Faut-il croire le classement de Shangaï ?

Une approche fondée sur l'aide multicritère à la décision [1]

Jean-Charles Billaut [2] Denis Bouyssou [3] Philippe Vincke [4]

31 mai 2010

1. Nous tenons à remercier Florence Audier, Ghislaine Filliatreau, Thierry Marchant, Michel Zitt et un arbitre anonyme pour leurs précieux commentaires sur une version antérieure anglaise de ce document. Ce texte est une version française révisée et actualisée de Billaut et al. (2009). Sauf mention contraire, les sites Internet cités dans ce texte ont été consultés à la date du 18 novembre 2009.

2. Laboratoire d'Informatique, Université François Rabelais, 64 Avenue Jean Portalis, F-37200 Tours, France, e-mail : jean-charles.billaut@univ-tours.fr

3. CNRS–LAMSADE FRE3234 & Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, F-75775 Paris Cedex 16, France, tel : +33 1 44 05 48 98, fax: +33 1 44 05 40 91, e-mail : bouyssou@lamsade.dauphine.fr

4. Université Libre de Bruxelles, 50 avenue F. D. Roosevelt, CP. 130, B-1050 Bruxelles, Belgique, e-mail : pvincke@ulb.ac.be

Résumé

Nous proposons dans cet article une analyse critique du classement mondial des universités publié chaque année par l'Institut de l'Enseignement Supérieur de l'Université Jiao Tong de Shangaï et plus connu sous le nom de « classement de Shangaï ». Après avoir rappelé comment le classement est construit, nous discutons de la pertinence des critères utilisés. Nous analysons ensuite la méthode d'agrégation proposée. Notre analyse se fonde sur les outils et concepts de l'« aide multicritère à la décision ». Nos conclusions principales sont que les critères utilisés ne sont pas pertinents, que la méthode d'agrégation présente des problèmes majeurs et que l'exercice souffre d'une réflexion insuffisante sur les questions liées à la structuration du problème. Le classement de Shangaï, malgré la grande couverture médiatique qu'il reçoit chaque année, n'est donc pas un outil pertinent pour juger de la « qualité » des institutions académiques, guider le choix des étudiants ou des familles, ou promouvoir des réformes du système d'enseignement supérieur.

Mots-clés : classement de Shangaï, aide multicritère à la décision, modèles d'évaluation, enseignement supérieur.

Table des matières

1 Introduction

En 2003, un groupe de personnes appartenant à l'Institut de l'Enseignement Supérieur (« *Institute of Higher Education* ») de l'Université Jiao Tong de Shangaï a publié sur Internet son premier classement académique international des universités (« *Academic Ranking of World Universities* », ARWU, 2003–09), plus connu sous le nom de « *classement de Shangaï* » [1]. Une description du classement est donnée dans Liu et Cheng (2005) et l'histoire de sa création est détaillée dans Liu (2009).

Ce classement a très vite fait l'objet d'une grande couverture médiatique. Ses résultats ont été utilisés par des décideurs politiques pour promouvoir des réformes de l'enseignement supérieur et par de nombreuses institutions académiques dans leur communication.

Pourtant, presque immédiatement après la sortie du premier classement, cette entreprise a fait l'objet de vives attaques. Une des premières est due à van Raan (2005a), qui a entamé un échange vigoureux avec les auteurs du classement (Liu et al., 2005, van Raan, 2005b). Depuis, les critiques ont été nombreuses et vigoureuses, à la fois dans la littérature académique (Buela-Casal et al., 2007, Dill et Soo, 2005, Gingras, 2008, Ioannidis et al., 2007, van Raan, 2006, Vincke, 2009, Zitt et Filliatreau, 2006) et dans des rapports et autres documents de travail (Bourdin, 2008, Brooks, 2005, Dalsheimer et Despréaux, 2008, Desbois, 2007, HEFCE, 2008, Kävelmark, 2007, Kivinen et Hedman, 2008, Marginson, 2007, Saisana et D'Hombres, 2008, Stella et Woodhouse, 2006). De plus, plusieurs numéros spéciaux du journal *Higher Education in Europe* (publié par l'OCDE) ont été consacrés aux discussions sur les classements des universités (HEE 2002, HEE 2005, HEE 2007, HEE 2008). En réaction à ces critiques, on aurait pu s'attendre à une baisse rapide de la popularité du classement de Shangaï. Cela aurait également pu inciter les auteurs du classement à stopper cette publication. Tout le contraire s'est produit. Chaque année, une nouvelle édition du classement est publiée et sa couverture médiatique semble augmenter. Des projets visant à transformer le système d'enseignement supérieur font souvent référence au classement de Shangaï. Par exemple, l'actuelle Ministre française de l'Enseignement Supérieur et de la Recherche en France s'est vue confiée en 2007 par le Premier Ministre et le Président de la République, la mission consistant en « l'amélioration du rang de nos

