Méthode de classification divisive sur intervalles d'estimation duale des quantiles de coûts spécifiques et de marges brutes

D. Desbois[1]

[1] INRAE-AgroParisTech, Paris Saclay Applied Economics

dominique.desbois@inrae.fr

Résumé

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 et des marges brutes en agriculture. Après avoir rappelé le cadre conceptuel de l'estimation des coûts et des marges en 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 les régions de douze États membres de l'UE des résultats issus de la classification hiérarchique divisive des estimations par intervalle, appliquée au blé.

Mots-clés

données d'intervalle, classification divisive, coûts spécifiques, marges brutes, régions européennes.

Abstract

This paper uses symbolic data clustering to explore the similarities between distributions of conditional quantile estimates, applying it to the problem of allocating specific costs and gross margins in agriculture. After recalling the conceptual framework of cost and margin estimation in agricultural production, the first part presents the empirical model, the quantile regression approach and the divisive clustering technique on interval data used. The second part presents the comparative analysis between the regions of twelve EU Member States of the results of the hierarchical divisive classification of interval estimates, applied to wheat.

Keywords

interval data, divisive clustering, specific costs, gross margins, European regions.

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 multiproductives [6]. 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 [8]. Puis, nous rappelons la méthodologie d'estimation permettant de prendre en compte l'hétérogénéité des exploitations agricoles, selon l'approche du quantile conditionnel proposée par [12]. Ensuite, pour explorer les distributions empiriques des intervalles d'estimation de quantiles conditionnels, nous présentons la procédure de classification utilisée [10] dans le cadre conceptuel de l'analyse symbolique de données [1]. 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 l'estimation du coût des fertilisants pour les productions végétales.

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 [8], l'affectation de la somme x_i des coûts des intrants pour l'exploitation agricole est réalisée par décomposition linéaire selon les produits bruts Y_i^j de l'exploitation agricole i pour chaque production j, où u_i est un vecteur aléatoire d'espérance nulle :

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

Comme [2], nous supposons que le processus générateur de données est un modèle linéaire à hétéroscédasticité multiplicative caractérisé par :

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

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 :

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

où F_ε^{-1} 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 :

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

Au moins deux types de modèle peuvent être dérivés de cette spécification [9] :

- $x = Y'\beta + u$ avec $u = K\varepsilon$, à résidus homoscédastiques ($V(\varepsilon \mid Y) = \sigma^2$), dénommé *modèle à translation simple*, *i.e.* un modèle linéaire à pentes homogènes ; puisque $Y'\alpha = K$ est constant, les quantiles conditionnels $\mu_q(x \mid Y,\beta,\alpha) = Y'\beta + K \times F_\varepsilon^{-1}(q)$ ont tous la même pente, mais diffèrent seulement d'un écart constant, croissant à mesure que l'ordre q du quantile augmente ;
- $x = Y'\beta + (Y'\alpha)\varepsilon$ avec $Y'\alpha > 0$ à résidus hétéroscédastiques, dénommé *modèle à changement d'échelle*, *i.e.* le modèle linéaire de quantiles conditionnels à pentes hétérogènes.

Suivant le modèle d'estimation des quantiles conditionnels pondérés par [13], la pondération Ω_I des observations, définie par $\{\omega_i; i = 1,\ldots,n\}$, est introduite dans la fonction de perte du problème de minimisation de la régression quantile comme suit :

$$\begin{aligned} &\Sigma_{(i;x_i \geq \beta y_i)}[\omega_i q \|x_i - \beta y_i\|] \\ &\qquad + \\ &\Sigma_{(i;x_i < \beta y_i)}[\omega_i (1-q) \|x_i - \beta y_i\|] \end{aligned} \tag{5}$$

conduisant à l'estimation des paramètres du modèle (2) comme solution optimale du problème de minimisation de la fonction de perte (5), soit :

$$\begin{aligned} \widehat{\beta_{\omega(q)}} = \\ argmin_{(\beta \in R^p)}\{&\Sigma_{(i;x_i \geq y_i'\beta)}[\omega_i q \|x_i - y_i'\beta\|] \\ &+ \\ &\Sigma_{(i;x_i < y_i'\beta)}[\omega_i (1-q) \|x_i - y_i'\beta\|]\} \end{aligned} \tag{6}$$

Les estimations pondérées des quantiles conditionnelles sont fournies par la procédure QUANTREG du logiciel SAS, version 9.2.

2.2 L'estimation duale, complète ou partielle

Le q^e quantile conditionnel possède la propriété d'équivariance, spécifique aux transformations monotones, impliquant les deux règles conditionnelles suivantes :

— si $\lambda \in [0;\infty]$ alors

$$\mu_q(\lambda \times X + C \mid Y) = C + \lambda \times \mu_q(X \mid Y) \tag{7}$$

— si $\lambda \in [-\infty;0]$ alors

$$\mu_q(\lambda \times X + C \mid Y) = C + \lambda \times \mu_{1-q}(X \mid Y) \tag{8}$$

Par re-paramétrisation en X de $M = Y - X$, la seconde règle permet de déduire l'estimation unitaire de marge brute à partir de l'estimation unitaire de coûts spécifique, suivant la séquence de transformations ci-après :

$$\mu_q(M \mid Y) = \mu_q(Y - X \mid Y) = 1 - \mu_{(1-q)}(X \mid Y) \tag{9}$$

L'estimation duale $\mu_q^{mb} = \mu_q(\widehat{M \mid Y})$ correspond à l'estimation $\mu_{(1-q)}^{cs} = \mu_{(1-q)}\widehat{(X \mid Y)}$.
Au terme de ce processus d'estimation, les distributions de paramètres sont *complètement estimées* si l'ensemble de leurs différentes estimations obtenues pour les p différents paramètres quantiles peuvent être considérées sur la base de leur variabilité comme significativement non-nulles. Dans le cas contraire, les distributions de paramètres sont *partiellement estimées*.

