La prise en compte de l'environnement par les agriculteurs : une analyse avec le package 'ClustOfVar'

V. Kuentz-Simonet[(a)], S. Lyser[(a)], M. Chavent[(b)], J. Saracco[(b)], J. Candau[(a)], P. Deuffic[(a)]

Une stratégie classique pour la ***typologie d'observations*** consiste à réaliser une analyse factorielle des données puis à appliquer une méthode de classification sur les scores des individus mesurés sur les composantes principales.

Certains auteurs (Vichi et Kiers, 2001 par exemple) ont souligné les effets néfastes de cette procédure (composantes qui contribuent peu à la détection d'une structure ou qui masquent l'information taxinomique) mais peu d'alternatives ont été proposées dans le cas de variables qualitatives.

Une approche par ***classification de variables*** est proposée ici comme ***alternative à la première étape d'analyse factorielle*** pour la typologie d'observations. Dans l'application considérée, cette démarche méthodologique permet de répondre à la question complexe de la prise en compte de l'environnement par les agriculteurs.

Le package R 'ClustOfVar'

- Développé spécifiquement pour la classification non supervisée de variables quantitatives, qualitatives ou un mélange des deux.
- Gestion des données manquantes.
- Fournit simultanément des classes de variables homogènes et des variables synthétiques des classes.
- Deux algorithmes de classification sont proposés : ascendant hiérarchique, type k-means.
- Aide au choix du nombre de classes de variables par une approche *bootstrap*.

Les données

- 544 agriculteurs français ont été interrogés en 2005 par une équipe de sociologues d'Irstea sur la « prise en compte de l'environnement ».
- Problématique abordée au travers de 2 aspects :
  - la perception des problèmes environnementaux,
  - les pratiques en faveur de l'environnement, via la conception :
    - du métier,
    - de l'environnement,
    - de la nature,
    - des mesures agro-environnementales.
- 67 variables qualitatives à 2 ou 3 modalités.

Contacts :
Vanessa Kuentz-Simonet, Sandrine Lyser
vanessa.kuentz-simonet@irstea.fr
sandrine.lyser@irstea.fr

Méthode

L'approche par classification de variables permet d'en réduire le nombre en « supprimant l'information redondante ». En effet, en réorganisant les variables en classes homogènes, elle construit des variables synthétiques (sans imposer de contraintes d'orthogonalité).

L'approche par classification de variables

- Elle maximise un critère d'**homogénéité** basé sur la notion de corrélation pour les variables quantitatives et de rapport de corrélation pour les variables qualitatives. L'homogénéité $H(C_k)$ de la classe C_k est une mesure d'adéquation entre les variables de la classe et la **variable synthétique quantitative** de la classe, notée $\mathbf{y}_k$ Elle est définie par : $H(C_k) = \sum_{\mathbf{x}_j \in C_k} r^2_{\mathbf{x}_j,\mathbf{y}_k} + \sum_{\mathbf{z}_j \in C_k} \eta^2_{\mathbf{y}_k|\mathbf{z}_j}$

où r^2 désigne la corrélation de Pearson au carré entre $\mathbf{y}_k$ et la variable quantitative $\mathbf{x}_j$ et η^2 désigne le rapport de corrélation entre $\mathbf{y}_k$ et la variable qualitative $\mathbf{z}_j$.

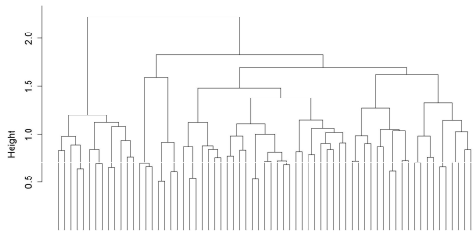
- La variable synthétique quantitative $\mathbf{y}_k$ est la variable « la plus liée » aux variables de la classe au sens du critère H qu'elle maximise.
- Il s'agit de la première composante principale issue de la méthode PCAMIX (voir par exemple Chavent et al., 2012) appliquée aux variables de la classe C_k.

Résultats

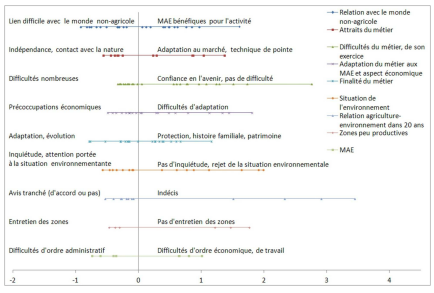
Le dendrogramme issu de la classification ascendante hiérarchique des variables permet d'analyser les agrégations successives de l'ensemble des variables et de visualiser les liaisons entre elles. Neuf classes de variables ont été retenues (par une approche *bootstrap*).

Labellisation des variables synthétiques des classes

Avec les coordonnées des modalités des **variables qualitatives** sur les variables synthétiques, on peut visualiser ces variables quantitatives comme une sorte de gradient. Il est alors possible d'interpréter et labelliser les variables synthétiques.



Dendrogramme issu de la classification des 67 variables qualitatives



Gradients des 9 variables synthétiques quantitatives

Une aide à la compréhension de la typologie des individus

- Les résultats sont plus intéressants et lisibles que ceux obtenus avec l'approche classique :
  - Caractérisation des classes d'individus par les 9 variables synthétiques et non les modalités des 67 variables initiales.
  - Interprétation simplifiée avec les variables synthétiques lues comme des gradients.
- 7 profils distincts : « intéressés par le changement », « préoccupés par les problèmes de l'environnement », « confiants en l'avenir », « attentifs à la protection de l'environnement », « rejetant les préoccupations environnementales », « adeptes de la déprise agricole », « soucieux de l'avenir ».

Conclusion

- La classification de variables apparaît comme une alternative intéressante à l'ACM pour réduire la dimension du tableau des données, avant de procéder à une typologie des observations.
- Au niveau de l'analyse des résultats, l'interprétation des variables synthétiques est intéressante. Elle apporte des premiers éléments de réponse sur les classes d'agriculteurs vis-à-vis de la perception environnementale.

Références bibliographiques

- Candau J., Deuffic P., Ginelli L., Lewis N., Lyser S., (2005), La prise en compte de l'environnement par les agriculteurs. Résultats d'enquête. CNASEA, 83 p.
- Chavent M., Kuentz V., Saracco J., (2012), Orthogonal rotation in PCAMIX, *Advances in Data Analysis and Classification*, 6(2), p 131-146.
- Chavent M., Kuentz V., Liquet B., Saracco J., (*A paraître*), ClustOfVar: An R Package for the Clustering of Variables, *Journal of Statistical Software.*
- Vichi M., Kiers H.A.L., (2001), Factorial k-means analysis for two-way data, *Computational Statistics and Data Analysis*, 37(1), p 49-64.

[(a)] **Irstea, UR ADBX Aménités et dynamiques des espaces ruraux,** 50 avenue de Verdun Gazinet Cestas, F-33612, France.
[(b)] **Inria Bordeaux Sud Ouest, Équipe CQFD Contrôle de qualité et fiabilité des données,** 351 cours de la Libération, 33405 Talence cedex, France.

www.irstea.fr

1[res] Rencontres R – 2 et 3 juillet 2012 – Bordeaux