LES CAHIERS DE L'ANALYSE DES DONNÉES

B. GHERMANI
C. ROUX
M. ROUX

Sur le codage logique des données hétérogènes : présentation de deux programmes permettant de rendre homogènes des données quelconques

Les cahiers de l'analyse des données, tome 2, n° 1 (1977), p. 115-118

<http://www.numdam.org/item?id=CAD_1977__2_1_115_0>

Les Cahiers de l'Analyse des Données
Vol. II - 1977 - n° 1 - p. 115-118

SUR LE CODAGE LOGIQUE DES DONNÉES HÉTÉROGÈNES : PRÉSENTATION DE DEUX PROGRAMMES PERMETTANT DE RENDRE HOMOGÈNES DES DONNÉES QUELCONQUES

par B. Ghermani, C. Roux et M. Roux

Etant donné un ensemble I d'observations (ou individus ou sujets) décrites par un ensemble Q de variables (ou questions, ou items) les programmes "DIBOUDI" et "STEAK", que nous présentons brièvement ici (1), ont tous deux pour effet de remplacer chaque variable $q \in Q$, qu'elle soit quantitative ou qualitative par un groupe de J_q nouvelles variables, appelées suivant l'usage, "modalités", ne pouvant prendre que les deux valeurs zéro ou un (ou éventuellement la valeur intermédiaire 1/2). L'intervalle de variation d'une variable quantitative est subdivisé en un certain nombre de tranches de valeurs correspondant aux modalités : une modalité sera égale à 1 si la valeur observée appartient à la tranche correspondante elle vaudra zéro dans le cas contraire. Un principe analogue sera suivi pour les variables qualitatives en faisant correspondre une modalité à un état de la variable originelle. Ainsi chaque observation, i, sera décrite après l'exécution de l'un ou de l'autre de ces programmes, par une suite de zéros ou de uns ;chacun de ces descripteurs (modalités) ayant un caractère qualitatif on aura de cette façon un tableau booléen homogène susceptible d'être soumis à l'Analyse factorielle des correspondances (2) ou à un programme de classification automatique.

Bien que fonctionnant de manières très voisines nos deux programmes n'en présentent pas moins des différences importantes pour l'utilisateur car elles conditionnent la présentation des données et l'introduction des divers paramètres du problème.

Des différences de structure entraînent des limitations inégales dans la taille des problèmes traités, enfin les dispositions de sortie des résultats sont aussi à prendre en considération.

Considéré dans toute sa généralité comme nous le faisons ici, le codage logique pose des problèmes d'une réelle complexité. Aussi souhaitons-nous que les utilisateurs comparent eux-mêmes les deux programmes et nous fassent part des avantages qu'ils y trouvent. De quelque manière qu'on s'y prenne, de nombreux paramètres doivent être fournis comme données préalables et celà est susceptible de causer des erreurs dont nous donnerons nous-même ci-dessous des exemples.

Nous comparerons nos programmes de trois points de vue différents : tout d'abord d'après l'état des données qu'ils acceptent, ensuite d'après les différents traitements qu'ils sont susceptibles de leur faire subir, enfin d'après les dispositions de sortie des résultats.

---

(1) *Vu le volume des programmes (1500 cartes environ) il ne nous paraît pas opportun de les publier intégralement ici ; les utilisateurs éventuels peuvent obtenir ces programmes auprès du Laboratoire de Statistique de l'Université Pierre et Marie Curie (PARIS 6) avec une notice d'utilisation détaillée.*

(2) *Pour l'analyse de tels tableaux voir : " Sur l'analyse des tableaux binaires associés à une correspondance multiple: [Bin. Mult.](*)" (Publication du Laboratoire de Statistique de l'Université Pierre et Marie Curie) pour des exemples d'utilisation cf Le genre Myosotis, TI C n 1, in l'Analyse des données" par J.-P. Benzécri et coll., tome I : La taxinomie, Dunod, Paris, 1973.*

(*) *Ce Cahier pp. 55-71.*

1. Les divers types de données acceptées et leur description

Les deux programmes reconnaissent essentiellement deux types de variables : les variables purement quantitatives pour lesquelles un découpage en classes de valeurs est nécessaire et les variables "qualitatives" ou considérées comme telles. On rangera en effet dans cette catégorie les variables en "échelles", c'est-à-dire des variables quantitatives mais ne prenant qu'un petit nombre de valeurs distinctes (e.g. les entiers compris entre 1 et 9).

On notera que, dans la notice du programme STEAK, ces variables qualitatives sont appelées "variables déjà en classes". Les deux programmes admettent les 2 types de variables mélangées de façon quelconque.