2.3 Analyse factorielle des distributions empiriques duales

Soit $\Delta = \{\delta_1,\ldots,\delta_i,\ldots,\delta_n\}$, l'ensemble des distributions empiriques de coûts spécifiques et de marges brutes, décrites par un ensemble de $2p$ estimations quantiles conditionnelles : $\hat{\Gamma} = \{\widehat{\mu_1^{cs}},\ldots,\widehat{\mu_j^{cs}},\ldots,\widehat{\mu_p^{cs}},\widehat{\mu_1^{mb}},\ldots,\widehat{\mu_j^{mb}}j,\ldots,\widehat{\mu_p^{mb}}\}$.
L'analyse factorielle des distributions empiriques est conduite par analyse en composantes principales normées (ACPn).En raison du centrage et de la standardisation des estimations, l'analyse montre que l'ACPn du tableau complet $\hat{\Gamma}$ des quantiles estimés (coûts spécifiques et marges brutes) et l'ACPn d'un sous-tableau (soit celui des coûts spécifiques, soit celui des marges brutes) sont structurellement équivalentes, à une constante signée près dans la définition des composantes principales. Cependant, l'utilisation conjointe des coûts et des marges permet d'enrichir conceptuellement l'interprétation.

Certains quantiles conditionnels des régions insuffisamment représentées n'étant pas estimables ou non significatifs, une affectation au plus proche barycentre, selon une norme quadratique des écarts, permet de décider de l'appartenance des régions partiellement estimées aux classes de la partition *P*, retenue comme référentiel typologique. La procédure d'imputation des estimations quantiles non significatives pour les régions partiellement estimées s'apparente aux méthodes de hot-deck métrique utilisées pour le traitement de la non-réponse.
Pour visualiser les distributions de coûts et de marges du référentiel typologique retenu (partition P), nous utilisons les intervalles d'estimation $[Inf; Sup]$, selon l'extension de l'analyse en composantes principales normalisée d'intervalles (ACPni) proposée par [3]. La localisation d'hyper-rectangles y est construite à partir de la projection des intervalles de confiance, arêtes des hyper-rectangles, et renseigne sur les différences significatives de niveau et de forme.

2.4 Classification par intervalles des distributions de coûts spécifiques

Pour un produit donné tel que le blé, le coûts spécifique ou la marge brute (j_0) et une région européenne (l), l'intervalle d'estimation des coefficients techniques de coût spécifique ou de marge brute

$$z_l^q = [Inf\{\widehat{\gamma_l^{j_0}(q)}\}; Sup\{\widehat{\gamma_l^{j_0}(q)}\}] \tag{10}$$

est obtenu par bootstrap marginal en chaînes de Markov [11]. Objets symboliques, les L distributions régionales $\Omega = \{\omega_1, \ldots, \omega_l, \ldots, \omega_L\}$,sont décrites par un ensemble de $Q = 5$ descripteurs qui sont les intervalles d'estimation des coefficients techniques pour les quantiles conditionnels $Z = \{z_1^{cs}, \ldots, z_q^{cs}, \ldots, z_Q^{cs}, z_1^{mb}, \ldots, z_q^{mb}, \ldots, z_Q^{mb}\}$.
Le choix d'un petit nombre de descripteurs, soit $Q = 5$, $\{z_{0,1}^{cs}, z_{0,25}^{cs}, z_{0,5}^{cs}, z_{0.75}^{cs}, z_{0,9}^{cs}, z_{0,1}^{mb}, z_{0,25}^{mb}, z_{0,5}^{mb}, z_{0,75}^{mb}, z_{0,9}^{mb}\}$ a été fait pour des raisons de comparabilité avec des approches graphiques plus classiques [5]. 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 ($Q = 99$).
Les dissimilarités locales entre la région l et la région l', associées aux intervalles d'estimation des coefficients techniques pour le q^e quantile conditionnel, sont calculées selon la distance euclidienne :

$$\begin{aligned} &\delta_M^2(z_l^q, z_{l'}^q) = \\ &(Inf\{\widehat{\gamma_l^{j_0}(q)}\} - Inf\{\widehat{\gamma_{l'}^{j_0}(q)}\})^2 \\ &+(Sup\{\widehat{\gamma_l^{j_0}(q)}\} - Sup\{\widehat{\gamma_{l'}^{j_0}(q)}\})^2 \end{aligned} \tag{11}$$

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

$$d_M(\omega_l, \omega_{l'}) = \left(\sum_{q=1}^{Q} \delta_M^2(z_l^q, z_{l'}^q)\right)^{\frac{1}{2}} \tag{12}$$

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, [4] propose un algorithme divisif de classification descendante hiérarchique sur données symboliques (DIVCLUST), valable pour les données d'intervalle et les données catégorielles.
Par la suite, nous détaillons pour les données 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 $\Omega = \{\omega_1, \ldots, \omega_l, \ldots, \omega_L\}$. À chaque partition $P_K = \{C_1, \ldots, C_k, \ldots, C_K\}$ en K classes symboliques, 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 $I(C_k) = \sum_{l \in C_k} \mu_l d_M^2(z_l, g(C_k))$ où μ_l est le poids du l^e pays et $g(C_k$ est le barycentre de classe définie par :

$$g(C_k) = \frac{1}{\sum_{l \in C_k} \mu_l} \sum_{l \in C_k} \mu_l z_l \tag{13}$$

