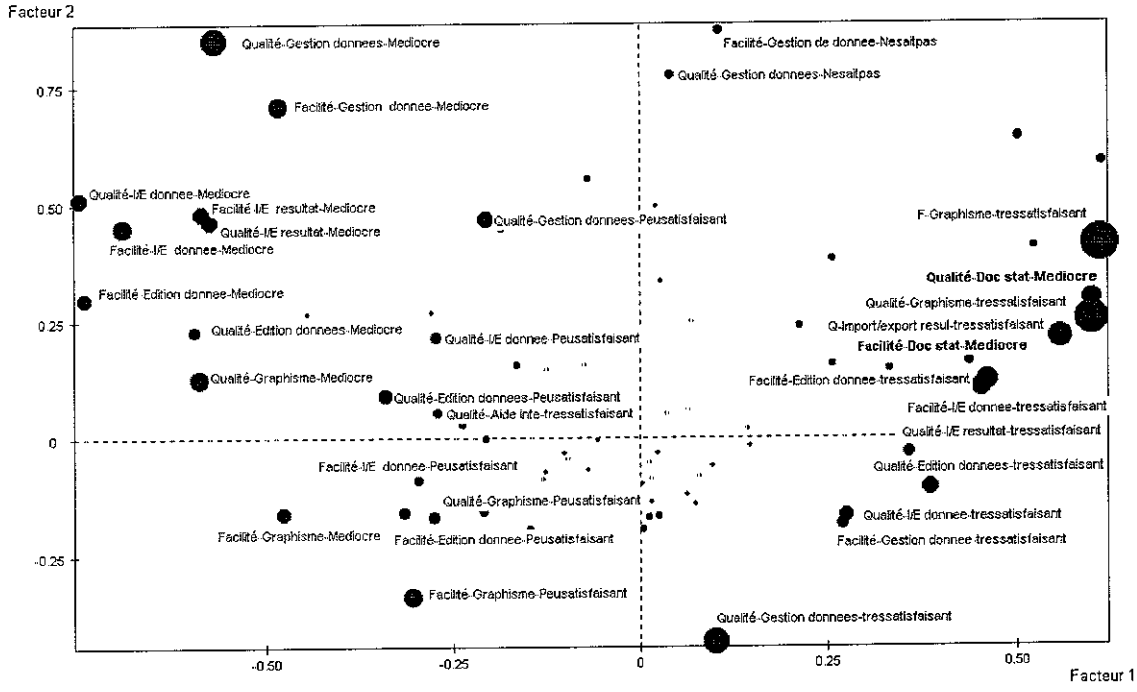


figure 3, la question L5 sur le premier plan factoriel

Nous avons une représentation assez différente que celle obtenue sur l'ensemble des logiciels décrits. Le premier axe est identique au premier axe de l'analyse précédente, on retrouve une opposition entre « mauvais » et « très satisfaisant » sauf pour la documentation statistique dont la modalité « médiocre » se retrouve avec les « très satisfaisant ». Cependant le second axe est très différent, on retrouve bien pour la modalité « médiocre » une séparation entre le graphique et la gestion des données (gestion, Import/export et édition des données) mais on trouve aussi une séparation pour la modalité « très satisfaisant » une séparation entre le graphique et la gestion de données. Sur le troisième axe on trouve les modalités « je ne sais pas » qui s'opposent à la modalité « médiocre » pour le graphisme.

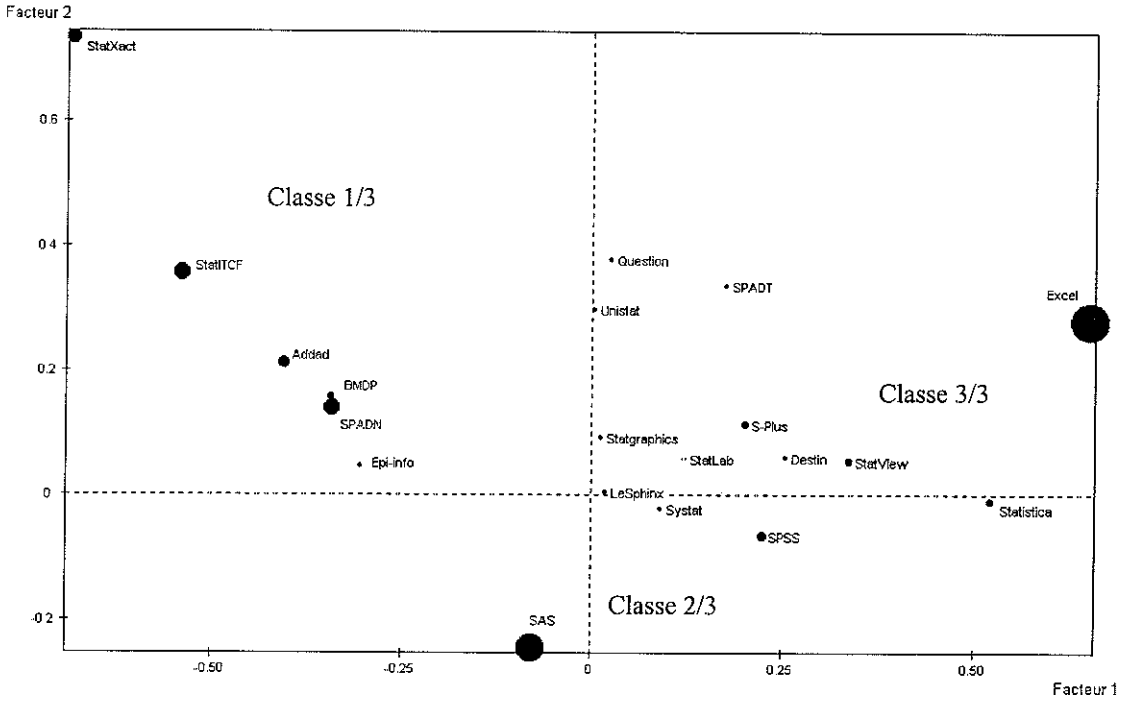


figure 4, les 20 logiciels les plus cités sur le premier plan factoriel

Sur ce plan factoriel on trouve trois grand ensemble de logiciels, l'un est caractérisé par SAS, l'autre par Excel et un ensemble de logiciels de statistique et enfin un groupe de logiciel d'analyse des données.

