REVUE DE STATISTIQUE APPLIQUÉE

M. BECUE
M.-K. DIALLO
D. GRANGÉ
L. HAEUSLER
Y. LECHEVALLIER
Y. MAJJAD
M. RINGENBACH
V. PERES
F. SERMIER

Enquête sur l'utilisation des logiciels de statistique. ASU 1992

Revue de statistique appliquée, tome 42, n° 3 (1994), p. 5-12
<http://www.numdam.org/item?id=RSA_1994__42_3_5_0>

Rev. Statistique Appliquée, 1994, **XLII** (3), 5-12

ENQUÊTE SUR L'UTILISATION DES LOGICIELS DE STATISTIQUE - ASU 1992 -

M. Becue(1), M.-K. Diallo(2), D. Grangé(3), L. Haeusler(4), Y. Lechevallier(5), Y. Majjad(3), M. Ringenbach(3), V. Peres(4), F. Sermier(6)

(1) Departament d'Estadistica i investigacio Operativa, Facultat d'Informatica de Barcelona Universitat Politècnica de Catalunya, c/Pau Gargallo, 5 08028 Barcelona.
(2) Université de Conakry, BP 1147 Conakry, Guinée.
(3) Unité Statistique Centre de Calcul du CNRS, 23, rue du Loess, 67200 Strasbourg.
(4) CISIA, 1, avenue Herbillon, 94160, Saint Mandé.
(5) INRIA, Rocquencourt, 78153, Le Chesnay.
(6) 9, rue Sophie Germain 75014 Paris.

RÉSUMÉ

En Mai 1992, le groupe «Logiciels et Statistique» de l'ASU faisait une enquête pour connaître l'opinion des utilisateurs de logiciels de statistique. 416 questionnaires ont été obtenus et traités par un groupe de travail. Cet article présente succinctement les résultats qui sont plus largement développés dans «La Revue de Modulad», n°11, p 45-88, juin 93. Ce document est disponible auprès de Danielle Grangé, CNUSC, 950 Rue de Saint-Priest, BP 7229, 34184 Montpellier Cedex 4.

Mots-clés : *enquête, logiciels de statistique, question ouverte, thémascope.*

SUMMARY

In May 1992, the group «logiciels et statistiques» from ASU sent a questionnaire to statisticians as well as to users of statistics, to enquire about their opinion on statistical software they used. 416 questionnaires were returned. This paper gives an abstract of results detailed in «La revue de Modulad», n°11, p 45-88, June 93,INRIA.

Keywords : *Sample Surveys, Statistical Software, Interviews, Cluster Analysis, Correspondence Analysis.*

1. Contexte de l'enquête

Dès sa création, en Janvier 1992, le groupe «Logiciels et Statistique» de l'ASU s'est intéressé à la comparaison de logiciels de statistique. Les travaux sont assez rares dans ce domaine ou commencent à dater. Plus nombreuses sont les publications commerciales mais celles-ci s'intéressent plus à l'aspect convivial et informatique du produit qu'à son contenu statistique. Au printemps 1992, le groupe «Logiciels et Statistique» décidait d'effectuer une enquête auprès des statisticiens et utilisateurs

de logiciels statistiques, 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 est fait.

Le groupe «Constitution du Questionnaire» se mettait alors en place. Une distribution de ce questionnaire a d'abord été effectuée aux XXIV èmes journées de l'ASU à Bruxelles en Mai 1992, puis au Congrès Distancia à Rennes en Juin 1992. Des sociétés ou clubs d'utilisateurs de logiciels (Addad, CISIA, Statgraphics, S-Plus et SAS/STAT) ont accepté de faire parvenir ce questionnaire à leurs clients. Ce mode de diffusion est, bien évidemment, le point faible de cette étude et entraînera un important biais d'enquête que nous retrouverons dans les résultats.

416 questionnaires étaient parvenus au 15 septembre 1992. Les membres du groupe ont décidé, dans un premier temps, de travailler sur les données selon leur intérêt. Ils étaient libres du choix des logiciels et du matériel. Ce mode de travail s'est révélé très motivant et donc fructueux. En effet, les membres du groupe provenant de divers horizons, n'avaient pas le même intérêt pour ces données et ont eu des regards complémentaires.

2. Les répondants

Les 416 personnes qui ont répondu au questionnaire, proviennent essentiellement d'organismes de la recherche (44,5%), de l'enseignement (16%), de l'industrie pharmaceutique (11%) ou des conseils et services (10%). En conséquence 62% travaillent dans des organismes de plus de 500 personnes.

