REVUE DE STATISTIQUE APPLIQUÉE

A. LECLERC

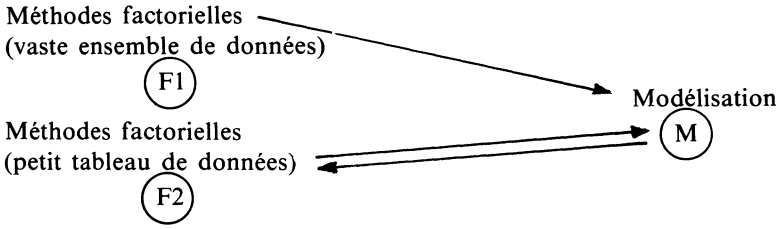
Discussion

Revue de statistique appliquée, tome 35, n° 3 (1987), p. 91-92
<http://www.numdam.org/item?id=RSA_1987__35_3_91_1>

A. LECLERC
Unité 88, INSERM, 91 Boulevard de l'Hôpital, 75013 Paris

Les articles publiés dans ce numéro constituent un ensemble très riche et complet sur la confrontation entre approches de type modélisation, et méthodes factorielles. Ceci n'aurait pas été possible sans un travail en commun important, entre équipes de formation différente. Ceux qui ont eu, pendant longtemps, à regretter l'existence d'un fossé entre ces deux approches ne peuvent que se réjouir : d'abord il est satisfaisant pour l'esprit de savoir que de nombreuses passerelles existent entre des méthodes habituellement considérées comme éloignées; ensuite, du point de vue du praticien, une sorte de « code » d'un usage complémentaire des méthodes semble se dessiner. C'est ce point que je voudrais développer ici : que peut-on conseiller à un praticien, à la lecture de ces articles, sur une stratégie d'analyse combinant approches de type modélisation, et méthodes factorielles ?

Schématiquement, les différentes propositions rentrent dans le cadre suivant :



Les articles de M. AITKIN, de A. BACCINI, et de leurs collaborateurs, concluent à l'intérêt d'une stratégie F1 → M , pour l'analyse de vastes ensembles de données.

Les articles de A. FALGUEROLLES, P.G.M. VAN DER HEIJDEN, et H. CAUSSINUS suggèrent des stratégies de type M → F2 , et insistent sur le fait que la modélisation est implicite au départ de l'application d'une méthode factorielle telle que l'Analyse des Correspondances. Enfin, WORSLEY suggère la possibilité d'une itération entre méthodes factorielles et modélisation, en donnant un exemple de stratégie de type F2 → M .

Il n'y a pas contradiction entre ces propositions, car elles concernent des situations différentes :

Dans le premier cas, les données sont nombreuses, la situation-type est celle d'une enquête où sont éventuellement distinguées variables « à expliquer » et « explicatives »; les méthodes factorielles utilisées seraient principalement l'Analyse des Correspondances multiples ou la classification. La stratégie d'analyse est bien résumée par les auteurs : « Les méthodes d'Analyse des données, telles que l'AFCM ou la Classification automatique, sont les mieux adaptées à l'étude globale et exploratoire de gros fichiers de données » (M. AITKIN *et al.*) »; Une fois les données simplifiées, diverses méthodes de modélisation peuvent alors être envisagées (A. BACCINI *et al.*). Notre propre expérience nous avait amené aux mêmes conclusions (cf. A. LECLERC et alt.). Ce qui n'est peut-être pas assez souligné par les auteurs, c'est que dans beaucoup de cas l'étape de modélisation est également indispensable : un modèle logistique ou linéaire permet de quantifier les effets propres des variables explicatives (autrement dit, de tenir compte du fait que les variables explicatives ne sont pas indépendantes entre elles); or ceci est un objectif à atteindre en présence de variables à expliquer, dans beaucoup de domaines d'étude, comme la Santé ou l'Éducation. Une méthode comme l'Analyse des Correspondances ne fournit pas de résultats assez précis; c'est aussi l'idée avancée par WORSLEY, concernant l'intérêt d'« un modèle qui peut fournir des conclusions quantitatives complétant utilement les représentations graphiques de l'Analyse des Correspondances ».

Dans le second cas, où l'approche de type « modèle » est première, il s'agit de tableaux de contingence à plusieurs dimensions, dont la taille n'est pas excessive. L'Analyse des Correspondances, ou une méthode proche, peut être utilisée pour décrire les résidus d'un modèle : « La modélisation est utile pour tenir compte de la problématique précise du cas étudié. Les méthodes d'analyse multidimensionnelle peuvent ensuite intervenir » (H. CAUSSINUS). Ceci correspond à une situation où la modélisation est possible d'emblée (la taille des données le permet), mais pas intéressante : il faudrait choisir entre deux modèles, l'un « trop simple », l'autre « trop riche » (modèle saturé, par exemple). La séquence M $\rightarrow$ F2 permet de commencer par un modèle sous-dimensionné, de décrire les résidus, et éventuellement de revenir à un modèle plus complexe, comme le suggère WORSLEY.

En conclusion, l'ensemble de ces articles clarifie bien la question de la complémentarité de différentes approches, et propose des stratégies opérationnelles d'analyse. Il est frappant, par ailleurs, de constater à quel point les distinctions classiques entre approches « exploratoires » ou descriptives, et approches « confirmatoires » (test d'hypothèses...) sont devenues floues. La pratique du statisticien, liée aux possibilités que lui offre l'ordinateur, ne lui permet plus toujours de se situer clairement par rapport à ce qu'il a appris en statistique. Ceci pourrait être un sujet intéressant à aborder dans la Revue de Statistique Appliquée.

Références

A. LECLERC, A. CHEVALIER, D. LUCE et M. BLANC (1985). — Analyses des Correspondances et modèle logistique : possibilités et intérêt d'approches complémentaires. *Revue de statistique appliquée,* Vol. 33, n° 1, 25-40.