Analyse de dissimilarités par arbre d'induction

Matthias Studer*, Gilbert Ritschard*, Alexis Gabadinho*, Nicolas S. Müller*

*Département d'économétrie et Laboratoire de démographie, Université de Genève
{matthias.studer, gilbert.ritschard, alexis.gabadinho, nicolas.muller}@unige.ch,
http://www.unige.ch/ses/metri/

Résumé. Dans cet article[1], nous considérons des objets pour lesquels nous disposons d'une matrice des dissimilarités et nous nous intéressons à leurs liens avec des attributs. Nous nous centrons sur l'analyse de séquences d'états pour lesquelles les dissimilarités sont données par la distance d'édition. Toutefois, les méthodes développées peuvent être étendues à tout type d'objets et de mesure de dissimilarités. Nous présentons dans un premier temps une généralisation de l'analyse de variance (ANOVA) pour évaluer le lien entre des objets non mesurables (p. ex. des séquences) avec une variable catégorielle. La clef de l'approche est d'exprimer la variabilité en termes des seules dissimilarités ce qui nous permet d'identifier les facteurs qui réduisent le plus la variabilité. Nous présentons un test statistique général qui peut en être déduit et introduisons une méthode originale de visualisation des résultats pour les séquences d'états. Nous présentons ensuite une généralisation de cette analyse au cas de facteurs multiples et en discutons les apports et les limites, notamment en terme d'interprétation. Finalement, nous introduisons une nouvelle méthode de type arbre d'induction qui utilise le test précédent comme critère d'éclatement. La portée des méthodes présentées est illustrée à l'aide d'une analyse des facteurs discriminant le plus les trajectoires occupationnelles .

1 Introduction

L'analyse des dissimilarités concerne un vaste ensemble de domaines. On y retrouve ainsi la biologie avec l'analyse des gènes et des protéines (alignement de séquences), l'écologie avec la comparaison d'écosystèmes, la sociologie, l'analyse de réseau dont la notion de similarité constitue la base ou encore l'analyse de textes pour n'en citer que quelques-uns. Lorsque les objets analysés sont complexes, des séquences ou des écosystèmes par exemple, il est souvent plus simple de réfléchir en termes de dissimilarités entre objets. Il est d'usage, lorsque l'on a su mesurer les dissimilarités, de procéder à une analyse en *cluster* qui facilite l'interprétation en réduisant la variabilité de ces objets. Une fois les groupes identifiés, on peut mesurer les liens entre ces objets et d'autres variables d'intérêt à l'aide de tests d'association ou de régression logistique sur la *clusterisation* obtenue.

[1]Travail réalisé dans le cadre d'un projet subventionné par le Fonds suisse de la recherche scientifique (FN-100012-113998). Les données ont été collectées par le Panel suisse de ménages, http://www.swisspanel.ch.