Analyse discriminante versus Forêts Aléatoires sur des données qualitatives : méthode d'évaluation contingente appliquée aux zones humides de l'estuaire de la Seine

Dimitri Laroutis, Salima Taibi

Analyse discriminante *versus* Forêts Aléatoires sur des données qualitatives : Méthode d'évaluation contingente appliquée aux zones humides de l'estuaire de la Seine

D. LAROUTIS[1] & TAÏBI S.[2]

[1]Laboratoire d'économie rurale (Lecor) et Care EA - 2260
dlaroutis@esitpa.org

[2] LMRS UMR CNRS 6085 - LAMSAD
staibi@esitpa.org

ESITPA 3, rue du Tronquet
BP 40118
76134 Mont-Saint-Aignan Cedex

Résumé :
La méthode d'évaluation contingente constitue une méthode économique quantifiant monétairement l'ensemble des valeurs que les individus attribuent à un bien environnemental donné. Au centre de cette méthode se trouve un questionnaire visant à révéler le consentement à payer (CAP) des individus pour la préservation, dans le cas de notre étude, des zones humides de l'estuaire de la Seine. Nous nous plaçons dans le cadre assez général de traitements de données d'enquêtes avec pour objectif d'une part, de prédire le consentement à payer et d'autre part, de pouvoir reproduire cette méthodologie pour des vagues d'enquêtes successives. Nous nous sommes limités au cas où la variable dépendante est binaire, la procédure pouvant être étendue au cas de variables qualitatives polytomiques.
Notre objectif est de construire un modèle permettant de pouvoir prédire le consentement à payer. Les prédicteurs étant pour la plupart des variables qualitatives (nominales ou ordinales), nous avons utilisé une procédure permettant de les transformer en variables quantitatives. Nous effectuons l'analyse des correspondances multiples des prédicteurs c'est-à-dire l'analyse des correspondances du tableau disjonctif. Les p variables explicatives sélectionnées X_1, X_2 , ..., X_p sont remplacées par les coordonnées des n individus sur les q axes factoriels ($q \leq p$) en opérant une pondération permettant de conserver l'importance des composantes.
Deux méthodes de classement ont été mises en œuvre sur une base de données d'enquête administrée auprès d'un échantillon représentatif de 300 individus : l'analyse discriminante et la méthode des forêts aléatoires. Le but est de comparer leurs performances en termes de classement.
Abstract :
The contingent valuation method constitutes an economic method quantifying monetarily the set of values which individuals allot to a given environmental good. At the center of this method we find a questionnaire aiming to reveal the willingness to pay of the

individuals for the preservation, in our study, of the seine estuary wetlands. Our objective is to build a model making it possible to be able to predict the CAP. The predictors for the majority are qualitative variables (nominal or ordinal), we used a procedure which allowed to transform them into quantitative variables. We carried out the analysis of the multiple correspondences of the predictors i.e. the analysis of the correspondences of the disjunctive table. The p selected explanatory variables X_1, X_2 ,..., X_p are replaced by the co-ordinates on q factorial axes ($q \leq p$) by weighting allowing to preserve the importance of the components.
Two methods of classification were implemented, the discriminating analysis and the method of the random forests. The goal is to compare their performances in terms of classification.

Mots clés : consentement à payer, variables qualitatives, forêts aléatoires, classement, analyse discriminante.

Key words : willingness to pay, qualitative variables, random forests, classification, discriminant analysis.

Remerciements
Les auteurs remercient le programme Seine Aval pour son soutien financier.

