



Classification sur distance de Wasserstein de distributions quantiles de coûts : application aux pays et régions européens

*Dominique Desbois*¹
desbois@agroparistech.fr

(1) UMR Paris-Saclay Applied Economics (PSAE),
INRAE-AgroParisTech

Plan

Introduction

- 1- Estimations de coûts spécifiques
- 2- Distance et Test de Wasserstein
- 3- AFTD, CAH, Test quadratique, DivClust : résultats

Perspectives

1- Estimations de coûts spécifiques

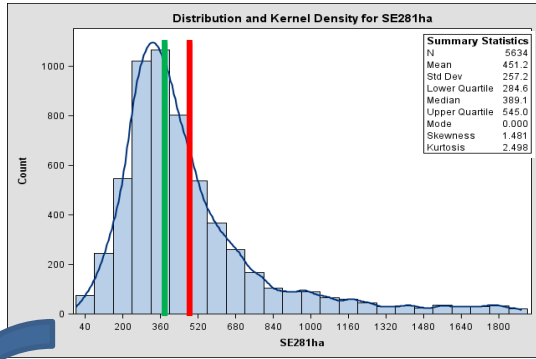
Estimations des coûts

Modèle économétrique des coûts spécifiques de production

$$X_{ih} = \sum_{k=1}^K \alpha_{ih}^k Y_{kh} + \varepsilon_{ih} \text{ with } \varepsilon_{ih} \text{ i.i.d.}$$

CHARGES	PRODUITS						TOTAL CHARGE
	Y_{1h}	...	Y_{kh}	...	Y_{Kh}		
X_{1h}	a_{1h}^1	...	a_{1h}^k	...	a_{1h}^K		$\sum X_{1h}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$
X_{ih}	a_{ih}^1	...	a_{ih}^k	...	a_{ih}^K		$\sum X_{ih}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$
X_{Ih}	a_{Ih}^1	...	a_{Ih}^k	...	a_{Ih}^K		$\sum X_{Ih}$
TOTAL PRODUIT	$\sum Y_{1h}$	...	$\sum Y_{kh}$	...	$\sum Y_{Kh}$		$\sum_k Y_{kh} = \sum_i X_{ih}$

Coûts spécifiques



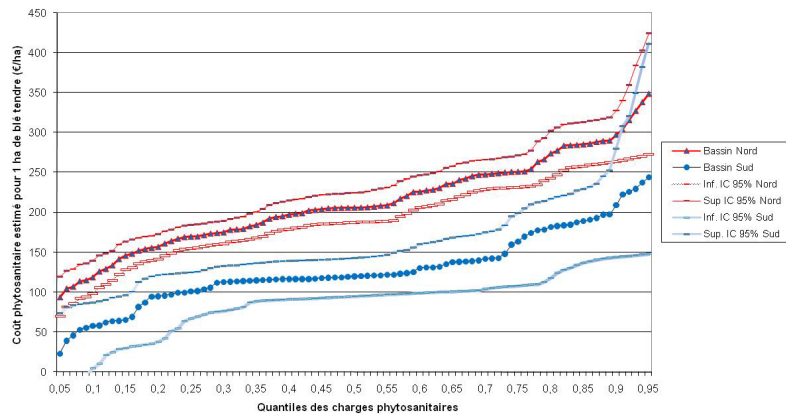
Méthode : régression quantile

$$\underset{\beta}{Min} \left\{ \sum_{\vec{i} x_i \geq y_i' \beta} q |x_i - y_i' \beta| + \sum_{\vec{i} x_i \leq y_i' \beta} (1 - q) |x_i - y_i' \beta| \right\}$$

Estimation: quantiles conditionnels

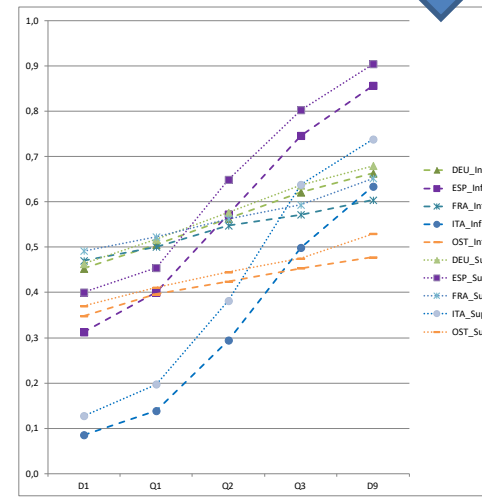
Processus quantile : blé, France, 2006

Blé tendre : processus quantile du coût phytosanitaire, bassins Nord et Sud
(RICA 2006, grandes cultures, estimations basées sur les surfaces)



Distributions

pays européens : porc, 2006



Desbois, Butault & Surry, 2013,2017)

Estimations conditionnelles de processus quantiles :

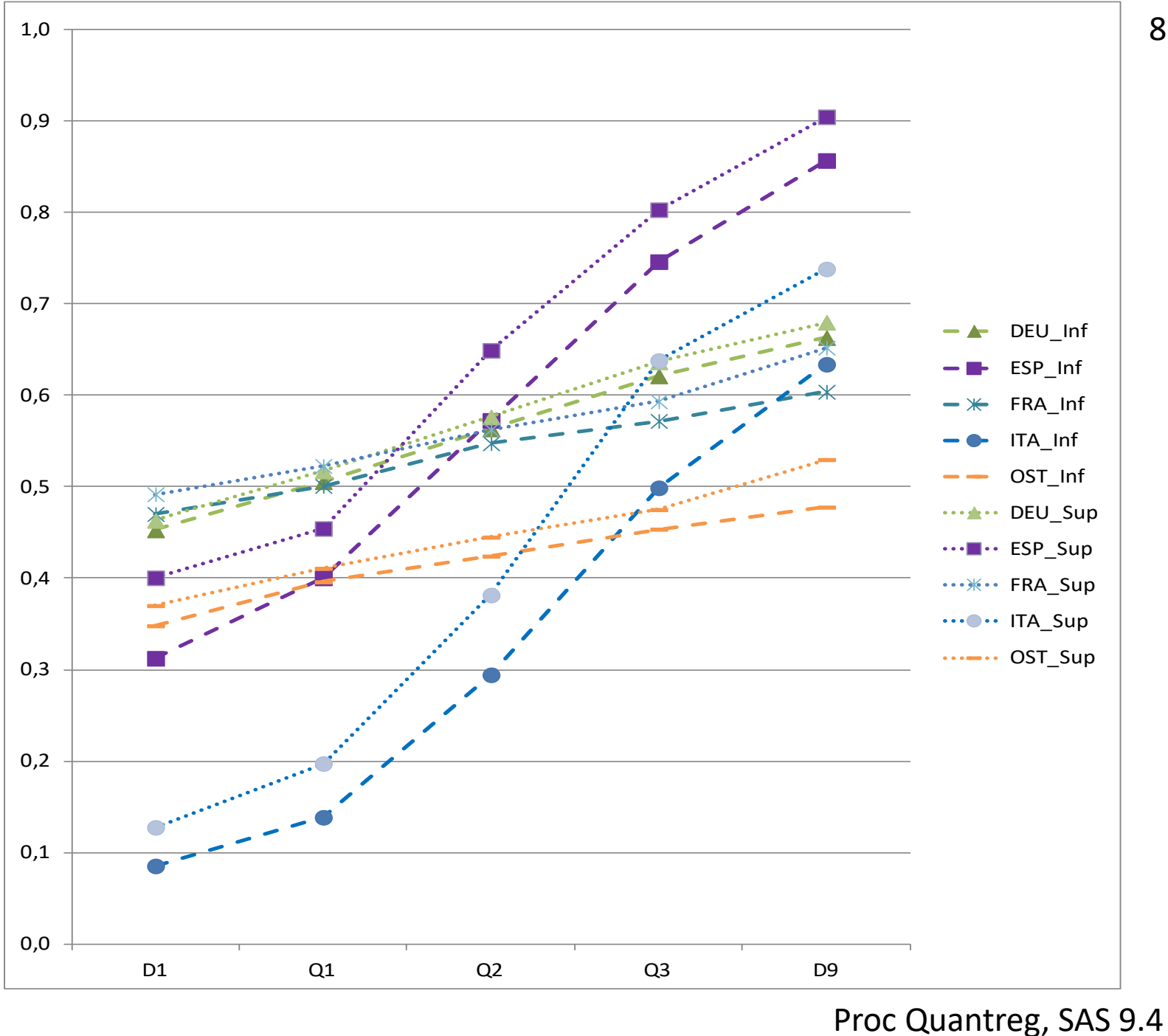
différences de niveaux
et
différences de structures

