## Commentaires sur l'article de G. Celeux et C. Robert

Gérard Govaert
*CNRS URA 817 - Université de Technologie de Compiègne*

Le texte de Celeux et Robert sur les effets parfois néfastes du recodage des variables continues en variables qualitatives appellent deux remarques de nature différente: la première concerne l'analyse des variables initiales qui aurait pu être un peu simplifiée et la seconde porte sur une analogie que l'on peut faire entre la discrétisation des images et le codage de variables continues en variables qualitatives.

### 0.1 L'analyse des variables continues

L'analyse préalable des variables continues, telle qu'elle est préconisée dans la conclusion du texte de Celeux et Robert peut être effectuée plus simplement et plus rapidement, sans faire appel notamment à l'analyse en composantes principales et en utilisant uniquement l'analyse des statistiques descriptives élémentaires (histogrammes, tableau de corrélations, ....):

- l'histogramme de la variable z4 met en évidence la présence du papillon aberrant,
- la corrélation très forte de 0.97 entre les variables z2 et z3 permet d'éliminer sans une grande perte d'information l'une des deux variables,
- enfin, la prise en compte de l'effet taille, assez classique pour ce type de données, conduit à ne retenir que les deux nouvelles variables t1=z1/z4 et t2=z2/z4.

Il suffit alors de représenter le nuage des papillons dans le plan (t1,t2) (figure 1) pour mettre en évidence très clairement la structure en trois classes de l'ensemble des papillons.

Remarquons que même une simple analyse des différents plans (zi,zj) des variables initiales aurait pu suffire: en effet, le plan (z3,z4) (figure 2) montre assez nettement une structure en trois classes ainsi que le papillon aberrant.

### 0.2 Discrétisation des images et recodage des variables continues

En ce qui concerne la perte d'information due au recodage des variables continues en variables qualitatives, il est intéressant de faire un parallèle avec le traitement d'images: si nous nous limitons, par exemple aux deux variables z1 et z2, le nuage des papillons peut être considéré comme une image dans le plan (z1,z2). Comme la variable z1 codée en entier de 21 à 31 peut prendre 11 valeurs différentes et z2 codée en entier de 30 à 42 peut prendre 13 valeurs différentes, l'image ainsi définie est constituée de 11x13 = 143 pixels ou unités élémentaires d'image.

Avec cette analogie, le recodage des variables z1, z2 en variables qualitatives à 3 modalités s'interprète alors comme une discrétisation de l'image qui passe de 143 pixels à 9 pixels.

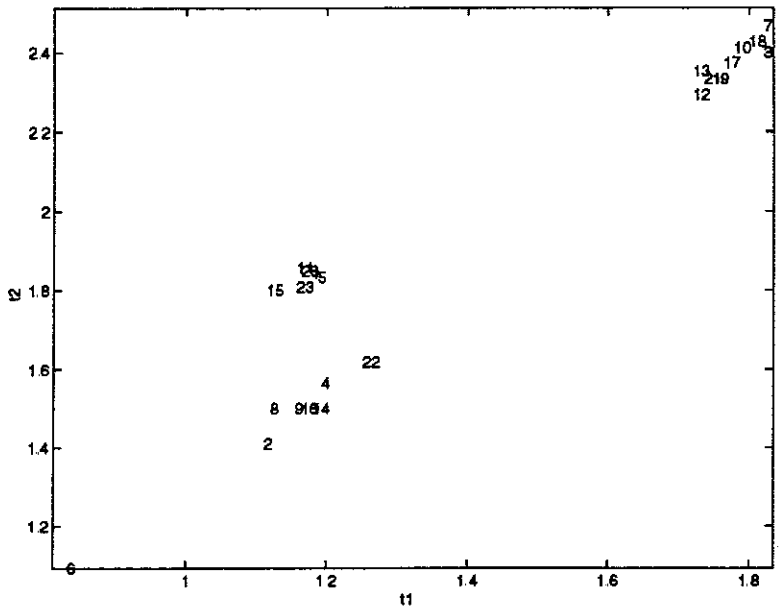


FIG. 1 - *Plan t1,t2*

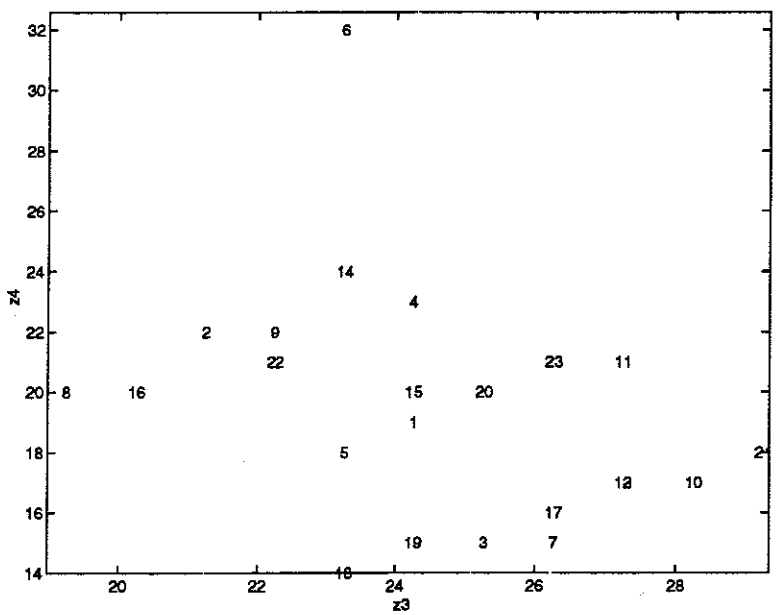


FIG. 2 - *Plan z3,z4*

Cette comparaison montre alors assez bien que ce qui parait assez naturel et classique en analyse des données l'est beaucoup moins en traitement d'images. Il parait en effet peu probable que s'il existait une forme intéressante et visible dans l'image initiale, elle soit encore visible dans l'image discrétisée.

Remarquons que pour notre problème complet, la diminution des pixels est encore plus importante: en effet, nous passons de 11x13x11x18=28314 unités élémentaires à 3x3x3x3=81 unités élémentaires. Dans ces conditions, pour que l'information pertinente soit conservée après une telle compression de l'information, il aurait fallu, comme le conclue le texte de Celeux et Robert, pour le moins que la discrétisation se soit faite en fonction de cette information (par exemple recherche d'une bonne rotation de l'image, ....), c'est-à-dire pour nous en prenant par exemple les meilleures combinaisons linéaires des variables avant d'en faire le recodage.