Introduction

La méthode d'évaluation contingente, réalisée dans le but de quantifier monétairement l'ensemble des valeurs que les individus attribuent à un bien environnemental, est apparue comme l'une des seules méthodes qui inclut conjointement les valeurs d'usage et de non-usage, y compris la valeur d'existence (Quiggin, 1998 ; Wei *et al*., 2007). Par le biais d'un questionnaire, elle permet de révéler le CAP des individus notamment pour des biens hors marché. Cette démarche a été suivie dans le cadre d'évaluations monétaires de forêts (Rekola, 2004), dans l'amélioration de la qualité de l'eau, dans la restauration de zones humides (Kosz, 1996) ou dans le cadre de programmes d'extensions agricoles (Roë *et al.*, 2004)… Le questionnaire se structure généralement autour de trois grandes parties :

- *La première partie* concerne la connaissance du bien environnemental (ici, les zones humides de l'estuaire de la Seine) par les habitants et leurs habitudes de fréquentation ;
- *La deuxième partie* tend à révéler le consentement à payer (CAP) des individus pour un programme de préservation du bien environnemental *via* un ensemble de questions ;
- *La troisième partie*, enfin, collecte les données socio-économiques standard telles que le nombre de personnes dans le foyer, le niveau d'études, le revenu net global du foyer…

Nous nous plaçons donc dans le cadre assez général de traitements de données d'enquêtes avec pour objectif de prédire d'une part le consentement à payer et d'autre part de pouvoir reproduire cette méthodologie pour des vagues d'enquêtes successives. Nous nous sommes limités au cas où la variable dépendante est binaire, la procédure pouvant être étendue au cas de variables
Notre objectif est de construire un modèle permettant de pouvoir prédire le CAP. Les prédicteurs étant pour la plupart des variables qualitatives (nominales ou ordinales), nous

avons utilisé une procédure permettant de les transformer en variables quantitatives en s'inspirant de la Méthode Disqual (Saporta, 1977). Nous effectuons l'analyse des correspondances multiples des prédicteurs c'est-à-dire l'analyse des correspondances du tableau disjonctif. Les p variables explicatives sélectionnées X_1, X_2 , …, X_p sont remplacées par les coordonnées des n individus sur les q axes factoriels ($q \leq p$) en opérant une pondération permettant de conserver l'importance des composantes.
Deux méthodes de classement ont été mises en œuvre, l'analyse discriminante et la méthode des forêts aléatoires. Le but est de comparer leurs performances en termes de classement.
L'identification des variables explicatives du CAP des individus dans le cadre de la méthode d'évaluation contingente se réalise généralement par un traitement économétrique de type Logit, Probit ou encore Tobit. La fiabilité de prédiction du CAP de ces modèles peut aujourd'hui être remise en perspective avec l'apparition de cette nouvelle technique statistique dite des Forêts Aléatoires (Random Forests) qui offre des résultats encourageants (Breiman, 2001). Les travaux récents sur le sujet soulignent clairement la supériorité prédictive de ce type de méthodes par rapport aux modèles de régression généralement utilisés (Iverson *et al*., 2004 ; Prasad *et al*., 2006 ; Peters *et al*., 2007).
Dans cet article, nous nous proposons dans le cadre de la méthode d'évaluation contingente d'appliquer la méthode des Forêts Aléatoires au CAP des individus pour la préservation des zones humides de l'estuaire de la Seine et de comparer ces résultats à ceux obtenus par une.
A notre connaissance une telle analyse comparative n'a jamais été réalisée dans ce domaine.

II- Méthodologie

L'analyse des données a permis de tester la cohérence des résultats. Un seul individu (atypique) a été éliminé (n=299). Nous n'avons gardé que les variables les plus corrélées au CAP, mais indépendantes entre elles. L'analyse des tests du Khi² a résulté en la constitution d'un groupe de 15 variables : Sexe, âge, situation familiale, niveau d'études, etc.
Au total les 34 facteurs issus de l'ACM ont été exploités afin de conserver toute l'information de la base de données initiale. Cette procédure permet aussi de pouvoir reconduire l'analyse pour d'autres vagues d'enquêtes sans sélection des facteurs. De plus en conservant tous les facteurs obtenus par l'analyse des correspondances multiples la quantification des variables Xj est celle qui donne la distance de Mahalanobis la plus grande entre les deux groupes.
Les coordonnées des individus ont été transformées en les pondérant par les pourcentages d'inertie.
L'échantillon total a été subdivisé en deux échantillons, l'échantillon d'apprentissage et l'échantillon test. La base de données comporte 299 observations, un seul individu (atypique) a été retiré de la base. La taille de l'échantillon d'apprentissage est de 200, celle de l'échantillon test est de 99. Les données ont été tirées de façon aléatoire en respectant la stratification selon la variable CAP.

