Analyse régionale des coûts spécifiques en production agricole : utilisation de la distance de Wasserstein

Dominique Desbois[1]

CORRESPONDANCE
dominique.desbois@inrae.fr

RÉSUMÉ

Dans le contexte d'une prochaine réforme de la Politique agricole commune, l'analyse des coûts de production agricole est une étape déterminante pour élaborer des politiques de prix équitables, en particulier au plan régional pour éviter d'éventuelles distorsions de concurrence. Cet article présente l'application de techniques statistiques basées sur la distance de Wasserstein à l'analyse régionale des coûts spécifiques de production agricole. Prenant en compte l'hétérogénéité et l'asymétrie des distributions, elle articule la méthodologie de l'estimation quantile conditionnelle à celles basées sur la distance de Wasserstein (analyse factorielle du tableau de distances, classifications non supervisées, test quadratique). La procédure est appliquée pour l'analyse régionale comparative des distributions de coûts spécifiques en fertilisants, pour les productions végétales sur la base du Réseau européen d'information comptable agricole, à l'échelle des régions de base européenne (NUTS2). Les résultats obtenus sur les coûts spécifiques en intrants apparaissent cohérents avec ceux de travaux antérieurement menés dans le cadre du 7e programme cadre de recherche de l'Union européenne.

MOTS-CLÉS

analyse factorielle du tableau de distances – classification divisive – distance de Wasserstein – estimation quantile – coûts spécifiques – production agricole – RICA – test quadratique

1. INRAE, AgroParisTech, UMR PSAE, 91123, Palaiseau, France

Regional analysis of specific costs in agricultural production: using the Wasserstein distance

Dominique Desbois[1]

CORRESPONDENCE
dominique.desbois@inrae.fr

ABSTRACT

In the context of the forthcoming reform of the Common Agricultural Policy, analysing agricultural production costs is a crucial step in developing fair pricing policies, particularly at regional level, to avoid potential distortions of competition. This article presents the application of Wasserstein distance-based statistical techniques to the regional analysis of specific costs in agricultural production. Taking into account the heterogeneity and asymmetry of the distributions, it combines the conditional quantile estimation methodology with those based on the Wasserstein distance (factorial analysis of the distance matrix, unsupervised divisive clustering, quadratic test). The procedure is applied to the comparative regional analysis of the distributions of specific fertiliser costs for crop production, based on the European Agricultural Accounting Network, at the level of the European basic regions (NUTS2). The results obtained for specific input costs and associated margins appear consistent with those of previous work carried out under the European Union's 7th Framework Programme for Research.

KEYWORDS

factor analysis of the distance matrix – unsupervised divisive clustering – Wasserstein distance – quantile estimation – specific costs – agricultural production – FADN – quadratic test

1. INRAE, AgroParisTech, UMR PSAE, 91123, Palaiseau, France

1. Introduction

Face à l'hétérogénéité des structures agricoles, la renationalisation de la Politique agricole commune et les spécialisations productives régionales en Europe, comment l'estimation des coûts spécifiques peut-elle apporter le maximum d'informations utiles pour optimiser la gestion des exploitations agricoles et adapter territorialement les instruments du soutien public ? Généralement, pour les petites et moyennes structures de production multiproduits, ces informations comptables sont agrégées au niveau de l'exploitation agricole et non pas imputées à chaque production, comme dans une comptabilité analytique disponible seulement pour les grandes structures et certaines commodités agricoles[1]. Pour la plupart des exploitations multiproduits, on est conduit à procéder alors par estimation suivant un modèle normatif, comme par exemple les normes de production d'ateliers laitiers ou céréaliers.

Pour s'affranchir de normes productives trop spécifiques, des solutions plus générales s'appuient sur des sources de données statistiques comme le RICA[2], pour estimer des coûts applicables à l'ensemble des exploitations agricoles (Bureau & Cyncynatus, 1991), s'inspirant de l'approche micro-économétrique de répartition des coûts de Divay et Meunier (1980) pour construire une matrice entrées-sorties au niveau national. Cependant, les estimateurs en moyenne généralement utilisés dans ces démarches statistiques masquent, dans un contexte hétérogène, certaines différences inter-structurelles. En outre, ils ont le défaut d'être sensibles à l'asymétrie ou aux valeurs extrêmes des distributions d'intérêt.

Afin de surmonter ces inconvénients, une solution est de générer pour chacun des principaux produits agricoles non seulement les estimations médianes de la distribution des coûts conditionnellement aux produits bruts, mais aussi les quantiles inférieurs ou supérieurs.

À cette fin, nous proposons une procédure globale en deux étapes pour repérer des différences de niveaux en termes de coût spécifiques pour chacun des quantiles de distribution (médiane, quartiles, déciles, centiles), déterminant des estimations plus précises pour chacune des sous-populations d'exploitations régionalement localisées. Dans une première étape, nous mobilisons la technique de régression quantile qui nous permet d'obtenir des estimations de quantiles de coûts spécifiques, conditionnellement aux produits bruts des productions des exploitations agricoles (Desbois *et al.*, 2017). Dans une seconde étape, afin de faire apparaître les différences et similitudes entre les nombreuses estimations régionales à comparer au niveau européen, nous calculons la distance de Wasserstein (Verde & Irpino, 2008) entre les distributions régionales des quantiles estimés, pour appliquer les procédures de statistiques adéquates à la description de ces similitudes et différences (analyse factorielle de tableaux de distances, classifications non-supervisées et test quadratique).

Cette procédure globale est appliquée à l'estimation du coût des fertilisants pour mettre en évidence les différences et les similitudes régionales entre distributions de coûts spécifiques en fertilisants, pour les productions végétales sur la base du Réseau européen d'information comptable agricole.

2. Méthodologie

Nous présentons d'abord la méthodologie d'estimation des coûts spécifiques, en particulier le modèle d'estimation quantile conditionnel des coûts spécifiques et ses procédures de test. Puis, pour présenter la méthodologie descriptive d'analyse comparative des distributions de quantiles empiriques, nous introduisons l'analyse comparative des distributions empiriques de quantiles basée sur la distance de Wasserstein (Verde & Irpino, 2008) entre distributions empiriques de quantiles.

2.1. Éstimation des coûts spécifiques

Nous présentons d'abord la méthodologie initiale d'estimation des coûts spécifiques, puis le modèle d'estimation quantile conditionnel des coûts spécifiques et les procédures de test de la régression quantile.

2.1.1. Le modèle classique d'estimation des coûts spécifiques de production

Conformément aux spécifications du modèle général d'estimation des coûts de production spécifiée dans (Desbois *et al.*, 2013), le total des charges spécifiques x_i, pour une exploitation agricole $i \epsilon I$, est lié aux p produits bruts Y_i^j pour $j= 1,...,p$ par le modèle de régression suivant :

$$x_i = \sum_{j=1}^{p} \left(\alpha_j \times Y_i^j\right) + \epsilon_i \text{ pour } \; i = 1, \cdots, n$$

où ϵ_i est un terme stochastique résiduel d'espérance nulle et à variance constante.