Par contre, parmi ces 416 répondants, on trouve peu de chercheurs en statistique (89). Mais ils sont plus nombreux (48%) à utiliser l'outil statistique pour faire de la recherche dans un autre domaine. La moitié des répondants ont une activité de Conseils et Services alors que les enseignants sont assez rares dans cet échantillon puisque 42% des répondants ont une activité d'enseignement occasionnelle et que pour 29% elle est inexistante ou non renseignée.

Parmi les principaux domaines d'activité, la rubrique «recherche ou enseignement scientifique» vient nettement en tête (34%), suivie de médecine, pharmacie, biologie (14%) puis de l'agro-alimentaire (9%). 52% des répondants ont un diplôme en statistique qui est essentiellement de niveau Bac + 5 .

Parmi les personnes qui n'ont pas de diplôme en statistique, 64% ont une formation en statistique inférieure à 2 ans. Mais l'ensemble des répondants utilise les méthodes statistiques depuis au moins 4 ans et en moyenne depuis 9 ans. 59% des répondants ont une autre formation ou diplôme.

Une très large majorité des répondants travaille dans un environnement micro PC. Moins nombreux sont ceux qui travaillent sur site central. 31% utilisent des logiciels qu'ils ont réalisés et 38% des logiciels réalisés par ou pour leur entreprise. Enfin 70% sont des hommes. Ils peuvent être décrits de la façon suivante : de façon significative, on trouve plus d'hommes que dans l'échantillon global dans les rubriques : préfèrent travailler par menu, ne sont pas membres de l'ASU, possèdent un ordinateur personnel, ne font pas partie d'un club d'utilisateurs, utilisent souvent un Mac, font occasionnellement des statistiques non paramétriques et sont responsables du choix de leurs logiciels.

Les femmes se trouvent en proportion plus importante que la moyenne dans les rubriques suivantes : membres de l'ASU, ne possèdent pas d'ordinateur personnel à domicile, proviennent de l'industrie pharmaceutique, utilisent très souvent les tests non paramétriques, ont une formation Bac + 4 et participent à un club d'utilisateurs.

3. Les logiciels

Dans la première partie du questionnaire il y avait les questions ouvertes suivantes :

- «noms des logiciels commercialisés que vous utilisez»

- «noms des logiciels dont vous disposez et que vous n'utilisez jamais et pourquoi?»

La deuxième partie demandait à l'utilisateur de décrire le mode d'utilisation et de donner son opinion sur les 3 logiciels commerciaux qu'il utilise le plus souvent. Dans cette partie, le répondant indique sur quelle machine il utilise le logiciel.

Nous obtenons donc plusieurs bases de dénombrement des réponses :

- dans la première partie il a été cité 152 logiciels différents , soit 1123 citations d'un logiciel utilisé (2,7 logiciels utilisés par répondant) et 229 citations d'un logiciel inutilisé (0,6 logiciels inutilisés en moyenne) :

- dans la deuxième partie, les répondants ont rempli 888 descriptions de logiciels (2,1 logiciels décrits par répondant) ce qui, en tenant compte des différents matériels, donnait 1059 associations de logiciels à un type de machine (Mac, PC, station, Mini ou site central).

- 26 logiciels ont été cités plus de 10 fois et seulement 5 logiciels plus de 50 fois (SAS, Addad, SPAD-SPADN, Stat-ITCF et Statgraphics). Mais la fréquence de citation de tel ou tel logiciel est très liée au mode d'enquête et ne peut être généralisée. Parmi les logiciels non utilisés mais achetés, on a 6 logiciels cités plus de 10 fois. Il est intéressant de noter que le caractère ouvert de cette question a permis l'apparition de logiciels que nous n'aurions sans doute pas proposés a priori dans une liste de logiciels statistiques : les tableurs, grapheurs ou gestionnaires de bases de données. Il est plus étonnant de voir Word cité 18 fois ce qui le met en 11 ème position des logiciels statistiques.

23,8% des répondants ne citent qu'un seul logiciel commercial utilisé. Il s'agit pour l'essentiel d'utilisateurs de SAS.

Enfin, lorsqu'on étudie les citations des logiciels utilisés en relation avec leur rang d'utilisation, on trouve, parmi les logiciels les plus fréquemment cités, SAS avec un rang moyen d'utilisation de 1,36. Les logiciels SPAD.N et Addad, avec un rang moyen de 2,40, apparaissent nettement comme des logiciels «de complément».