1. Depuis 2007, les auteurs du classement de Shangaï proposent un classement des institutions en distinguant cinq champs disciplinaires (Sciences naturelles et Mathématique, Ingénierie / Technologie et Informatique, Sciences de la vie et Agriculture, Médecine et Pharmacie, Sciences sociales), voir `http://www.arwu.org/`. Depuis 2009, ils proposent de plus un classement des institutions en 5 domaines (Mathématique, Physique, Chimie, Informatique, Économie / Gestion). Dans la mesure où la méthodologie utilisée pour ces classements par discipline et par domaine est similaire à celle utilisée pour le classement global, nous ne l'analyserons pas dans ce texte.

établissements d'enseignement supérieur dans les classements internationaux, avec l'objectif de classer au moins deux établissements français parmi les 20 premiers et 10 parmi les 100 » [2].

Le classement de Shangaï cherche apparemment à répondre à la question suivante : « Quelle est la meilleure université au monde ? ». Pour certains de nos lecteurs, cette question pourra sembler assez peu sérieuse et donc de peu d'intérêt. Néanmoins, ces lecteurs doivent garder à l'esprit qu'il peut exister dans le paysage politique des décideurs « paresseux » qui pourraient vouloir utiliser les résultats d'un classement simplement car « ils existent ». Des décideurs plus stratèges pourraient souhaiter utiliser ces résultats afin de promouvoir leur propre vision d'une meilleure organisation du système d'enseignement supérieur [3]. Enfin, il faut savoir — comme c'est souvent le cas avec les outils de gestion — que la seule existence d'un classement va contribuer à modifier le comportement des acteurs impliqués, créant des changements qui ne sont parfois pas désirés. C'est pourquoi nous pensons qu'il n'est pas inutile de passer un peu de temps sur la question.

Ce texte se veut être une contribution à l'analyse des forces et des faiblesses du classement de Shangaï. Notre point de vue sera celui de spécialiste de la Recherche Opérationnelle ayant travaillé dans le domaine des modèles d'évaluation et de décision en présence de plusieurs critères (Bouyssou et al., 2000, 2006, T'kindt et Billaut, 2006), alors que la plupart des analyses précédentes du classement de Shangaï se sont concentrées sur des aspects bibliométriques, sous l'impulsion de l'importante contribution de van Raan (2005a). Ce texte est organisé de la façon suivante. La section 2 décrit brièvement le fonctionnement du classement de Shangaï. La section 3 est consacrée à l'étude des critères utilisés. La section 4 analyse la méthode d'agrégation utilisée au prisme de l'aide multicritère à la décision. Une dernière section présente nos conclusions.

2. Voir `http://www.elysee.fr/elysee/elysee.fr/francais/interventions/2007/juillet/lettre_de_mission_adressee_a_mme_valerie_pecresse_ministre_de_l_enseignement_superieur_et_de_la_recherche.79114.html`, lettre de mission du 5 juillet 2007.

3. Citons ici la Ministre de l'Enseignement Supérieur et de la Recherche lors du débat au Sénat en juillet 2007 qui a amené à l'adoption de la LRU, relative aux Libertés et Responsabilités des Universités : « M. [. . .] a rappelé que le classement de Shanghai était certes critiquable mais que puisqu'on ne pouvait changer les indicateurs dont nous n'étions pas maîtres, il valait mieux les retourner en notre faveur. Lorsqu'ils choisissent leur future université, les étudiants américains, australiens, chinois, indiens regardent ce classement. C'est la mondialisation. On ne peut s'en abstraire et nous devons donc gagner des places, ce qui n'est pas contraire à l'exigence d'excellence de l'université française. », voir `http://www.senat.fr/cra/s20070711/s20070711_8.html`.

