Statistique et analyse des données

Roger Astier

Utilisation du logiciel GENSTAT au D.E.A. de statistique de l'université Paris-Sud

Statistique et analyse des données, tome 7, n° 1 (1982), p. 1-12

<http://www.numdam.org/item?id=SAD_1982__7_1_1_0>

Statistique et Analyse des données
1982 - 1 - 1-12

UTILISATION DU LOGICIEL GENSTAT AU D.E.A. DE STATISTIQUE DE L'UNIVERSITE PARIS-SUD

Roger ASTIER

Bâtiment 425, Université PARIS-XI
91405 - ORSAY

Résumé : *Après avoir donné les raisons conduisant au choix de Genstat, l'utilisation de ce logiciel au D.E.A. de statistique est abordée; l'intérêt pédagogique pour les étudiants et pour les enseignants, apporté par le logiciel dans l'enseignement des statistiques est présenté ensuite.*

Abstact :*Some qualities of Genstat explain why this statistical package has been chosen to help the understanding and practice of statistics at the post-graduation.*

Mots clés : *logiciel Genstat, pédagogie.*

0 - INTRODUCTION.

Ce logiciel a été utilisé pendant quatre année par les enseignants suivants: Marie Cottrel, Jean Coursol, Elisabeth Lesquoy et Roger Astier. Ce texte présenté au 2ème Congrés GENSTAT (7.9 ocotbre 1981-Wageningen-Pays Bas) résume leurs réflexions.

La pratique du logiciel, avec son langage propre est enseignë pendant la première année du D.E.A. de Statistiques Appliquées. Cet enseignement fait intervenir des connaissances dans les deux domaines, de la statistique (la plupart des étudiants ont suivi un certificat de statistique pendant leur maîtrise), et de l'informatique ; l'auditoire est relativement hétérogène :

15 % des étudiants savent écrire un programme (en FORTRAN) ;
50 % des étudiants ont été initiés à un langage mais n'ont pas de pratique ;
les 35 % restant affrontent un ordinateur pour le première fois.

Quant au corps enseignant intervenant au D.E.A., on peut remarquer que :

4 enseignants savent écrire un programme ;
6 enseignants n'ont pas de pratique informatique.

Avant d'aborder la pratique pédagogique de Genstat présentons le logiciel lui-même.

1- POURQUOI GENSTAT ?

1.1. Quelques points d'histoire.

Nous avons découvert Genstat en 1975 grâce au laboratoire de Biométrie de CNRZ (INRA, Jouy-en-Josas). Un cours d'apprentissage est organisé l'année suivante pour les statisticiens de l'Université. En octobre 1977 Genstat est introduit au D.E.A. de Statistique, en même temps que sont abandonnés les programmes FORTRAN utilisés jusque là. En 1978, une thèse compare plusieurs logiciels statistiques (BERNARD, 1978) ; de ce travail Genstat ressort comme le logiciel le plus adapté à nos besoins, tant techniques que pédagogiques. L'introduction d'un cours sur les séries chronologiques, centré sur les méthodes de Box et Jenkins, renforce ce point de vue, la modélisation ARMA étant présente dans le logiciel. Disons aussi que Genstat apparaît actuellement comme le logiciel statistique ayant le meilleur rapport qualité/prix. Enfin il est possible de rencontrer les auteurs anglais du logiciel, de discuter avec eux ; ceci est absolument impossible avec les logiciels américains concurrents.

1.2. Qualités propres de Genstat.

1. Genstat est un logiciel statistique et non pas un package "calculs scientifiques" permettant des traitements statistiques. La différence, dont plusieurs exemples témoignent dans cet exposé, nous paraît très importante pour la formation des statisticiens.

2. L'utilisateur choisit lui-même les noms des éléments manipulés (variables, facteurs, matrices ...), ces noms étant automatiquement rappelés dans les impressions. Ceci permet un interprétations immédiate des résultats et une relecture aisée des programmes écrits.

