MÉTHODES DE CLASSIFICATION SUPERVISÉE APPLIQUÉES AUX TERRES AGRICOLES

Tedjani TARAYOUN

Insee, Direction de la méthodologie et de la coordination statistique et internationale

tedjani.tarayoun@insee.fr

Mots-clés : classification, prédiction, apprentissage, data mining

Résumé

Le Service de la Statistique et de la Prospective (SSP) du Ministère de l'Agriculture et de l'Alimentation publie, conjointement avec la Fédération nationale des sociétés d'aménagement foncier et d'établissement rural (FNSAFER), le prix des terres et prés (en euros courants à l'hectare). Ces prix reflètent l'état du marché de vente des terres agricoles. Ils sont calculés sur la base des valeurs des transactions. Deux modes de calcul coexistent suivant le niveau géographique considéré :

- Des moyennes triennales sont calculées au niveau des petites régions agricoles regroupées (zonage infra-départemental) et des départements.
- Des prix régionaux et nationaux sont calculés à partir des indices annuels régionaux définis par des modèles hédoniques.

Dans les deux cas, le prix que l'on cherche à mesurer est celui des terres et prés agricoles destinés à conserver, au moment de la transaction, leur vocation agricole. La difficulté centrale de cette mesure est que la variable relative à la destination de la terre est mal codée. Jusqu'à présent, comme approximation, on réduisait le champ au marché de vente des terres et prés de plus de 70 ares, en faisant l'hypothèse que les plus grandes terres conservent un usage agricole. En réalité, ce critère ne garantit pas la destination agricole de la terre (plusieurs sociétés d'aménagement foncier et d'établissement rural ont noté que les prix publiés ne correspondaient pas tous à des prix agricoles). Par ailleurs, en faisant cela, le marché des petites terres agricoles est entièrement exclu du champ alors qu'il correspond à environ 60 % des transactions, soit près de 8 % de la surface vendue chaque année.

C'est ainsi qu'il a été jugé nécessaire de remplacer le filtre des 70 ares et de construire à la place un indicateur pour classer les terres en deux groupes, entre les « terres à destination agricole » et les « terres à destination non agricole ». Une étude via des techniques d'apprentissage supervisé a été menée en ce sens. L'objectif de ces techniques est de construire un modèle de classification à partir de données où l'on connaît de façon sûre la vocation de la terre, puis d'utiliser ce modèle pour extrapoler les résultats aux terres dont la vocation est inconnue ou incertaine. Pour tenir compte de l'hétérogénéité des départements en termes de marchés fonciers, l'étude a été conduite département par département, de façon indépendante.

L'objet de cette communication est de présenter les résultats de la classification des terres agricoles. Les méthodes d'apprentissage les plus courantes ont été mises en œuvre. Parmi elles, la traditionnelle régression logistique *logit*, ainsi que des méthodes plus récentes et plus sophistiquées : les arbres de décision CART, les *Support Vector Machine* (séparateurs à vaste marge) et des méthodes d'agrégation de modèle (les forêts aléatoires et le *Boosting*). Une analyse comparative mettra en perspective le pouvoir prédictif et les spécificités de chaque méthode.

Bibliographie

[1] Tufféry, « Data mining et statistique décisionnelle: l'intelligence des données », 2017.
[2] Tufféry, « Modélisation prédictive et apprentissage statistique avec R », 2017.
[3] J. Friedman, T. Hastie, R. Tibshirani, « The Elements of Statistical Learning », 2009.
[4] A. Karatzoglou, D. Meyer « Support Vector Machines in R » *Journal of Statistical Software*, 2006.