

Uncovering excellence in academic rankings: a closer look at the Shanghai ranking.

C. Dehon, A. McCathie & V. Verardi

Université libre de Bruxelles, ECARES - CKE

September 2010

- ▶ Popularité croissante des classements d'universités
- ▶ Mais que mesurent-ils réellement?
- ▶ Notre objectif: identifier les facteurs sous-jacents mesurés par le classement de Shanghai.



Analyse en composantes principales.

$$SCORE_i = 0.1 * Alumni_i + 0.2 * Award_i + 0.2 * HiCi_i \\ + 0.2 * N\&S_i + 0.2 * PUB_i + 0.1 * PCP_i$$

- ▶ **Alumni**: Alumni ayant un Prix Nobel/une Médaille Fields.
- ▶ **Award**: Professeurs ayant un Prix Nobel/une Médaille Fields.
- ▶ **HiCi**: Chercheurs "hautement cités" selon une liste de l'ISI.
- ▶ **N&S**: Articles publiés dans les revues *Nature* et *Science*.
- ▶ **PUB**: Articles référencés dans le *Science Citation Index-expanded*, et le *Social Science Citation Index*.
- ▶ **PCP**: Score moyen obtenu dans les 5 catégories ci-dessous divisé par le nombre d'académiques temps-plein.

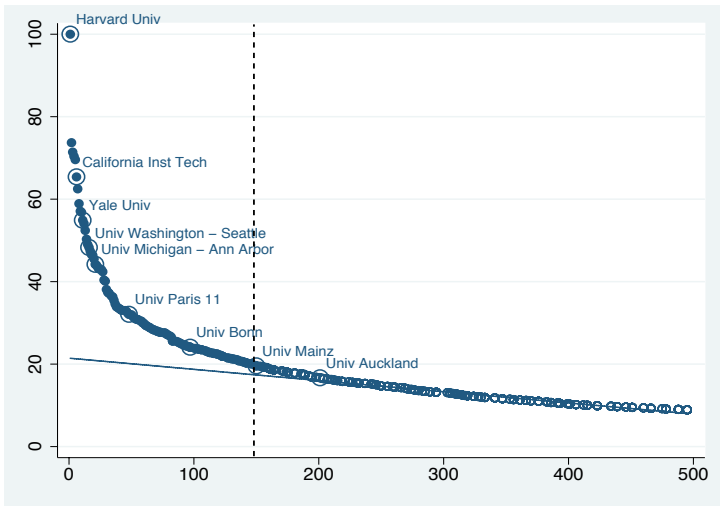


Figure: Overall score relative to rank

Critiques formulées à l'encontre du classement de Shanghai:

- ▶ Ignore en grande partie le caractère complexe de l'université.
- ▶ Favorise les universités de langue anglaise.
- ▶ Est fortement biaisé dans le sens des sciences et des technologies.
- ▶ Est fortement influencé par la taille de l'institution.
- ▶ Est fortement influencé par la normalization choisie.

Analyse en composantes principales des universités du top 150:

- ▶ On retient la première composante principale: celle-ci capture 64% de l'inertie totale de l'échantillon.
- ▶ Forte corrélation positive avec chacune des 5 variables.

Corr(.,.)	Alumni	Award	HiCi	N&S	PUB
ϕ_1	0.78	0.81	0.89	0.92	0.70

- ▶ MAIS: 18% de Φ_1 est dû uniquement à Harvard, et le top 10 a lui seul contribue 60% de cette composante.

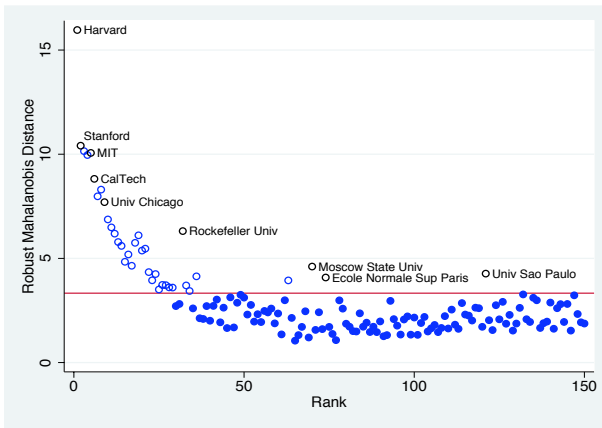
⇒ Présence de valeurs aberrantes.