4. Relations entre les techniques, les logiciels et les utilisateurs

4.1. Quels logiciels pour quelles techniques ?

Chaque description des trois logiciels les plus utilisés comprenait :

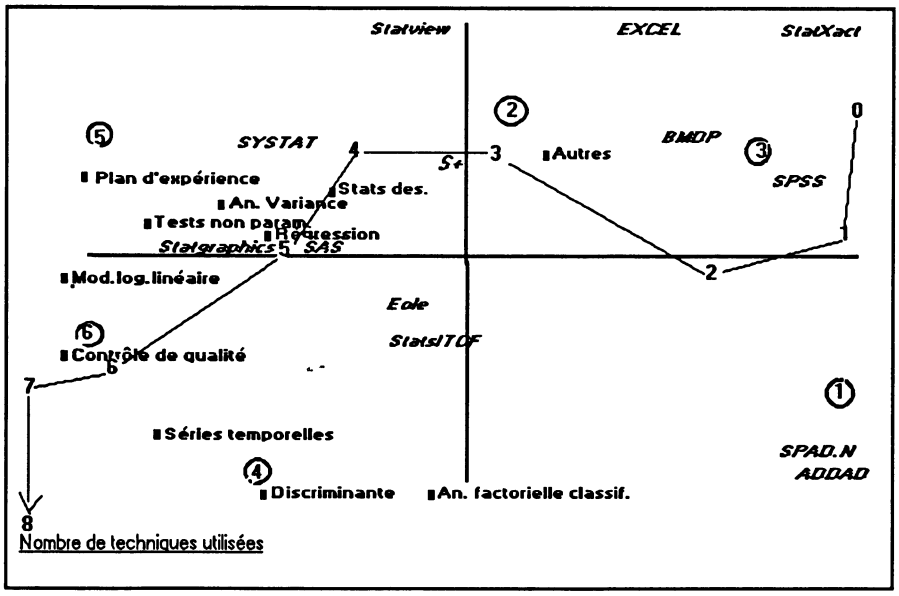
- les techniques utilisées par la personne dans ce logiciel

- des éléments d'évaluation de la qualité de ce logiciel dans divers domaines, notamment l'assistance, la documentation et son interfaçage avec les autres logiciels.

Pour chaque utilisateur, on a donc la description de sa propre utilisation d'au plus trois logiciels. Pour analyser cette information, nous avons considéré comme unité statistique de base une description de logiciel., soit au total 888 descriptions de logiciels.

Le tableau des «descriptions de logiciels» a été soumis dans un premier temps à une analyse des correspondances multiples, les variables actives étant les techniques utilisées dans les logiciels. Dans un deuxième temps, des classes ont été construites regroupant des utilisations de logiciels proches quant aux techniques employées.

plan factoriel (1,2)



Sur le plan factoriel (1,2), les logiciels utilisés ont été positionnés en éléments illustratifs ainsi que le nombre de techniques utilisées dans chaque logiciel et les classes de descriptions de ces logiciels. Ces classes ont été construites à partir des coordonnées factorielles.

Le premier axe du plan factoriel oppose les logiciels qui ne sont utilisés que pour une seule technique, à ceux qui sont utilisés pour une armada de méthodes différentes, et essentiellement des techniques de modélisation, modèle log-linéaire, contrôle de

qualité, séries temporelles, plan d'expérience. SAS, Statgraphics et Systat sont les logiciels les mieux représentés du côté de la modélisation et de la multi-utilisation. Le deuxième axe oppose analyses factorielles et classifications, analyse discriminante et séries chronologiques aux autres techniques.

4.2. Qui utilise quoi?

La première partie du questionnaire comprend 22 questions portant sur la connaissance de l'utilisateur : son identification, ses activités professionnelles ou associatives, les techniques statistiques, l'environnement informatique et les logiciels utilisés.

Le thème analysé ici est le thème du traitement statistique qui contient les informations relatives aux méthodes statistiques utilisées; il est bâti sur la question suivante :

«Quelles sont les techniques statistiques que vous utilisez et avec quelle fréquence?».

Q12-1 statistiques descriptives	Q12-2 régressions
Q12-3 plan d'expérience	Q12-4 analyse de la variance
Q12-5 séries chronologiques, prévision	Q12-6 contrôle de qualité
Q12-7 analyses factorielles et classifications	Q12-8 tests non paramétriques
Q12-9 analyse discriminante	Q12-10 modèle log-linéaire, survie

La variable associée à chacun de ces thèmes est une variable qualitative ayant les 5 modalités suivantes : «Jamais», «Occasionnellement», «Souvent», «Très Souvent», «Non Réponse».