Les coefficients s'interprètent comme un coefficient unitaire d'allocation de charges spécifiques proportionnel au produit brut des productions d'une exploitation agricole multi-productive.

Les valeurs des coefficients $\hat{\alpha}_j$, pour $j= 1,...,p$, estimées par la régression des moindres carrés ordinaires (MCO), fournissent les espérances conditionnelles des taux unitaires de coûts spécifiques pour l'ensemble J des produits.

2.1.2. Le modèle quantile conditionnel d'estimation des coûts spécifiques

Proposée par Desbois *et al.* (2017), l'estimation des quantiles conditionnels des coûts spécifiques reprend, sur la base de la « régression quantile » (Koenker & Bassett, 1978), le modèle classique des coûts spécifiques prenant en compte l'hétérogénéité de la valeur totale x_i des charges spécifiques pour l'exploitation i modélisée linéairement par la somme pondérée des charges proportionnellement allouées à chacun des produits Y_i^j. Par analogie avec la régression MCO, la régression quantile des coûts spécifiques est formulée comme suit selon Cameron et Trivedi (2005) :

$$x = Y'\beta^{(q)} + u^{(q)} \text{ avec } \; u^{(q)} = Y'\delta \times \epsilon^{(q)} \text{ et } Y'\delta > 0$$

où $\beta^{(q)} = (\beta_1^{(q)}, \cdots, \beta_j^{(q)}, \cdots, \beta_p^{(q)})$ est le vecteur des coefficients des p produits bruts, et $\epsilon^{(q)}$ un vecteur d'aléas distribués identiquement et indépendamment, d'espérance nulle et de variance constante σ^2.

1. Commodité agricole : produit destiné à l'approvisionnement standard de l'industrie agroalimentaire, qui ne porte pas spécification de l'origine territoriale ou du signe de qualité (AB, AOP, etc)
2. Réseau d'Information Comptable Agricole : pour plus de précisions, cf. https://agreste.agriculture.gouv.fr/agreste-web/methodon/S-RICA/methodon/

Sous cette hypothèse, le q^e quantile conditionnel du coût de production x, conditionné par Y et par les paramètres β et δ, s'en déduit avec la forme analytique suivante :

$$\mu_q(x \vee Y, \beta, \delta) = Y'\beta + \delta \times F_\epsilon^{-1}(q)$$

où $F_\epsilon(q)$ est la fonction de répartition de $\epsilon^{(q)}$.

Suivant le modèle général présenté par Haultfoeuille et Givord (2014), le modèle peut être exprimé selon des spécifications correspondant à la nature de l'hétérogénéité résiduelle, soit :

$$x = Y'\beta^{(q)} + \epsilon^{(q)} \text{ avec } \quad \mu_q\left(\epsilon^{(q)} \vee Y\right) = 0$$

i) $x = Y'\beta + u$ avec $u = K\epsilon$ désigné comme le modèle linéaire de quantile conditionnel à pente homogène (résidus homoscédastiques) ou « modèle à translation simple » (TS) ;

ii) $\mu_q\ (x \vee Y, \beta, \delta) = Y'\beta + \delta \times F_\epsilon^{-1}\ (q)$, avec $Y'\delta > 0$, désigné comme le modèle linéaire de quantile conditionnel à pente hétérogène (résidus hétéroscédastiques) ou « modèle à translation-échelle » (TE).

Suivant le modèle pondéré d'estimation des quantiles conditionnels proposé par Koenker et Zhao (1994) introduisant la pondération des observations, $\Omega_i = \{\omega_i;\ i=1,\cdots,n\}$, l'estimation des paramètres du modèle est solution optimale du problème de minimisation de la fonction de perte pondérée :

$$\hat{\beta_\omega}(q) = \underset{\beta \in R^p}{arg_{min}}\{\sum_{i:x_i \geq y_i'\beta}[\omega_i q|x_i - y_i'\beta|] + \sum_{i:x_i < y_i'\beta}[\omega_i(1-q)|x_i - y_i'\beta|]\}$$

Les estimations pondérées des quantiles conditionnelles sont fournies par la procédure *QUANTREG* du logiciel *SAS*, version 9.2 (SAS Institute Inc., 2009).

Appliquée aux exploitations agricoles, cette méthode économétrique fournit une distribution estimée des coûts spécifiques pour les principaux produits agricoles, complétant ainsi l'espérance conditionnelle de ces coûts fournis par les estimations MCO. En outre, par le biais d'une procédure de ré-échantillonnage non-paramétrique (cf. infra § 2.1.3), le processus quantile fournit une distribution empirique des estimations sans formuler d'hypothèse inadaptée sur la distribution des charges.

2.1.3. Les procédures de test de la régression quantile

Pour estimer la variance de l'estimateur $\hat{\beta}\omega$, nous optons pour la technique du rééchantillonnage, en *bootstrap* marginal sur chaînes de Markov (*MCMB - Markov Chain Marginal Bootstrap*), sans hypothèse à priori sur les distributions d'erreurs, suivant He et Hu (2002). Afin d'obtenir une inférence fiable, nous appliquons la règle de Kocherginsky *et al.* (2005), soit $n.min\{[q,1-q]\} > 5p$ pour la taille des échantillons bootstrappés du quantile q.

L'hypothèse d'hétéroscédasticité est testée selon une procédure de rejet de l'hypothèse nulle d'égalité des coefficients $\beta^{(q)}$ entre k différents quantiles selon une statistique suivant une distribution du Chi-Deux à $(k-1) \times p$ degrés de liberté, d'après Koenker et Bassett (1982).

Sur la base du processus empirique suivant pour l'estimateur $\hat{\beta}\omega$:

$$v_n(\tau) = \sqrt{n} \cdot \left(\hat{\beta_n}(\tau) - \beta_0(\tau)\right) \text{ où } \tau \in [\varepsilon, 1-\varepsilon]$$

les tests de Khmaladze (1982) pour l'hypothèse du modèle à pente homogène (TS) et pour le modèle à pente hétérogène (TE) sont mis en œuvre par la procédure *KhmaladzeTest* du package R *quantreg*, d'après Koenker et Xiao (2002). Une combinaison de bande passante de Bofinger (1975) et de bande passante de Hall et Sheater (1988) est utilisée pour le test TS et pour le test TE, suivant Koenker et Xiao (2002).

La significativité des coefficients est testée suivant une statistique de Wald distribuée sous l'hypothèse de nullité selon une loi du Chi-Deux dont le nombre de degrés de libertés est fonction du nombre *p* de coefficients testés, selon Koenker et Machado (1999).

Au terme de ce processus d'estimation, les distributions de paramètres sont dites « totalement estimées » si toutes les estimations obtenues pour les *p* paramètres quantiles peuvent être testées comme significativement non-nulles ou pas sur la base de leur variance estimée. Dans l'éventualité où ce test ne peut être conduit en raison de variances non-estimables de certains paramètres pour des combinaisons particulières de régions et de produits, les distributions de paramètres sont dites « partiellement estimées » .