3. L'écriture des opérations d'entrée-sortie, est très facile, ce qui simplifie la programmation et les vérifications. Par exemple la séquence suivante donne une idée de cette simplicité :

```
'UNIT' $ 95
'VARIATE' TEMPS = 1...95          déclaration des temps de mesure ;
'READ' NB-PASSG                   lecture des valeurs d'une série chro-
                                  nologique, ici nb mensuel de passa-
                                  gers AIR INTER ;
'GRAPH' NB-PASSG ; TEMPS          représentation graphique pour visua-
                                  liser une tendance ;
'SCALAR' NB-COR = 30              déclaration d'un scalaire ;
'DERIVE/LAG = NB-COR' AUTOCOR  =  ACF(NB-PASSAG)
                                  calcul des autocorrélations de la
                                  série jusqu'au décalage 30, mises
                                  dans la structure AUTOCOR ;
'VARIATE'DECALAGE = 0,1...30
'PRINT/FORM = P' DECALAGE, AUTOCOR $ 10, 10.3
                                  impression, "en parallèle" des décala-
                                  ges et des autocorrélations.
```

Les impressions peuvent aussi s'obtenir de façon automatique dans la plupart des procédures.

4. L'existence d'un grand choix de fonctions mathématiques et statistiques, permet d'effectuer tout changement de variables désiré et l'on

peut effectuer directement les opérations matricielles usuelles: produit, inverse déterminant, valeurs-vecteurs propres (matrice symétrique), décomposition en valeurs singulières. Prenons par exemple le modèle linéaire $Y = X\theta + \varepsilon$ et soit $Y = XM\eta + \varepsilon$ un sous-modèle; l'estimation de η dans le sous-modèle vaut :

$$\hat{\eta} = \underline{(M^T X^T XM)^{-1}}\ \underline{M^T X^T X}\ \hat{\theta} .$$

Ce calcul peut être programmé de façon claire, en utilisant les 3 fonctions suivantes :

PDT	(Z = PDT(X;Y)	correspond a "Z = X.Y")
TPDT	(Z = TPDT(X;Y)	correspond à "Z = X^T.Y")
INV	(Z = INV(X)	correspond à "Z = X^{-1}3)

et les 2 matrices intermédiaires MTXTX (pour les produit $M^T X^T X$) et INVERSE (pour $(M^T X^T XM)^{-1}$):

```
'CALCULATE' MTXTX = TPDT(M;TPDT(X;X)
          :  INVERSE = INV(PDT(MTXTX;M))
          :  ETA = PDT(INVERSE;PDT(MTXTX;TETA)) .
```

Ces instructions peuvent aussi être condensées en une seule (... de lecture plus difficile toutefois).

5. Grâce à des procédures standards bien choisies, Genstat est facilement utlisable en apprentissage. Ainsi la directive suivante :

```
'ESTIMATE' NB-PASSG ; MOD-ARMA
```

provoque l'estimation des paramètres d'un modèle Arma (appelé ici MOD-ARMA) ajusté aux valeurs d'une série chronologique (référencée par NB-PASSG).

Mais on peut modifier cette procédure à l'aide "d'options", si un traitement plus complet est nécessaire ; par exemple l'écriture :

'ESTIMATE/PRINT = PC,RECYCLE = Y,REPORT = 2,CRIT = 0.001'
NB-PASSG ; MOD-ARMA

permet - d'obtenir des impressions supplémentaires (mot clé PRINT);
- de poursuivre, le processus itératif d'estimation (mot clé : RECYCLE) .
- de récupérer une trace écrite des valeurs intermédiaires des paramètres, toutes les deux itérations (mot clé : REPORT) .
- d'imposer la valeur 0.001 pour le critère de convergence (mot clé : CRIT).

6. Genstat dispose d'un bon choix des algorithmes intervenant dans les procédures statistiques, accompagné d'une bonne fiabilité des résultats, tout en travaillant de façon interne en simple précision dans la plupart des cas ; une exception cependant : les résultats des analyses de variance dans les modèles très déséquilibrés sont moins précis, mais ces modèles peuvent être étudiés de façon satisfaisante par régression.