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

$$W(P_k) = \sum_{k=1,\ldots,K} I(C_k) \tag{14}$$

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

$$\begin{aligned} W(P_k) = &\sum_{k=1,\ldots,K} \mu_k d_M^2(g(C_k, g)) \\ &\text{où } \mu_k = \sum_{l=1,\ldots,L} \mu_l \end{aligned} \tag{15}$$

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

$$I(\Omega) = W(P_K) + B(P_K) \tag{16}$$

Ainsi, minimiser l'hétérogénéité (mesurée par W) est équivalent à maximiser l'homogénéité (mesurée par B).
Générée 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 défini par le ratio suivant :

$$D(\Psi) = \frac{B^q(A_k, \overline{A_k})}{I^j(C_k)} = 1 - \frac{W^j(A_k, \overline{A_k})}{I^q(C_k)} \tag{17}$$

où l'inertie inter-classes $B^q(A_k, \overline{A_k})$ et l'inertie $I^q(C_k)$ sont calculées par rapport au q^e quantile conditionnel. Aussi, minimiser l'inertie intra-classes $W(A_k, \overline{A_k})$ équivaut à maximiser l'inertie inter-classes $B(A_k, \overline{A_k})$ et, par

conséquence, le critère discriminant $D(\Psi)$.
Comme dans la méthode de Ward, la « hiérarchie supérieure » [14] à la partition P_K est indexée par l'indice h de la classe C_K, définie par son inertie inter-classes comme suit :

$$h(C_k) = B(A_k, \overline{A_k}) = \frac{\mu(A_k)\mu(\overline{A_k})}{\mu(A_k) + \mu(\overline{A_k})} d^2(g(A_k), g(\overline{A_k})) \quad (18)$$

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

$$W(P_{K+1}) = W(P_K) - h(C_K^*) \quad (19)$$

D'après les deux règles conditionnelles de l'équivariance (§2.2), le critère discriminant (15) appliqué aux quantiles conditionnels demeure invariant par la transformation monotone correspondant à une reparamétrisation de X de $M = Y - X$. Ainsi, les hiérarchies de classification divisive obtenues sont en dualité au sens de (7) par leurs seuils divisifs si l'on passe des quantiles conditionnels de coûts spécifiques aux quantiles conditionnels de marges brutes, et inversement. Comme pour le *biplot*, cette propriété justifie d'étiqueter le dendrogramme résultant de façon soit alternative, soit simultanée par les seuils estimées de coûts spécifiques ou de marges brutes pour en faciliter l'interprétation.

3 Application aux régions de l'Union européenne

L'estimation des quantiles de marge brute et de coûts spécifiques pour p=5 quantiles, déciles extrêmes (D1 et D9), quartiles (Q1 et Q3) et médiane (Q2) est effectuée conditionnellement au produit brut (cf. figure 1).

3.1 Estimations duales des coûts spécifiques et des marges brutes

Ces estimations vérifient empiriquement la relation de dualité (9), déduite de l'équivariance conditionnelle des estimateurs quantiles conditionnels par application des propriétés (7) et (8). Ainsi, pour le Schleswig-Holstein ($A010$), le premier décile estimé la marge brute unitaire ($D1mb$) est le complément à 1 000 € du neuvième décile estimé des coûts spécifiques ($D9cs$) : $\widehat{\mu_{0,1}^{mb}} = 458 = 1000 - \widehat{\mu_{0.9}^{cs}} = 1000 - 542$. La relation de complémentarité similaire entre le quartile inférieur de marge brute ($Q1mb$) et le quartile supérieur de coûts spécifiques ($Q3cs$) est également vérifiée quel que soit la région, par exemple pour le Schleswig-Holstein : $\widehat{\mu_{0,2}^{mb}} = 646 = 1000 - \widehat{\mu_{0,75}^{cs}} = 1000 - 354$. Cette propriété de dualité des estimations a été vérifiée par l'application conjointe de procédures d'estimation portant sur les coûts spécifiques d'une part et d'autre part sur les marges brutes.
Signalons le caractère très asymétrique des distributions,