2 Comment fonctionne le classement de Shangaï ?

Cette section décrit brièvement la manière dont procèdent les auteurs du classement d'après ARWU (2003–09) et Liu et Cheng (2005). Nous nous concentrons sur la dernière édition du classement publié en novembre 2009 [4], bien que la méthodologie ait évolué dans le temps.

2.1 Qui sont les auteurs du classement ?

Les auteurs du classement sont un petit groupe de personnes appartenant à l'Institut de l'Enseignement Supérieur de l'Université Jiao Tong de Shangaï, groupe dirigé par le Professeur Nian Cai Liu. Les auteurs du classement reconnaissent (Liu et al., 2005, p. 108) qu'ils n'avaient au départ de leur travail aucune connaissance particulière en bibliométrie (Nian Cai Liu est un chimiste spécialisé dans les polymères). Selon Liu (2009), les auteurs du classement ont commencé à travailler sur le classement des universités en 1998, à la suite d'une impulsion donnée par le gouvernement chinois (le « projet 985 » mentionné dans Liu, 2009).

Les auteurs du classement insistent (Liu et Cheng, 2005, p. 135) sur le fait qu'ils ne reçoivent aucun financement particulier pour réaliser le classement et qu'ils ne sont guidés que par des considérations académiques. La situation est différente avec d'autres classements, comme celui produit par le *Times Higher Education Supplement* (Times Higher Education Supplement, 2008).

2.2 Quels sont leurs objectifs ?

Au cours des années 1990, la question de la rénovation du système d'enseignement supérieur en Chine est devenue une question importante. Ceci n'est pas surprenant, compte tenu de la situation politique de la Chine et de l'augmentation de sa puissance économique.

L'objectif annoncé des auteurs du classement est de bâtir un outil permettant de comprendre et d'analyser l'écart existant entre les universités chinoises et les universités dites de « classe mondiale », avec le souci très légitime de réduire cet écart. Les auteurs du classement ne donnent pas de définition précise de ce qu'ils entendent par « université de classe mondiale ». En raison de la difficulté d'obtenir des données comparables sur le plan international, ils ont décidé de classer les universités sur la seule base de leurs performances académiques (Liu et Cheng, 2005, p. 133).

4. Voir en particulier `http://www.arwu.org/Methodology2009.jsp`.

2.3 Comment les universités ont-elles été sélectionnées ?

Les auteurs du classement indiquent (Liu et Cheng, 2005, p. 127–128) avoir analysé environ 2000 institutions du monde entier. Ceci est supposé inclure toutes les institutions ayant reçu des Prix Nobel et des Médailles Fields, celles ayant un nombre significatif de publications dans *Nature* ou *Science*, ayant des personnes référencées par *Thomson Scientific* (plus connu sous son ancien nom de *ISI*) comme étant des « *Highly Cited Researchers* », ayant un nombre significatif de publications indexées dans les bases de données de *Thomson Scientific*. Les auteurs du classement indiquent que cela inclut les plus grandes universités de chaque pays. Le classement qui est publié ne comprend que 500 institutions. Les 100 premières sont classées. Les suivantes sont classées par tranches de 50 (jusqu'au rang 201) puis de 100 [5].

2.4 Les critères utilisés

Les auteurs du classement utilisent six critères regroupés en quatre domaines.

2.4.1 Qualité de l'enseignement

Ce domaine n'utilise qu'un seul critère : le nombre d'anciens élèves de l'institution ayant reçu un prix Nobel (sauf Paix et Littérature et y compris le prix de la Banque de Suède en Science Économique) ou une médaille Fields. Un ancien élève est défini comme étant une personne ayant reçu une licence, un master ou un doctorat dans cette institution. Si un lauréat a reçu un diplôme de plusieurs institutions, chacune en reçoit une part. Rappelons que les prix Nobel sont attribués (chaque année) depuis 1901 et que les médailles Fields sont attribuées (tous les quatre ans) depuis 1936. Tous les prix et toutes les médailles n'ont pas le même poids : ils sont « actualisés » suivant un schéma linéaire (une récompense après 1991 compte 100%, une récompense entre 1981 et 1990 compte pour 90%, etc.). Quand plusieurs personnes reçoivent un prix ou une médaille, chaque institution en reçoit une part [6]. Ceci définit le premier critère, noté ALU.