7. On bénéficie de la grande portabilité du logiciel : nous travaillons sur UNIVAC-1100, IBM-370, IRIS-80 sans aucune modification pour les programmes écrits en Genstat ; Genstat est d'ailleurs implanté sur une dizaine de marques d'ordinateurs.

8. La programmation en Genstat est très rapide : c'est une aide pour les enseignants désireux de préparer des exemples, et c'est un bon outil pour les étudiants, leur permettant de chercher par eux-mêmes de bons modèles et de travailler sur des données concrètes.

1.3. Défauts de Genstat.

Certains défauts peuvent disparaitre dans le futur, d'autres seront plus difficilement supprimés.

1. Documentation et manuel d'utilisation écrit en anglais, n'existant

pas encore en français. Le manuel d'utilisation est d'accès difficile ; une version française, plus facile, sera disponible au moment du Congrès Compstat 82 de Toulouse. Deux livres d'initiation à Genstat, écrits en anglais seront disponibles à Pâques 1982.

2. Messages et diagnostics sont difficilement compréhensibles et exploitables pour les débutants en informatique ; la traduction des messages en français n'est pas suffisante pour corriger l'erreur, bien que Genstat indique le numéro de la ligne de programme où l'erreur est rencontrée.

3. Liaison difficile, et très peu de documentation à ce propos, entre Genstat et d'autres programmes au niveau de l'exécution. Tous les résultats imprimés par le logiciel peuvent aussi être écrits sur un fichier, sous une forme permettant de constituer des données pour une autre programme; cela est une façon d'enchaîner Genstat et d'autres programmes.

Cependant l'utilisateur peut désirer que Genstat et d'autres programmes travaillent à tour de rôle sur des informations restant en mémoire centrale ; ce genre de liaison est pour le moment très difficile.

4. Genstat est peu répandu en France (4 centres de calcul : INRA Jouy-en-Josas, Paris-Sud Univac, CIRCE, Air Saint-Cyr) alors qu'en Angleterre 79 centres de calcul (dont 49 non universitaires) l'utilisent.

2 - APPRENTISSAGE SIMULTANE DES STATISTIQUES ET DU LANGAGE GENSTAT

Cet apprentissage a lieu pendant les quatre premiers mois de la scolarité du D.E.A. ; la charge d'enseignement des étudiants par semaine est alors de :

- 4 heures sur le Modèle Linéaire ;
- 2.30 heures sur Genstat ;
- 4 heures sur la Modélisation ou 4 heures sur les Séries Chronologiques.

Genstat est également utilisé, au second semestre, dans le module Plan d'Expérience - Modèle Mixte. Nous nous limitons, ici, à la description de l'apprentissage de Genstat.

2.1. Déroulement d'un séance de 2.30 heures.

Approximativement découpée en :

- 1.30 heure sur le langage Genstat, sur l'intérêt statistique des résultats obtenus et sur la description des algorithmes utilisés par le logiciel.

- 1 heure pour répondre aux question personnelles des étudiants relatives aux erreurs, aux solutions possibles et l'interprétation des résultats

2.2. Découpage du semestre en 12 séances.

- trois séances d'introduction au langage Genstat : structures, entrées-sorties, représentations graphiques, bloc de programme ...

- une séance sur la régression dans le cas d'une matrice de rang plein ;

- deux séances sur l'analyse de la variance : facteurs, écriture d'un modèle, modèles orthogonaux et non orthogonaux, changement de l'ordre des facteurs dans l'écriture d'un modèle, covariables ...

- une séance sur l'utilisation des facteurs en régression : variables muettes introduites par Genstat et contraintes d'estimation...

- trois séances sur le cas multivariate : analyse en composantes principales, analyse discriminante, programmation de l'analyse de variance multivariate et des tests usuels ;

- une séance sur les possibilités qu'offre Genstat pour lire des données situées sur un autre support que le programme, pour utiliser des séquences d'instructions Genstat (macros) ;

- une séance sur l'analyse des corrélations canoniques

2.3. Travail demandé aux étudiants.

A chaque séance, des exercices d'apprentissage, du langage et des statistiques sont proposés sous forme de programme à écrire avec interprétation des résultats.