2.2. Analyse comparative des distributions empiriques de quantiles

Définissant la distance de Wasserstein entre distributions empiriques de quantiles conditionnels dans le formalisme de l'analyse symbolique de données (Billard & Diday, 2007), nous proposons comme outils d'analyse comparative l'analyse factorielle sur tableau de distances de Wasserstein et des procédures de classification non supervisée sur distance de Wasserstein au niveau régional. Nous présentons également un test de Wasserstein associé à une procédure graphique pour expliciter les écarts entre distributions empiriques de quantiles.

2.2.1. Distance de Wasserstein entre distributions de quantiles empiriques

Soit $\Delta = \{\partial_1, \cdots, \partial_l \cdots \partial_L\}$, l'ensemble des distributions empiriques régionales de coûts spécifiques estimés, décrites par un ensemble $\hat{\mu}_p$ de p estimations quantiles conditionnelles :

$$\hat{\mu}_p = \left\{\hat{\mu}_1{}^{cs}, \ldots, \hat{\mu}_j{}^{cs} \ldots, \hat{\mu}_p{}^{cs}\right\}$$

La distance quadratique (i.e. d'ordre $k=2$) de Wasserstein entre la distribution région l et la région l', basée sur les intervalles d'estimation des coefficients techniques pour le q^e quantile conditionnel, sont calculées selon la formule suivante :

$$d_W^2(\delta_l, \delta_{l'})^2 = \frac{1}{M}\sum_{m=1}^{M} d^2\left(I_l^m, I_{l'}^m\right)$$

où M est le nombre d'intervalles inter-quantiles de la distribution empirique, m l'indice de l'intervalle inter-quantile entre les quantiles q et q', et $d(I_l^m, I_{l'}{}^m)$, la distance entre les m^e intervalles inter-quantiles homologues des deux distributions régionales δ_l et $\delta_{l'}$, définie par :

$$d_W^2\left(I_l^m, I_{l'}^m\right) = \left[\left(c_l^m - c_{l'}^m\right)^2 + \frac{\left(r_l^m - r_{l'}^m\right)^2}{3}\right]$$

où $c_l^m = \left(\underline{\gamma_l^m} + \overline{\gamma_l^m}\right)/2$ est le centre,

$r_l^m = \left(\underline{\gamma_l^m} - \overline{\gamma_l^m}\right)/2$ le rayon,

$\underline{\gamma_l^m} = \underline{\hat{\beta}_\omega^q}$ l'infimum,

$\overline{\gamma_l^m} = \overline{\hat{\beta}_\omega^q}$ le supremum,

de l'intervalle inter-quantiles $I_l^m = \left[\underline{\gamma_l^m}, \overline{\gamma_l^m}\right]$, de modalité *m* entre les quantiles q et q'.

Notons que la distance quadratique de Wasserstein suit l'inégalité triangulaire pour les régions *l*, *l'*, et *l''* :

$d_W(\delta_l,\delta_{l'}) \leq d_W(\delta_l,\delta_{l''}) + d_W(\delta_{l''},\delta_{l'})$ quelque soient les distributions δ_l, $\delta_{l'}$ et $\delta_{l''}$, comme l'ont démontré Clement et Desch (2007) quel que soit l'ordre *k* de la distance de Wasserstein.

Selon la théorie du transport optimal (Bogachev & Kolesnikov, 2012), la distance de Wasserstein constitue la valeur minimale du coût de transport entre deux mesures de probabilité, utilisant comme coût élémentaire la distance entre les points représentant les quantiles des distributions. Économiquement, elle peut s'interpréter comme le coût minimal de l'effort à consentir pour passer d'une distribution d'intrants à une autre, en l'espèce le coût minimal théorique d'une transition entre deux techniques culturales.

2.2.2. Analyse factorielle du tableau des distances de Wasserstein

L'analyse factorielle d'un tableau de distances (AFTD) permet de générer une projection multidimensionnelle à partir d'un ensemble de distances mutuelles entre différents taxons. Le principe de cette méthode est d'extraire les valeurs propres et vecteurs propres d'une matrice dérivée de la matrice des distances, afin de déterminer une projection optimale de l'ensemble d'objets dans un sous-espace de dimension fixée. Intitulée « analyse du triple » par Benzécri *et al.* (1973), cette méthode est équivalente au positionnement multidimensionnel classique proposé par Torgerson (1958). Combinant les projections des variables et observations sur les deux premières dimensions factorielles, le *biplot* (Gabriel, 1971) constitue le graphique utilisé pour visualiser la répartition des distributions quantiles.

L'AFTD de Wasserstein entre distributions quantiles estimées de coûts spécifiques est réalisée selon la méthode à l'aide de la procédure *cmdscale* de *R* (positionnement multidimensionnel classique).

2.2.3. Classifications sur distance de Wasserstein et validation des hiérarchies et partitions

Les méthodes de classification automatique à partir du tableau des distances quadratiques de Wasserstein entre distributions empiriques de quantiles conditionnels de coûts sont des adaptations de l'algorithme des nuées dynamiques initialement proposé par Diday (1971), de la classification ascendante hiérarchique (CAH) selon le critère proposé par Ward (1963) et de l'algorithme de classification divisive (ACD) proposé par Chavent *et al.* (2007). Selon Irpino *et al.* (2006) puis Verde et Irpino (2008), le critère de Ward est cohérent et optimal pour la distance de Wasserstein en tant que métrique euclidienne sur les intervalles munis de la mesure uniforme. L'ACD comme l'algorithme des nuées dynamiques sont optimaux localement, au niveau des partitions retenues.

La procédure *clusterCrit* proposée par Desgraupes (2017) du logiciel R est utilisée pour calculer le logarithme du rapport de la variance totale à la somme des variances intraclasses, choisi comme critère de validation des partitions *P*, obtenues par classifications non-supervisées.

2.2.4. Test de Wasserstein et graphique de comparaison entre distributions

Le test de Wasserstein compare deux distributions empiriques en utilisant la distance de Wasserstein d'ordre k, W_k, entre deux fonctions de répartition E et F:

$$w_k(E,F) = \left(\int_{x\in R} |E(x) - F(x)|^k dx\right)^{1/k}$$

Le calcul de la *p-valeur* est basé sur l'inégalité de Dvoretzky *et al.* (1956) qui fournit une limite maximale à la distance entre une fonction de distribution empirique G_n d'un échantillon de taille *n* par rapport à la fonction de distribution théorique G de la population

$$Pr(_{x\in R}|G_n(x) - G(x)| > \varepsilon) \leq Ce^{-2n\varepsilon^2}, \forall\varepsilon$$

Confirmant la conjecture de Birnbaum et McCarty (1958), Massart (1990) a prouvé cette inégalité pour la constante $C=2$.

Contrairement aux tests ne requérant que des données ponctuelles, le test quadratique ($k=2$) de Wasserstein tire parti des hypothèses distributionnelles sur les intervalles inter-quantiles (e.g. distribution uniforme).