L'arbre hiérarchique de la figure 2 nous donne une classification en 3 classes de nos 20 logiciels. Cette classification est conforme à la représentation factorielle de la figure 4.

La classe 1/3 représente les 7 logiciels qui se trouvent sur la partie gauche du premier plan factoriel de la figure 4. Cette classe se caractérise par les modalités « médiocre » et « peu satisfaisant » sur l'ensemble des fonctionnalité avec une appréciation « peu satisfaisant » pour l'appréciation d'ensemble sauf pour l'aide à l'interprétation où le jugement est très bon.

La classe 2/3 représente 3 logiciels dont SAS. On trouve ici une qualité d'ensemble très satisfaisante ainsi que pour la gestion des données et la documentation statistique, la facilité d'utilisation est jugée un peu plus sévèrement. Par contre le graphisme est jugé « médiocre » en facilité et « peu satisfaisant » en qualité.

La classe 3/3, représentant 10 logiciels, se trouve sur la partie droite du plan factoriel de la figure 4. Le graphisme et jugé « très satisfaisant », ce résultat est aussi obtenu pour la facilité en gestion, édition et Import /Export des données par contre le jugement sur la qualité est plus sévère.

9. La question ouverte de l'enquête

Mónica Bécue-Bertaut

Universitat Politècnica de Catalunya. Barcelona.

A la fin du questionnaire, on demandait aux répondants de s'exprimer plus librement sur ce qu'ils considéraient être un bon logiciel de statistique au moyen de la question suivante:
L 6 - Vous vous êtes exprimés sur les logiciels de statistiques, qualités et défauts. Pourriez-vous, pour conclure, exprimer ce que serait pour vous de façon générale un bon logiciel de statistique :

Parmi les répondants, 341 (soit 63% des répondants) ont formulé une réponse à cette question ouverte et leur réponse est, en moyenne, assez longue (table 1). On peut noter que le nombre de mots différents est plutôt élevé, ainsi que le nombre de hapax. Il semblerait donc que cette question n'est pas sans intérêt pour les utilisateurs de logiciels de statistique

Nombre total de réponses:	341
Nombre total de mots:	10452
Nombre de mots distincts:	1889
Pourcentage de mots distincts:	18.1
Longueur moyenne des réponses:	30.7

Table 1. Bilan du traitement

9.1 Les thèmes présents dans les réponses

Une partie du glossaire des mots est reproduite à la table 2. De sa lecture, se dégagent les principales qualités recherchées dans un logiciel de statistiques. Pour ces utilisateurs *un logiciel de statistique* a pour fonction de traiter des *données* et produire des *résultats*, ce qui n'est guère surprenant. Mais, quelles qualités demande-t-on à ces logiciels?

Il semble que les répondants insistent davantage sur les qualités dont ils ont souffert le manque, que ce soit dans les logiciels actuellement ou antérieurement utilisés. Et, en général, puisqu'on leur demande les qualités désirées, ils en demandent beaucoup! Certains hiérarchisent leurs demandes, mais ils ne sont qu'une minorité et les décomptes effectués ne tienent pas compte de cette hiérarchisation. On peut aussi noter que le "ne" (29 citations) est assez fréquent; dans ce cas, le répondant définit le logiciel par ce qu'il ne doit pas être ou offrir, en opposition à, très certainement, des mauvaises expériences passées.

On peut résumer les demandes formulées, en regroupant les mots prononcés dans les réponses en "catégories" qui représentent d'une certaine façon les différents "thèmes" auxquels se réfèrent les répondants.

FORMES LEXICALES PAR ORDRE DE FREQUENCE		
450	DE	575
482	DES	322
644	ET	286
441	D	237
1799	UN	227
986	LES	219
22	A	206
1016	LOGICIEI	195
965	L	177
603	EN	157
1800	UNE	148
967	LA	137
176	AVEC	125
1342	POUR	119
1318	PLUS	118
556	DONNEES	117
984	LE	90
1825	UTIIISAIION	84
641	ESI	83
1676	STAIISIIQUE	83
1543	RESUITATS	81
1577	SAS	76
443	DANS	75
1678	STATISIIQUES	75
856	IL	68
1713	SUR	66
1237	PAR	66
1448	QUE	60
1459	QUI	60
1219	OU	59
1262	PAS	58
211	BON	56
540	DOCUMENIAIION	55
1017	LOGICIEIS	54
249	CE	53
69	AIDE	51
1096	METHODES	49
828	GRAPHIQUES	49
939	INTERPRETAIION	48
1785	IRES	40
1149	N	38
726	FACIIE	38
1439	QU	37
934	INIERFACE	37
1631	SONT	37
405	CONVIVIAI	37
655	ETRE	36
566	DU	35
729	FACILIIE	34
1880	WINDOWS	33
1620	SIMPLE	31
99	ANALYSE	31
827	GRAPHIQUE	31
1042	MAIS	31
212	BONNE	30
1604	SERAII	29
1822	UTILISAIEUR	29
712	EXPORT	29
1408	PROGRAMMAIION	29
1151	NE	29
547	DOIT	27
1762	TOUI	27
1114	MISE	27
408	CONVIVIAIIIE	27
206	BIEN	26
146	AU	26
865	IMPORI	26
767	FONCTIONNAIIIES	25
1198	ON	25
1166	NON	25