modèles à « translation d'échelle »
(ESP, ITA)
pentes hétérogènes

versus

modèles à « homothétie d'échelle »
(DEU, FRA, OST)
pentes homogènes

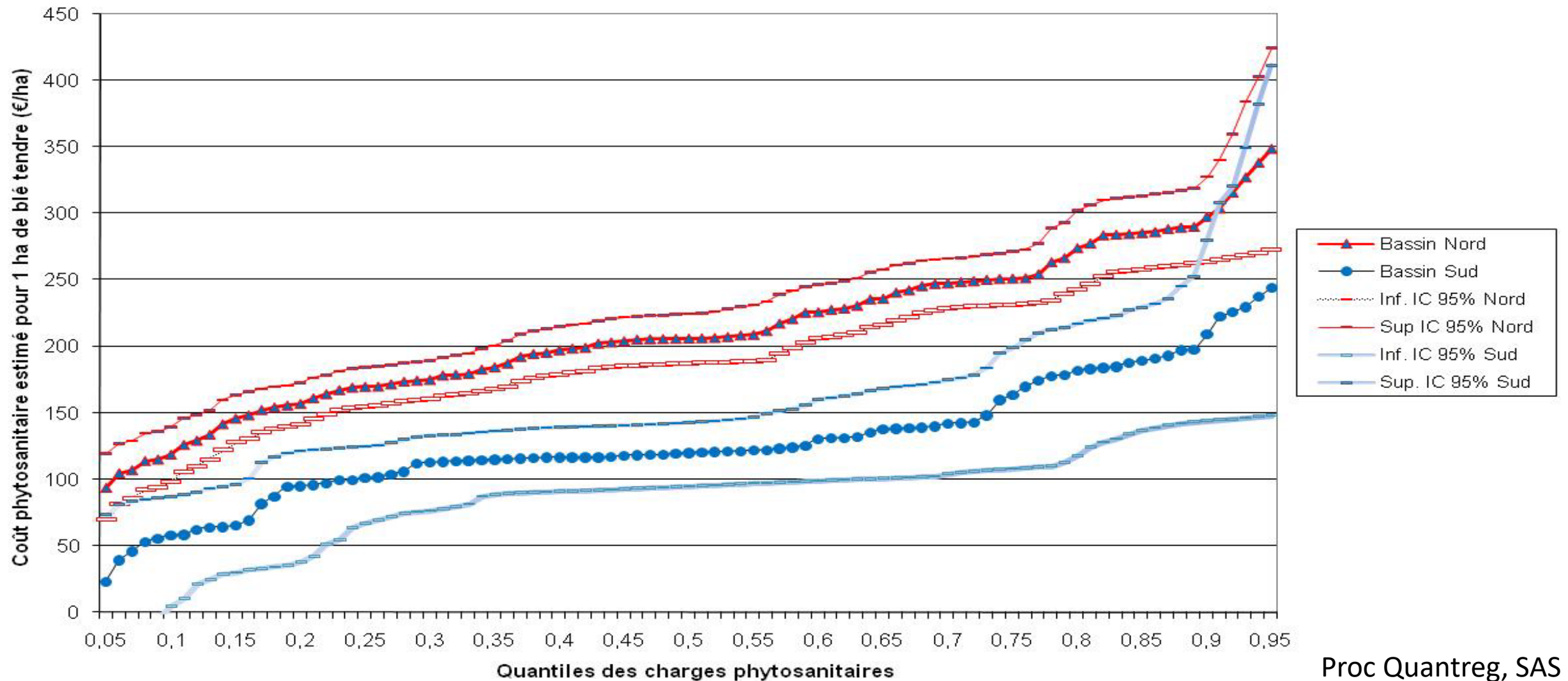
Proc Quantreg, SAS 9.4



Estimations conditionnelles de processus quantiles : test graphique de dominance stochastique entre deux distributions

9

**Blé tendre : processus quantile du coût phytosanitaire, bassins Nord et Sud
(RICA 2006, grandes cultures, estimations basées sur les surfaces)**



Proc Quantreg, SAS 9.4

Les statistiques d'ordre sont équivariantes par transformation monotone

$$\text{si } \lambda \in [0; \infty] \quad \text{alors} \quad \mu_q(\lambda \cdot X + C \mid Y) = C + \lambda \cdot \mu_q(X \mid Y)$$

$$\text{si } \lambda \in [-\infty; 0] \quad \text{alors} \quad \mu_q(\lambda \cdot X + C \mid Y) = C + \lambda \cdot \mu_{(1-q)}(X \mid Y)$$

Par re-paramétrisation en X de $M = Y - X$

On obtient :

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

2- Distance et Test de Wasserstein

Distance de *Wasserstein*

- Sur l'espace de Lebesgue $\mathcal{L}^k$ des fonctions dont la puissance d'ordre $k \in [0, +\infty[$ est intégrable pour la mesure de Lebesgue, la distance de Wasserstein d'ordre k peut être définie par :

$$W_k(\mu, \nu) = \inf \left\{ \left(\mathbb{E}[\|X - Y\|^k] \right)^{1/k} \mid \mathbb{P}_X = \mu \wedge \mathbb{P}_Y = \nu \right\}$$

- La distance de Wasserstein W_k vérifie sur $\mathcal{L}^k$ tous les axiomes d'une métrique : séparation, symétrie et inégalité triangulaire (Clement & Desch, 2008)
- Pour les distributions X et Y , on peut construire une distance de Wasserstein d'ordre 2 (Irpino & Verde, 2006) sur l'espace $\mathcal{L}^2$ des fonctions de carré intégrables, à partir de leurs fonctions quantiles F^{-1} :

$$W_2(X, Y) = \sqrt{\int_0^1 |F_X^{-1}(t) - F_Y^{-1}(t)|^2 dt}$$

Intervalles d'estimation quantile conditionnelle

- Le processus quantile conditionnel partitionne la distribution empirique du paramètre d'intérêt (coût unitaires des intrants) en intervalles d'estimation quantile conditionnelle contigus I_h associés à leur poids π_h :

$$I_h = [\underline{y}_h; \overline{y}_h] \quad \text{tel que} \quad \underline{y}_h \leq \overline{y}_h$$

- Les intervalles du processus d'estimation quantile conditionnelle forment une tribu sur $[0;1]$:

$$\begin{aligned} I_h \cap I_{h'} &= \emptyset \\ \bigcup_{h=1}^H I_h &= [0; 1] \end{aligned}$$

- La pondération π_h vérifie les axiomes probabilistes d'une mesure positive finie :

$$\begin{aligned} \pi_h &\geq 0 \\ \text{et} \\ \sum_{h=1}^H \pi_h &= 1 \end{aligned}$$

- Sous hypothèse de distribution locale uniforme, chaque intervalle d'estimation peut être codé par son centre c_h et son rayon r_h :

$$c_h = (\underline{y}_h + \overline{y}_h)/2 \quad r_h = (\overline{y}_h - \underline{y}_h)/2$$

Distance de Wasserstein entre processus d'estimations quantiles conditionnels

- La distance quadratique univariée entre deux processus quantiles d'estimation conditionnelle Z et Z' :

$$w_2(Z, Z')^2 = \sum_{h=1}^H \pi_h^* d_w(I_h, I'_h)^2 = \sum_{h=1}^H \pi_h^* \left[(c_h - c'_h)^2 + \frac{1}{3}(r_h + r'_h)^2 \right]$$

munis de la pondération $\pi_h^* = (\pi_h + \pi'_h)/2$

peut être étendue sans difficultés à un espace multivarié de dimension finie Δ , (Verde & Irpino, 2006):

$$W_2(Z, Z')^2 = \sum_{\delta=1}^{\Delta} \left\{ \sum_{h_{\delta}=1}^{H_{\delta}} \pi_{h_{\delta}}^* \left[(c_{h_{\delta}} - c'_{h_{\delta}})^2 + \frac{1}{3}(r_{h_{\delta}} + r'_{h_{\delta}})^2 \right] \right\}$$

Décomposition de la distance quadratique de Wasserstein

- Décomposition de la distance quadratique de Wasserstein entre deux distributions u et v de fonctions quantiles respectives F_u^{-1} et F_v^{-1}

$$d_W(u, v)^2 = d_W^2(F_u, F_v) = \int_0^1 (F_u^{-1}(t) - F_v^{-1}(t))^2 dt$$

$$= (\mu_u - \mu_v)^2 + (\sigma_u - \sigma_v)^2 + 2\sigma_u \sigma_v - 2Cov_{QQ}(u, v)$$

$$= (\mu_u - \mu_v)^2 + (\sigma_u - \sigma_v)^2 + 2\sigma_u \sigma_v (1 - Cor_{QQ}(u, v))$$

$$\text{où } Corr_{QQ}(u, v) = \int_0^1 (F_u^{-1}(t) - \mu_u)(F_v^{-1}(t) - \mu_v) dt / \sigma_u \sigma_v$$

est interprétable comme la somme de trois composantes :

$$= \textit{tendance} + \textit{dispersion} + \textit{forme}$$

(Irpino et Romano, 2007)

Test de *Wasserstein*

- 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 .