La procédure *cdfCompare* de *R*, est utilisée pour comparer empiriquement les fonctions de répartition associées aux distributions quantiles estimées.

3. Application aux pays et régions de l'Union européenne

La méthodologie d'estimation micro-économétrique des coûts de production spécifiques en fonction des quantiles conditionnels est appliquée aux exploitations de l'échantillon 2006 (n = 56 180) du Réseau européen d'information comptable agricole (RICA), afin de démontrer sa faisabilité et sa pertinence au niveau régional (niveau II) de la nomenclature statistique européenne des unités territoriales (NUTS) pour douze pays de l'UE, sur la base constituée dans le cadre du programme *Farm Accountancy Cost Estimation and Policy Analysis of European Agriculture* (*FACEPA*) du 7e Programme Cadre de la Communauté européenne.

Dans une première étape, l'analyse comparative des distributions de quantiles estimés est effectuée pour quinze produits (blé, autres céréales, légumineuses sèches, cultures industrielles, oléagineux, horticulture, fruits, vin, autres cultures, bovins, porcs, œufs et volailles, lait de vache, autres animaux d'élevage, autres activités) au niveau des douze pays européens, principaux producteurs agricoles et de l'année 2006, choisis comme références au niveau du programme *FACEPA*.

Puis, l'analyse comparative des distributions empiriques de quantiles conditionnels de coûts est également menée au niveau de l'ensemble des régions européennes (NUTS II), totalement estimées sur ces quinze produits pour 38 régions et partiellement estimées pour 28 régions.

Enfin, nous mobilisons également les estimations conditionnelles de quantiles sur cinq classes de productions (cultures annuelles, cultures permanentes, productions animales sur sol, productions animales hors-sol, autres productions) pour 98 régions NUTS II.

Aux fins de comparaison avec des travaux antérieurs (Desbois, 2023), nous retenons pour cette application empirique de la distance de Wasserstein aux distributions de coûts, les estimations de coûts spécifiques en fertilisants des productions végétales (cultures annuelles et cultures permanentes) pour les régions européennes et du blé pour les pays.

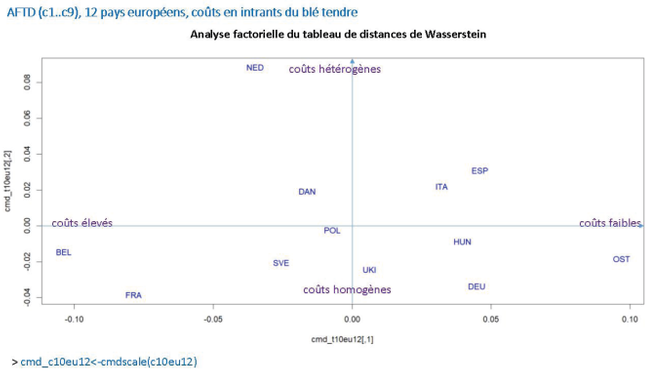


Figure 1. AFTD des distributions de déciles estimés de 12 pays européens

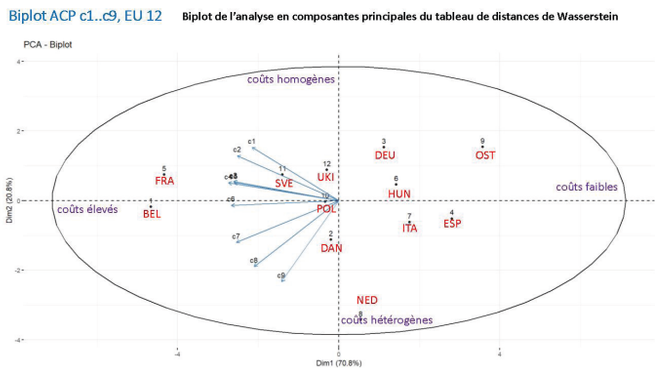


Figure 2. Biplot des distributions de déciles estimés de 12 pays européens

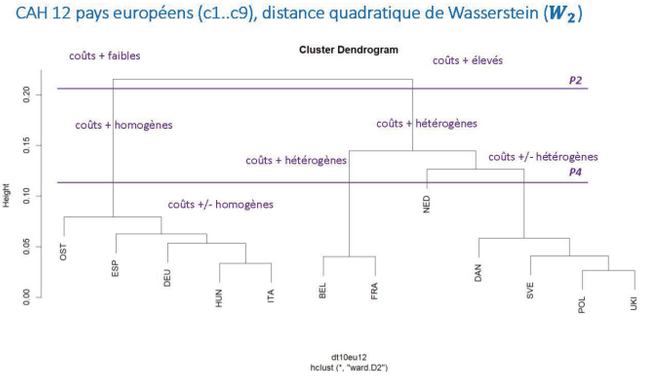


Figure 3. Dendrogramme de la CAH sur tableau des distances de Wasserstein entre les distributions de déciles estimés pour 12 pays européens

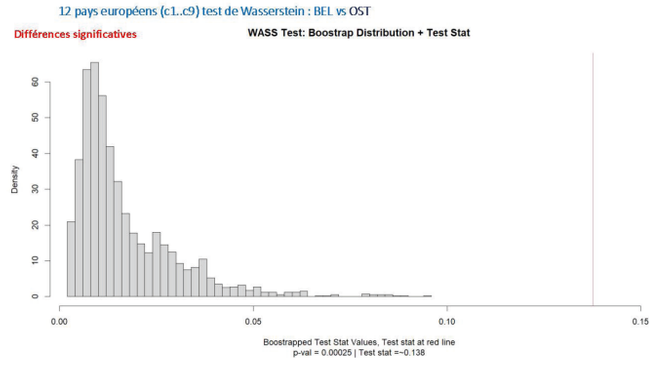


Figure 4. Test quadratique de Wasserstein entre la distribution des déciles de la Belgique et celle de l'Autriche

3.1. Analyse comparative des distances distributionnelles entre pays

L'analyse factorielle du tableau des distances (AFTD) de Wasserstein (figure 1) entre les distributions empiriques d'estimations quantiles conditionnelles des coûts du blé en intrants illustre une structure en deux dimensions, avec un axe horizontal (*F1*) opposant les coûts unitaires faibles (*F1>0*), comme la distribution de l'Autriche (*OST*) pour les plus faibles, aux coûts unitaires élevés (*F1<0*), comme la distribution de la Belgique (*BEL*) ou celle de la France (*FRA*) pour les coûts unitaires les plus élevés. L'axe vertical (*F2*) de cette image euclidienne de la matrice de distances de Wasserstein oppose les distributions de déciles les plus hétérogènes (*F2>0*), comme celle des Pays-Bas (*NED*) représentatifs du modèle à translation- échelle (TE), aux distributions de déciles les plus homogènes (*F2<0*), comme celle du Royaume-Uni (*UKI*) ou celle de l'Allemagne (*DEU*), représentatifs du modèle à translation simple (TS).

