Application de la classification symbolique à l'estimation des coûts de production agricoles

Dominique Desbois

Application de la classification symbolique à l'estimation des coûts de production agricoles[i]

Dominique Desbois *

* Economie publique, Inra, 16 rue Claude Bernard, F-75231 PARIS CEDEX 05.
dominique.desbois@inra.fr
https://www6.versailles-grignon.inra.fr/economie_publique_eng/PersonalPages2/Dominique-Desbois

Réumé. Cette communication utilise la classification des données symboliques pour explorer les similitudes entre distributions d'estimations quantiles conditionnelles, en l'appliquant au problème de l'allocation des coûts spécifiques en agriculture. Après avoir rappelé le cadre conceptuel de l'estimation des coûts de production agricole, la première partie présente le modèle empirique, l'approche de régression quantile et la technique de classification des données d'intervalle utilisée. La seconde partie présente l'analyse comparative entre douze États membres européens des résultats issus de la classification hiérarchique divisive des intervalles d'estimation.

1. Introduction

L'intégration de l'agriculture dans les 28 États membres résultant de l'élargissement de l'Union européenne (UE) a suscité des besoins récurrents d'estimation des coûts de production des principaux produits agricoles, tout au long des réformes de la politique agricole commune (PAC), sur les marchés concurrentiels comme réglementés. L'analyse des coûts de production agricole est un outil d'analyse des marges des agriculteurs : elle permet d'évaluer la compétitivité prix des exploitations agricoles, l'un des éléments majeurs du développement et de la durabilité des chaînes alimentaires dans les régions européennes. Pour répondre aux besoins de simulations et d'analyses d'impact pour les différentes organisations communes de marchés, nous devons fournir des informations sur l'ensemble de la répartition des coûts de production afin d'évaluer les options de politique agricole publique. En se basant sur le constat de l'asymétrie et de l'hétérogénéité au sein de la distribution empirique des intrants agricoles, nous avons proposé une méthodologie adaptée à l'estimation de la distribution empirique des coûts de production spécifiques des principaux produits agricoles dans un contexte européen où les exploitations agricoles restent principalement multi-productives (Desbois, Butault et Surry, 2017).

À partir de cette approche, nous présentons le modèle empirique d'estimation des coûts de production spécifiques, inspirée d'une approche micro-économétrique de répartition des coûts pour construire une matrice entrées-sorties au niveau national (Divay et Meunier, 1980). Puis, nous rappelons la méthodologie d'estimation

Comment citer ce document :
Desbois, D. (2019). Application de la classification symbolique à l'estimation des coûts de production agricoles. In: Actes des 26èmes Rencontres, Société Francophone de Classification (SFC) (p. 65-70). Presented at XXVI-èmes Rencontres de la Société francophone de Classification, Nancy, FRA (2019-09-03 - 2019-09-05). 147

permettant de prendre en compte l'hétérogénéité des exploitations agricoles, selon l'approche du quantile conditionnel proposée par Koenker et Bassett (1978). Ensuite, pour explorer les distributions empiriques des intervalles d'estimation de quantiles conditionnels, nous présentons la procédure de classification utilisée (Chavent *et al.*, 2007) dans le cadre conceptuel de l'analyse symbolique de données (Billiard et Diday, 2006). Nous introduisons alors le graphique des résultats de la procédure de classification appliquée aux intervalles d'estimation des quantiles conditionnels. Enfin, nous concluons sur la pertinence de cette approche appliquée à la production de porc.

2. Cadre conceptuel et aspects méthodologiques

Nous présentons d'abord la méthodologie d'estimation des coûts spécifiques. Puis, nous introduisons l'outil de classification des intervalles d'estimation dans le formalisme de l'analyse symbolique de données.