5. Pour la bonne interprétation des résultats donnés sur le site `http://www.arwu.org` en 2009, il faut garder en tête que les universités sont citées au delà du rang 100 au sein de chaque « bloc » par ordre *alphabétique* selon le nom anglais adopté par les auteurs du classement. Ce point semble avoir été oublié par certains commentateurs de la version 2009 du classement dans la presse grand public.

6. Pour ce qui concerne les prix Nobel, le comité précise le pourcentage revenant à chaque lauréat.

2.4.2 Qualité du corps académique

Ce domaine comporte deux critères. Le premier compte le nombre de membres du « corps académique » de l'institution ayant reçu un prix Nobel ou une médaille Fields. Les conventions utilisées pour déclarer qu'une personne fait partie du personnel académique restent floues. On actualise les récompenses de la façon suivante : 100% pour les lauréats après 2001, 90% pour les lauréats entre 1991 et 2000, 80% pour les lauréats entre 1981 et 1990, etc. Si le prix est attribué à plusieurs lauréats, il est partagé comme pour le critère ALU. Quand une personne a plusieurs affiliations, chaque institution en reçoit une part (il semble que la répartition se fait équitablement entre les institutions des lauréats). Ceci définit le critère AWA.

Le second critère de ce domaine est le nombre de personnes appartenant au personnel académique de l'institution figurant dans la liste des chercheurs les plus cités, dans chacun des 21 domaines de la Science tels que donnés par *Thomson Scientific*. Dans chaque domaine, ces chercheurs forment une liste de 250 personnes[7] dont les travaux ont reçu le plus grand nombre de citations dans les bases de données de *Thomson Scientific* (source `http://hcr3.isiknowledge.com/popup.cgi?name=hccom`.). Pour ce calcul, les citations sont comptabilisées sur une période de 20 ans. Ceci définit le critère HiCi (« *Highly Cited* »).

2.4.3 Production scientifique

Ce domaine comprend deux critères. Le premier est le nombre d'articles publiés dans *Nature* et *Science* par les membres du corps académique de l'institution pendant les cinq dernières années. Cela soulève le problème des articles ayant plusieurs auteurs. La règle dans ce cas est d'attribuer le poids de 100% à l'affiliation de l'auteur « correspondant » (celui qui a soumis l'article), 50% à l'affiliation du premier auteur (celle du deuxième auteur si le premier auteur est aussi l'auteur correspondant), 25% pour l'affiliation du prochain auteur, et 10% pour les affiliations des autres auteurs. Ceci définit le critère N&S. Pour les institutions spécialisées en Sciences Humaines et Sociales (SHS), ce critère est jugé peu significatif et il est « neutralisé ».

Le second critère compte le nombre d'articles publiés par les membres du personnel académique de l'institution. Ce comptage est effectué en utilisant les bases de données de *Thomson Scientific* (*Science Citation Index-Expanded* et *Social Science Citation Index* en retenant seulement les publications de type « *Article* » et « *Proceeding Paper* ») sur une période d'un an. Puisqu'il est bien connu que ces bases de données ne couvrent pas de façon satisfaisante les SHS, un coefficient de

7. Dans certains domaines, *Thomson Scientific* donne plus de 250 noms. Nous supposons que c'est en raison d'ex aequo.

2 est attribué à chaque publication indexée dans le *Social Science Citation Index*. Ceci définit le critère PUB.

2.4.4 Productivité

Ce dernier domaine n'a qu'un seul critère. Il s'agit du score total des cinq premiers indicateurs divisé par la taille (en équivalent temps plein) du corps académique [8] de l'institution (Liu et Cheng, 2005, p. 129). Ce critère est « ignoré » quand ce nombre n'a pas pu être obtenu [9]. Ceci définit le critère PY.

2.5 Source des données

Hormis pour la taille du corps académique de chaque institution, les données sont collectées sur Internet. Cela inclut le site officiel des prix Nobel (`http://nobelprize.org/nobel_prizes/`), le site officiel des médailles Fields de la *International Mathematical Union* (`http://www.mathunion.org/general/prizes`) et plusieurs sites de *Thomson Scientific* (`http://www.isihighlycited.com` et `http://www.isiknowledge.com`). Les auteurs du classement ne donnent pas leurs sources exactes pour ce qui concerne la taille du corps académique de chaque institution [10]. Les données brutes utilisées par les auteurs du classement ne sont pas communiquées.

