## Commentaire sur la note de Gilles Celeux et Claudine Robert

Jean-Christophe Turlot
STID Pau, URA CNRS 1104

Un exemple à l'appui, les auteurs présentent avec acuité les conséquences "malheureuses" d'une discrétisation de variables numériques "continues". Une telle transformation est probablement nécessaire dès lors que nous sommes en présence de variables de nature mixte (quantitatives et qualitatives) ; elle a pour effet de dénaturer les données, que l'on se place d'une point de vue probabiliste (le passage d'une densité sur la droite réelle à une densité d •iscrète) ou d'un point de vue géométrique (les points d'une droite réelle sont transformés en sommets d'un simplexe). C'est le point de vue géométrique qui est abordé ici : l'analyse en composantes principales est fondée sur la matrice des covariances entre variables continues ; elle vise à décrire l'essentiel des liaisons entre variables de manière linéaire au moyen d'un petit nombre de facteurs. Cette technique peut être "étendue" à des variables de nature mixte par discrétisation des variables numériques, le tableau "disjonctif-complet" étant alors traité par l'analyse des correspondances multiples (ACM). L'essentiel des liaisons entre les variables qualitatives naturelles ou construites est décrit par les premières composantes principales de l'ACM. La liaison entre deux variables qualitatives est mesurée ici par la statistique du F2 de Pearson.

Le point que je souhaite aborder est le suivant : cette technique d'h Üomogénéisation des données engendre-t-elle une perte de précision ou une perte d'information ? Une réponse sybilline serait la suivante : oui, il y a perte de précision (l'utilisateur est souvent réticent de prime abord à une opération de discrétisation); quant à la seconde question, on ne peut y répondre. Cependant, les applications montrent que les choses ne sont pas si simples.

i) la perte de précision :
lorsque le médecin établit un diagnostic, celui-ci dispose d'un ensemble de paramètres dont certains sont de nature quantitative. Or, le même niveau d'un paramètre sera décrété comme normal ou au contraire anormal selon le niveau d'a ~utres critères quantitatifs ou qualitatifs.
Une telle démarche montre que les mesures sont interprétées de manière essentiellement qualitative, mais aussi que cette interprétation peut être fortement dépendante du niveau de variables concomitantes. La présence d'interactions nécessite une discrétisation adaptée des données et l'on peut penser que, pour bien des problèmes, l'adéquation entre la structure des données et le modèle qualitatif retenu rend évanescente la perte de précision liée à la discrétisation.

ii) la perte d'information :
tout statisticien jugera, comme les auteurs, que l'ACP normée (non pas au sens usuel, mais en considérant les ratios z1/z4, z2/z4, z3/z4) conduit à une discrimination éclatante sur le plan principal (fig.5). A contrario, l'analyse en composantes principales (fig.3) fondée sur une information exhaustive (en regard des mesures effectuées) donne une ä image beaucoup plus confuse des dissemblances entre espèces.
On peut aisément imaginer des situations où la discrétisation de variables numériques conduirait à une représentation plus pertinente du point de vue géométrique que celle fournie par l'ACP. Aussi ne peut-on parler formellement de perte d'information liée à un codage nécessaire de variables numériques. Dans la situation présente, l'information utile est attachée à la forme des papillons : l'ACP non normée (fig.3) exhibe à l'évidence un facteur "taille" faiblement discriminant associé à l'axe 1, la variable z4 engendrant par sa variabilité une grande dispersion intra-classe comme semble le montrer la figure 5 en regard de la figure 3. On doit reconnaître que le plan principal de l'ACP normée par z4 contient "l'essentiel" de l'information discriminante.
L'analyse des correspondances effectuée ne peut être pertinente au sens de la disc rimination entre espèces : le découpage en classes effectué conduit à un effet taille

"noyant" les composantes de forme. Cependant la discrétisation des profils numériques (z1/z4, z2/z4, z3/z4) doit donner une image informative au sens de la discrimination des espèces par leur forme. La confrontation des plans principaux de l'ACP normée par z4 (fig.5) et de l'AFC fondée sur les profils doit permettre une meilleure perception de l'effet du découpage des variables en classes dans un objectif de discrimination.

A l'évidence la procédure standard de discrétisation par découpage en classes des observations est ici inadaptée. Il nous semble que certaines considérations en amont et en aval d'un découpage en classes doivent être nécessairement prises en compte :

- en amont : le choix des variables numériques pertinentes en regard du problème posé. Ces variables ne sont pas nécessairement des mesures observées, elles peuvent être issues d'une transformation des données (comme dans l'exemple des papillons) justifiée par le spécialiste (l'information utile à la reconnaissance des trois espèces de papillons se retrouve dans leur forme et non dans leur taille qui peut constituer un facteur de perturbation).

- en aval : la manière de traiter l'information issue du tabl eau des données mis sous forme "disjonctive-complète". L'analyse des correspondances multiples utilisée de manière systématique n'est certainement pas adaptée à toute situation. Ce fait peut être illustré par l'application suivante menée en collaboration avec un géographe podologue. Il s'agissait de voir si des mesures simples de terrain pouvaient expliquer le niveau d'érosion du sol en vue d'un aménagement. La discrétisation des variables - telles que la densité du couvert végétal, l'orientation des sols, la pente, la nature du sol (lithologie) en surface et en profondeur - suivie de l'ACM, donnait une image confuse du degré d'érosion en regard de ces prédicteurs. Or le principe de discrétisation semblait acceptable dans la situation présente. En réalité, l' un des discriminant majeur est constitué de la combinaison de la lithologie en surface et en profondeur. La constuction d'une nouvelle variable, ou variable-produit, distinguant les associations de lithologie présentes sur le terrain a engendré une image alors très claire du degré d'érosion en fonction des prédicteurs. Il existe, bien entendu, d'autres techniques à même de mesurer le niveau de l'interaction entre deux prédicteurs qualitatifs.

En conclusion, je serais tenté de dire que, si la discrétisation des données numériques est nécessaire, elle peut engendrer une perte d'information substantielle pour certains types de problèmes (il est facile de construire un exemple où la perte de précision engendre une perte d'information); mais aussi que la connaissance permet d'éviter une procédure automatique de discrétisation suivie de l'ACM du tableau disjontif occultant bien souvent certains éléments importants de la structure naturelle des données. Le choix des variables et la manière de les traiter constituent une protection contre une trop grande perte d'information non liée à la discrétisation en soi.