2.1 Le modèle d'estimation des coûts spécifiques de production

Inspiré de Divay et Meunier (1980), l'affectation de la somme x_i des coûts des intrants pour l'exploitation agricole i est réalisée par décomposition linéaire le long des produits bruts Y_i^j de l'exploitation agricole i pour chaque production j, où u_i est un vecteur aléatoire d'espérance nulle :

(1) $$x_i = \sum_{j=1}^{p} \beta_j Y_i^j + u_i$$

Comme Cameron et Trivedi (2005), nous supposons que le processus générateur de données est un modèle linéaire à hétéroscédasticité multiplicative caractérisé par :

(2) $$x = Y'\beta + u \text{ avec } \quad u = Y'\alpha \times \varepsilon \quad \text{et} \quad Y'\alpha > 0$$

où $\varepsilon \sim iid[0, \sigma]$ est un vecteur aléatoire identique et indépendant à moyenne nulle et variance constante σ^2. Sous cette hypothèse, $\mu_q(x|Y, \beta, \alpha)$, le q^e quantile conditionnel du coût de production *x*, conditionné par *Y* et les paramètres, α et β se déduit analytiquement comme suit :

(3) $$\mu_q(x|Y, \beta, \alpha) = Y'[\beta + \alpha \times F_\varepsilon^{-1}(q)] = Y'\gamma.$$

où F_ε est la distribution cumulée des erreurs.

Le coefficient technique du q^e quantile de coûts spécifiques pour le j^e produit est défini par le j^e composant du vecteur de pente :

(4) $$\gamma^j(q) = [\beta + \alpha \times F_\varepsilon^{-1}(q)]^j$$

Au moins deux types de modèle peuvent être dérivés de cette spécification (D'Haultfœuille et Givord, 2014) :

i) $x = Y'\beta + u$ avec $u = K\epsilon$, à résidus homoscédastiques ($V(\epsilon|Y) = \sigma^2$), dénommé modèle à *translation simple*, *i.e.* soit un modèle linéaire à pentes

Comment citer ce document :
Desbois, D. (2019). Application de la classification symbolique à l'estimation des coûts de production agricoles. In: Actes des 26èmes Rencontres, Société Francophone de Classification (SFC) (p. 65-70). Presented at XXVI-èmes Rencontres de la Société francophone de Classification, Nancy, FRA (2019-09-03 - 2019-09-05). 147

homogènes ; puisque $Y'\alpha = K$ est constant, les quantiles conditionnels $\mu_q(x|Y,\beta,\alpha) = Y'\beta + KF_e^{-1}(q)$ ont tous la même pente β, mais diffèrent seulement d'un écart constant, croissant à mesure que l'ordre q du quantile augmente ;

ii) $x = Y'\beta + (Y'\alpha)\varepsilon$ et $Y'\alpha > 0$ à résidus hétéroscédastiques, dénommé modèle à *translation-échelle*, *i.e.* le modèle linéaire de quantiles conditionnels à pentes hétérogènes.

2.2 Classification par intervalles des distributions de coûts spécifiques

Pour un produit donné j_0 tel que le porc et le l^e pays européen, l'intervalle d'estimation des coefficients techniques pour les coûts spécifiques

(5) $$z_l^q = \left[Inf\_\hat{\gamma}_l^{j_0}(q); Sup\_\hat{\gamma}_l^{j_0}(q)\right]$$

est obtenu par *bootstrap* marginal par chaînes de Markov (He et Hu, 2002). Objets symboliques, les L distributions nationales $\Omega = \{\omega_1, \cdots, \omega_l, \cdots \omega_L\}$ sont décrites par un ensemble de $Q = 5$ descripteurs[1], qui sont les intervalles d'estimation des coefficients techniques pour les quantiles conditionnels $Z = \{z^1, \ldots, z^q, \ldots, z^Q\} = \{z^{0.10}, z^{0.25}, z^{0.50}, z^{0.75}, z^{0.90}\}$.