Blé		Coûts spécifiques (€) pour 1 000 € de produit brut						Marges brutes (€) pour 1 000 € de produit brut					
Code	Régions RICA	D1cs	Q1cs	Q2cs	Q3cs	D9cs	MCOcs	D1mb	Q1mb	Q2mb	Q3mb	D9mb	MCOmb
A010	Schleswig-Holstein	389	407	364	354	542	464	458	646	636	593	611	536
A030	Niedersachsen	214	229	295	351	327	305	673	649	705	771	786	695
A050	Nordrhein-Westfalen	325	245	222	224	233	204	767	776	778	755	675	796
A060	Hessen	383	410	343	356	333	428	667	644	657	590	617	572
A090	Bayern	283	354	418	561	669	487	331	439	582	646	717	513
A115	Sachsen-Anhalt	245	214	290	284	289	188	711	716	710	786	755	812
F131	Champagne-Ardenne	379	514	537	530	752	492	248	470	463	486	621	508
F132	Picardie	314	299	376	465	465	332	535	535	624	701	686	668
F134	Centre	303	317	376	490	568	367	432	510	624	683	697	633
F135	Basse-Normandie	699	646	617	525	687	507	313	475	383	354	301	493
F136	Bourgogne	475	442	543	657	779	491	221	343	457	558	525	509
F141	Nord-Pas-de-Calais	321	370	452	508	687	498	313	492	548	630	679	502
F151	Lorraine	524	508	488	398	531	452	469	602	512	492	476	548
F152	Alsace	589	721	817	886	1001	864	-1	114	183	279	411	136
F153	Franche-Comté	689	759	680	625	832	706	168	375	320	241	311	294
F162	Pays de la Loire	428	421	435	562	555	403	445	438	565	579	572	597
F163	Bretagne	368	400	456	421	483	396	517	579	544	600	632	604
F192	Rhône-Alpes	609	640	904	935	911	903	89	65	96	360	391	97
F193	Auvergne	336	384	528	792	804	434	196	208	472	616	664	566
I260	Emilia-Romagna	187	270	374	437	481	346	519	563	626	730	813	654
I282	Umbria	316	305	327	358	359	376	641	642	673	695	684	624
B341	Flandre	305	363	446	496	1085	552	-85	504	554	637	695	448
B343	Wallonie	397	306	342	407	407	404	593	593	658	694	603	596
D370	Danemark	213	261	363	462	552	426	448	538	637	739	787	574
U411	England-North	332	348	346	471	564	375	436	529	654	652	668	625
U412	England-East	264	302	339	421	475	323	525	579	661	698	736	677
U413	England-West	363	343	294	308	406	349	594	692	706	657	637	651
U431	Scotland	338	405	522	443	490	478	510	557	478	595	662	522
E570	Extremadura	182	206	218	373	498	299	502	627	782	794	818	701
O690	Österreich	201	240	266	261	344	200	656	739	734	760	799	800
S710	Stattbyrdslan	341	331	388	522	567	323	433	478	612	699	659	677
H761	Közép-Dunántúl	382	465	419	348	638	563	362	652	581	535	618	437
H762	Nyugat-Dunántúl	361	357	393	388	509	759	491	612	607	643	639	241
H765	Eszak-Alfoid	302	238	339	371	341	197	659	629	661	762	698	803
P785	Pomroze & Mazury	300	353	395	505	562	418	438	495	605	647	700	582
P790	WielkoPolska &Slask	252	280	344	430	536	424	464	570	656	720	748	576
P795	Mazowsze & Podlasie	266	263	302	349	458	336	542	651	698	737	734	664
P800	Malopolska & Pogórze	254	290	372	331	383	347	617	669	628	710	746	653

FIGURE 1 – Estimations des coûts spécifiques et des marges brutes pour 1000 € de produit brut.

ainsi les écarts d'estimation de coûts spécifiques les plus importants (supérieurs à 20% en valeur absolue) entre l'estimation conditionnelle de la médiane ($Q2cs$ et $Q2mb$) et celle de la moyenne ($MCOcs$ et $MCOmb$) sont pour les coûts spécifiques l'Extremadure (128%) et la Flandre (143%), tandis que pour les marges brutes, il s'agit de la Basse-Normandie (29%) et de la région hongroise du Nyugat-Dunàntùl (60%).
La distribution des valeurs propres de l'ACPn conduit à retenir les deux premières composantes principales selon la règle de Kayser. Ces deux premières composantes principales rassemblent 91% de l'information disponible. Le *biplot* de ces deux premières composantes principales est reproduit en figure 2. La première composante principale ($Dim1$) représente 80% de l'inertie. L'ensemble des quantiles conditionnels de coûts spécifiques (cs) est corrélé négativement ($Dim1 < 0$) tandis que l'ensemble des quantiles conditionnels de marge brute (mb) est corrélé positivement ($Dim1 > 0$). Les médianes conditionnelles ($Q2cs$ et $Q2mb$ de signes opposés) sont les quantiles estimés les mieux corrélés à cette première composante principale, qui reflète ainsi le niveau global de la distribution des quantiles estimés. L'angle entre l'orientation du vecteur $Q2cs$ et du vecteur $MCOcs$, identique à celui entre $Q2mb$ et $MCOmb$, fournit une indication intéressante sur l'asymétrie des distributions : plus cet angle est ouvert, plus l'asymétrie est forte. L'asymétrie des estimations quantiles de coûts supérieurs est portée par les quantiles d'ordre supérieur ($Q3cs$ et $D9cs$) tandis que celle des estimations quantiles de marge brute par les quantiles d'ordre inférieur ($Q1mb$, $D1mb$).
La seconde composante principale ($Dim2$), représentant 11% de l'inertie, restitue la dispersion de la distribution en opposant, dans le demi-plan $Dim1 < 0$, les quantiles su-

périeurs de coûts $Q3cs$ et $D9cs$ ($Dim2 < 0$) aux quantiles inférieurs de coûts spécifiques $Q1cs$ et $D1cs$ (demi-plan $Dim2 > 0$). De façon similaire, la seconde composante principale oppose dans le demi-plan $Dim1 > 0$ les quantiles inférieurs de marges brutes $D1mb$ et $Q1mb$ ($Dim2 > 0$) aux quantiles supérieurs de marge brute $Q3mb$ et $D9mb$ ($Dim2 < 0$).

3.2 Organisation de la variabilité des régions totalement estimées

Le *biplot* permet d'étiqueter chaque quadrant des plans factoriels de projection des observations par des orientations croissantes (e.g. $Q2^{+}_{cs}$ pour la zone d'estimations élevée des coûts spécifiques) ou décroissantes (e.g. $Q2^{-}_{mb}$ pour la zone d'estimations faibles des marges brutes) pour faciliter la lecture et l'interprétation des graphiques factoriels (cf. figures 4 et 5). La hiérarchie divisive obtenue avec l'option de distance euclidienne (figure 3) montre que, à l'exception du quantile *D9*, l'ensemble des estimations des quantiles *D1*, *Q1*, *Q2* et *Q3* est utilisé par les valeurs discriminantes, ce qui conduit à conserver ces paramètres pour décrire la distribution.