FORMES LEXICALES PAR ORDRE DE FREQUENCE		
1441	QUALITE	25
1645	SOUS	24
742	FAIRE	24
974	LANGAGE	23
1795	TYPE	23
155	AUSSI	23
1369	PRESENIATION	18
1771	TRAITEMENTS	18
101	ANALYSES	17
1731	TECHNIQUE	17
1090	MENUS	17
1160	NIVEAU	17
184	AYANT	17
1635	SORTIES	16
952	J	16
1661	SPSS	16
182	AVOIR	16
199	BESOINS	16
288	CHOIX	16
1395	PROCEDURES	16
1351	POUVOIR	16
1823	UTIIISAIEURS	16
1827	UTIIISE	16
1670	SIATGRAPHICS	15
346	COMPLET	15
1745	TESTS	15
1848	VERSION	15
1091	MES	14
1721	TABLEAUX	14
1628	SOIT	14
1888	Y	14
818	GRANDE	14
1567	S	14
679	EXCEL	14
1715	SURTOUI	14
1616	SI	14
686	EXEMPIES	13
1788	TROP	13
328	COMME	13
829	GRAPHISME	13
794	FRANCAIS	13
773	FONCTIONS	13
1131	MOINS	12
685	EXEMPIE	12
1649	SPAD	12
1765	TOUTES	12
163	AUTRE	12
1723	TABLEUR	12
1285	PERMETTANI	12
1747	IEXTE	12
1301	PEU	12
1674	STAIISIICIEN	11
347	COMPIEIE	11
1622	SIMPIES	11
745	FAIT	11
867	IMPORIANI	11
1338	POSSIBIE	11
300	CLAIRE	11
508	DEVRAII	11
1593	SE	11
1442	QUALIIES	11
1464	RAPIDE	11
234	CALCULS	10
1257	PARTICUIIER	10
1089	MENU	10
1761	TOUS	10
231	CALCUI	10
1083	MEME	10
1302	PEUT	10
1095	METHODE	10
1222	OUTILS	10

Table 2. Glossaire des mots prononcés au moins 10 fois

Thèmes des réponses

Pour retrouver les thèmes des réponses, employés par les répondants, la lecture du glossaire des mots (élargi aux mots répétés au moins 5 fois) est enrichi par la lecture du glossaire des segments répétés et, dans certains cas, des concordances de certains mots, pour retrouver le contexte dans lequel ceux-ci sont employés quand cela se révèle nécessaire pour mieux appréhender leur signification. Dans ce qui suit, nous énumérons ces thèmes, en précisant les mots employés qui s'y réfèrent, ainsi que leur fréquence.

Facilité d'emploi

Un point très important est la facilité d'emploi, simplicité de maniement ou convivialité du logiciel. Certains, peu nombreux, insistent aussi sur l'aide en ligne. Ainsi on trouve *facile(s)* (38+9), *convivial* (37), *facilité* (34), *facilement* (9), *simple(s)* (31+11), *convivialité* (27), *simplicité* (5), *agréable* (5). On réclame un *"logiciel simple et convivial d'utilisation"*, *"d'abord facile et riche"*, *"simple d'emploi"*, *"à la fois simple et éclectique"*. Bien sûr, cette facilité est extensible aux divers phases du traitement, on veut pouvoir faire un "*import-export de données très facile*", des *"graphiques plus faciles à faire"*, obtenir des "*résultats qui sont facilement mis en forme*" et avoir "*la possibilité de faire des calculs sophistiqués en toute simplicité*".

Méthodes statistiques

Les allusions aux méthodes et techniques statistiques sont, cela n'est guère surprenant, nombreuses: *méthode(s)* (10+49), *technique(s)* (17+21), *traitement(s)* (23+18), *choix* (16), *tests* (15), *calcul (s)* (10+10), *analyse(s)* (31+17), *outil(s)* (7+10), *toutes* (12), *nombreuses* (10), *classiques* (8), *nouvelles* (6). Certains mots sont assez vagues, et font référence tant aux possibilités statistiques qu'à celles plus liées à des problèmes de gestion ou de présentation; ainsi on trouve *fonctionalités* (25) *possibilité(s)* (22+23), *fonction(s)* (9+13), *options* (7), *complet* (15). Les usagers demandent un "*vaste choix de méthodes*", "*une palette complète de méthodes*", "*une panoplie d'outils statistiques étoffée*" ou même réclament de pouvoir appliquer "*toutes les méthodes ou modèles statistiques possibles*" ou au moins une "*grande variété de méthodes*", sans oublier les "*méthodes nouvelles*" que les logiciels devraient intégrer rapidement. Le logiciel *"doit être complet"*. On peut désirer que le logiciel fasse le choix des méthodes, oriente le choix, ou en laisse le contrôle à l'utilisateur: ainsi on trouve *"il lui est demandé de faire des choix pertinents"*, "*choix stratégique des méthodes explicite"*, *"le logiciel devrait orienter les choix des tests*" ou, au contraire, "*possibilité de choix par l'utilisateur*". Bien sûr, on aime avoir "*la documentation précise des méthodes utilisées*".