2.6 Normalisation et agrégation

Une institution est repérée sur chacun des six critères par un nombre positif. Chaque critère est alors normalisé de la façon suivante. Un score de 100 est attribué à l'institution ayant le meilleur score, les autres reçoivent un score proportionnel. Cela conduit à un score entre 0 et 100 pour chaque institution et pour chaque critère. Ces valeurs sont, elles, fournies par les auteurs du classement.

Les auteurs du classement indiquent (Liu et Cheng, 2005, p. 129) que des ajustements sont réalisés, quand l'analyse révèle des « anomalies statistiques ». La

8. Notre traduction de « number of full-time equivalent academic staff ».

9. Dans ARWU (2003–09), les auteurs du classements indiquent que : « ce nombre a pu être obtenu pour les institutions des pays suivants : USA, UK, France, Canada, Japon, Italie, Chine, Australie, Pays-Bas, Suède, Suisse, Belgique, Corée du Sud, République Tchèque, Slovénie, Nouvelle Zélande, etc. », notre traduction de l'anglais. Nous ignorons si cela signifie que ce nombre a pu être obtenu pour *toutes* les institutions de ces pays.

10. Plus précisément, les auteurs mentionnent dans ARWU (2003–09) que ce nombre a été obtenu d'institutions telles que « le ministère de l'enseignement supérieur, le service national de statistiques, l'association nationale des universités, la conférence nationale des recteurs », notre traduction de l'anglais.

nature et l'objet de ces ajustements ne sont pas rendus publics. Florian (2007) montre que ces ajustements sont néanmoins importants.

Les auteurs du classement utilisent ensuite une somme pondérée pour agréger les scores normalisés. Les poids des six critères sont ALU : 10%, AWA : 20%, N&S : 20%, HiCi : 20%, PUB : 20%, et Py : 10%. Ainsi, chaque institution reçoit un score « global » entre 0 et 100. Ces scores sont normalisés, de sorte que la meilleure institution reçoive le score de 100. Ce score final normalisé est utilisé pour classer les institutions [11].

2.7 Les résultats de 2009

Le tableau 1 présente la liste des 20 meilleurs universités mondiales d'après l'édition 2009 du classement de Shangaï. Le tableau 2 présente la liste des 20 meilleures universités européennes.

Un examen rapide du tableau 1 révèle une domination flagrante des universités américaines dans le classement : seules 3 institutions parmi les 20 premières ne sont pas américaines. Ceci explique peut-être pourquoi autant de décideurs européens ont réagi fortement au classement de Shangaï. En Europe, la domination du Royaume-Uni est manifeste. De « petits » pays comme la Suisse, la Norvège, la Suède ou les Pays-Bas, semblent mieux tirer leur épingle du jeu dans le classement que de plus grands pays comme l'Italie ou l'Espagne.

La figure 1 présente la distribution des scores globaux normalisés pour les 500 institutions répertoriées dans le classement de Shangaï. Notons que la courbe devient très plate dès que les 100 premières institutions sont dépassées.

Concluons cette brève présentation du classement de Shangaï par trois affirmations extraites de Liu et Cheng (2005, p. 135). Le classement de Shangaï « *utilise des critères objectifs sélectionnés avec soin* », « *est fondé sur des données internationales comparables que tout le monde peut vérifier* » et « *ne comporte aucune mesure subjective* » [12].

3 Une analyse critique des critères utilisés

Nous commençons notre analyse du classement de Shangaï par un examen détaillé des six critères utilisés. Cette analyse est principalement une synthèse de la, déjà importante, littérature à ce sujet mentionnée à la section 1. En particulier, on s'appuiera sur les travaux de van Raan (2005a) et de Ioannidis et al. (2007). Cependant, tandis que cette littérature s'appuie principalement sur des considérations

11. Il est clair que cette dernière normalisation n'a pas d'impact sur le classement.

12. Notre traduction de « *carefully selected objective criteria* », « *based on internationally comparable data that everyone can check* » et « *[is such that] no subjective measures were taken* »