1.1 Les paramètres descriptifs des données

Outre les paramètres indiquant le nombre de variables de chaque type, le nombre de classes de chaque variable il faut encore préciser de quel type est chaque variable. Ce problème est résolu de façon différente par chacun des deux programmes : STEAK lit la liste des variables qualitatives désignées par leur sigle ; il découvre leur rang par référence à la liste de toute les variables fournie préalablement. DIBOUDI lit une carte-filtre (LIST) ne comportant que des zéros ou des uns, les uns correspondant aux variables qualitatives ; celà implique donc de connaître les rangs de ces dernières, rangs qui peuvent varier quand on supprime certaines d'entre elles.

STEAK fait l'inventaire des modalités qu'il rencontre pour chaque variable qualitative et signale les cas où ces modalités sont en nombre supérieur au nombre de classes prévues (ce qui pourraît être l'indice d'une erreur dans les données) tandis que DIBOUDI exige qu'on lui fournisse la liste de ces modalités ; en revanche il accepte que celles-ci soient numériques ou alphabétiques.

1.2 Les différentes présentations pratiques des données

Pour les deux programmes les données peuvent être perforées sur cartes ou enregistrées sur support magnétique avec un FORMAT ; chaque observation doit comporter un identifieur alphabétique ou numérique placé, de préférence, avant les valeurs des variables, lesquelles peuvent avoir un FORMAT quelconque, mais constant pour tout le jeu des données.

2. Les options de traitements et leurs paramètres de commande

Comme nous l'avons dit, le but principal de nos deux programmes est d'obtenir une description booléenne des différentes observations. Cependant chacun d'eux offre quelques possibilités supplémentaires de traitement. Nous allons examiner maintenant comment sont atteints ces divers objectifs.

2.1 Les différentes façons d'opérer le découpage en classes

On peut distinguer, selon-nous, trois façons de réaliser des classes de valeurs pour une variable quantitative : soit faire des classes d'amplitudes égales, soit faire des classes d'effectifs à peu près égaux, soit encore faire des classes dont les bornes sont choisies "arbitrairement" par l'utilisateur.

STEAK admet le mélange de ces trois catégories de traitement en lisant les listes des sigles des variables entrant dans chacune d'elles.

DIBOUDI n'a pas autant de souplesse puisqu'il n'admet que les mélanges suivants :

1°) Bornes choisies et amplitudes égales

2°) Bornes choisies et effectifs égaux

Il ne permet pas de mélanger amplitudes égales et effectifs égaux, ce qui n'est d'ailleurs pas grave car on choisit généralement l'une ou l'autre des stratégies pour l'ensemble des variables quantitatives.

Comme pour les variables qualitatives DIBOUDI sélectionne les variables devant subir l'un ou l'autre des traitements à l'aide d'une carte-filtre où les uns désignent les variables à traiter, dont il faut connaître le rang.

2.2 Le tableau de contingence

Le sous- produit le plus intéressant de chacun de nos deux programmes est sans doute la possibilité de dresser un tableau de contingence (encore appelé tableau de BURT) (1)"croisant" deux groupes de variables ; il s'agit d'un tableau dont les lignes et les colonnes représentent des modalités de variables et dont la case (i,j) contient le nombre d'observations présentant à la fois la modalité i et la modalité j.

L'analyse factorielle d'un tel tableau permet d'étudier les liaisons entre les deux groupes de variables à la fois dans leur intensité et dans leur forme. La mise en "éléments supplémentaires" des observations initiales dans leur description booléenne par rapport à l'un des deux sous-ensembles de variables, est une aide précieuse pour l'interprétation de ces liaisons.

DIBOUDI présente ici la souplesse d'emploi la plus grande par la possibilité qu'il a de sélectionner (par leur rang) les variables de chaque groupe dans n'importe quel ordre, alors que STEAK exige que les enregistrements des données comportent d'abord les valeurs des variables du premier groupe suivies des valeurs des variables du deuxième groupe.

En revanche STEAK permet de mettre en éléments supplémentaires certaines observations (nécessairement placées à la fin des données) non prises en compte pour l'élaboration du tableau de contingence ; ceci est particulièrement indiqué pour des observations quelque peu aberrantes dont on veut voir comment elles se placent dans la relation générale entre les 2 groupes de variables.

2.3 Quelques traitements supplémentaires

Les deux programmes ont les possibilités supplémentaires suivantes :

1°) Transposition du tableau à analyser si le nombre de classes est plus grand que le nombre d'observations.

2°) Analyse du tableau des classes, c'est-à-dire du tableau obtenu par remplacement des valeurs initiales par le numéro de la classe de ces valeurs.

3°) Edification de nouveaux noms de variables obtenus par adjonction des numéros des classes aux deux premières lettre du sigle fourni par l'utilisateur.

Le programme DIBOUDI peut également construire un tableau où les modalités prennent les 3 valeurs 0, 1/2 et 1, 1/2 étant affecté aux valeurs de variables quantitatives proches des frontières des intervalles de classes (2). Cependant ce programme requiert une place mémoire importante car la plupart de ses options exige d'avoir le tableau originel en mémoire centrale. Cette disposition n'est d'ailleurs pas forcément un inconvénient, car le programme STEAK, qui nécessite moins de place en mémoire, est vraisemblablement plus coûteux en temps de calcul du fait de ses nombreuses interventions sur mémoire périphérique.