Documentation complète et claire

Une bonne documentation, complète, technique et statistique est fondamentale: *documentation* (55) *interprétation* (48), *exemples* (13), *français*(e) (13+4), *complète* (11), *documenté* (5), *aide(s)*

(51+5), *claire* (5). La documentation doit être, *"technique et statistique exhaustive"*, *"complète"*, *"claire"*, *"détaillée"*, *"en français"* ou *"traduite"*, *"bien pensée"*, *"rapide d'utilisation"*. Elle doit comporter des exemples, et offrir une *"aide à l'interprétation"*.

Gestion des données, import-export

La gestion des données, ainsi qu'une communication aisée avec les autres logiciels et les bases de données, est une demande très répandue, comme le montre la fréquence des mots suivants: *données* (117), *interface(s)* (37+6), *export* (29), *import* (26), *gestion* (19), *tableur* (12), *saisie* (7), *importation* (7), *exportation* (5), *volumes* (6), *gros* (6), *ouvert* (6), *souple* (5), *recodage(s)* (4+5). On demande un *"accès facile aux données"*, une *"gestion des données facile"*, un logiciel *"convivial et performant pour la gestion des données"*; on veut traiter *"de grands volumes de données"*, et se préoccupe de "l'*importation, exportation des données*", quelquefois du codage et recodage.

Présentation des résultats

Bien sûr, une exploitation confortable des résultats et une intégration aisée des sorties dans les rapport sont souhaitées: *édition* (19), *présentation* (18), *sortie(s)* (6+16), *tableaux* (14), *rapports* (5), *graphique(s)* (31+49), *graphique* (31), *graphisme(s)* (13+7). L'utilisateur recherche *"des résultats qui sont facilement mis en forme pour l'édition des rapports"*, *"avoir plus de possibilités d'édition des graphiques"*, désire un logiciel *"avec un graphisme permettant une édition de qualité"*: L'un d'eux affirme même *"l'enjeu, c'est la présentation des résultats"* et un autre remarque que l'important *"dans l'utilisation quotidienne, est la souplesse de présentation des sorties"*.

Fiabilité du logiciel

Les qualités les plus fondamentales - essentielles, même, ne sont pas très mentionnées, peut être parce qu'on les suppose toujours présentes. Ainsi insister sur le fait que les résultats doivent être exacts, les méthodes fiables et les traitements suffisamment rapides est peu fréquent. On trouve néanmoins *"rigoureux sur le plan des principes des méthodes et du déroulement des calculs"*, *"méthodes statistiques fiables"*, *"fiabilité des méthodes proposées"*. Mais *fiabilité* n'est employé que trois fois, *fiable* et *fiables* respectivement trois et deux fois Il faut des *"méthodes statistiques fiables"* et la *"transparence des méthodes de calcul"*. Les mots *précis, précise* et *précision* sont prononcés chacun trois fois, *précises*, une seule fois. On trouve *juste* et *justesse* une fois chacun *"précis (justesse des résultats)"*. Un répondant affirme que le logiciel *"ne doit pas être buggé"*. On trouve aussi les mots *puissant* (8), *puissance* (9), *performant* (7), *rapide(s)* (11+2), *efficace* (6), *idéal* (6): *"un bon logiciel est aussi un logiciel puissant"*,

Programmation/ développement

Pour certains utilisateurs, la possibilité d'intégrer ou de développer ses propres méthodes, en disposant de la possibilité de programmer est un atout. On trouve cette affirmation *"Pour faire ses études statistiques, par contre, les statisticiens ont besoin d'un logiciel pour travailler par*

programmation ou par langage de commande". Les mots *programmation* (29), *langage* (23), *procédures* (16) sont présents avec ces fréquences

2. Vocabulaire et caractéristiques des répondants

Déterminer les caractéristiques des répondants qui influencent le choix du vocabulaire, et ainsi, dans une certaine mesure, mettre en relief les thèmes privilégiés par certaines catégories, est l'un des objectifs poursuivis.

Dans le cas de cette enquête, la taille réduite de l'échantillon et son homogénéité, quant au niveau d'études et d'environnement de travail, le déséquilibre des effectifs des catégories ou sous-catégories (non représentatif de ce qui existe dans la population), le grand nombre de mots différents employés - correspondant au niveau culturel des répondants, mais aussi le fait que ces mots soient quasi "d'usage obligé", d'où l'extrême importance du contexte des mots (ainsi *méthodes* peut être employé pour demander que le logiciel permette à l'utilisateur de conserver le choix des méthodes ou, au contraire, pour que le logiciel guide ce choix) ne permettent de mettre en évidence que des tendances assez fragiles.

Non expérimentés

	LIBELLE DE LA FORME GRAPHIQUE	---POURCENTAGE--- INTERNE	GLOBAL	FREQUENCE INTERNE	GLOBALE	V.TEST	PROBA
1	TECHNIQUE	1.30	30	5.	17	2.616	.004
2	SORTIE	.78	11	3.	6	2.551	.005
3	DOCUMENTATION	2.34	.98	9.	55.	2.287	.011
4	BASE	78	.16	3	9.	2.068	.019
5	NOMBREUSES	78	.18	3	10.	1.939	.026

Expérimentés, diplômés en statistique

	LIBELLE DE LA FORME GRAPHIQUE	---POURCENTAGE--- INTERNE	GLOBAL	FREQUENCE INTERNE	GLOBALE	V.TEST	PROBA
1	SERAIT	.82	51	18.	29	2.340	010
2	SORTIES	.50	28	11.	16	2.166	.015
4	POUR	1.59	2.11	35	119.	-2.084	.019
3	SUR	73	1.17	16.	66.	-2.393	008
2	PROCEDURES	.05	28	1.	16	-2.639	004
1	N	.27	67	6.	38	-2.913	002