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

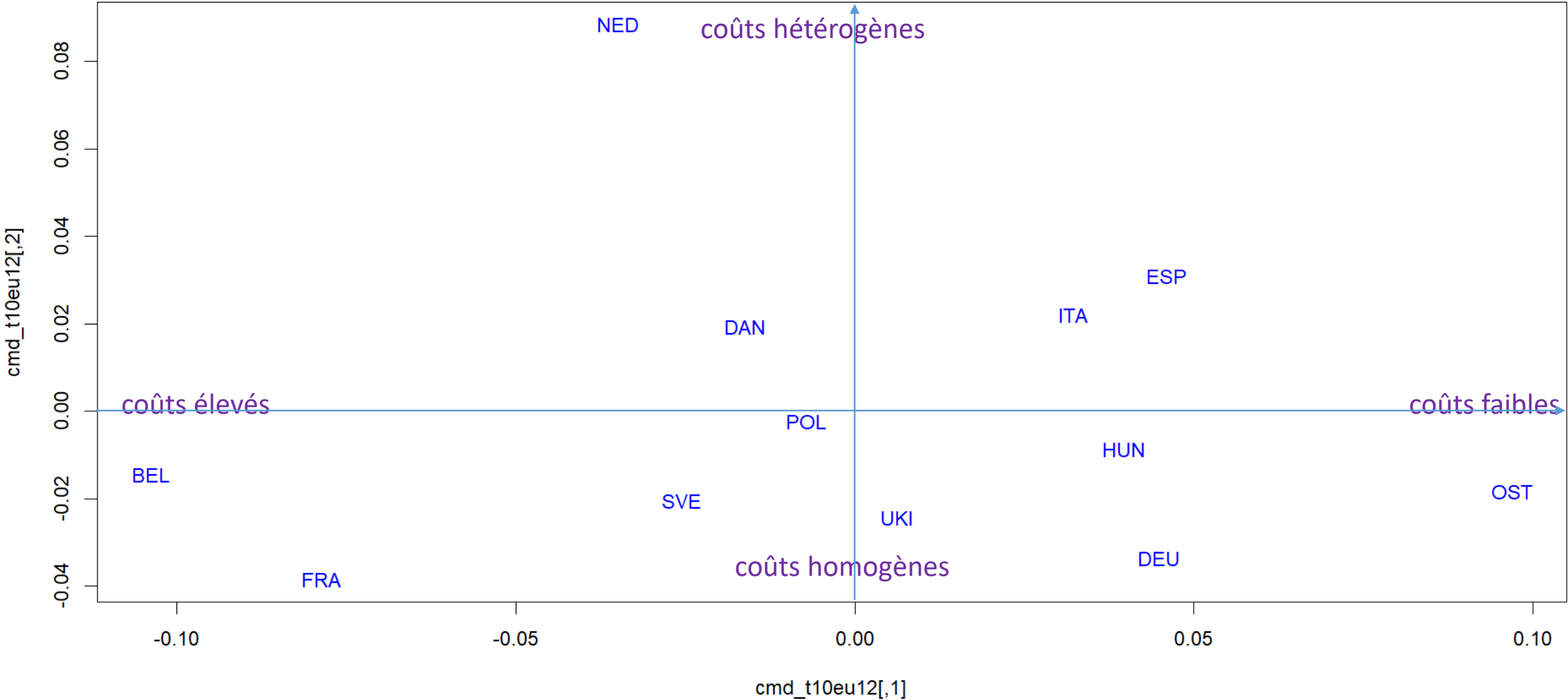
- Le calcul de la p -valeur est basé sur l'inégalité de **Dvoretzky–Kiefer–Wolfowitz** (1956) qui fournit une limite maximale à la distance entre une fonction de distribution empirique G_n d'un échantillon de taille n par rapport à la fonction de distribution théorique G de la population

$$Pr(\sup_{x \in \mathbb{R}} |G_n(x) - G(x)| > \varepsilon) \leq C e^{-2n\varepsilon^2}, \forall \varepsilon$$

- 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 de Wasserstein tire parti des hypothèses distributionnelles sur les intervalles interquantiles (e.g. distribution uniforme).

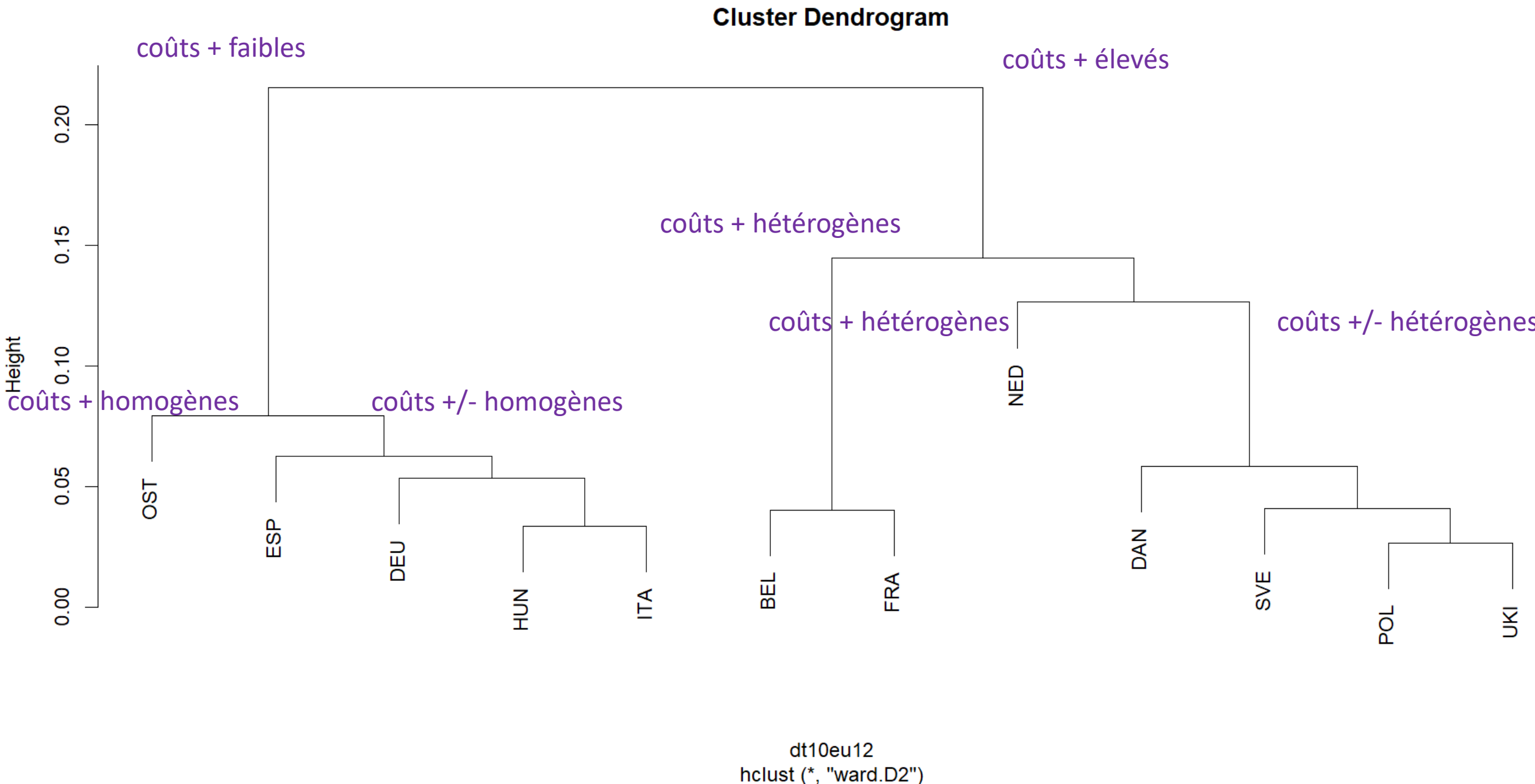
3- Résultats : AFTD, CAH, Test 2-W, DivClust

Analyse factorielle du tableau de distances de Wasserstein



```
> cmd_c10eu12<-cmdscale(c10eu12)
```

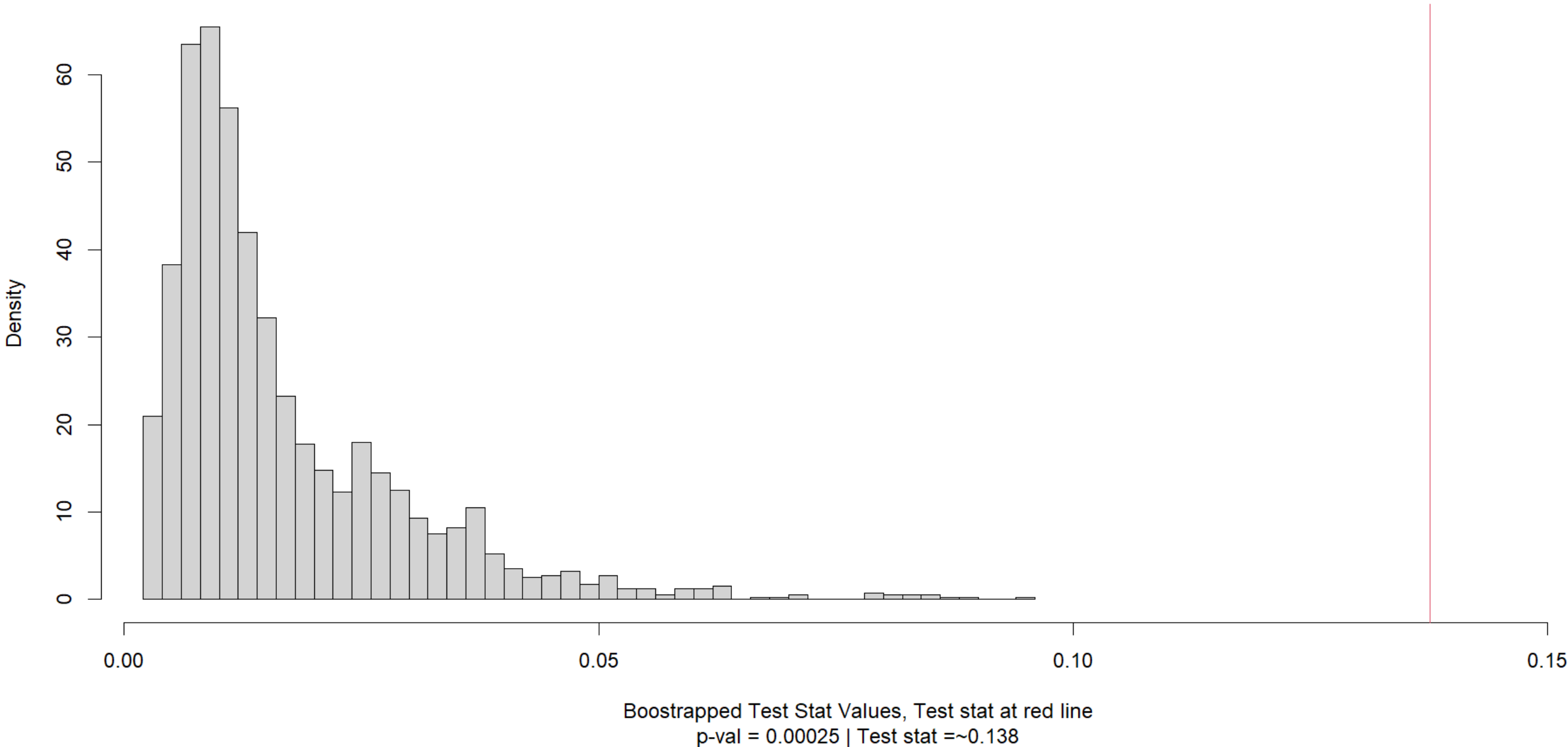
CAH 12 pays européens (c1..c9), distance quadratique de Wasserstein (W_2)