Pendant les deux derniers mois un sujet plus important est proposé aux étudiants ; comme nous disposons en moyenne d'une dizaine de

sujets, chacun est traité par 3 étudiants. Par exemple, des étudiants ont étudié les données suivantes : lors d'un séjour en centre de vacances, des examens médicaux sur des adolescents ont été effectués ; on dispose alors :

sur 25 filles de 9 à 11 ans, des mesures suivantes :
- profession de chaque parent, âge
- (taille, poids, capacité respiratoire, débit respiratoire ...)
- (taille, poids, capacité respiratoire, débit respiratoire ...) mesurés 30 jours plus tard.

sur 35 garçons de 9 à 11 ans des mesures analogues.

Une analyse de données est demandée, et les questions suivantes sont posées:

- Y-a-t-il des différences sensibles, d'après ces données entre filles et garçons de 9-11 ans ?

- La profession des parents paraît-elle influencer le développement physique ?

3 - GENSTAT ET LA PRATIQUE PEDAGOGIQUE.

Les enseignants de l'Université d'Orsay mentionnés eu début, ont utilisé Genstat pour d'autres cours de statistique que celui du D.E.A. sans pour autant l'enseigner ; l'aide que leur a apporté le logiciel est présentée ici. Dans le cadre du D.E.A., l'apprentissage de Genstat fournit en plus un outil aux étudiants, favorisant une attitude active vis-à-vis des statistiques et permettant une approche concrète des traitements statistiques par l'intermédiaire de l'informatique. Décrivons l'aide à la compréhension de la statistique.

3.1. Possibilité de faire les calculs nécessités par un traitement statistique.

A propos d'analyse en composantes principales, par exemple, on peut utiliser la directive correspondante, appelée PCP, comme une boîte noire délivrant des résultats quand on lui fournit des données Une démarche plus pédagogique consiste à effectuer les calculs, formation de la matrice de covariance, diagonalisation, obtention des composantes principales, variance de chaque composantes ...; Genstat

permet de programmer aisèment ces calculs. Lorsque cette démarche a été vue sur un ou deux exemples, on présente un raccourci, l'utilisation de PCP, l'étudiant connaissant alors, les étapes intermédiaires.

3.2. Résumé d'un problème statistique.

Par exemple, un résumé de la démarche à suivre dans un problème de régression peut être :

- Quelles sont les variables concernées ?
- Quelle est la variable dépendante ?
- Comment choisir les régresseurs ?

La description des directives Genstat, utilisées en régression est analogue à ce résumé :

- la directive 'REGRESS' introduit les variables concernées;
- la directive 'Y' introduit la variable dépendante ;
- les directives 'FIT', 'ADD', 'DROP' ... permettent différents choix de régresseurs.

On peut également aller plus loin car Genstat permet de désigner les variables concernées par la régression, par la matrice de covariance empirique construite sur ces variables ; on souligne alors que l'information nécessaire pour la régression est uniquement contenue dans la matrice de covariance empirirque (modèle gaussien). Ce logiciel suit de très près le raisonnement statistique.

3.3. Adaptation à l'enseignant.

Un logiciel statistique permet à l'enseignant de présenter des exemples sur données actuelles et réelles, ce qui est attrayant pour les étudiants. Par sa souplesse de programmation Genstat permet à l'enseignant de souligner ce qui lui tient à coeur : dans une analyse de variance du taux de croissance de veaux (TAUX-CR) en fonction d'un facteur 'mode d'élevage' (appelé ETABLE), d'un facteur 'alimentation' (appelé ALIMENT) et d'un troisième facteur SEXE, j'ai présenté le programme suivant :

```
1       'UNIT' $ 72
2       'FACTOR' ETABLE $ 3 : ALIMENT $2 : SEXE $ 2
3       'READ' TAUX-CR, ETABLE, ALIMENT, SEXE
4       'PRINT' TAUX-CR, ETABLE, ALIMENT, SEXE
              pour insister sur la vérification des données
5       'TREATMENT' ALIMENT*SEXE
6       'AOVA' TAUX-CR
              pour étudier un premier modèle
7       'TREATMENT' ALIMENT*SEXE+ETABLE
              pour étudier un deuxième modèle et pouvoir tester le
              sous-modèle précédent ;
8       'ANOVA'  TAUX-CR;FVAL = AJUSTMNT ;RES = RESIDUS
9       'GRAPH' RESIDUS;AJUSTMNT
              pour effectuer la représentation des résidus en fonction
              des valeurs ajustées (calculées ligne 8) et avoir une
              indication graphique de la validité du modèle étudié
              et de la qualité de l'ajustement.
```