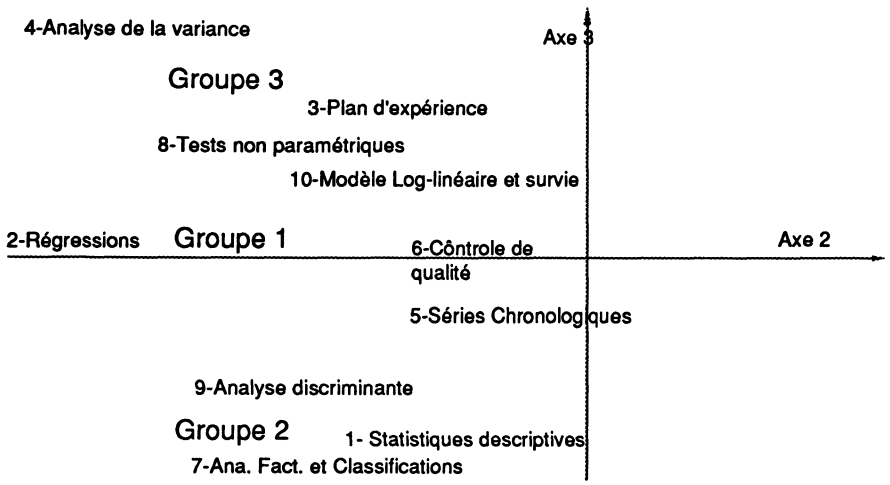
En appliquant une méthode de classification automatique sur ces variables, nous avons retenu une typologie en 5 classes dont les principales caractéristiques sont les suivantes :

- la classe la plus importante est caractérisée par la polyvalence des méthodes statistiques utilisées et elle représente 30% de la population. Beaucoup de logiciels sont cités car les personnes de cette classe pratiquent plus de 7 méthodes statistiques différentes alors qu'un logiciel, en moyenne, ne couvre que 4 types de techniques statistiques. Ceci montre qu'aucun logiciel de statistique n'est perçu comme universel par ces utilisateurs.

- les autres classes de cette partition sont beaucoup plus spécialisées. Les classes associées à la pratique de l'analyse factorielle, de la classification et de la discrimination utilisent les logiciels SPAD.N et ADDAD. Les classes associées à la pratique de l'analyse de variance, des plans d'expérience, des tests non paramétriques et des méthodes log-linéaires utilisent les logiciels SAS, BMDP, Systat, SPSS et Statgraphics.

Après avoir regroupé les modalités «Occasionnellement», «Souvent», «Très Souvent» en une seule modalité «Oui» nous avons réalisé une nouvelle Analyse Factorielle des Correspondances. Sur le plan factoriel (2-3) suivant nous avons uniquement représenté la modalité «Oui» de chacune des variables, les modalités

Plan factoriel (2-3)



«Non Réponse» se trouvent au centre car elles caractérisent l'axe 1 de cette analyse. Sur le plan (2-3) apparaît trois grandes familles de variables.

La première famille se situe sur le deuxième axe factoriel et elle comprend les variables associées aux méthodes de régression et de séries chronologiques.

La deuxième famille représente les méthodes d'analyse factorielle, de classification, de discrimination et les techniques descriptives.

La troisième famille est composée des méthodes d'analyse de variance, des plans d'expérience, des tests non paramétriques et des méthodes log-linéaires.

En résumé les techniques descriptives et les méthodes de régression sont utilisées dans toutes les pratiques statistiques. Les méthodes d'analyse factorielle, de classification, de discrimination forment une famille de méthodes utilisées dans les secteurs de la recherche, de l'assurance et de la médecine. L'autre famille représente les méthodes d'analyse de variance, les plans d'expérience, les tests non paramétriques et les modèles log-linéaires. Ces méthodes sont surtout utilisées dans le secteur de l'industrie pharmaceutique. La pratique du contrôle de qualité et des séries chronologiques est très ponctuelle et non liée à un secteur d'activité.

5. Qu'est ce qu'un bon logiciel de statistique?

Après avoir recueilli les opinions concernant les qualités et les défauts des logiciels utilisés, on a tenté de faire s'exprimer les souhaits des enquêtés à travers une question ouverte. Celle-ci était libellée de la manière suivante :

«Dans ce questionnaire, vous vous êtes exprimé sur le contenu des logiciels de statistique et leurs qualités ou défauts. Certains points importants pour vous n'ont peut être pas été évoqués. Pouvez-vous conclure en exprimant ce que serait pour vous, d'une façon générale, un bon logiciel de statistique?»

Près de la moitié des répondants ont laissé cette réponse en blanc. Le corpus formé par les 216 réponses exprimées, a une longueur de 7376 mots et est formé par 294 mots distincts.

La lecture de ce glossaire, enrichie de celle des concordances de nombreux mots, permet de dégager les principaux thèmes de réponse, que nous présentons ci-dessous.