12 pays européens (c1..c9) test de Wasserstein : BEL vs OST

Différences significatives

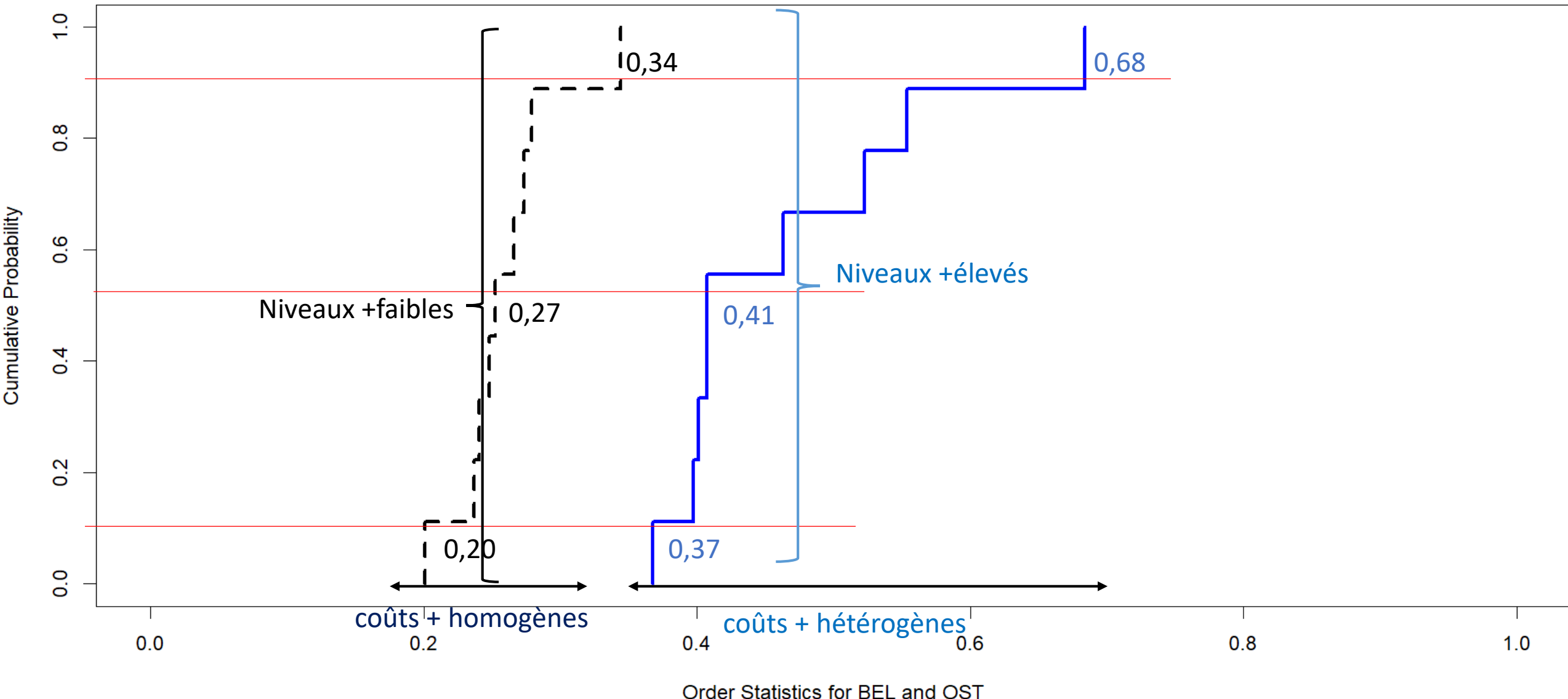
WASS Test: Bootstrap Distribution + Test Stat



12 pays européens (c1..c9), FdR empiriques: BEL vs OST

```
> BEL<-as.numeric(c10eu12[1,2:10])  
> OST <-as.numeric(c10eu12[9,2:10])  
> cdfCompare(x=BEL,y=OST,xlim=c(0,1),ylim=c(0,1), discrete=TRUE)
```

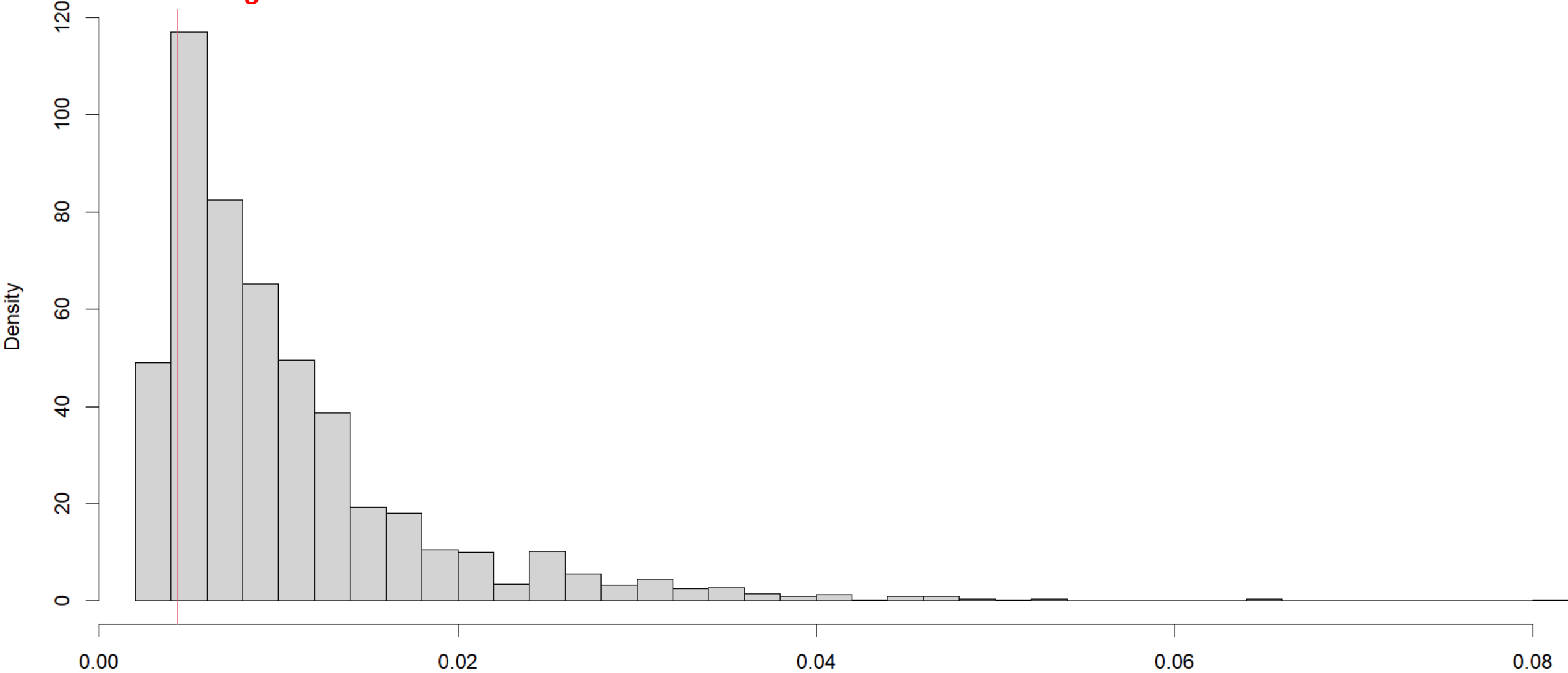
**Empirical CDF for BEL (solid line)
with Empirical CDF for OST (dashed line)**