Expérimentés, non diplômés en statistique

	LIBELLE DE LA FORME GRAPHIQUE	---POURCENTAGE--- INTERNE	GLOBAL	FREQUENCE INTERNE	GLOBALE	V.TEST	PROBA
1	SUR	1.71	1.17	34.	66.	2.593	.005
2	RAPPORTS	25	09	5.	5.	2.546	005
3	SOUPLE	25	09	5.	5	2.546	005
4	UTILISATEUR	.85	51	17	29.	2.385	.009
5	ACCESSIBLE	.25	.11	5	6.	1.993	.023
4	DOCUMENTATION	60	.98	12.	55	-2.012	.022
3	COMMANDES	00	16	0.	9	-2.057	020
2	PAR	.70	1.17	14.	66.	-2.350	009
1	III	.75	1.33	15	75.	-2.769	.003

Très expérimentés, diplômés en statistique ou non

	LIBELLE DE LA FORME GRAPHIQUE	---POURCENTAGE--- INTERNE	GLOBAL	FREQUENCE INTERNE	GLOBALE	V.TEST	PROBA
1	PROCEDURES	96	28	9.	16	3.390	000
2	PAR	2.35	1.17	22	66	3.220	.001
3	DEMANDE	.43	09	4	5.	2.717	.003
4	ETUDE	53	.14	5	8.	2.611	.005
5	SELON	.43	.11	4.	6.	2.384	009
6	JE	85	.37	8.	21	2.167	015
7	N	1.28	67	12.	38	2.126	017
8	MMM	.43	12	4.	7	2.111	.017
9	POUR	3.09	2.11	29	119.	2.092	.018
10	MES	.64	.25	6	14.	2.082	.019
1	EI	3.63	5.04	34.	284.	-2.139	.016

Table 3 Mots sur et sous représentés dans chacun des groupes

Il est néanmoins possible d'affirmer que les caractéristiques dont l'effet est le plus marqué sont le nombre d'années d'expérience en statistique, et le type de formation (principalement, la possession d'un diplôme de statistique ou non). A posteriori, le retour aux réponses telles qu'elles ont été exprimées, permet de voir plus clairement en quoi se différencient les demandes. En tenant compte des effectifs, et des analyses exploratoires effectuées, les quatre groupes suivants sont formés: "Moins d'un an d'expérience, diplômé en statistique ou non", "Expérience entre un an et quinze ans, diplômé en statistique", "Expérience entre un an et quinze ans, non diplômé en statistique" et "Seize ans d'expérience et plus, diplômé en statistique ou non".

Pour chacun de ces groupes, les mots sur et sous représentés sont extraits (table 3); pour cette sélection, les mots dont la fréquence est supérieure ou égale à 5 sont retenus. Le nombre de mots

caractéristiques n'est pas très important, excepté pour le groupe des très expérimentés, ce qui corrobore les affirmations antérieures: les différences de vocabulaires ne sont pas très marquées entre les catégories, les préoccupations étant assez semblables. L'édition des réponses caractéristiques de chacun de ces groupes (suivant le critère de distance du chi-deux entre la réponse et la réponse moyenne du groupe), permet néanmoins de détecter des différences intéressantes. La table 4 reproduit les quatre réponses les plus caractéristiques de chacun des groupes, et nous résumons dans les lignes qui suivent le contenu tel qu'il peut être appréhendé au moyen de la lecture des réponses des groupes.

Non expérimentés:
Il n'y a pas de doute, ils insistent sur la nécessité de disposer d'une documentation de qualité: "*logiciel dont le support de documentation est puissant*", "*une documentation technique plus approfondie*", "*une documentation technique et statistique*", La convivialité et simplicité d'utilisation sont aussi demandées: "*logiciel qui allie convivialité (ergonomie), performances statistiques, exemples concrets, conseils*".

Expérimentés, diplômés en statistique:
Ceux-ci demandent des logiciels complets, fiables, programmables et capables de traiter de gros volumes de données. Ils aiment disposer de nombreuses méthodes et pouvoir aussi en développer: "*ce serait un logiciel offrant une gamme complète de traitements*" Ces utilisateurs font mention d'aspects statistiques particuliers: "*tenir compte des plans de sondage*", "*choix varié et largement paramétrable dans des modules de statistiques classiques inférentielles et d'analyse de données*", "*pour l'analyse statistique, choix de la méthode, puissance et convergence des estimateurs*". On peut aussi noter qu'ils comparent les logiciels et définissent les qualités requises à partir de ceux qu'ils utilisent *"comme XXX en plus convivial*", "*mêmes possibilités statistiques que XXX*", et effectuent des demandes assez concrètes. Et bien sûr, ils ne dédaignent pas le confort d'utilisation.

Expérimentés, non diplômés en statistique:
L'aide au choix des méthodes et à l'interprétation des résultats sont les thèmes suremployés par ce groupe. On veut "*un logiciel qui guide l'utilisateur dans l'analyse*", le logiciel être "*facile à utiliser pour quelqu'un qui n'est pas expert en statistiques*", "*abordable pour tous*", "*accessible à des non statisticiens*", *"bien documenté sur l'utilisation mais aussi les méthodes utilisées*", doit *"faciliter l'accès aux statistiques*", donner "*une réponse claire sur l'applications des différents tests au problème de l'utilisateur*", conseiller "*sur les types d'analyses les mieux appropriées*", "*démystifier les statistiques*".