L'algorithme divisif permet d'interpréter les différences les

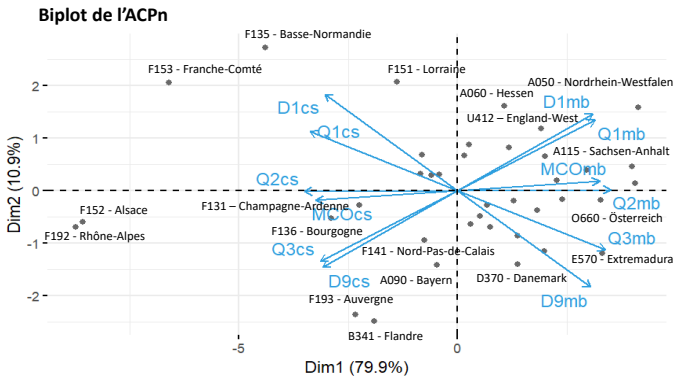

FIGURE 2 – Biplot de l'ACPN des estimations duales par quantiles de coût et de marge, régions de l'UE12. Source : traitement de l'auteur, RICA-UE 2006.

plus marquantes entre classes au sien d'une même partition en fonction de seuils de coûts et de marges. La première partition $P2$ en deux classes $C1$ et $C2$ où les régions se répartissent selon les valeurs du quartile inférieur de coûts ($Q1c$) et du quartile supérieur de marges ($Q3m$). La classe $C1$ rassemble les régions pour lesquelles $Q1c$ est supérieur ou égal à 577 € ($Q1c \geq 577$) et $Q3m$ est inférieur à 423€ ($Q3m < 423$). La classe $C2$ regroupe les régions pour lesquelles $Q1c$ est inférieur à 577€ ($Q1c < 577$) et $Q3m$ est supérieur ou égal à 423€ ($Q3m \geq 423$). Ainsi, au sommet de la hiérarchie divisive, la procédure de classification permet d'identifier deux modèles contrastés pour les distributions empiriques des coûts et des marges unitaires du blé pour 1€ de produit brut :

— d'une part, le modèle à résidus homoscédastiques identifiable à la classe $C1$ (coûts élevés et faibles marges aux distributions relativement homogènes) ;
— d'autre part, le modèle à résidus hétéroscédastiques identifiable à la classe $C2$ (coûts plus faibles et marges plus élevées aux distributions relativement hétérogènes).

La seconde partition $P3$ en trois classes divise la classe $C2$ en deux agrégats $C2.1$ et $C2.2$ selon les quantiles médian de coûts ($Q2c$) et de marge ($Q2m$). L'agrégat $C2.1$ rassemble les régions dont le quantile médian de coûts est supérieur ou égal à $609,5$ € ($Q2c \geq 609,5$) et le quantile médian de marge est inférieur à $390,5$ € ($Q2m < 390,5$). L'agrégat $C2.1$ regroupe les régions dont le quantile médian de coûts est inférieur à $609,5$€ ($Q2c < 609,5$) et le quantile médian de marge est supérieur ou égal à $390,5$ € ($Q2m \geq 390,5$). Cette césure correspond à une opposition de leurs projections selon la première composante principale ($Dim1$).

Affichée en figure 3, la partition $P4$ en quatre classes est optimale pour la différence minimale dans le logarithme du rapport des déterminants, fournie par le paquet *clusterCrit* [7], qui constitue une règle cohérente avec le critère de l'algorithme DIVCLUS-T [10]. Au niveau de cette partition $P4$, la classe $C1$ se scinde en deux agrégats $C1.1$ et $C1.2$ selon le décile supérieur de marge ($D9m$) et le décile inférieur de coûts ($D1c$). L'agrégat $C1.1$ regroupe les régions dont le neuvième décile est supérieur ou égal à 351 € ($D9m \geq 351$) et le premier décile de coûts est inférieur à 649€ ($D1c < 649$). L'agrégat $C1.2$ rassemble les régions dont le neuvième décile est inférieur à 351€ ($D9m < 351$) et le premier décile de coûts est supérieur ou égal à 649€ ($D1c \geq 649$). Cette césure correspond à une opposition de leurs projections selon la seconde composante principale ($Dim2$).

La figure 4 projette la partition en treize classes (*P13*) sur le plan principal de l'ACPn distributionnelle et permet de prendre en compte davantage d'information structurelles apportées par les quantiles conditionnels. En effet, l'axe *F2* constitue le facteur de dispersion intraclasse lié aux niveaux relatifs des quantiles conditionnels supérieurs (*D9* et *Q3*) vis-à-vis des quantiles conditionnels inférieurs (*Q1* et *D1*). Selon l'axe *F2*, la classe *C1* aux estimations quantiles les plus extrêmes est scindée en deux agrégats bien distincts. D'une part dans le quadrant $F1 > 0$ & $F2 > 0$, l'aîné $C1.1 = \{$*Rhône-Alpes*, *Alsace*$\}$ aux quantiles supérieurs de coûts parmi les plus élevés ($Q3^{+}_{cs} = 911$€, $D9^{+}_{cs} = 956$€) et aux quantiles inférieurs de marge parmi les plus faibles ($Q1^{-}_{mb} = 89$ €, $D9^{-}_{mb} = 44$€). D'autre part, dans le quadrant $F1 > 0$ & $F2 < 0$, le benjamin $C1.2 = \{$*Franche-Comté*, *Basse-Normandie*$\}$ présentant un décile supérieur de coûts équivalent mais dont les autres estimations quantiles sont des extrema de second rang, avec des quantiles inférieurs de coûts parmi les plus élevés ($Q1^{+}_{cs} = 703$€, $D1^{+}_{mb} = 694$€) et des quantiles supérieurs de marges parmi les plus faibles ($Q3^{-}_{mb} = 297$€, $D9^{-}_{mb} = 306$€).