III- Résultats

Pour comparer les méthodes d'analyse discriminante et de forêts aléatoires, nous nous intéressons à leur taux de classement.
L'analyse discriminante consiste à chercher des axes sur lesquels on projette les observations de telle sorte que :

- les centres des k groupes soient projetés avec la dispersion maximale ;
- les projections des observations de chaque groupe soient en moyenne peu dispersées.

Les résultats de l'analyse factorielle discriminante sur un échantillon (randomisé) de taille 200 montrent que l'axe a un bon pouvoir discriminant, la fonction discriminante est significative (Pv < 0,0005). La fonction score aboutit à un bon taux de classement puisque le pourcentage d'observations bien classées est de 77% pour les deux groupes.

Tableau de classement (éch. d'apprentissage)

Observé	*Groupe*	*Prévu*	*CAP*
CAP	*Taille*	*0*	*1*
0	*118*	*91*	*27*
		(77,12%)	*(22,88%)*
1	*82*	*19*	*63*
		(23,17%)	*(76,83%)*

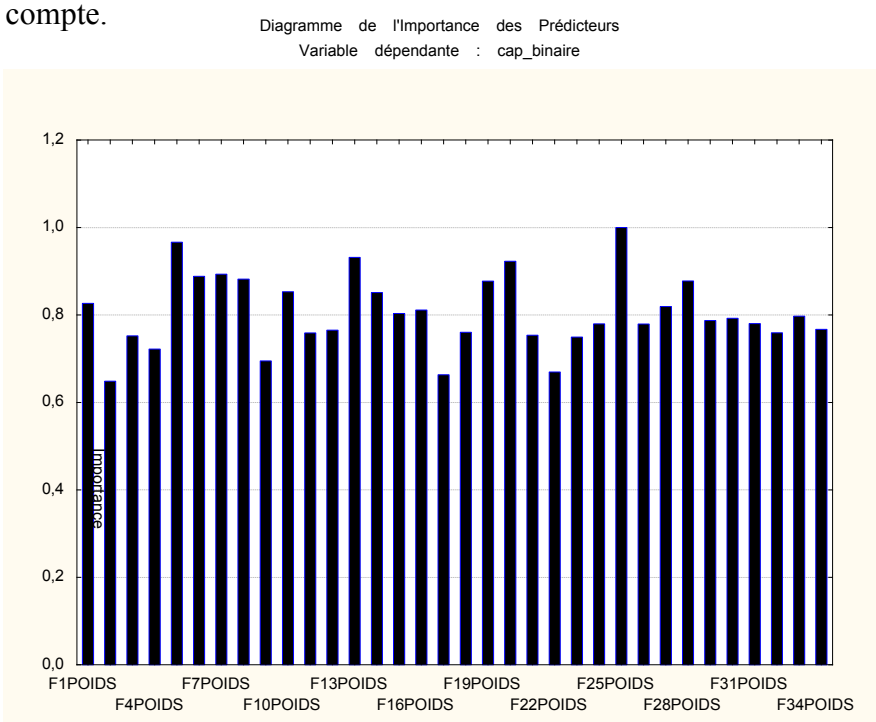
Les résultats déduits de l'analyse discriminante pour l'échantillon « test » montrent également une bonne discrimination (Pv<0,0005). Le pourcentage d'observations bien classées dans chaque groupe dépasse 80%.

Tableau de classement (éch. test)

Observé	*Groupe*	*Prévu*	*CAP*
CAP	*Taille*	*0*	*1*
0	*62*	*52*	*10*
		(83,87%)	*(16,13%)*
1	*37*	*5*	*32*
		(13,51%)	*(86,49%)*

Les résultats sur les échantillons de base, d'apprentissage et test nous montrent tous des taux de classement stables.