Cette image euclidienne est très proche (corrélation élevée des axes, au changement de signe près pour la seconde composante principale) de celle donnée par le *biplot* de l'ACP du tableau des déciles des distributions empiriques pour les douze pays européens (figure 2) dont l'interprétation est très similaire en termes de niveaux (composante principale *Dim1*) et d'homogénéité/hétérogénéité (composante principale *Dim2*).

3.2. Classifications non-supervisées des distributions nationales de coûts

La classification ascendante hiérarchique (CAH) basée sur la matrice de distances de Wasserstein produit une hiérarchie de partitions (figure 3, dendrogramme) que nous étiquetons en fonction des résultats aux tests de Wasserstein (figure 4), confirmés par les comparaisons graphiques entre fonctions de répartition (figure 5). Ainsi, le test quadratique de Wasserstein de l'écart entre la distribution des déciles de la Belgique et celle de l'Autriche montre que leurs différences sont statistiquement significatives au risque de première espèce *p=0,025 %* pour la statistique de test *DW=0,138* (figure 4). La comparaison graphique entre les fonctions de répartition correspondantes des coûts unitaires montre que celle de l'Autriche présente des niveaux de déciles estimés plus faibles et plus homogènes que celle de la Belgique.

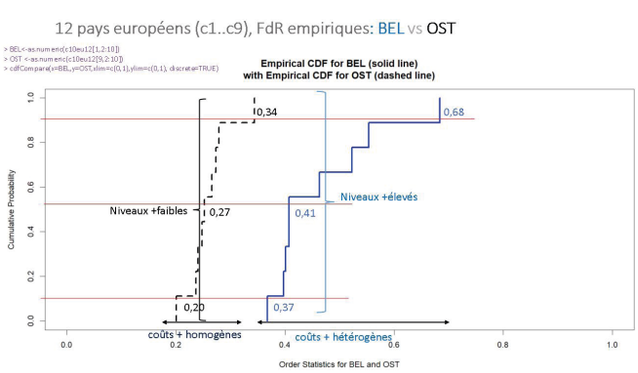


Figure 5. Comparaison graphique entre la distribution des déciles de la Belgique et celle de l'Autriche

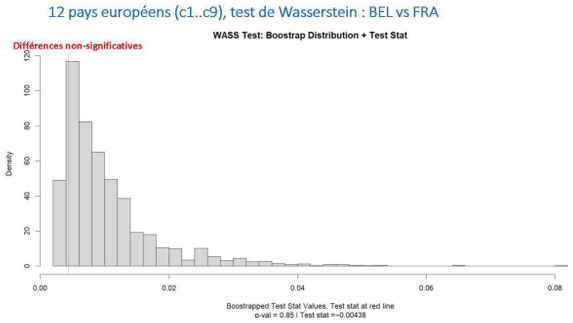


Figure 6. Test quadratique de Wasserstein entre la distribution des déciles de la Belgique et celle de la France

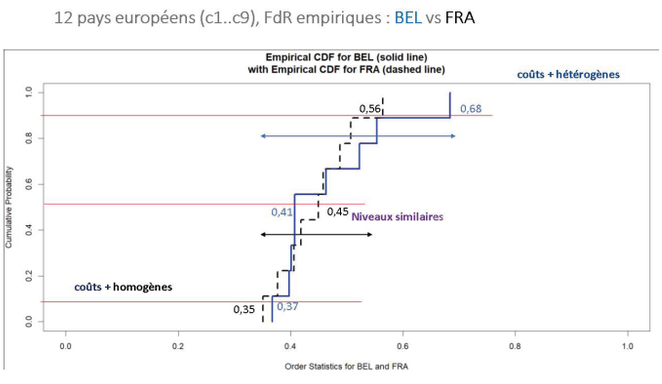


Figure 7. Comparaison graphique entre les fonctions de répartition des déciles estimés de coûts unitaires, Belgique versus France

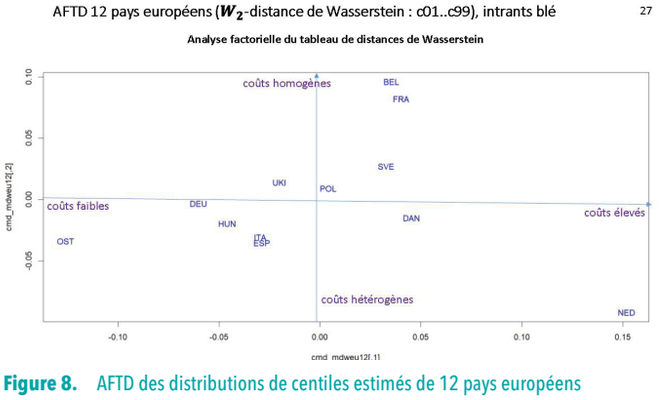


Figure 8. AFTD des distributions de centiles estimés de 12 pays européens

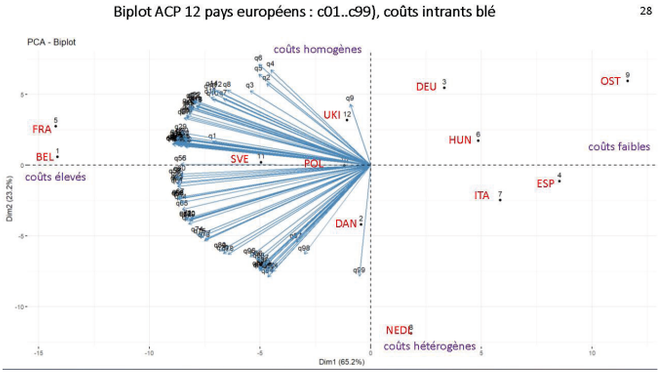


Figure 9. Biplot de l'ACP du tableau des centiles estimés de 12 pays européens

Observons que la partition *P2* en deux classes de la CAH oppose les classes de coûts élevées aux classes à faible coûts, tandis que l'opposition entre classes de coûts homogènes (modèle TS) versus classes de coûts hétérogènes (modèle TE) concerne les partitions de rang supérieur (*P4* et supra).

Les comparaisons effectuées avec le même critère d'agrégation (Ward, 1963) pour la distance euclidienne classique ne révèlent que des différences d'ordre d'agrégation aux nœuds inférieurs de la hiérarchie sans remettre en cause la composition des classes au niveau des partitions en deux classes (*P2*), en trois classes (*P3*), ou en quatre classes (*P4*).

Inversement, le test quadratique de Wasserstein ne permet pas selon la statistique *DW=0,00438* de rejeter l'hypothèse d'égalité entre les distributions estimées de quantiles de coûts unitaires de la Belgique et de la France, le risque de première espèce correspondant (*p=85 %*) étant trop élevé (figure 6), ce que tend à confirmer la comparaison graphique entre les fonctions de répartition des déciles estimés de coûts unitaires en Belgique et en France (figure 7).