Selon l'axe $F2$, la classe $C2.1$ est formée par la réunion de deux agrégats :

— d'une part, situé dans le quadrant $F1 < 0$ & $F2 > 0$), l'agrégat $C2.1.1 = \{$*Extremadura*, *Wielkopolska*

& Slask, *Danemark*, *Emilia-Romagna*, *England-East*, *Picardie*} présente des estimations de quantiles inférieurs de coûts plus faibles ($Q1^-_{cs} = 273$ €, $D1^-_{cs} = 238$ €) et de quantiles supérieurs de marges plus élevées ($Q3^-_{mb} = 727$€, $D9^-_{cs} = 762$€) que l'ensemble des régions actives ;

— d'autre part, situé essentiellement dans le quadrant $F1 < 0$ &$F2 < 0$ l'agrégat $C2.1.2$ ={*Mazowsze & Podlasie*, *Malopolska & Pogórze*, *Eszak-Alfoid*, *Umbria*, *England West*, *Wallonie*, *Hessen*, *Nordrhein-Westfalen*, *Österreich*, *Sachsen-Anhalt*, *Niedersachsen*} se distingue du précédent par des estimations de quantiles inférieurs de marges plus fortes ($Q1^+_{mb} = 673$€, $D9^+_{cs} = 647$€) et des estimations de quantiles supérieurs de coûts plus faibles ($Q3^-_{cs} = 327$€, $D9^-_{cs} = 353$€).

Au niveau des coûts faibles (demi-plan $F1 < 0$), les agrégats $C2.1.1.2.1$ ={*Emilia-Romagna*, *Danemark*, *Wiekopolska & Slask*} et $C2.1.2.1.2$ ={*Hessen*, *Wallonie*, *England-West*}, situés au même niveau médian de coûts spécifiques ($Q2^-_{cs} = 360$€ versus $Q2^-_{cs} = 326$€ respectivement), se différencient selon l'axe $F2$ par leurs profils de distribution en quantiles inférieurs de coûts (soit $D1^-_{cs} = 217$€ versus $D1^-_{cs} = 383$€, et respectivement $D1^-_{cs} = 217$€ versus $D1^-_{cs} = 383$€) relativement moins élevés pour l'agrégat $C2.1.1.2.1$ par rapport à ceux de l'agrégat $C2.1.2.1.2$ alors qu'ils appartiennent tous les deux à la classe $C2.1$ de la partition $P4$. Enfin, à un niveau plus élevé de coûts (demi-plan $F1 > 0$), les agrégats $C2.2.2.1.2$ ={Bayern, Nord-Pas-de-Calais} et $C2.2.2.2.1$ ={Lorraine, Bretagne, Pays de la Loire, Scotland}, issus de l'éclatement de la classe $C2.2.2$ de la partition $P4$, se distinguent tant selon les quantiles supérieurs de marge (avec des estimations plus élevées pour $C2.2.2.1.2$, soit $Q3^+_{mb} = 638$€ et $D9^+_{mb} = 698$€, que pour $C2.2.2.1.2$, soit $Q3^+_{mb} = 567$€ et $D9^+_{mb} = 586$€) que selon les quantiles inférieurs de coûts (avec des estimations plus faibles pour $C2.2.2.2.1$ soit $Q1^-_{cs} = 362$€ et $D1^-_{cs} = 302$€, que pour $C2.2.2.1.2$, soit $Q1^+_{cs} = 434$€ et $D1^+_{cs} = 415$€).

Ainsi, dans l'allocation spécifique des charges aux produits, l'estimation en quantiles conditionnels nous permet de conserver l'information distributionnelle relative à l'hétérogénéité des coûts et des marges pour un niveau donné de dépenses spécifiques au blé, contrairement à l'estimation des moindres carrés ordinaires.

Le graphique suivant (figure 5) visualise les résultats de l'analyse en composantes normée sur intervalles (ACPni) sur la base des intervalles d'estimations par intervalles $[Inf, Sup]$ de quantiles conditionnels des barycentres des 13 agrégats du référentiel typologique $P13$ choisi pour l'affectation pseudobarycentrique.

Le premier axe factoriel $F1$ de l'ACPni donne un gradient décroissant (inversé par rapport aux axes factoriels des ACP classiques précédentes) des estimations de coûts spécifiques allant des estimations supérieures de coûts médians et inférieures de marges médianes (agrégats $C1.1$ et $C1.2$ projetés aux extrêmes du pôle $F1 < 0$, $Q2^+_{cs} = 755$€ et $Q2^-_{mb} = 246$€, en moyenne) aux estimations inférieures

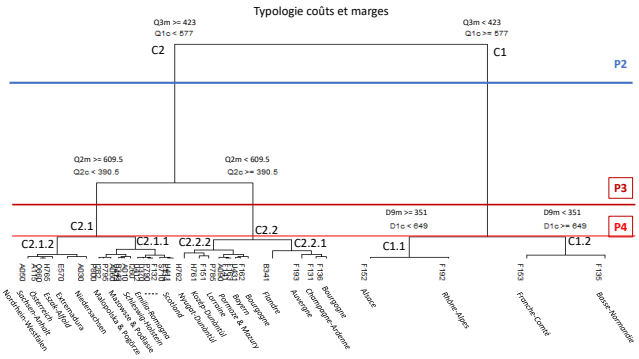

FIGURE 3 – Classification symbolique divisive en distance euclidienne pour les estimations duales en quantiles de coûts et de marges, régions de l'UE12.