Le programme STEAK a en outre la possibilité de dédoubler les colonnes du tableau des classes en adjoignant à de fortes valeurs initiales (resp. de faibles valeurs) une faible valeur dans la colonne dédoublée (resp. une forte valeur) afin de donner même poids à toutes les observations.

Il a aussi l'avantage de compter le nombre exact d'observations fournies, l'utilisateur pouvant se contenter de donner un majorant de ce nombre d'observations.

3. Les dispositions de sortie des résultats

La présentation des résultats a été prévue par les deux programmes de façon à pouvoir enchaîner aisément le programme BENTAB d'analyse factorielle des correspondances. Nous examinerons successivement quels résultats on peut attendre puis la facilité de l'enchaînement de l'analyse factorielle.

---

(1) *cf* [*Bin. Mult.*], *ce Cahier pp. 55-71.*

(2) *Plus généralement afin d'éviter toute perte d'information, on souhaiterait attribuer à tout individu les deux modalités dont les centres l'encadrent de valeurs non nulles de somme 1 : (e.g. 0,3 et 0,7 si les distances aux centres respectifs sont dans le rapport de 7 à 3, et non seulement 0,5 et 0,5.*

3.1 Les différentes sorties de résultats

Outre le tableau des variables éclatées, transposé à la demande si les modalités sont plus intéressantes que les observations, et le tableau de contingence inscrits en mémoire périphérique, les deux programmes peuvent fournir un tableau dit des rangs ou des classes ; ce tableau, de mêmes dimensions que le tableau originel, est obtenu en remplaçant dans celui-ci les valeurs par le numéro de la classe à laquelle elles appartiennent, pour les variables quantitatives, et par les variables qualitatives on obtient le rang de la modalité considérée.

Les deux programmes fournissent également des statistiques élémentaires fort utiles : bornes des classes pour les variables quantitatives et effectifs de chaque classe, pour toutes les variables.

Les deux programmes construisent, à partir des identifieurs à 2 caractères des variables brutes, des identifieurs à 4 caractères pour les modalités obtenues ; les deux premiers caractères rappellent le sigle de la variable tandis que les 2 derniers indiquent le numéro de la modalité. Mais DIBOUDI présente l'avantage de pouvoir inscrire ces identifieurs soit sur le même fichier que les résultats, soit sur un fichier différent suivant le traitement ultérieur.

Signalons encore que DIBOUDI prévoit une sortie éventuelle des tableaux sur cartes perforées (celà peut aussi être obtenu par STEAK, au prix d'une astuce de carte-contrôle, mais celà interdit alors l'enchaînement de l'analyse factorielle).

3.2 Souplesse d'enchaînement de l'analyse factorielle

Bien que nos deux programmes soient conçus en vue de l'utilisation de leurs résultats par le programme BENTAB d'analyse factorielle, l'enchaînement de celui-ci se fait de façon plus ou moins heureuse. Il a fallu, en effet, modifier BENTAB en particulier pour lire les identifieurs des modalités sur mémoire périphérique.

Dans STEAK cette modification est transparente pour l'utilisateur du seul BENTAB. C'est-à-dire que l'on peut, avec la version modifiée, traiter un tableau brut, sans codage par découpage en classes, comme par le passé. Autre avantage de STEAK : il n'est pas nécessaire de connaître le nombre exact d'observations pourvu qu'on en connaisse un majorant, de même il n'est pas nécessaire de connaître le nombre de modalités formées : ces deux paramètres sont calculés puis transmis à l'analyse factorielle.

En revanche DIBOUDI possède une possibilité intéressante : celle de pouvoir choisir l'ordre des lignes du tableau des résultats lorsque celui-ci est demandé transposé. On peut ainsi mettre à la fin du tableau certaines modalités qui pourront ensuite être mises en éléments supplémentaires dans BENTAB.

4. Conclusion

Pour résumer d'une phrase ce qui vient d'être exposé sur la comparaison de nos deux programmes on peut dire que du point de vue du traitement, proprement dit, ils font la même chose, du point de vue des "entrées" STEAK est d'un emploi assez commode tandis que DIBOUDI possède plus de souplesse quant aux dispositions de "sorties".

Il va de soi que dans un avenir que nous espérons aussi proche que possible on tentera de faire un programme qui conjugue les avantages des deux programmes actuels... et évite l'inconvénient d'être très délicat à manipuler.

Quoi qu'il en soit, il restera toujours indispensable pour les utilisateurs qui se servent de ce type de programme pour la première fois, de procéder en deux étapes : construction du fichier des données recodées puis exploitation de ce fichier par analyse factorielle ou tout autre programme ; la deuxième étape n'étant lancée que si la première s'est déroulée avec succès. Précaution particulièrement opportune quand on fait exécuter ces calculs pour la première fois.