Les dissimilarités locales entre pays *l* et pays *l'*, associées aux intervalles d'estimation des coefficients techniques pour le q^e quantile conditionnel, sont calculées selon la distance euclidienne :

(6) $$\delta_M\left(z_l^q, z_{l'}^q\right) = \sqrt{\left(Inf\_\hat{\gamma}_l^{j_0}(q) - Inf\_\hat{\gamma}_{l'}^{j_0}(q)\right)^2 + \left(Sup\_\hat{\gamma}_l^{j_0}(q) - Sup\_\hat{\gamma}_{l'}^{j_0}(q)\right)^2}$$

Pour cette métrique *M*, une dissimilarité globale entre pays *l* et pays *l'* basée sur les différences entre distributions nationales des intervalles d'estimation des coefficients techniques est calculée selon le critère quadratique suivant :

(7) $$d(\omega_l, \omega_{l'}) = \left(\sum_{q=1}^{Q} \delta_M^2\left(z_l^q, z_{l'}^q\right)\right)^{1/2}.$$

Étant donné la matrice des dissimilarités entre distributions nationales de coûts spécifiques issues des calculs précédents, nous pouvons utiliser les méthodes de classification non supervisée. De façon similaire à la méthode de Ward, Chavent *et al.* (2007) proposent un algorithme de classification hiérarchique par division hiérarchique sur données symboliques (DIVCLUS-T), valable pour les données d'intervalle et les données catégorielles. Par la suite, nous détaillons pour les données

1 Le choix d'un petit nombre de descripteurs a été fait pour des raisons de comparabilité avec des approches graphiques plus classiques (Desbois, 2015). Cependant, si les objectifs de l'analyse l'exigeaient, il pourrait être étendu sans inconvénient aux ensembles de descripteurs de cardinalité supérieure : déciles (*Q = 9*), voire centiles (*Q = 99*).

d'intervalle les principes opérationnels de cette procédure de classification non supervisée.

L'algorithme divisif de classification hiérarchique partage récursivement chaque classe en deux sous-classes, à partir de l'ensemble des pays en tant qu'objets symboliques $\Omega = \{\omega_1, \cdots, \omega_l, \cdots \omega_L\}$. À chaque partition en k classes symboliques $\mathrm{P}_K = \{C_1, \cdots, C_k \cdots, C_K\}$, une classe doit être scindée pour obtenir une partition P_{K+1}, à $K+1$ classes, optimisant le critère de sélection basé sur l'inertie.

L'inertie de la k^{e} classe est définie par $I(C_k) = \sum_{l \in C_k} \mu_l d_M^2 \left(z_l, g(C_k)\right)$ où μ_l est le poids du l^{e} pays et $g(C_l)$ est le barycentre de classe définie par :

(8) $$g(C_k) = \frac{1}{\sum_{l \in C_k} \mu_l} \sum_{l \in C_k} \mu_l z_l.$$

L'inertie intra est définie par la somme des inerties des classes à leurs barycentres :

(9) $$W(P_K) = \sum_{k=1,\ldots,K} I\,(C_k)$$

L'inertie inter est définie par l'inertie des barycentres relatives au barycentre global g de l'ensemble Ω , comme suit:

(10) $$B(P_K) = \sum_{k=1,\ldots,K} \mu_k d_M^2\,(g(C_k), g) \text{ where } \mu_k = \sum_{l=1,\ldots,k} \mu_l.$$

Pour une partition P_K, l'inertie totale regroupe l'inertie intra avec l'inertie inter :

(11) $$I(\Omega) = W(P_K) + B(P_K).$$

Ainsi, minimiser l'hétérogénéité (mesurée par *W*) est équivalent à maximiser l'homogénéité (mesurée par *B*).