L'AFTD de Wasserstein entre les distributions estimées de centiles de coûts unitaires pour les douze pays européens retenus dans cette analyse montre une image euclidienne (figure 8) structurellement similaire à l'AFTD de Wasserstein des distributions déciles (figure 1), malgré l'inversion de signe des deux premiers axes factoriels assimilable à un artefact numérique. Les positions relatives des pays sont globalement conservées à deux exceptions : ainsi, l'ordre du premier axe factoriel est inversé entre les Pays-Bas (*NED*) aux coûts les plus élevés et le sous-groupe {Belgique (*BEL*), France (*FRA*)} se rapprochant du coût médian ; les coordonnées de l'Allemagne (*DEU*) et de la Hongrie (*HUN*) sur le second axe factoriel changent de signe, se rapprochant des distributions hétérogènes. Ces inversions sont attribuables aux informations plus détaillées apportées par les centiles par rapport aux déciles.

La structure *biplot* de l'ACP du tableau des distributions centiles estimées (figure 9) apparaît globalement cohérente avec l'AFDT de Wasserstein des distributions centiles (figure 8). Notons cependant la nette transition des Pays-Bas (NED) vers les coûts médians sur la première composante. Hormis cette transition attribuable aux valeurs extrêmes des Pays-Bas, l'ordre relatif des pays sur les deux composantes principales (figure 9) apparaît globalement similaire à celui observé sur les deux premiers axes factoriels (figure 8). Notons également l'inversion de l'ordre des projections des premiers centiles estimés (e.g. *q3* versus *q9*) sur le biplot de l'ACP due à des valeurs extrêmes.

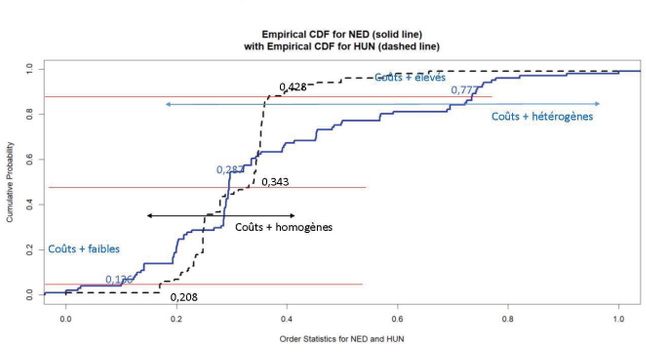


Figure 10. Comparaison graphique entre les distributions quantiles des Pays-Bas et de la Hongrie

Cependant, l'interprétation globale des deux axes factoriels de l'AFTD est conservée en passant des déciles aux centiles. Par exemple, l'opposition en figure 10 sur l'axe vertical entre les Pays-Bas (*NED*) et la Hongrie (*HUN*) est significative selon le test quadratique de Wassertstein (*DW= 0,0174* et *p=0,025 %*), la Hongrie présentant une distribution de coûts plus homogène (modèle TS) que les Pays-Bas dont les estimations de coûts unitaires sont très hétérogènes (modèle TE).

Les analyses effectuées sur la base de la distance de Wasserstein confirment les résultats obtenus par des analyses selon la distance euclidienne effectuées dans (Desbois, 2023) : la confrontation de ces deux types d'analyse confirme que le contexte national des douze pays producteurs étudiés demeure un facteur de différentiation principal pour les coûts spécifiques, des coûts faibles aux coûts élevés dans la dimension horizontale (*Dim1*), tandis qu'au sein des coûts moins extrêmes l'hétérogénéité de forme des distributions, coûts homogènes (modèle TS) versus coûts hétérogènes (modèle TE), caractérise la dimension verticale (*Dim2*) de différentiation secondaire.

L'algorithme de classification divisive sur la matrice des distances de Wasserstein DivClustW pour les douze pays européens (figure 11) permet de préciser les seuils de quantiles de coûts unitaires responsables de la division en partitions basées sur les centiles de coûts unitaires du blé en intrants. Ainsi, la partition en deux classes (P2) est-elle expliquée par un seuil de coûts en intrants *S1* de 35 % du produit brut pour le 37e centile : d'une part, la Suède (*SVE*), la Belgique (*BEL*) et la France (*FRA*) présentent un 37e centile de coûts unitaires supérieurs à 35 % du produit brut, d'autre part, l'Espagne (*ESP*), l'Autriche (*OST*), l'Italie (*ITA*), la Hongrie (*HUN*), l'Allemagne (*DEU*), le Royaume-Uni (*UKI*), la Pologne (*POL*), le Danemark (*DAN*) et les Pays-Bas (*NED*) présentent un 37e centile de coûts unitaires inférieur à 35 % du produit brut. Cette partition *P2* traduit l'opposition sur la première composante principale (*Dim1*) entre pays à coûts élevés (*Dim1<0*) et pays à coûts faibles (*Dim1>0*), illustrée par le *biplot* de l'ACP du tableau des centiles (figure 9). De même, la partition P3 recouvre l'opposition entre la classe hétérogène {Pays-Bas, Danemark} (modèle TE) et les classes plus homogènes (modèle TS) que sont {Suède, Belgique, France} et {Espagne, Autriche, Italie, Hongrie, Allemagne, Royaume-Uni, Pologne}.

3.3. Analyse comparative distributionnelle entre régions européennes

L'analyse factorielle du tableau des distances de Wasserstein entre les distributions estimées de centiles de coûts unitaires pour les 38 régions européennes complètement estimées sur les quinze produits retenus pour l'estimation quantile conditionnelle (figure 12) met en évidence une dimension horizontale (*F1*) liée à l'opposition entre les distributions à niveaux de coûts unitaires élevés (*F1>0*), comme pour l'Alsace, et les distributions à niveau de coûts unitaires faibles (*F1<0*), comme pour la Saxe-Anhalt. Dans la dimension verticale (*F2*) s'opposent les distributions à niveaux de coûts homogènes (*F2<0*, modèle TS), comme pour la Lorraine, aux distributions à niveaux de coûts hétérogènes (*F2>0*, modèle TE), comme pour l'Auvergne.

Notons qu'à l'exception de la Picardie, l'ensemble des régions françaises (Nord Pas de Calais, Lorraine, Champagne-Ardenne, Auvergne, Bourgogne, Basse-Normandie, Franche-Comté, Rhône-Alpes, Alsace) sont projetées dans la zone des coûts élevés (*F1>0*). Cette structure

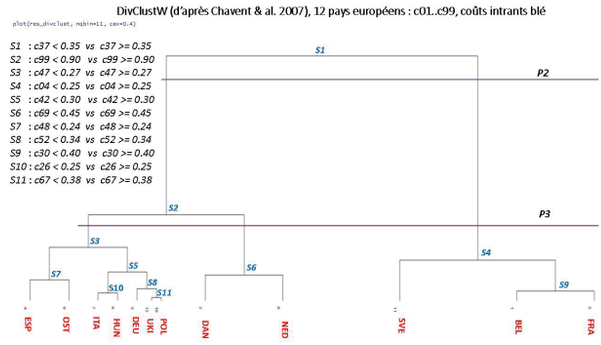


Figure 11. Classification divisive sur distances de Wasserstein (DivClustW) entre 12 pays européens, estimées sur cinq groupes de produits