Un point important apparaît être la documentation (53 citations de «documentation» et 15 citations de «doc» dans les 216 réponses), avec une forte demande d'aide statistique, de documentation conçue comme une formation à la statistique, offrant une aide méthodologique, en particulier au moyen d'exemples.

La documentation doit être de qualité, si possible en français («français» 19 citations). Une aide en ligne, à l'écran, est appréciée.

- Les méthodes statistiques sont souvent mentionnées («méthodes» est cité 48 fois, «méthode» 10 fois et «méthodes statistiques» 15 fois). «Une bonne documentation sur les méthodes» est demandée, contenant plus d'information sur les méthodes employées - et éventuellement des précisions sur les algorithmes mis en oeuvre pour les implanter. Certains demandent la possibilité d'effectuer des analyses statistiques nombreuses et diversifiées, et de pouvoir, éventuellement, utiliser les méthodes récentes.

La convivialité («convivial», 21, «convivialité» 19, «simple»14, «facile»25, «facilement» 11, «facilité» 10 citations respectivement) est un thème très récurrent. On recherche un logiciel simple à utiliser, convivial, d'apprentissage rapide et facile, qui offre un traitement clair des erreurs.

Des thèmes plus secondaires, mais néanmoins d'une fréquence non négligeable sont :

-La présentation soignée des sorties («sorties» 28 citations), en particulier la possibilité de sorties graphiques («sorties graphiques» 10 citations), est recherchée.

-La possibilité d'introduire ses propres programmes dans le logiciel, et de programmer des calculs complémentaires et/ou des enchaînements spécifiques («programmation» 23, «programmes» 11 citations respectivement) est demandée.

Il est intéressant de mettre en relation les caractéristiques des utilisateurs et les mots employés dans les réponses. Pour cela, on réalise une typologie des utilisateurs en fonction de leur profil personnel d'activité, de leur formation en statistique, du nombre d'années d'utilisation des méthodes statistiques et du nombre d'années d'utilisation d'au moins un logiciel de statistique. Cette classification nous permet d'obtenir 6 classes : 3 classes de statisticiens, 2 classes d'utilisateurs non statisticiens et une classe de non-réponses.

Classe 1 : Statisticiens peu expérimentés, conseil et industrie (19% des utilisateurs)

Classe 2 : Statisticiens expérimentés, conseil et industrie (17%)

Classe 3 : Statisticiens expérimentés, recherche et enseignement (18%)

Classe 4 : Non réponses (6%)

Classe 5 : Non statisticiens, peu expérimentés (15%) :

Classe 6 : Non statisticiens expérimentés (25%)

Puis a été réalisé une analyse factorielle sur le tableau de contingence contenant la fréquence d'apparition des mots dans chaque classe. Un premier axe de différenciation des réponses est lié à l'expérience. Les personnes non expérimentées en statistique insistent davantage sur l'aide à l'écran, les aides à l'interprétation et la présence d'exemples dans la documentation. Les personnes expérimentées en statistique évoquent la documentation de base, la qualité des méthodes, et surtout les informations sur les techniques utilisées.

Les statisticiens de la classe 2 (conseil/industrie) comparent les logiciels et les utilisateurs :«... Il y a des logiciels généraux et des logiciels spécialisés. Un logiciel peut être bien s'il est spécialisé. Il ne peut pas y avoir de logiciel général, universel, qui soit universellement bien...». Les statisticiens de l'industrie (classes 1 et 2) se montrent de manière générale très exigeants : logiciel complet, de manipulation simple, permettant une bonne manipulation des fichiers.

Au fur et à mesure qu'augmente le niveau de formation en statistique, la demande est plus focalisée sur les méthodes statistiques et leur implantation. Convivialité et meilleure documentation sont des thèmes abordés par tous, même si ce n'est pas exactement dans les même termes.

6. Conclusion

Bien que cette étude ne puisse être considérée comme représentative de la situation des logiciels de statistique en France elle donne malgré tout des informations intéressantes sur l'opinion et la pratique d'utilisateurs de logiciels de statistiques.

L'analyse des deux approches de l'utilisation des techniques statistiques, l'une par le choix des méthodes, l'autre par la sélection du logiciel, nous permet de définir trois grandes familles de méthodes statistiques. Ces familles, se retrouvant aussi bien du côté des utilisateurs que du côté des logiciels, donnent une mesure de l'influence des logiciels sur la méthodologie employée par les utilisateurs.

Le traitement de la question ouverte montre que l'attente de l'utilisateur est très dépendante de son niveau de formation, de sa pratique des statistiques ce qui induira ainsi des opinions très variées sur les logiciels.