Très expérimentés, diplômés en statistique ou non:
Ces utilisateurs se sentent à l'aise avec les logiciels, et les statistiques, qu'ils soient diplômés ou non. Ce qui ne veut pas dire qu'ils n'ont pas de demande à formuler! Néanmoins, ils ont renoncé, en plus grande proportion, à trouver le logiciel idéal "*le logiciel idéal n'existe pas*", "*il ne saurait y avoir un unique logiciel de statistiques*". C'est le groupe dont les réponses sont les plus longues (33.4 mots en moyenne, pour une moyenne générale de 30.7 mots). Ils parlent à la première personne "*J'utilise*", *"adapté à mes besoins*", "*le bon logiciel statistique est un logiciel que j'utilise bien*". Leurs demandes recouvrent souvent celles des deux groupes antérieurs, elles

peuvent aussi demander des possibilités très précises: *"des deux logiciels que j'utilise, aucun ne permet de faire du bootstrap"*, *"rajouter des logiciels de statistiques spatiales et géographiques"*. Notons aussi ce commentaire d'un utilisateur qui a suivi les logiciels durant de nombreuses années: *"les utilisateurs n'exploitent pas, faute de formation, les nouveautés introduites"*. Un bon résumé de leurs réponses serait *"adapté à mes besoins"* et *"selon l'étude à mener, un logiciel sera plus ou moins adapté"*

Non expérimentés

1 Logiciel dont le support de documentation est puissant (technique et abordable, francais si possible), dont les fonctionnalités de base (exportation, importation de données, graphisme) sont simples Interêt de disposer de tests pré programmés avec le détail de ce qu'ils calculent Un plus interprétation à l'issue de sortie de résultats

2 1) Une documentation technique plus approfondie pour ceux qui ont fait des statistiques mais ont passablement oublié 2) Des exemples précis de leur utilisation de façon à les mettre en oeuvre dans d'autres domaines 3) Une amélioration des qualités graphiques 4) La possibilité d'attacher à des points sur un graphiques des vecteurs exprimant une tendance

3 Convivialité = menus déroulants plus langage L4G. Aide en ligne statistique (plus mise à jour sur CD-ROM) plus interprétation. Utilisation aisée des formats usuels = DB, tableau, fichier plat, SGBD, utilisable sur des fichiers de gros volume Technique de visualisation des données rapide et multiple. Sortie directement utilisable ou insérée dans un document Word Intégration rapide des nouvelles techniques statistiques Editeur de pages HTML

4 Logiciel qui puisse gérer des grosses bases de données Convivialité (menu et programmation). Bonne représentation graphique avec possibilité de manier les graphiques comme on les dessine (liberté) Documentation détaillée et si possible en français (évite des interprétations erronnées)

Expérimentés, diplômés en statistique:

1 Ce serait un logiciel offrant: une gamme complète de traitements, du plus simple au plus pointu, des possibilités variées de recodages, un mode programmation évolué et une interface graphique, -permettant un choix en fonction de la répétitivité de l'étude et de sa complexité, -un module d'exportation des données et des résultats simple et réellement efficace, un grapheur efficace type Excel 5 0 ou Cricket Graph 3 0, bref tout ce qui permet a la fois un bon travail des données statistiques, une présentation des résultats directement intégrable à un rapport client

2 Un bon logiciel statistique serait un logiciel simple d'accès (pour des utilisateurs non statisticiens) avec des menus déroulants par exemple, puissant et rapide, capable de traiter une grande quantité de données et étant compatible avec les autres logiciels du marché Par ailleurs, il devrait comporter un module de programmation afin de personnaliser les méthodes

3 Un bon logiciel serait un logiciel évolutif (nouvelles méthodes) convivial (interface en menus) permettant l'accès à des non informaticiens contenant un niveau de programmation pour les plus initiés Et ce qu'il nous manque le plus sont les traitements intégrant une structure liant les unités statistiques Dans notre cas, il s'agit de données localisées, surtout liées par une structure de graphisme

4 1) Un logiciel qui permette d'être maître de la méthode de l'analyse statistique en proposant des options. 2) Convivial et performant pour la gestion de données (import/export/recodage) 3) Résultats systématiquement consultables en différé dans divers traitements de texte et à l'écran dans l'édition du logiciel 4) Documentation qui commente les sorties, aide à l'interprétation, donne indications et références sur la méthode

Table 4. Réponses les plus caractéristiques des groupes de répondants

Expérimentés, non diplômés en statistique:

1 Un bon logiciel de statistique serait un logiciel qui permettrait d'importer des données sans trop de modifications sur les données brutes Le logiciel serait d'autant performant s'il permettait d'avoir une aide sur l'interprétation des résultats statistiques en orientant beaucoup plus les résultats vers des graphismes

2 Un logiciel sous windows 95, avec menus déroulants et interface utilisateur me paraît utile pour faciliter l'accès aux statistiques De plus, une aide a l 'interprétation des résultats pourrait s'imaginer

3 Il n'y a pas un bon logiciel statistique et il n'y en aura pas non plus un seul dans le futur Il en faut plusieurs, chacun étant adapté à une catégorie d'utilisateurs en fonction de leur besoin Mon souhait est donc qu'une réflexion soit effectuée sur une typologie des utilisateurs et que les logiciels s'adaptent aux besoins de ces derniers sans chercher à faire un logiciel généraliste censé convenir à tout le monde, mais plusieurs versions plus spécialisées

4 Il est très difficile, voire impossible de transférer les données sur les logiciels grand public tel que windows Les logiciels ne sont utilisables que pour des spécialistes et non pour des personnes juste formées Le vocabulaire est souvent très complexe, il n'y a aucune aide a la décision (ex: vous avez x% de chance d'avoir ce défaut)

Très expérimentés, diplômés en statistique ou non:

1 SAS: personnellement je trouve que SAS-windows constitue une regression pour l'utilisateur habitué à SAS-DOS pour tout ce qui concerne l'impression des résultats c'est catastrophique Quant à SAS sous UNIX on n'en voit pas les avantages et c'est bien compliqué, quand ça marche Ceci pour dire qu'un bon logiciel de statistiques, ce serait, pour mes besoins SAS-PC, plus des impressions faciles, plus des graphiques corrects sans la complexité de Graph, plus des échanges de données avec les autres logiciels qui soient expliqués dans une brochure SAS

2 Des deux logiciels que j'utilise, aucun ne permet de faire du bootstrap (par exemple) ou du rééchantillonage, ou des boucles de traitement Je me mets a S+ cet été, peut être est-ce mieux? SAS est évidemment très bon, je regrette le caractère massif et indigeste de sa documentation, très discrète par ailleurs sur les aspects algorithmiques

3 Dépend de l'esprit du logiciel, cela veut dire si c'est pour l'enseignement ou pour la recherche Aussi du public visé, les étudiants d'économie, ils veulent un logiciel plus simple (par menu) Pour faire ses études statistiques, par contre, les statisticiens ont besoin d'un logiciel pour travailler par programmation ou par langage de commande

4 J'ai choisi ce logiciel Epi-info parce qu'il correspond exactement a mes besoins qui sont simples De plus, ce logiciel diffusé par l'OMS (organisation mondiale de la santé) est libre de droits, il peut ainsi être diffusé à des clients Le bon logiciel statistique est un logiciel que j'utilise bien et qui répond rapidement à mes questions courantes.

Table 4. Réponses les plus caractéristiques des groupes de répondants

(suite)

9.2 En guise de conclusion

On peut effectuer une brève comparaison avec les réponses obtenues lors de l'enquête précédente, en 1992, tout en rappelant que le mode d'échantillonage ne permet guère d'assurer la réelle existence de tendances. Remarquons que le taux de réponse à la question ouverte est plus élevé (il augmente de 52% à 63%), et l'on passe ainsi de 216 réponses ouvertes (sur 416 personnes enquêtées) à 351 (sur 542). Evidemment, le corpus augmente de taille (de 7376 à 10452 mots), et toute comparaison doit tenir compte de cette différence de longueur. Sans faire de comparaison systématique, ni tests statistiques, on peut voir que l'on réclame davantage de simplicité d'utilisation (la fréquence de *simple* passe de 14 à 31, celle d'*utilisation* de 50 à 84) et, fait plus marqué et intéressant, l'aide à l'interprétation des résultats semble constituer une demande plus importante (*aide* obtient 51 citations au lieu de 27, *interprétation* 48 au lieu de 12), alors que la proportion de diplômés en statistique est approximativement égale à la moitié dans les deux enquêtes.

Utilisation des logiciels statistiques - ASU 1996 Id |_|_|_|_|8| PJ |_|

PREMIERE PARTIE : L'utilisateur et son environnement

Sexe : H |1| F |2| Age : |_|_| ans

L 'Entreprise / Organisme où vous travaillez :

Q.1- Secteur d'activité (entourer 1 seule case) :

Administration / Etabliseements publics	|1|	Collectivité locale	|2|
Entreprise publique	|3|	Secteur privé	|4|

Q.2- Domaine d'activité (entourer 1 seule case) :

Administration	|1|	Industrie pharmaceutique	|2|
Enseignement / Recherche	|3|	Industrie agro-alimentaire	|4|
Transport / Energie	|5|	Industrie (autre)	|6|
Banque /Assurances	|7|	Instituts de sondages	|8|
Economie / Finances	|9|	Autres Etudes & Services	|10|

Autre préciser (9) : . . . . . .

Q.3- Ancienneté dans l'entreprise

Moins d'un an |0| ou Nombre d'années : |_|_|

Q 4- Localisation de votre lieu de travail : Département |_|_| ou Pays : . . . . . .

Q 5- Description de votre activité (en clair)

. . . . . .

. . . . . .

L'aspect informatique de votre travail statistique :

Q.6- Quelles machines utilisez vous (entourer 1 ou plusieurs cases) ?

P.C |1| Macintosh |2| Station de travail |3| Mini |4| Site central |5|

Q.7- Accédez-vous à vos données ou logiciels à travers un réseau ? oui |1| non |2|

Q.8- Utilisez-vous Internet ?

Q8.1- Pour votre travail statistique	oui |1|	non |2|
Q8.2- Pour autre chose	oui |1|	non |2|

Q.9- Etes-vous :

Q9.1- Responsable du choix des logiciels statistiques pour votre société ?	oui |1|	non |2|
Q9.2- Responsable du choix de vos logiciels statistiques ?	oui |1|	non |2|
Q9.3- Utilisateur des logiciels statistiques disponibles dans votre société ?	oui |1|	non |2|

Q 10- Quels SGBD utilisez-vous pour accéder à vos données statistiques (entourer 1 ou plusieurs cases) ?

INGRES |1| ACCESS |2| ORACLE |3| SYBASE |4| DB2 |5| aucun |0|

Autre préciser: (9) .....

Q 11- Quels logiciels bureautiques utilisez-vous pour la mise en forme de vos résultats statistiques (entourer 1 ou plusieurs cases) ?

Tableur |1| Grapheur |2| Traitement de texte, PAO |3| aucun |0|

DEUXIEME PARTIE : L'utilisateur et sa culture statistique

Q 12- Formation générale :

Bac |1| DEUG |2| IUT |3| Licence/maîtrise |4| Grande école |5| 3ème cycle |6|

Q 13- Formation en statistique

Q 13.1- Avez-vous un diplôme de statistique ?	oui |1|	non |2|
Q 13.2- Au cours de vos études, avez-vous eu un enseignement de statistique ?	oui |1|	non |2|
Q 13.3- Avez-vous suivi un stage de formation en statistique ?	oui |1|	non |2|

Q 14- Nombre d'années d'utilisation des méthodes statistiques dans votre activité professionnelle : |__|__| an(s)

Q 15- Etes-vous en relation avec des statisticiens (entourer 1 ou plusieurs cases) ?