12 pays européens (c1..c9), test de Wasserstein : BEL vs FRA

WASS Test: Boostrap Distribution + Test Stat

Différences non-significatives

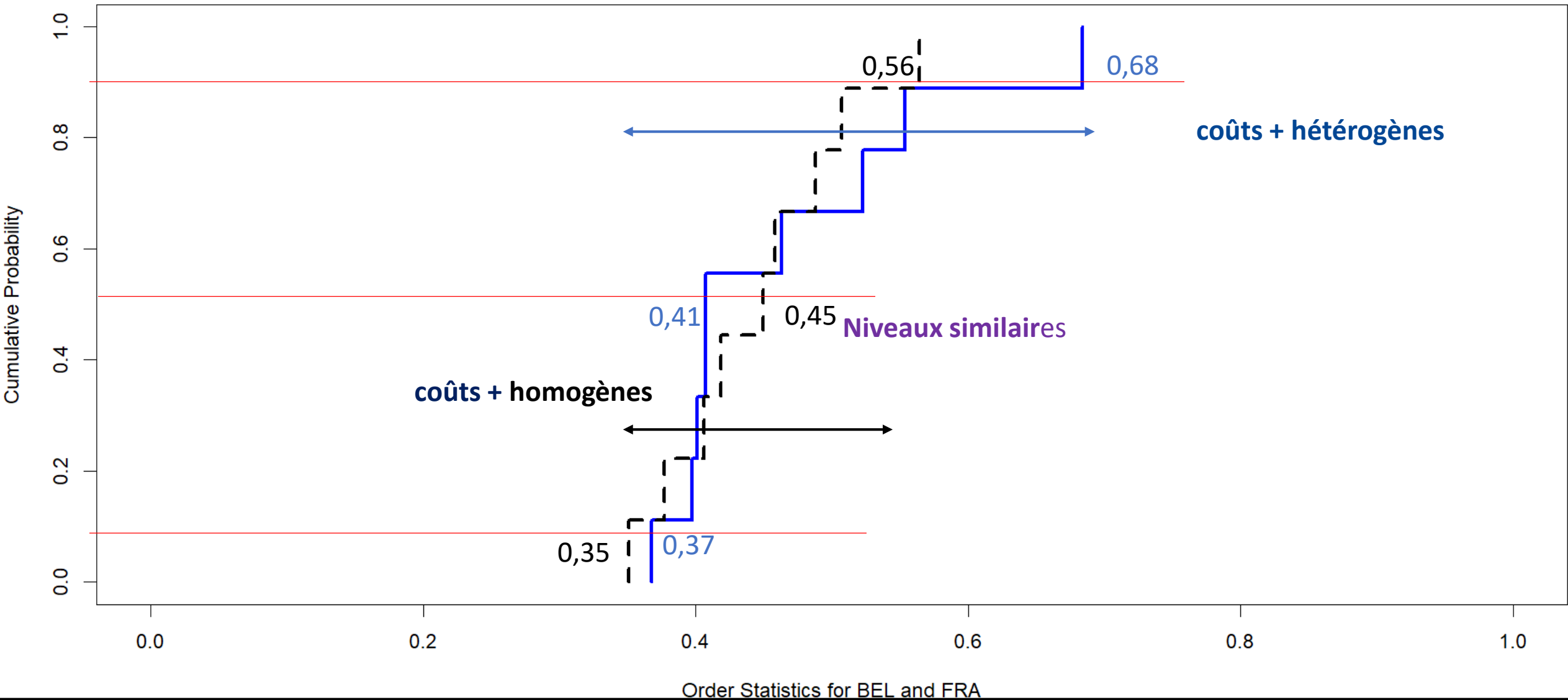


Boostrapped Test Stat Values, Test stat at red line

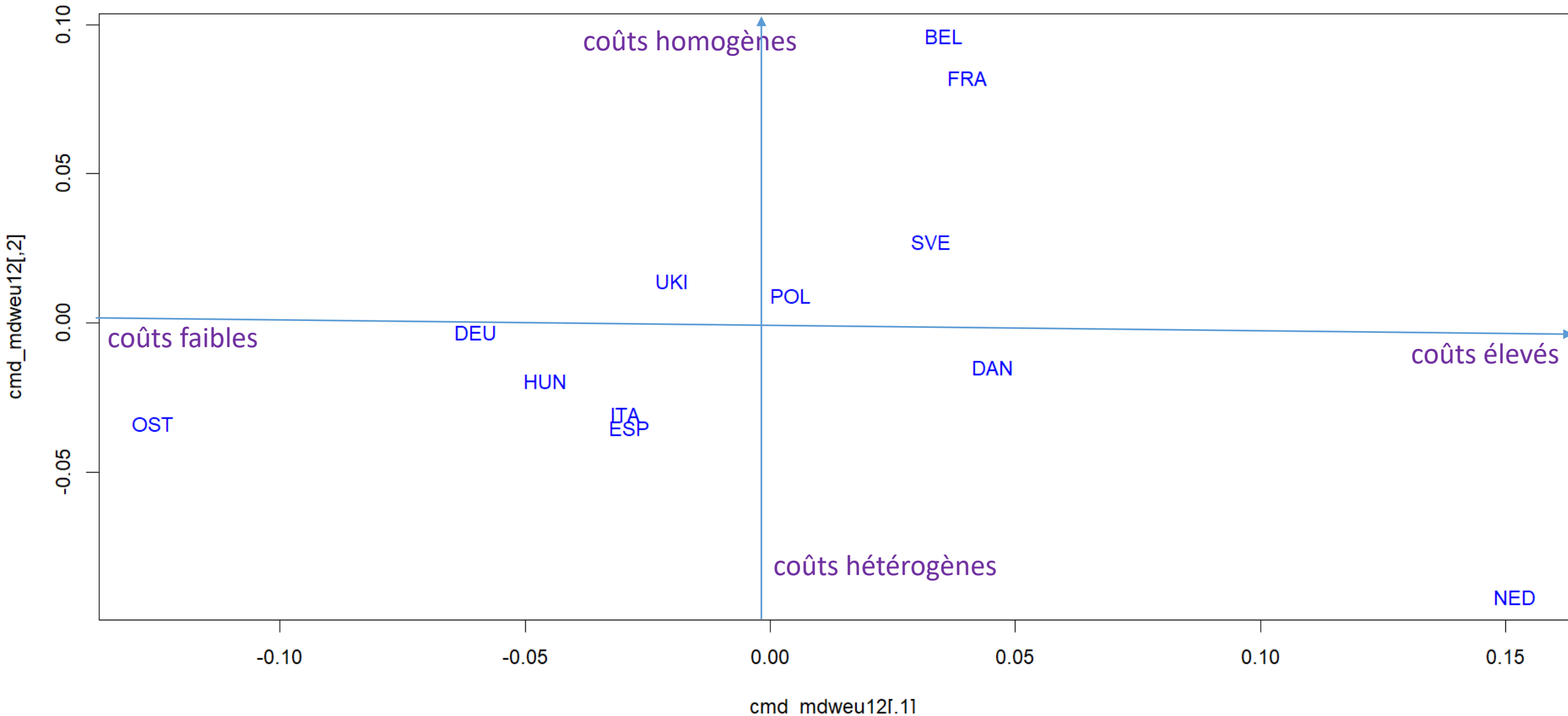
p-val = 0.85 | Test stat =~0.00438

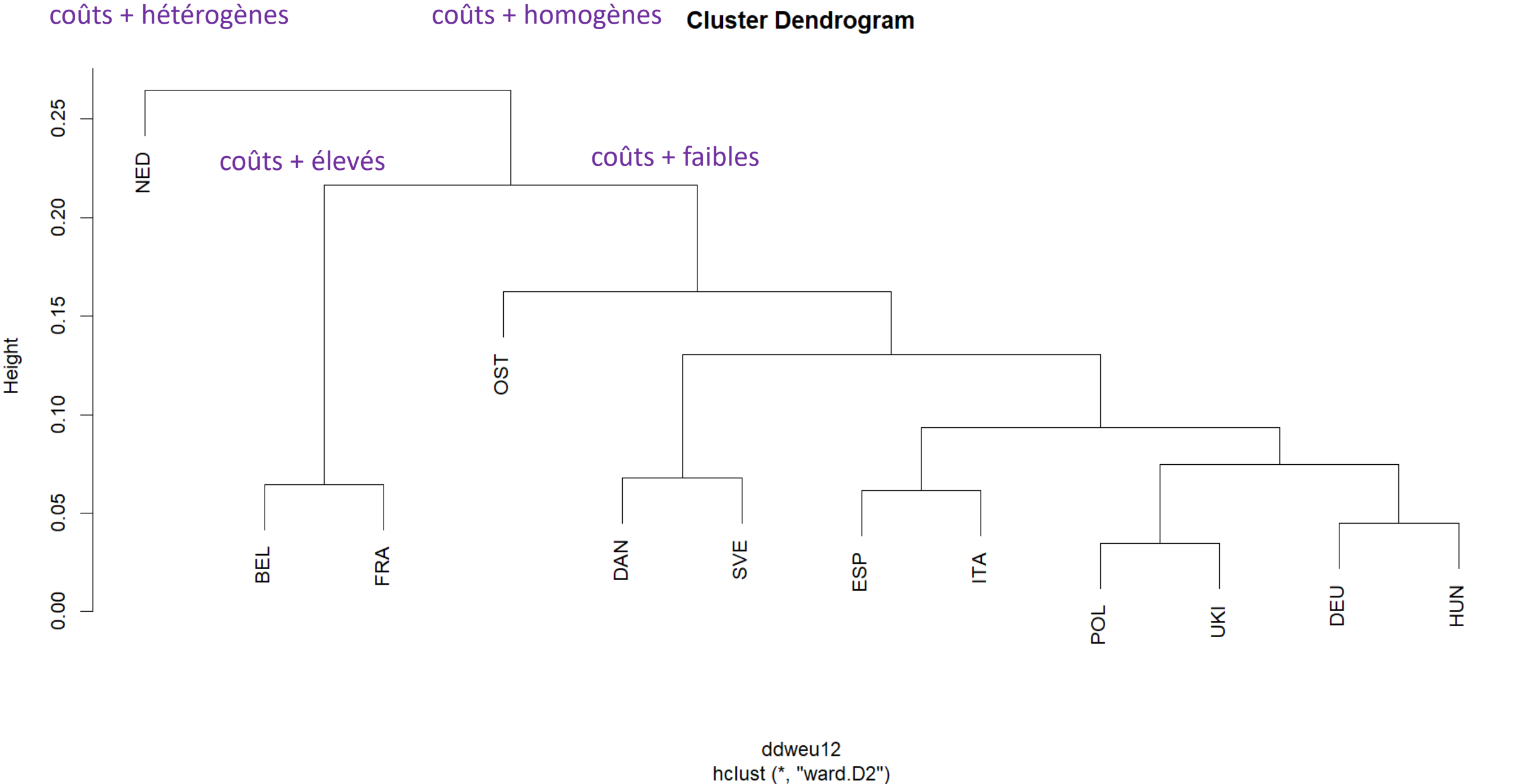
12 pays européens (c1..c9), FdR empiriques : BEL vs FRA