Rang	Institution	Pays	ALU	AWA	HiCi	N&S	PUB	Py	Score
1	Harvard	USA	100,0	100,0	100,0	100,0	100,0	74,1	100,0
2	Stanford	USA	40,0	78,7	86,6	68,9	71,6	66,9	73,7
3	UC Berkeley	USA	69,0	77,1	68,8	70,6	70,0	53,0	71,4
4	Cambridge	UK	90,3	91,5	53,6	56,0	64,1	65,0	70,4
5	MIT	USA	71,0	80,6	65,6	68,7	61,6	53,9	69,6
6	CalTech	USA	52,8	69,1	57,4	66,1	49,7	100,0	65,4
7	Columbia	USA	72,4	65,7	56,5	52,3	70,5	46,6	62,5
8	Princeton	USA	59,3	80,4	61,9	40,5	44,8	59,3	58,9
9	Chicago	USA	67,4	81,9	50,5	39,5	51,9	41,3	57,1
10	Oxford	UK	59,0	57,9	48,4	52,0	66,0	45,7	56,8
11	Yale	USA	48,5	43,6	57,0	55,7	62,4	48,7	54,9
12	Cornell	USA	41,5	51,3	54,1	52,3	64,7	40,4	54,1
13	UC Los Angeles	USA	24,4	42,8	57,4	48,9	75,7	36,0	52,4
14	UC San Diego	USA	15,8	34,0	59,7	53,0	66,7	47,4	50,3
15	U Pennsylvania	USA	31,7	34,4	58,3	41,3	69,0	39,2	49,0
16	U Wash Seattle	USA	25,7	31,8	53,1	49,5	74,1	28,0	48,3
17	U Wisc Madison	USA	38,4	35,5	52,6	41,2	68,1	28,8	47,4
18	UC San Francisco	USA	0,0	36,8	54,1	51,5	60,8	47,5	46,6
19	Tokyo Univ	Japon	32,2	14,1	43,1	51,9	83,3	35,0	46,4
20	Johns Hopkins	USA	45,8	27,8	41,3	48,7	68,5	24,8	45,5

Tableau 1 – Les 20 meilleures universités au monde dans le classement de Shangaï (2009). Source : ARWU (2003–09).

Rang	Institution	Pays	ALU	AWA	HiCi	N&S	PUB	Py	Score
4	Cambridge	UK	89,4	91,5	53,8	53,9	65,4	65,5	70,2
10	Oxford	UK	57,6	57,9	48,9	49,8	66,1	45,7	56,3
21	U Coll London	UK	30,4	32,2	40,4	45,2	66,0	36,1	44,6
23	Swiss Fed Inst Tech - Zurich	Suisse	35,0	36,3	35,9	41,8	52,6	57,0	43,6
26	Imperial Coll	UK	18,1	37,4	40,4	35,9	61,7	39,4	41,9
40	U Paris 06	France	35,7	23,6	22,9	27,7	59,3	21,7	33,3
41	U Manchester	UK	23,8	18,9	28,1	28,5	58,3	29,9	33,0
43	U Copenhagen	Danemark	26,8	24,2	26,2	24,8	54,5	33,2	32,7
43	U Paris 11	France	32,5	46,2	14,5	20,1	50,3	23,9	32,7
50	Karolinska Inst Stockholm	Suède	26,8	27,3	31,6	19,4	49,6	25,8	31,7
52	U Utrecht	Pays-Bas	26,8	20,9	28,1	30,5	48,2	25,1	31,5
53	U Edinburgh	UK	19,7	16,7	27,1	32,2	50,1	30,5	31,0
54	U Zurich	Suisse	10,9	26,8	24,6	28,5	49,0	32,5	30,9
55	U Munich	Allemagne	32,3	22,9	16,2	24,5	52,5	31,4	30,4
57	TU Munich	Allemagne	40,1	23,6	25,1	18,1	45,2	30,4	30,2
61	U Bristol	UK	9,5	17,9	29,0	28,6	46,7	33,7	29,5
63	U Heidelberg	Allemagne	17,3	27,2	17,8	22,5	49,0	29,5	28,7
65	King's College	UK	14,5	23,1	29,0	15,0	49,2	30,4	28,5
65	U Oslo	Norvège	22,5	33,4	17,8	16,1	45,6	29,2	28,5
70	ENS Paris	France	52,1	24,5	12,6	17,9	27,6	57,1	28,1

Tableau 2 – Les 20 meilleures universités européennes dans le classement de Shangaï (2009). Source : ARWU (2003–09).