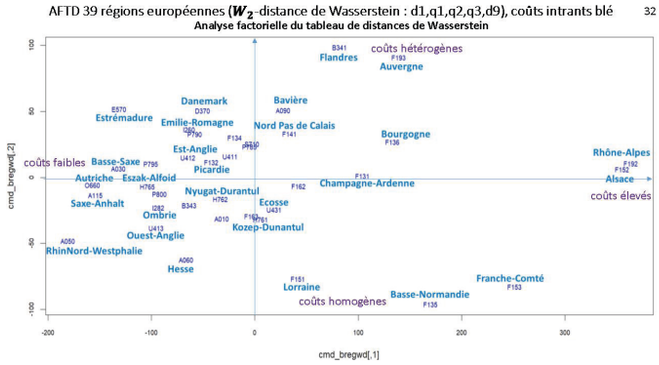


Figure 12. AFTD des distances de Wasserstein entre 29 régions européennes, estimées sur quinze produits

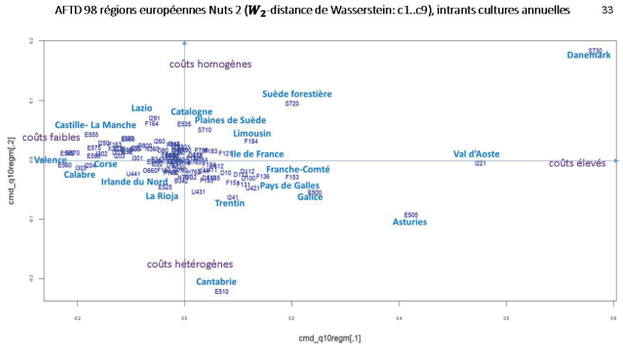


Figure 13. AFTD des distances de Wasserstein entre 98 régions européennes, estimées sur cinq groupes de produits

explique la position de la distribution nationale (*FRA*) sur le premier axe factoriel de la figure 1, dans la zone des coûts élevés, comme barycentre pondéré des distributions régionales françaises. Inversement, l'ensemble des régions allemandes (RhinNord-Westphalie, Saxe-Anhalt, Basse-Saxe, Hesse) sont projetées en zone des coûts faibles (*F1<0*), à l'exception de la Bavière (*F1>0*). Ceci induit la projection de l'Allemagne (*DEU*) dans la zone des coûts faibles de la figure 1, en tant que barycentre pondéré des régions allemandes.

On retrouve une structure similaire pour l'AFTD des distances de Wasserstein entre les 98 distributions régionales de coûts unitaires estimés à partir de cinq groupes de produits (figure 13). La première dimension horizontale (*F1*) oppose les coûts faibles (*F1<0*), comme pour la province de Valence, aux coûts élevés (*F1>0*), comme pour le Val d'Aoste. La seconde dimension verticale (*F2*) oppose les coûts homogènes (*F2<0*, modèle TS), comme pour la Catalogne, aux coûts hétérogènes (*F2>0*, modèle TE), comme pour la Cantabrie.

Par contre, l'agrégation des paramètres, passant de quinze produits à cinq groupes de produits, élargit notablement l'échantillon des exploitations agricoles prises en compte pour l'estimation quantile, modifiant les estimations quantiles de coûts spécifiques et partant le tableau des proximités relatives entre régions. En effet, l'extension à l'ensemble des régions européennes modifie notablement les échelles de coûts spécifiques, en raison de la présence des régions méditerranéennes introduisant davantage d'hétérogénéité de par leur plus grande différenciation productive en termes de systèmes de culture et d'intensification.

Dès lors, une comparaison avec les résultats précédents n'est plus pertinente : le positionnement bi-dimensionnel des régions de l'AFTD (figure 13) s'effectue par rapport à un nouveau barycentre, celui de l'ensemble des régions européennes. Toutefois, notons que l'interprétation structurelle des axes factoriels en termes de niveau de coût (axe horizontal) et de gradient d'hétérogénéité (axe vertical) demeure.

Ainsi, les différences enregistrées entre les estimations régionalisées des quantiles conditionnels de coûts en fertilisants pour les productions végétales montrent qu'il n'y a pas « un » coût spécifique de production qui pourrait être estimé à l'échelle nationale par une moyenne conditionnelle mais des classes régionales de distribution des coûts spécifiques.

En outre, les classes régionales, associées soit au modèle homogène (TS) soit au modèle hétérogène (TE), s'inscrivent dans un schéma bidimensionnel stable au regard du nombre choisi d'estimations quantiles conditionnelles. Ce schéma bidimensionnel croisant niveau et hétérogénéité des estimations de coût spécifique apparaît compatible avec les résultats obtenus pour le blé dans des analyses antérieures (Desbois, 2023).

4. Conclusions

Ce travail met en évidence la pertinence d'outils d'analyse comparative basés sur la distance distributionnelle de Wasserstein pour analyser différences et similitudes entre les distributions empiriques d'estimations quantiles conditionnelles de coûts unitaires pour les productions agricoles végétales, que ce soit au niveau des pays européens ou de leurs régions NUTS II.

Comme instrument spécifique d'analyse comparative, nous incorporons la distance de Wasserstein entre distributions aux analyses statistiques exploratoires suivantes : l'analyse factorielle du tableau des distances de Wasserstein (AFTDW), la classification ascendante hiérarchique sur distances de Wasserstein (CAHDW), la classification divisive sur distances de Wasserstein (DivClustW), et le test quadratique de Wasserstein (TQW) d'égalité entre distributions empiriques.

Les analyses comparatives effectuées sur les coûts spécifiques en intrants à partir du Réseau européen d'information comptable agricole, présentent des conclusions robustes au nombre de quantiles estimés : dans leur dimension principale de variabilité, les distributions aux échelles de coûts faibles sont opposées à celles aux coûts élevés ; dans leur dimension secondaire, parmi les distributions de coûts aux échelles équivalentes, les distributions homogènes en coûts (modèle TS) s'opposent aux distributions hétérogènes en coûts (modèle TE).

La différenciation des distributions nationales et régionales selon ce schéma bidimensionnel de coûts en intrants apparaît cohérente avec celle observée dans les travaux antérieurement menés dans le cadre du programme *FACEPA* (Desbois *et al.*, 2017). La prégnance de ce schéma bidimensionnel, tant au niveau régional que national, invite à en examiner les conséquences pour les exploitations agricoles, tant en termes d'approvisionnement que d'impact environnemental.

Nous envisageons d'étendre ce type d'estimation quantile et d'analyse comparative des distributions empiriques de coûts spécifiques aux systèmes de catégories structurelles décrivant l'espace européen des exploitations agricoles : d'une part, les classes de dimension économique et, d'autre part, les orientations technico-économiques.■

Remerciements

Seul responsable d'éventuelles erreurs ou omissions, l'auteur remercie les professeures Rosana Verde et Paula Brito pour leurs commentaires au sujet de l'inégalité triangulaire pour la distance de Wasserstein, lors des Rencontres 2024 de la Société francophone de Classification à Marseille.