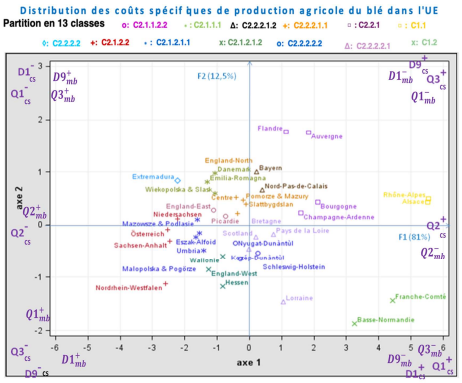

FIGURE 4 – blé, projection de la partition en treize classes des régions complètement estimées.

de coûts médians et aux estimations supérieures de marges médianes (agrégats $C2.1.2.2$ et $C2.1.1.1$ projetés aux extrêmes du pôle $F1 > 0$, $Q2^-_{cs} = 243$€ et $Q2^+_{mb} = 757$€, en moyenne).

Le second axe factoriel $F2$ de l'ACPni oppose l'agrégat $C1.1$ ={*F192-Rhône-Alpes*, *F152-Alsace*} dans le quadrant $F1 < 0$ & $F2 > 0$, présentant un écart d'estimation élevé entre le premier et le dernier décile de coûts ($[D9\text{-}D1]^+_{cs} = 357$€, en moyenne). Dans le quadrant $F1 < 0$ & $F2 < 0$, l'agrégat $C1.2$ ={*F135-Basse-Normandie*, *F153-Franche-Comté*} dont les écarts d'estimation entre le premier et le dernier décile sont parmi les plus faibles ($[D9\text{-}D1]^-_{cs} = 66$€, en moyenne). Le second axe constitue également l'axe majeur de dispersion intra-classe des autres agrégats de la partition $P13$, en particulier celui de l'agrégat $C2.2.1$ ={*B341-Flandre*, *F193-Auvergne*, *F136-Bourgogne*, *F131-Champagne-Ardenne*}. Cette analyse peut servir de test graphique de séparation : les agrégats $C1.1$ et $C1.2$ de la classe $C1$ présentent entre

eux des différences de coûts et de marge à la fois en termes de niveau (selon l'axe $F1$ corrélé au niveau médian $Q2$) et de structure (selon l'axe $F2$, corrélé à l'écart inter-décile $D9 - D1$).

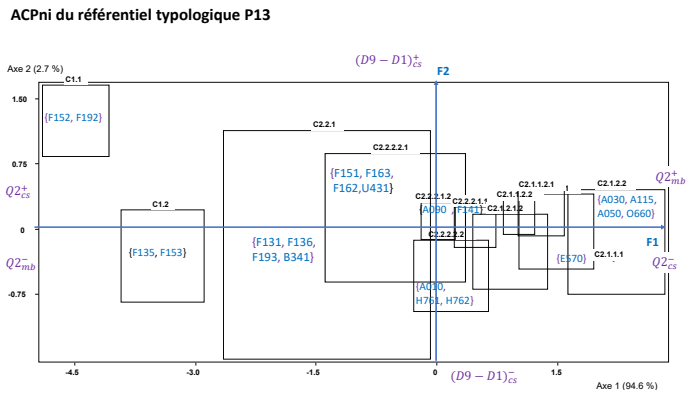

FIGURE 5 – ACPni du référentiel typologique en 13 classes, estimations duales en quantiles de coûts et de marges, régions de l'UE12.

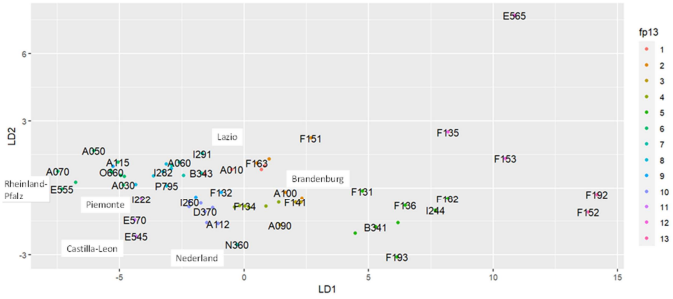

FIGURE 6 – blé, ADL de la partition P13 pour les régions complètement estimées et classement des régions partiellement estimées

Une affectation au plus proche barycentre, selon une norme quadratique des écarts, permet de décider de l'appartenance des régions partiellement estimées aux treize classes de la partition $P13$, retenue comme référentiel typologique. Quasiment équivalente à une analyse discriminante linéaire (ADL, figure 6) réalisée à partir des estimations complètes de distributions régionales (échantillon d'apprentissage) et appliquée aux régions partiellement estimées (échantillon-test), l'affectation réalisée par la procédure FASTCLUS, à partir des estimations barycentriques de quantiles conditionnels, est projetée dans le premier plan factoriel (figure 7) de l'ACPn des barycentres d'agrégats représentant 98% de l'inertie interclasses de la partition $P13$.

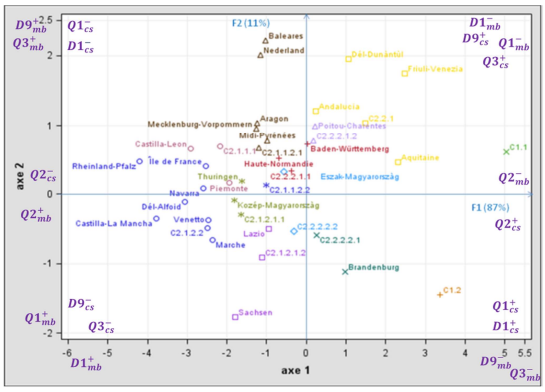

FIGURE 7 – ACPn des barycentres de la partition P13 et imputation des régions partiellement estimables.