Au sein de votre entreprise |1| A l'extérieur |2| Pas du tout |3|

Q 16- Etes-vous membre d'un club d'utilisateurs de logiciels statistiques ? oui |1| non |2|

Q 17- Etes-vous membre d'une association de statisticiens ? oui |1| non |2|

Q 18- Avec quelle fréquence utilisez vous les techniques statistiques suivantes ?
(Pour chaque technique, coder : 1=Jamais, 2=Ponctuellement, 3=Souvent, 4=très Souvent)

	Fréquence
Comptages	
Statistiques descriptives	
Régressions linéaires	
Analyse de la variance	
Plans d'expériences	
Séries chronologiques	
Contrôle de qualité	
Analyses factorielles	

	Fréquence
Classifications	
Segmentation	
Tests non paramétriques	
Analyse discriminante	
Modèles log-linéaires	
Analyses de survie	
Méthodes neuronales	
Techniques de marketing	

Autres techniques statistiques utilisées, préciser : .....

TROISIEME PARTIE : Environnement Logiciels Statistiques

Q 19- Citer tous les logiciels statistiques commercialisés dans l'ordre d'importance de l' utilisation que vous en faites

Rang	Nom du logiciel
Logiciel 1	
Logiciel 2	
Logiciel 3	
4	

Rang	Nom du logiciel
5	
6	
7	
8	

Nous allons maintenant nous intéresser aux trois logiciels commercialisés que vous utilisez le plus souvent, c'est à dire ceux cités précédemment du rang 1 au rang 3. Dans la suite du questionnaire ils seront référencés par logiciel 1, logiciel 2, logiciel 3

L 1--Précisez l'environnement dans lequel vous utilisez habituellement ces logiciels (entourer 1 ou plusieurs cases)

logiciel 1	P.C. [1]	Macintosh [2]	Station de travail [3]	Mini [4]	Site central [5]
logiciel 2	P.C. [1]	Macintosh [2]	Station de travail [3]	Mini [4]	Site central [5]
logiciel 3	P.C. [1]	Macintosh [2]	Station de travail [3]	Mini [4]	Site central [5]

L 2- Avec quelle fréquence utilisez vous les techniques statistiques dans ces logiciels ? (Pour chaque technique, coder : 1=Jamais, 2=Ponctuellement, 3=Souvent, 4=Très souvent)

Techniques utilisées	logiciel 1	logiciel 2	logiciel 3
Comptages			
Statistiques descriptives			
Régressions linéaires			
Analyse de la variance.			
Plans d'expérience			
Séries chronologiques			
Contrôle de qualité			
Analyses factorielles			
Méthodes neuronales			

Techniques utilisées	logiciel 1	logiciel 2	logiciel 3
Classifications			
Segmentation			
Tests non paramétriques			
Analyse discriminante			
Modèles log-linéaires			
Analyses de survie			
Techniques de marketing			

	logiciel 1	logicjel 2	logiciel 3
L 3- Comment travaillez-vous avec ces logiciels? (entourer 1 seule case par logiciel)			
Par menu ou interface graphique	|1|	|1|	|1|
Par langage de commande	|2|	|2|	|2|
Par macro ou programmation	|3|	|3|	|3|

L 4- Avez-vous reçu une formation pour utiliser ces logiciels ? (entourer1 seule case par logiciel)	logiciel 1	logiciel 2	logiciel 3
Non	|1|	|1|	|1|
Oui, en interne	|2|	|2|	|2|
Oui par une formation externe	|3|	|3|	|3|

L 5- Donner votre avis sur ces logiciels : Pour chaque rubrique, **coder** :
1=Médiocre, 2=Peu satisfaisant, 3=Assez satisfaisant, 4=Très satisfaisant, 5=Ne sais pas

QUALITE DES FONCTIONNALITES	logiciel 1	logiciel 2	logiciel 3
Appréciation d'ensemble	|__|	|__|	|__|
Gestion de données (recodage, sélection)	|__|	|__|	|__|
Import / Export de données	|__|	|__|	|__|
Edition de données	|__|	|__|	|__|
Import / Export de résultats	|__|	|__|	|__|
Aide à l'interprétation statistique	|__|	|__|	|__|
Graphisme	|__|	|__|	|__|
Documentation technique	|__|	|__|	|__|
Documentation statistique	|__|	|__|	|__|

FACILITE D'UTILISATION	logiciel 1	logiciel 2	logiciel 3
Appréciation d'ensemble	|__|	|__|	|__|
Gestion de données (recodage, sélection)	|__|	|__|	|__|
Import / Export de données	|__|	|__|	|__|
Edition de données	|__|	|__|	|__|
Import / Export de résultats	|__|	|__|	|__|
Aide à l'interprétation statistique	|__|	|__|	|__|
Graphisme	|__|	|__|	|__|
Documentation technique	|__|	|__|	|__|
Documentation statistique	|__|	|__|	|__|

L 6- Vous vous êtes exprimé sur les logiciels de statistiques, qualités et défauts. Pourriez-vous, pour conclure, exprimer ce que serait pour vous de façon générale un bon logiciel de statistique : (si vous le souhaitez, rajoutez une feuille agrafée au questionnaire)

.....

.....

.....

.....