Crédits

Cette étude a été en partie financée par le projet ANR CLAND (16-CONV-0003) et par le 7e Programme-Cadre de la Communauté européenne (FP7/2007-2013) sous l'agrément n° 212292 pour le projet Farm Accountancy Cost Estimation and Policy Analysis of European Agriculture (FACEPA).

In Memoriam

L'auteur dédie ce travail à la mémoire d'une part d'Yves Surry, professeur émérite de la Swedish University of Agricultural Sciences, son co-directeur de thèse décédé en 2021, et d'autre part d'Edwin Diday, professeur émérite de l'Université Paris-Dauphine, décédé en 2023, dont les remarques et suggestions ont nourri cette recherche.

Références

Benzécri, J.-P. et al. (1973). *L'analyse des données. Tome 1: La taxinomie. Tome 2: L'analyse des correspondances.* Dunod.

Billard, L., & Diday, E. (2007). Symbolic Data Analysis : Conceptual Statistics and Data Mining (Wiley Series in Computational Statistics). Dans *John Wiley & Sons, Inc. eBooks*. http://dl.acm.org/citation.cfm?id=1206598

Birnbaum, Z. W., & McCarty, R. C. (1958). A Distribution-Free Upper Confidence Bound for $ \Pr \{Y. *The Annals Of Mathematical Statistics, 29*(2), 558-562. https://doi.org/10.1214/aoms/1177706631

Bofinger, E. (1975). Non-Parametric Estimation of Density for Regularly Varying *Distributions. Australian Journal Of Statistics, 17*(3), 192-195. https://doi.org/10.1111/j.1467-842x.1975.tb00957.x

Bogachev, V. I., & Kolesnikov, A. V. (2012). The Monge–Kantorovich problem: Achievements, connections, and perspectives. *Russian Mathematical Surveys, 67*(5), 785-890. https://doi.org/10.1070/RM2012v067n05ABEH004837

Bureau, J.-C., & Cyncynatus, M. (1991). *Estimation de coûts de production et de coefficients input-output à partir de données comptables: Méthodes et applications aux produits agricoles sur la base du RICA* (Notes et Documents, n° 38). INRA, Département Économie et Sociologie Rurales.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chavent, M., Lechevallier, Y., & Briant, O. (2007). DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, *52*(2), 687-701. https://doi.org/10.1016/j.csda.2007.03.013

Clement, P., & Desch, W. (2007). An elementary proof of the triangle inequality for the Wasserstein metric. *Proceedings Of The American Mathematical Society, 136*(1), 333-339. https://doi.org/10.1090/s0002-9939-07-09020-x

Desbois, D. (2023). Coûts spécifiques et marges brutes du blé en Europe : Pratiques innovantes d'estimations pour l'analyse des changements d'échelles régionaux ou structurels. *NOV'AE*, (8), 1-23. https://doi.org/10.17180/novae-2023-NO-art08

Desbois, D., Butault, J., & Surry, Y. (2013). Estimation des coûts de production en phytosanitaires pour les grandes cultures. Une approche par la régression quantile. *Économie Rurale, 333*, 27-49. https://doi.org/10.4000/economierurale.3857

Desbois, D., Butault, J., & Surry, Y. (2017). Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la régression quantile. *Économie Rurale, 361*, 3-22. https://doi.org/10.4000/economierurale.5320

Desgraupes, B. (2017). *Clustering indices* [Vignette R]. CRAN. https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

Diday, E. (1971). *Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques*. https://www.numdam.org/item/RSA_1971__19_2_19_0/

Divay, J.-F., & Meunier, F. (1980). Deux méthodes de confection du tableau entrées-sorties. *Annales de l'Insee, 37*, 59-108.

Dvoretzky, A., Kiefer, J., & Wolfowitz, J. (1956). Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals Of Mathematical Statistics, 27*(3), 642-669. https://doi.org/10.1214/aoms/1177728174

Gabriel, K. R. (1971). The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika, 58*(3), 453. https://doi.org/10.2307/2334381

Hall, P., & Sheather, S. J. (1988). On the Distribution of a Studentized Quantile. *Journal Of The Royal Statistical Society Series B (Statistical Methodology), 50*(3), 381-391. https://doi.org/10.1111/j.2517-6161.1988.tb01735.x

Haultfoeuille, X. d', & Givord, P. (2014). La régression quantile en pratique. *Economie et Statistique / Economics And Statistics, 471*(1), 85-111. https://doi.org/10.3406/estat.2014.10484

He, X., & Hu, F. (2002). Markov Chain Marginal Bootstrap. *Journal Of The American Statistical Association, 97*(459), 783-795. https://doi.org/10.1198/016214502388618591

Irpino, A., Verde, R., & Lechevallier, Y. (2006). Dynamic clustering of histograms using Wasserstein metric. In A. Rizzi & M. Vichi (Eds.), *COMPSTAT 2006: Proceedings in computational statistics* (pp. 869-876). Physica-Verlag.

Khmaladze, E. V. (1982). Martingale Approach in the Theory of Goodness-of-Fit Tests. *Theory Of Probability And Its Applications, 26*(2), 240-257. https://doi.org/10.1137/1126027

Kocherginsky, M., He, X., & Mu, Y. (2005). Practical Confidence Intervals for Regression Quantiles. *Journal Of Computational And Graphical Statistics, 14*(1), 41-55. https://doi.org/10.1198/106186005x27563

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33-50. https://doi.org/10.2307/1913643

Koenker, R., & Bassett, G. (1982). Robust Tests for Heteroscedasticity Based on Regression Quantiles. *Econometrica, 50*(1), 43-61. https://doi.org/10.2307/1912528

Koenker, R., & Machado, J. A. F. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal Of The American Statistical Association, 94*(448), 1296-1310. https://doi.org/10.1080/01621459.1999.10473882

Koenker, R., & Xiao, Z. (2002). Inference on the Quantile Regression Process. *Econometrica, 70*(4), 1583-1612. https://doi.org/10.1111/1468-0262.00342

Koenker, R., & Zhao, Q. (1994). L -estimation for linear heteroscedastic models. *Journal Of Nonparametric Statistics, 3*(3 4), 223-235. https://doi.org/10.1080/10485259408832584$

Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals Of Probability, 18*(3), 1269-1283. https://doi.org/10.1214/aop/1176990746

SAS Institute Inc. (2009). The QUANTREG procedure. In *SAS/STAT 9.2 User's Guide* (2nd ed., Chapter 72, pp. 5352-5425). SAS Institute Inc.

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.

Verde, R., & Irpino, A. (2008). Comparing histogram data using a Mahalanobis–Wasserstein distance. In *COMPSTAT 2008: Proceedings in computational statistics* (pp. 77-89). Physica-Verlag.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal Of The American Statistical Association, 58*(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

Cet article est publié sous la licence Creative Commons (CC BY-SA). https://creativecommons.org/licenses/by-sa/4.0/.

Pour la citation et la reproduction de cet article, mentionner obligatoirement le titre de l'article, le nom de tous les auteurs, la mention de sa publication dans la revue « NOV'AE », la date de sa publication et son URL.