Empirical CDF for BEL (solid line)
with Empirical CDF for FRA (dashed line)



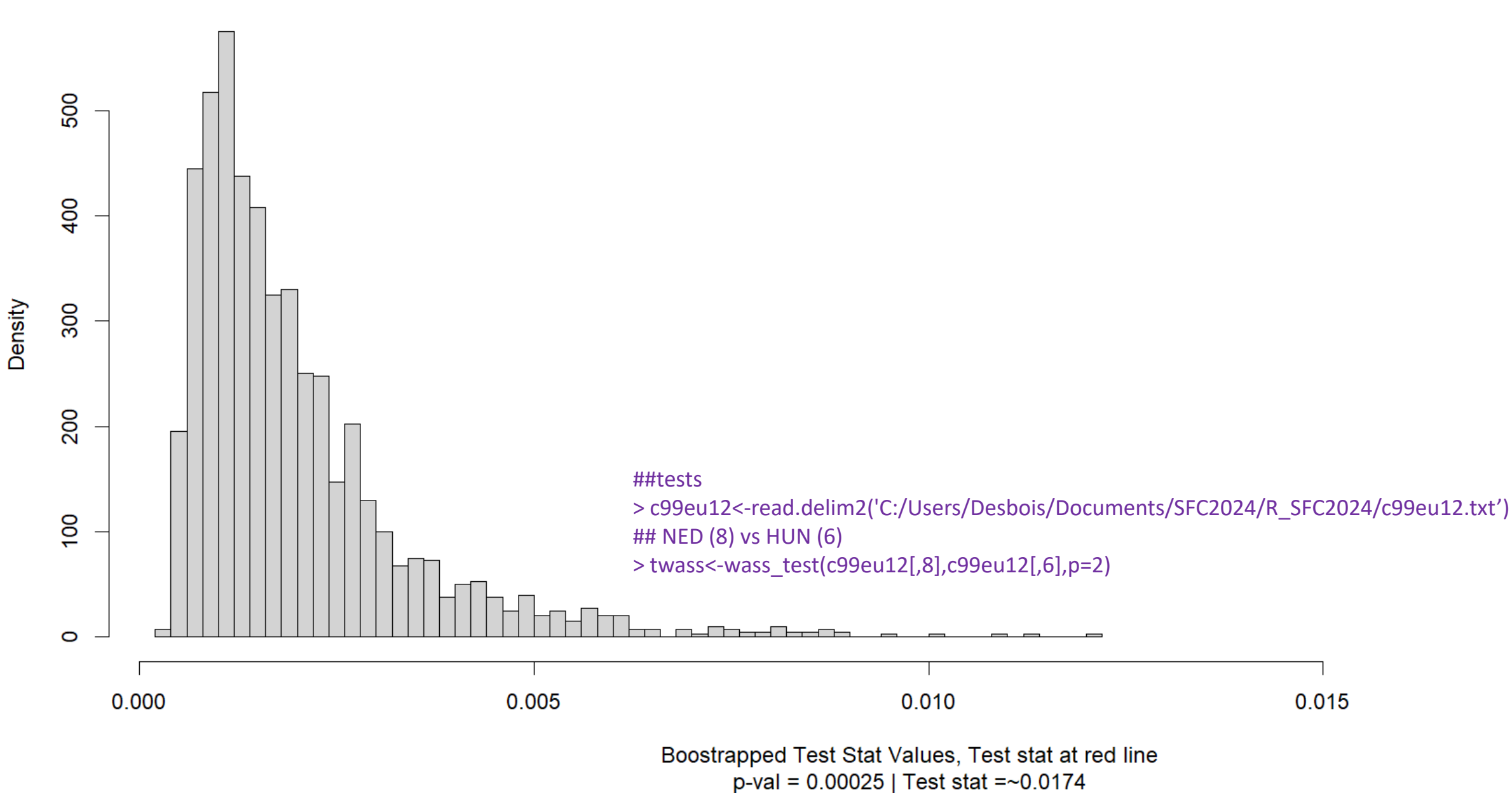
Analyse factorielle du tableau de distances de Wasserstein