La méthode des forêts aléatoires a été appliquée à la même base de données. Cent arbres ont été utilisés. La méthode des forêts aléatoires présente l'avantage de donner un indice (indice de Gini) qui permet de juger de l'importance des prédicteurs. Le diagramme ci-dessous montre que tous les prédicteurs ont quasiment la même importance et sont à prendre en compte.



La matrice de classement ci-dessous permet de montrer que le taux d'observations bien classées est de plus de 75% pour chaque groupe pour l'échantillon d'apprentissage et de plus de 78% pour l'échantillon test.

Tableau de classement (éch. d'apprentissage)

Observé	*Groupe*	*Prévu*	*CAP*
CAP	*Taille*	*0*	*1*
0	*118*	*100*	*18*
		(84,75%)	*(15,25%)*
1	*82*	*20*	*62*
		(24,4%)	*(75,6%)*

Tableau de classement (éch. test)

Observé	*Groupe*	*Prévu*	*CAP*
CAP	*Taille*	*0*	*1*
0	*62*	*55*	*7*
		(88,71%)	*(11,29%)*
1	*37*	*8*	*29*
		(21,63%)	*(78,37%)*

Enfin pour affecter un nouvel individu à une classe (0 ou 1), on considérera celui-ci comme un individu supplémentaire. On projetant cet individu sur les axes factoriels, ses coordonnées après pondérations seront utilisées pour déterminer son consentement à payer.

IV- Conclusion

Les deux méthodes mises en œuvre montrent une même performance pour prédire le classement d'un individu connaissant ses attributs. Le passage par l'analyse des correspondances multiples a amélioré la qualité des résultats de prédiction et permettra notamment d'utiliser d'autres modèles de prédiction dans ce domaine.

Bibliographie

Breiman (L.), 2001, « Random Forests », *Machine Learning* 45, pp. 5-32.

Iverson (L.R.), Prasad (A.M.), Liaw (A.), 2004, « New machine tools for predicitive vegetation mapping after climate change: Bagging an Random Forest perform better than Regression Tree Analysis », *Landscape ecology of trees forest*, pp. 317-320.

Kosz (M.) 1996, « Valuing riverside wetlands : the case of the « Donau-Auen » national park », *Ecological Economics* 16, pp.109-127.

Leng (W.), He (H.S.), Bu (R.), Dai (L.), Hu (Y.), Wang (X.), 2008, « Predicting the distributions of suitable habitat for three larch species under climate warning in Northeastern China », *Forest Ecology and Management* 254, pp. 420-428.

Peters (J.), De Baerts (B.), Verhoest (N.E.C.), Samson (R.), Degroeve (S.), De Becker (P.), Huybrechts (W.), 2007, « Random forests as a tool for ecohydrological distribution modelling », *Ecological Modelling* 207, pp. 304-318.

Prasad (A.M.), Iverson (L.R.), Liaw (A.), 2006, « Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction », *Ecosystems* 9, pp. 181-199.

Quigging (J.), 1998, « Existence value and the contingent valuation method », *Australian Economic Papers* 37, pp. 312-329.

Rekola (M.), 2004, « Incommensurability and uncertainty in contingent valuation: willingness to pay for forest and nature conservation policies in Finland », doctoral dissertation, University of Helsinki, p. 108.

Roë (B.), Haab (T. C.), Sohngen (B.), 2004, « The Value of Agricultural Economics Extension Programming: An Application of Contingent Valuation », *Review of Agricultural Economics* 26, pp. 373-390.

Saporta G. (1977). "Une méthode et un programme d'analyse discriminante pas à pas sur variables qualitatives", INRIA Analyse des données et informatique, Vol. 1, pp. 201-210.

Wei (Y.), Davidson (B.), Chen (D.), White (R.), Li (B.), Zhang (J.), 2007, « Can Contingent Valuation be Used to Measure the in Situ Value of Groundwater on the North China Plain? », *Water Ressource Management* 21, pp. 1735-1749.