Généré par la réponse binaire (*oui*/*non*) à une question $\Psi = [z^q \leq c\,?]$, notons $\{A_k, \overline{A_k}\}$ la bipartition induite de la classe C_k formée de n_k objets. Afin de choisir parmi les $n_k - 1$ bipartitions possibles de la classe C_k, le critère discriminant est defini par le ratio suivant :

(12) $$D(\Psi) = \frac{B^q(A_k, \overline{A_k})}{I^j(C_k)} = 1 - \frac{W^j(A_k, \overline{A_k})}{I^q(C_k)},$$

où l'inertie inter $B^q\left(A_k, \overline{A_k}\right)$ et l'inertie $I^q(C_k)$ sont calculées par rapport au q^e quantile conditionnel. Ainsi, minimiser l'inertie intra $\mathrm{W}\{A_k, \overline{A_k}\}$ équivaut à maximiser l'inertie inter $\mathrm{B}\{A_k, \overline{A_k}\}$ et, par conséquence, le critère discriminant $D(\Psi)$.

Comme dans la méthode de Ward, la « *hiérarchie supérieure* » (Mirkin, 2005) à la partition P_K est indexée par l'indice h de la classe C_K , définie par son inertie inter comme suit:

(13) $$h(C_k) = B\left(A_k, \overline{A_k}\right) = \frac{\mu(A_k)\mu(\overline{A_k})}{\mu(A_k)+\mu(\overline{A_k})} d^2\left(g(A_k), g\left(\overline{A_k}\right)\right)$$

L' algorithme DIVCLUS-T scinde la classe C_K^* qui maximise $h(C_K)$, assurant que la partition suivante $P_{K+1} = P_K \cup \{A_K, \overline{A_K}\} - C_K^*$ présente la valeur minimum de l'inertie intra, conformément à l'équation suivante

(14) $$W(P_{K+1}) = W(P_K) - h(C_K^*).$$

3. La hiérarchie divisive des estimations de coûts

Nous analysons les résultats obtenus en coûts spécifiques pour le porc (figure 1), l'un des produits sélectionnés dans le cadre du projet FACEPA, sur la base 2006 du

réseau européen d'information comptable agricole (UE-RICA). Les coûts spécifiques du porc comprennent principalement les intrants alimentaires, vétérinaires, et énergétiques.

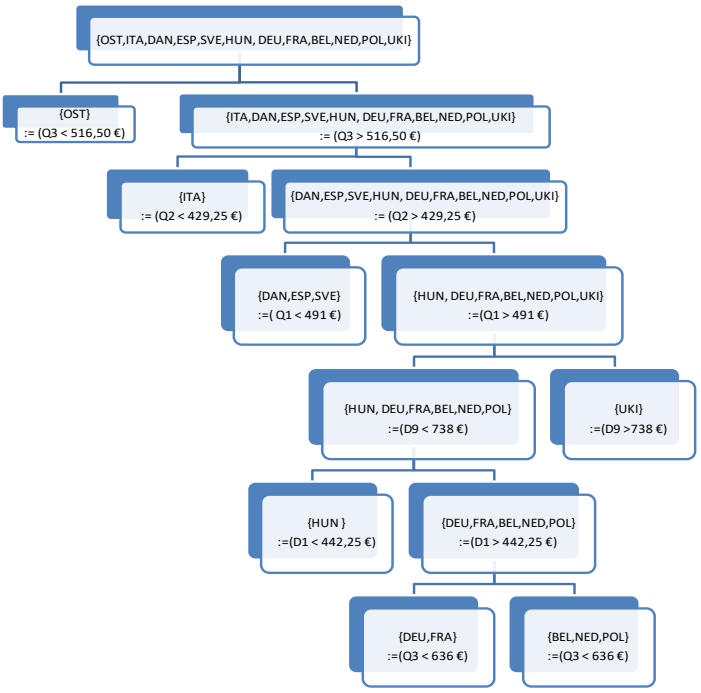

FIG. 1 - Classification de 12 pays européens sur la base de coûts spécifiques pour 1 000 € de produit brut porcin. Source : traitement de l'auteur, selon UE-RICA 2006.