3.4. Genstat ne décide par lors d'un test.

Les statistiques, et leurs degrés de liberté, des tests les plus classiques sont évaluées, mais rien n'indique si un résultat est significatif; l'utilisateur doit donc connaître la distribution de la statistique intervenant, se référer à une table et décider si le résultat est significatif. En multidimensionnel, les différents tests sont faciles à programmer (valeurs propres, trace, déterminant accessibles en une ligne de programme) mais ne sont pas choisis par le logiciel. Genstat force donc nos étudiants à une attitude active vis-à-vis des tests.

4 - CONCLUSION

Après 8 heures d'enseignement les étudiants écrivent des programmes simples en GENSTAT ; apprenant par la suite à traiter des problèmes d'analyse de variance et de régression, ils étudient par eux-même des cas concrets. Le temps consacré initialement au langage s'avère utile, car il permet, lors de la présentation des directives nouvelles de voir uniquement la partie statistique. Bien que ne parcourant pas

tous les aspects du logiciel, pendant le premier semestre, il apparaît que si un étudiant a besoin de directives nouvelles, il est capable en fin de semestre, d'en acquérir par lui-même le fonctionnement, ce qui est souvent le cas lors du stage de D.E.A.. Ainsi nous munissons les étudiants d'un bon outil pour la pratique des statistiques appliquées. Il me paraît également souhaitable de promouvoir la diffusion des Genstat vers d'autres utlisateurs des statistiques.

L'aide à la présentation des problèmes statistiques, et à leur compréhension, qu'apporte Genstat est très précieuse pour les enseignants ; dans ce domaine il n'est pas absolument nécessaire d'enseigner le langage mais il apparaît que l'auditoire est attiré par le langage, concis et facilement compréhensible, et une demande d'enseignement du langage apparaît alors.

La connaissance du langage autorise également l'enchaînement de plusieurs traitements, en récupérant les résultats d'un traitement pour les injecter dans le traitement suivant ; ainsi, chaque étude peut être traitée de façon particulière ce qui est indispensable lorsqu'on travaille sur des données réelles, ne s'analysant jamais par un traitement standard. La souplesse d'utilisation de Genstat permet une pratique sérieuse des statistiques appliquées, ce qui touche bien d'autres personnes en dehors des statisticiens. Il est enfin satisfaisant de penser que cet enseignement, dans le cadre du D.E.A. représente pour nos étudiants un investissement dont les résultats apparaîtront encore bien après leur départ de l'Université.

5 - REFERENCES:

ASTIER R., BOUVIER A., DENIS J.B., Jolivet E., PONS O., TOMASSONE R. VILA J.P.(1982) : Un langage statistique : GENSTAT. INRA, Publications Versailles (à paraître).

BERNARD G.(1977) : Ecriture d'un algorithme en langage Genstat - Exemple de l'analyse des correspondances, in Analyse des données et Informatique, INRIA, Versailles, pp. 851-856.

- BERNARD G.(1977) : Comparaison de trois logiciels spécialisés pour l'analyse statistique : Genstat, Bmdp, Spss - Thèse 3ème cycle, Université Paris XI.

BERNARD G. (1978) : A comparison of three statistical packages: GENSTAT, BMDP, SPSS, in Compstat 78, Corsten L.C.A. et Hermans J., éd., Wien, pp. 445-451

NELDER J. and all; (1977) : GENSTAT : A general statistical Program. Rothamsted Experimental Station, Harpenden - Hert.

NELDER J. (1976) : Intelligent program : the next stage in statistical computing in Recent development in statistics, Barra J.R. et al., ed., Grenoble, pp. 79-96.