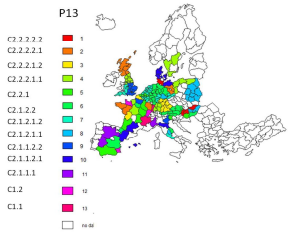

FIGURE 8 – blé, projection cartographique de la partition P13 des régions complètement et partiellement estimées.

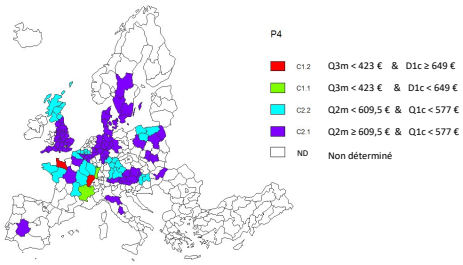

FIGURE 9 – blé, projection cartographique de la partition P4 des régions complètement estimées.

Une cartographie des régions complètement estimées et de l'affectation des régions partiellement estimées (figure 8) situe la localisation des classes du référentiel typologique en treize classes et leur répartition au sein de l'espace agricole de l'Union européenne.Cette carte typologique en treize classes précise et complète, la cartographie effectuée à partir de la partition $P4$ en quatre classes (figure 9) par une catégorisation plus fine et respectivement une couverture territoriale plus étendue. Comme affiché en figure 9, il est loisible de communiquer la genèse duale de chaque classe fournie par la hiérarchie divisive en termes de seuils de coûts et de marges pour en faciliter l'interprétation.

4 Conclusion

Sur la base du Réseau d'information comptable européen, nous avons testé sur une base régionale la faisabilité et la pertinence pour le blé de notre méthodologie d'estimation micro-économétrique des coûts spécifiques de production et des marges brutes selon les quantiles conditionnels. Cette méthodologie est complétée par une procédure d'imputation pour les régions partiellement estimées.
Compte-tenu de la nature duale des estimations de coûts spécifiques et de marge brute, la représentation en *biplot* constitue une aide à l'interprétation pertinente. En cohérence, les noeuds du dendrogramme de la classification divisive basée sur le pourcentage de variabilité interclasses sont étiquetés de façon duale par les seuils estimés de coûts et de marge facilitant l'interprétation de la hiérarchie des partitions.
L'analyse en composantes principales sur intervalle permet d'interpréter les composantes de la variabilité de la distribution des agrégats de la typologie choisie comme référentiel pour l'imputation pseudo-barycentrique des régions partiellement estimées.
Grâce à ce type d'analyse, nous confirmons qu'il n'y a pas *un* coût spécifique national de production qui pourrait être estimé par une moyenne conditionnelle mais des classes régionales européennes de distribution des coûts spécifiques et des marges brutes qui peuvent être positionnées dans un schéma bidimensionnel stable selon un nombre déterminé d'estimations quantiles conditionnelles.
Pour mieux distinguer les différences entre certaines des distributions régionales, il est loisible d'étendre l'analyse à une échelle de quantiles plus fine si nécessaire.
In Memoriam : l'auteur dédie ce travail à la mémoire d'***Edwin Diday***, Professeur émérite de l'Université Paris Dauphine, récemment disparu.

Références

[1] L. Billard et E. Diday, *Symbolic Data Analysis : Conceptual Statistics and Data Mining,* Wiley–Blackwell, 2006.

[2] A. Cameron et P. K. Trivedi, *Microeconometric. Methods and Applications,* University Press, 2005.

[3] P. Cazes, A. Chouakria, E. Diday et Y. Schektman, Extension de l'analyse en composantes principales à des données de type intervalle, *Revue de Statistique Appliquée*, Vol. 45(3), pp. 5-24, 1997.

[4] M. Chavent, Y. Lechevallier et O. Briant, Divclus-t : A Monothetic Divisive Hierarchical Clustering Method, *Comput. Statist. Data Anal.*, Vol. 52(2), pp. 687-701, 2007.

[5] D. Desbois, *Estimation des coûts de production agricoles : approches économétriques*, Thèse de doctorat, Université Paris Saclay, 2015.

[6] D. Desbois, J.P. Butault et Y. Surry, Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile, *Économie rurale*, Vol. 361, pp. 3-22, 2017.

[7] B. Desgraupes, Clustering indices, *Vignette R* , CRAN, 2017.

[8] J.F. Divay et F. Meunier, Deux méthodes de confection du tableau entrées-sorties, *Annales de l'INSEE*, Vol. 37, pp. 59-109, 1980.

[9] X. D'Haultfoeuille et P. Givord, La régression quantile en pratique, *Économie et Statistique*, Vol. 471(1), pp. 85-111, 2014.

[10] M. Fuentes et M. Chavent, Clustering divisif monothétique, *Vignette R*, 4^e Rencontre *R*, 2015.

[11] X. He et F. Hu, Markov Chain Marginal Bootstrap, *Journal of the American Statistical Association*, Vol. 97(459), pp. 783-795, 2002.

[12] R. Koenker et G. Bassett, Regression quantiles, *Econometrica*, Vol. 46, pp. 33-50, 1978.

[13] R. Koenker et Q. Zhao, L-estimation for linear heteroscedastic models, *Journal of Nonparametric Statistics*, Vol. 3, pp. 223-235, 1994.

[14] B. Mirkin, *Clustering for Data Mining. A Data Recovery Approach,* CRC Press, 2005.