Détection des valeurs aberrantes: distance de Mahalanobis

$$RD_i = \sqrt{((x_i - T(X))'C(X)^{-1}(x_i - T(X)))}$$

- ▶ Il nous faut trouver des estimateurs **robustes** de T_n et C_n :
Minimum Covariance Determinant Estimator
(Rousseeuw, 1985)
⇒ Cherchons un sous-échantillon de notre échantillon de départ qui:
 - ▶ contienne 50% des observations de l'échantillon de départ.
 - ▶ ait la plus faible variance généralisée, définie comme $\det(\Sigma)$ où Σ est la matrice de covariance du sous-échantillon.
- ▶ Les estimateurs de T_n et C_n sont calculés sur ce sous-échantillon.

Détection des valeurs aberrantes:



ACP robuste: basée sur le *Reweighted MCD estimator*
(Croux and Haesbroeck, 2000)

⇒ ACP réalisée sur l'échantillon obtenu en attribuant un poids de zéro aux valeurs aberrantes détectées.

- ▶ Les deux premières composantes principales capturent 68% de l'inertie totale.
- ▶ Chacune est corrélée avec un sous-ensemble différent des 5 variables:

Corr(.,.)	Alumni	Award	HiCi	N&S	PUB
ϕ_1^R	-0.04	0.03	0.87	0.85	0.70
ϕ_2^R	0.82	0.85	-0.05	0.16	-0.13

⇒ Deux facteurs sous-jacents non-corrélés: d'une part, les chercheurs "superstars" et d'autre part, la production de recherche.

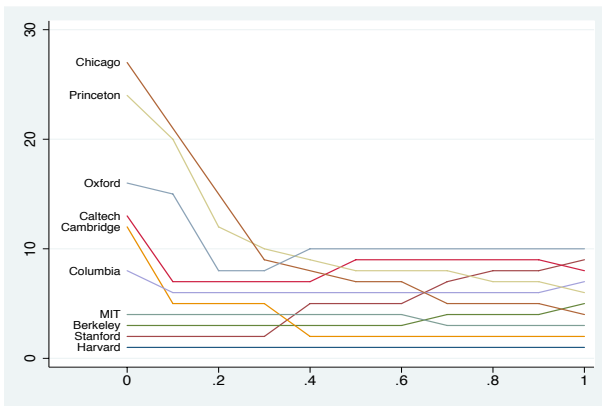
Classements alternatifs réalisés en variant les pondérations:

$$SCORE_i = w_i * (Alumni + Award) + (1 - w_i) * (HiCi + N\&S + PUB)$$

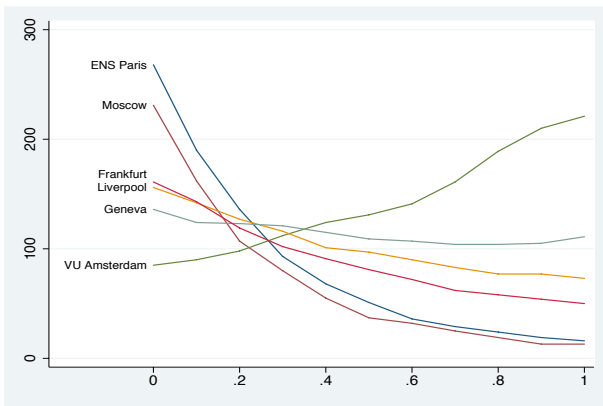
avec $w_i = 0, 0.1, \dots, 1$

- ▶ Le haut du classement reste essentiellement inchangé.
- ▶ Pour la majorité des universités du top 150, le rang est fortement influencé par la pondération choisie.

Exemple 1: Le top 10



Exemple 2: Quelques universités européennes



En conclusion: cette étude nous a permis de...

- ▶ rappeler l'importance des méthodes robustes en statistique, dont l'enseignement est un complément indispensable à celui de la statistique classique.
- ▶ attirer l'attention sur la nécessaire prudence qui doit accompagner toute interprétation ou utilisation des classements d'universités.