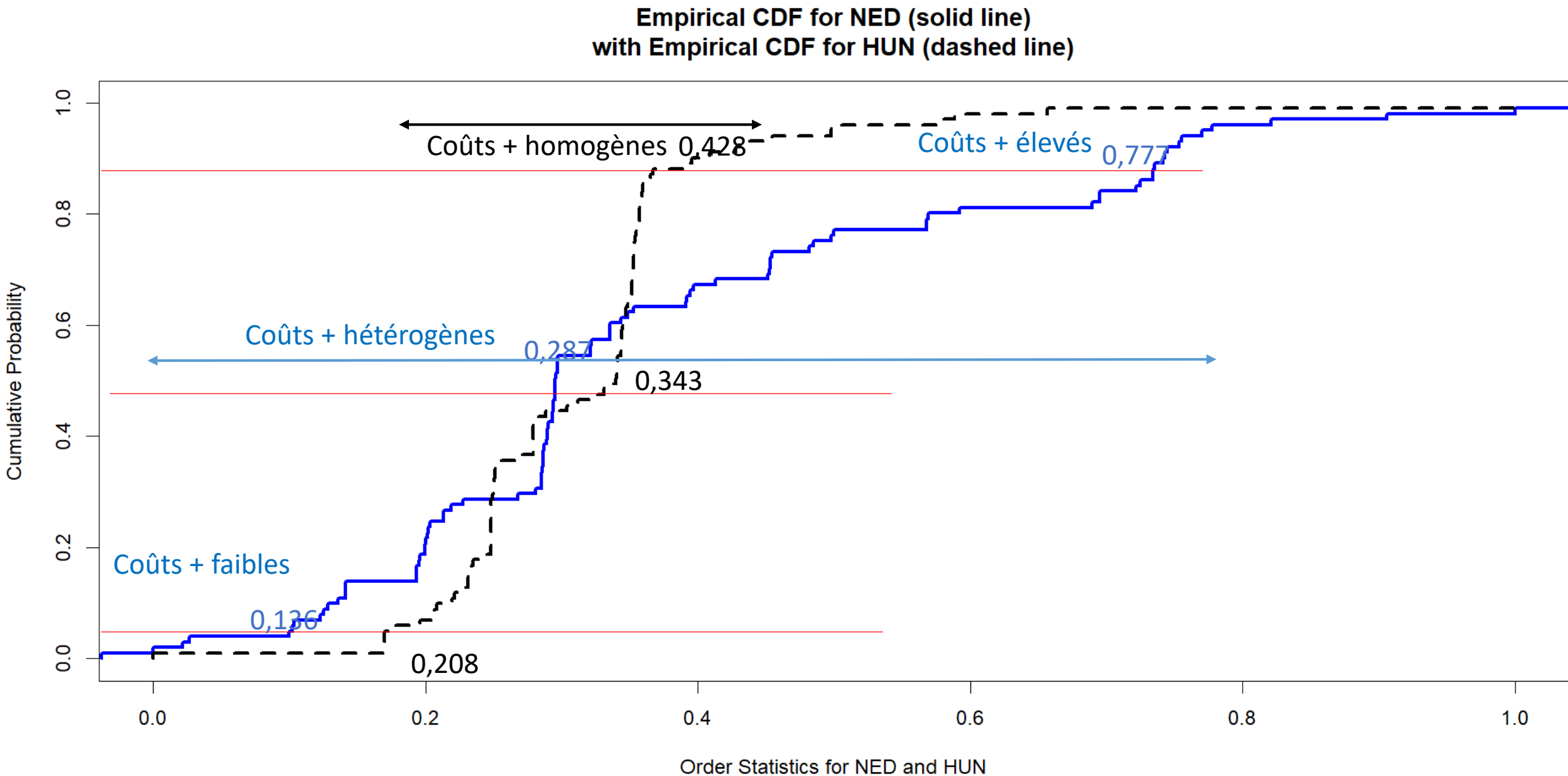




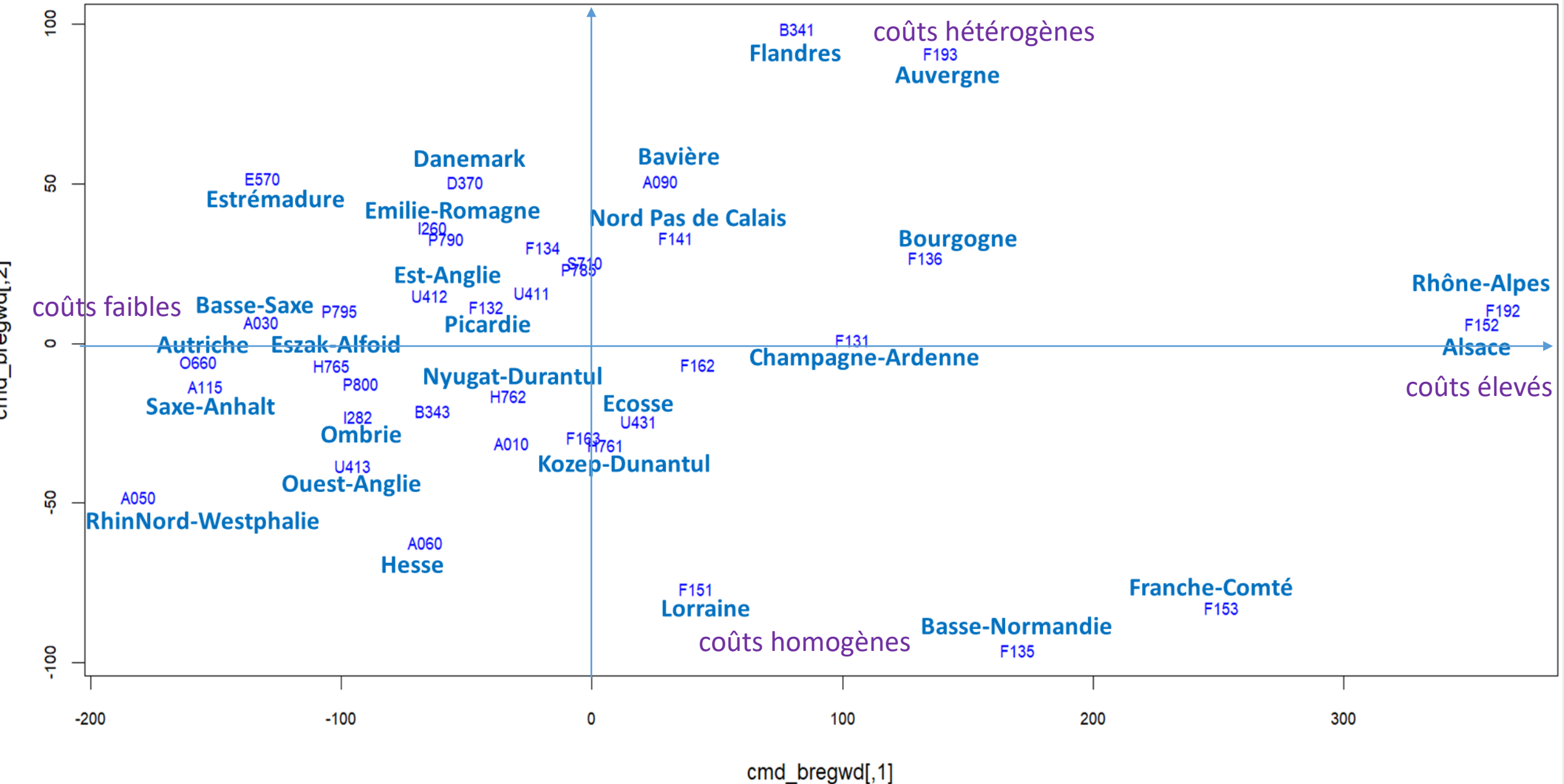
WASS Test: Bootstrap Distribution + Test Stat