Au sommet de la hiérarchie divisive, la procédure de regroupement permet d'identifier deux modèles contrastés de distributions empiriques des coefficients techniques du porc pour des coûts de production spécifiques : d'une part, l'Autriche (OST), caractérisée par son niveau de quartile supérieur (Q3 < 516,50 €), est la distribution la plus plate représentant le modèle à *translation simple* basé sur l'hypothèse de producteurs homogènes en leurs coûts spécifiques ; d'autre part, l'Italie (ITA), séparée par son niveau médian (Q2 < 429,25 €), présente la deuxième distribution la plus pentue illustrant le modèle à *translation-échelle*, formalisant l'hypothèse de producteurs hétérogènes en leurs coûts spécifiques.

4. Conclusions

L'analyse des intervalles d'estimation à l'aide d'une classification hiérarchique divisive permet d'identifier différents types de distributions nationales de coûts spécifiques pour le porc. Les différences entre les groupes de pays sont délimitées par des seuils exprimés selon les quantiles conditionnels en termes unitaires du produit brut. Ces analyses identifient deux modèles d'échelle de répartition des coûts, celui de la translation simple opposé à celui de la translation-échelle. Ces seuils peuvent être utilisés pour segmenter les populations d'exploitations agricoles afin d'analyser

Comment citer ce document :
Desbois, D. (2019). Application de la classification symbolique à l'estimation des coûts de production agricoles. In: Actes des 26èmes Rencontres, Société Francophone de Classification (SFC) (p. 65-70). Presented at XXVI-èmes Rencontres de la Société francophone de Classification, Nancy, FRA (2019-09-03 - 2019-09-05). 147

ultérieurement les impacts différentiels des mesures de politique agricole. L'application de cette méthodologie est envisagée au deuxième niveau de la Nomenclature européenne des unités territoriales statistiques (NUTS 2), soit 281 régions.

Références

Billard L., Diday E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, 321 p.

Chavent, M., *et al*. (2007). DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Comput. Statist. Data Anal.*, doi: 10.1016/j.csda.2007.03.013.

Desbois D. (2015). *Estimation des coûts de production agricoles : approches économétriques*. Thèse ABIES-AgroParisTech, dirigée par J.C. Bureau et Y. Surry, 249 p.

Desbois D., Butault J.-P., Surry Y. (2017). Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile, *Economie rurale*, n° 361, pp. 3-22.

Divay J.F., Meunier F. (1980). Deux méthodes de confection du tableau entrées-sorties. *Annales de l'INSEE*, n°37, pp. 59-109.

D'Haultfœuille X., Givord P. (2014). La régression quantile en pratique. *Économie et Statistique*, n°471, pp. 85-111.

He X., Hu F. (2002). Markov Chain Marginal Bootstrap, *Journal of the American Statistical Association*, n°97, pp. 783-795.

Koenker R., Bassett G. (1978). Regression quantiles. *Econometrica*, vol. 46, pp. 33-50.

Mirkin, B. (2005*). Clustering for Data Mining. A Data Recovery Approach*. Chapman & Hall, CRC Press, London, Boca Raton, FL, 366 p.

Summary

This communication analyses the similarities between distributions of conditional quantile estimates, applying it to the problem of cost allocation in agriculture. The first part presents the empirical model, the quantile regression approach and the interval data clustering technique used. The second part presents the comparative analysis of the results between twelve European Member States.

[i] Cette communication développe certains travaux de l'auteur réalisés lors de la préparation de sa thèse (Desbois, 2015), co-dirigée par Y. Surry et JC. Bureau, dans le cadre du projet FACEPA (*Farm Accountancy Cost Estimation and Policy Analysis*) du 7e programme-cadre de la Communauté européenne (FP7 / 2007 2013, approbation n ° 212292). Cette mention n'implique aucune approbation par les personnes et organismes cités du texte placé sous l'entière responsabilité de l'auteur.