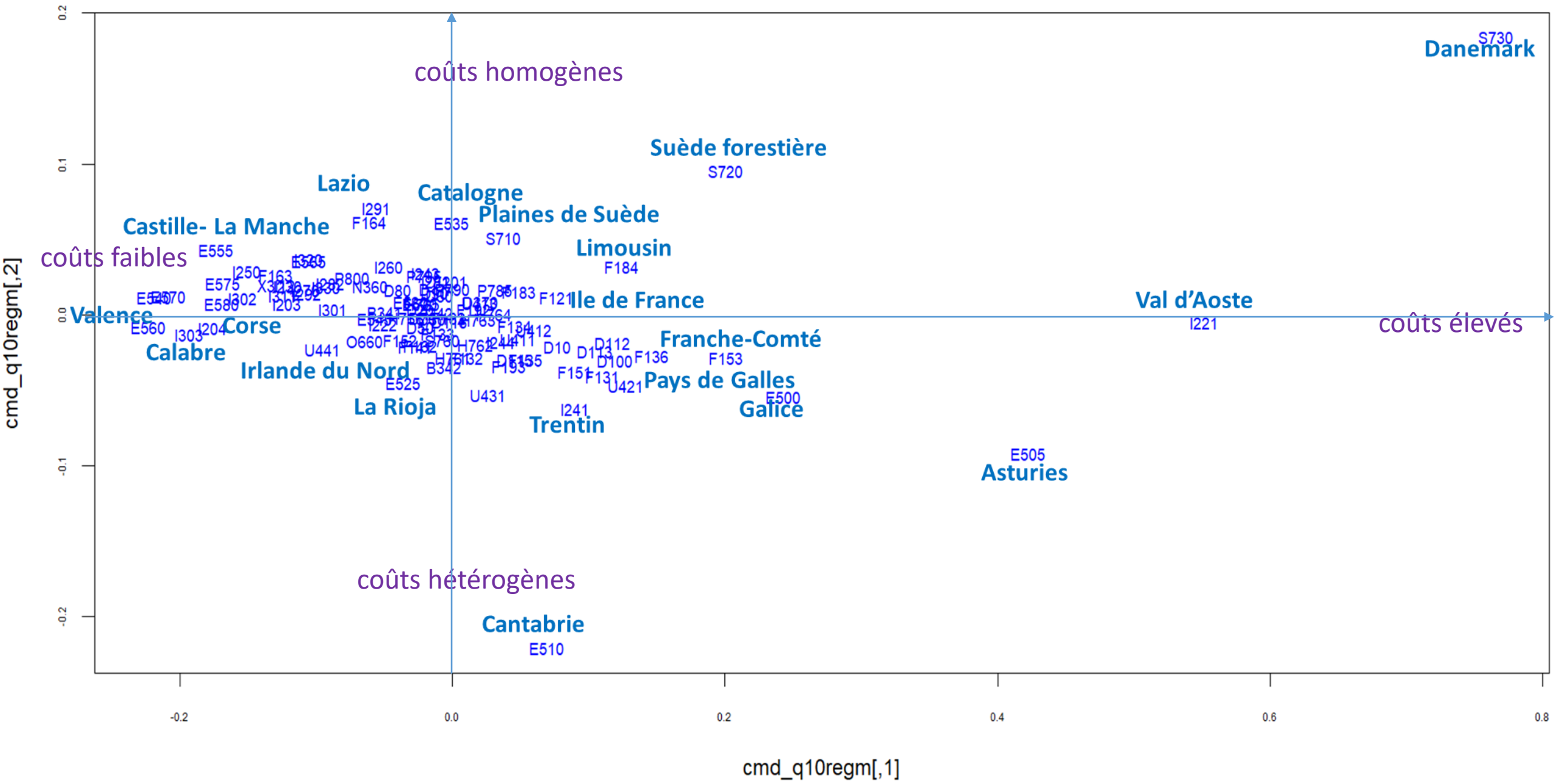
Différences significatives





Analyse factorielle du tableau de distances de Wasserstein

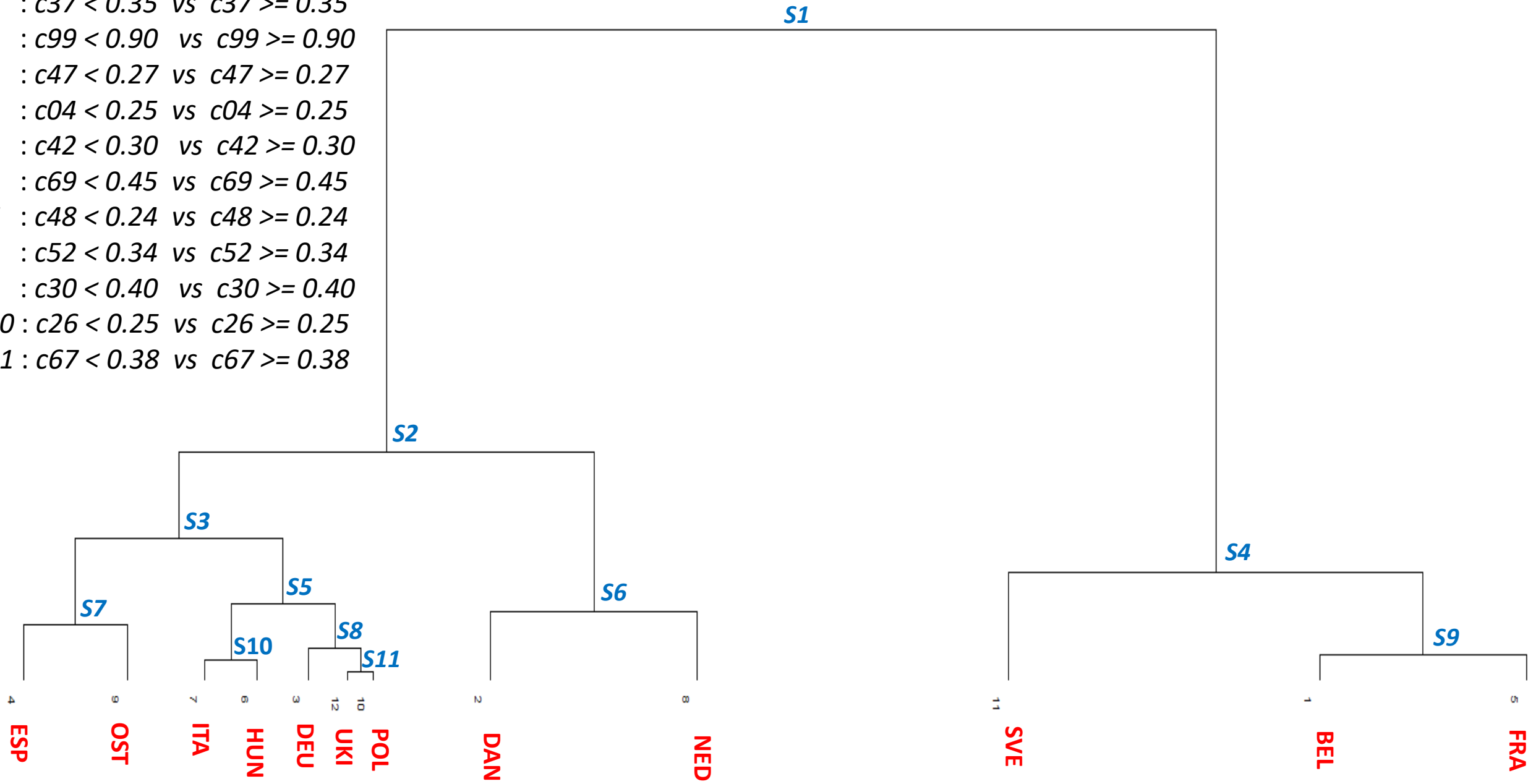




DivClustw (d'après Chavent & al. 2007), 12 pays européens : c01..c99, coûts intrants blé

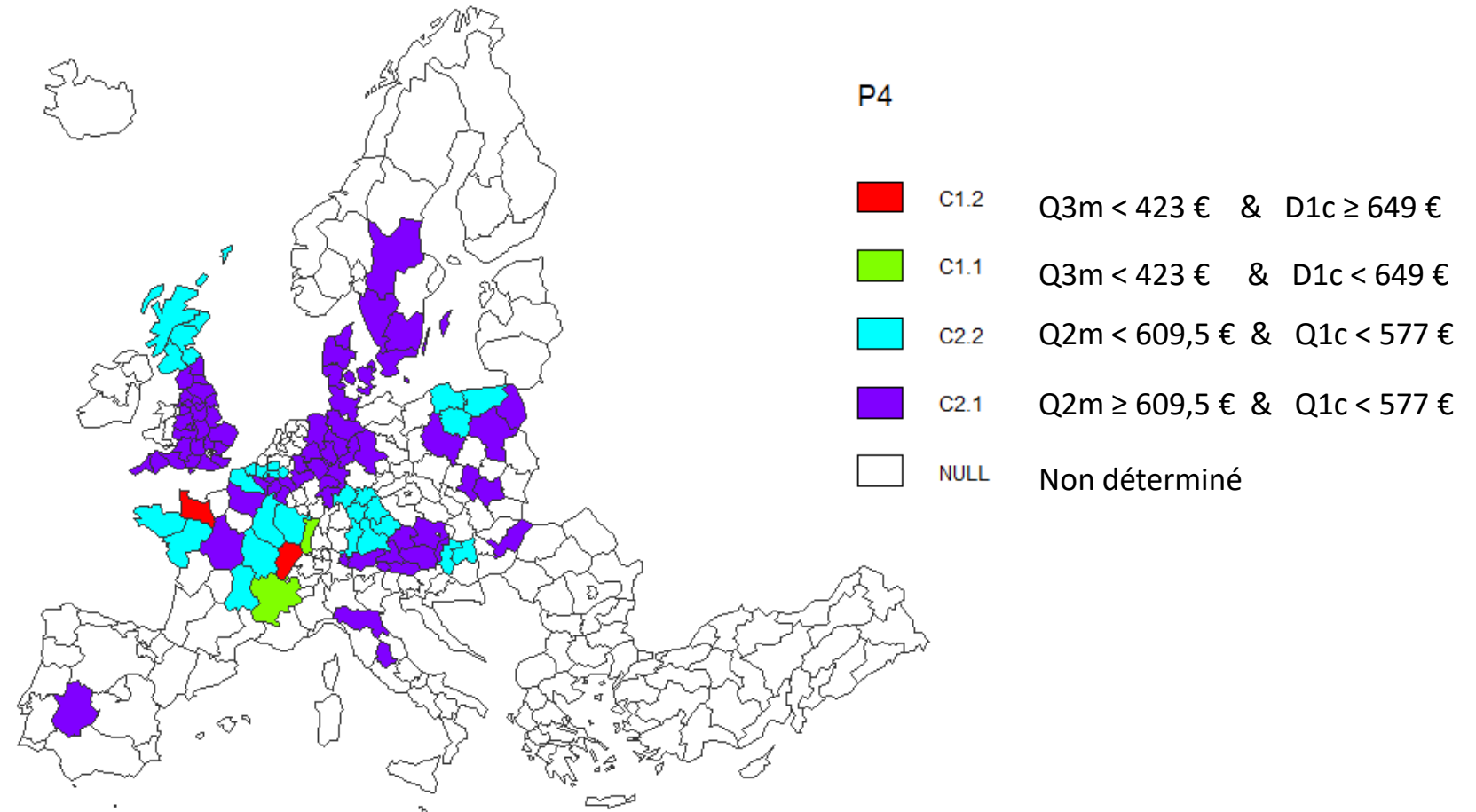
plot(res_divclust, nqbin=11, cex=0.4)

- S1 : $c_{37} < 0.35$ vs $c_{37} \geq 0.35$
- S2 : $c_{99} < 0.90$ vs $c_{99} \geq 0.90$
- S3 : $c_{47} < 0.27$ vs $c_{47} \geq 0.27$
- S4 : $c_{04} < 0.25$ vs $c_{04} \geq 0.25$
- S5 : $c_{42} < 0.30$ vs $c_{42} \geq 0.30$
- S6 : $c_{69} < 0.45$ vs $c_{69} \geq 0.45$
- S7 : $c_{48} < 0.24$ vs $c_{48} \geq 0.24$
- S8 : $c_{52} < 0.34$ vs $c_{52} \geq 0.34$
- S9 : $c_{30} < 0.40$ vs $c_{30} \geq 0.40$
- S10 : $c_{26} < 0.25$ vs $c_{26} \geq 0.25$
- S11 : $c_{67} < 0.38$ vs $c_{67} \geq 0.38$



Perspectives

Vers une partition des coûts spécifiques et des marges brutes pour les exploitations agricoles en Europe



❑ Une méthodologie d'estimation quantile des coûts spécifiques

- ✓ Basée sur un modèle général comptable d'allocation des intrants aux extrants
- ✓ Adaptée à la nature duale des concepts comptables de coûts spécifiques et de marges brutes
- ✓ Adaptée à la nature asymétrique des distributions de charges et de revenus
- ✓ Réalisable au niveau II de la NUTS (régions européennes)

❑ Des outils exploratoires et décisionnels pour étudier les distributions quantiles empiriques

- ✓ L'analyse factorielle du tableau des distances de Wasserstein (AFTDW)
- ✓ La classification ascendante hiérarchique sur distances de Wasserstein (CAHDW)
- ✓ Le test quadratique de Wasserstein (T2W) d'égalité de distributions empiriques
- ✓ Une extension à la classification divisive (DivClustW)

❑ L'élaboration d'un référentiel typologique de coûts spécifiques

- ✓ Un référentiel typologique de coûts spécifiques au niveau régional européen

Références

- Birnbaum, W. R. & McCarty Z. C. (1958) A Distribution-Free Upper Confidence Bound for $\Pr\{Y < X\}$, Based on Independent Samples of X and Y, *Ann. Math. Statist.* 29(2): 558-562 (juin). DOI: 10.1214/aoms/1177706631
- Chavent, M.; Lechevallier, Y.; Briant O. (2007) DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics and Data Analysis* 52: 687-701.
- Clement, P. ; Desch W. (2008) An Elementary Proof of the Triangle Inequality for the Wasserstein Metric, *Proceedings of the American Mathematical Society*, Volume 136, n°1, janvier 2008, pp. 333–339
- 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.
- Dvoretzky, A.; Kiefer, J.; Wolfowitz, J. (1956), "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator", *Annals of Mathematical Statistics*, 27 (3): 642–669.
- Gabriel, K. R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, **58**, 453--467. 10.2307/2334381.
- Gower, J.C.; Hand D. J. (1996). *Biplots*. Chapman & Hall.
- Irpino A. & Romano E. (2007) Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation, *EGC*, vol. RNTI-E-9, pp. 99-110
- Massart, P. (1990), "The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality", *Annals of Probability*, 18 (3): 1269–1283.
- Irpino, A., & Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: V. Batanjeli, H. H. Bock, A. Ferligoj, & A. Ziberna, (Eds.), *Data science and classification, IFCS 2006* (pp. 185–192). Berlin: Springer.