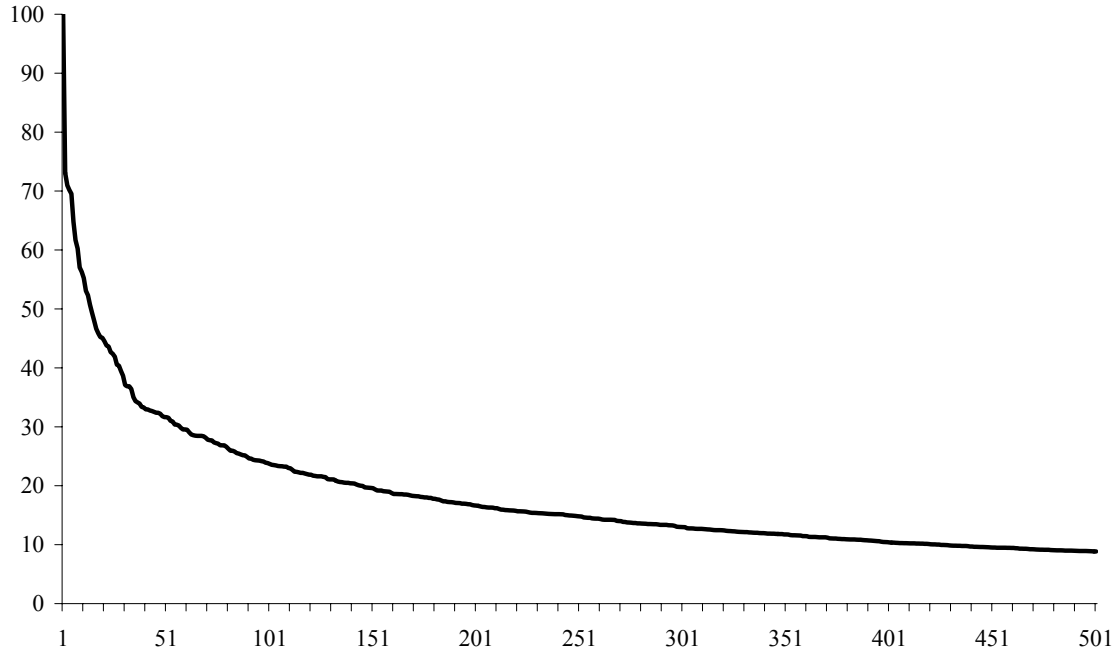

FIGURE 1 – Distribution des scores normalisés des 500 universités du classement de Shangaï (2009). Source : ARWU (2003–09).

bibliométriques, on fera également référence ici à la littérature en aide multicritère à la décision concernant le structuration des objectifs, et la construction de critères.

3.1 Critères relatifs aux prix Nobel et médailles Fields

Les critères ALU et AWA comptent des prix Nobel et des médailles Fields. Ces deux critères sont particulièrement problématiques.

Observons tout d'abord que pour le critère AWA, les prix Nobel et les médailles Fields sont attribués à l'institution à laquelle les lauréats appartiennent *au moment de l'annonce de la récompense*. Ce n'est pas un problème majeur pour les médailles Fields qui ne récompensent que des personnes de moins de 40 ans. C'est, en revanche, une difficulté sérieuse pour les prix Nobel. Un examen minutieux de la liste de ces prix révèle en effet que la règle générale consiste à récompenser une personne longtemps après que la recherche qui fait l'objet de la récompense a été conduite. Un exemple classique de cela est Albert Einstein : il a conduit sa recherche alors qu'il était employé par l'office suisse des brevets à Zurich ; il a reçu le prix Nobel longtemps après, alors qu'il était affilié à l'université de Berlin (nous reviendrons plus loin sur le cas d'Albert Einstein). De fait, il ne semble pas injustifié

de dire que le lien entre AWA et la qualité de la recherche conduite dans l'institution semble, au mieux, très approximatif. Même en supposant que le lauréat n'a pas changé d'institution, le temps qui sépare l'époque à laquelle était conduite la recherche et celui de l'annonce du prix, est tel que ce critère reflète davantage les qualités passées de l'institution que son potentiel actuel de recherche. Cette remarque est également vraie pour le critère ALU. Au mieux, ce critère indique que l'institution a eu, à une époque, la possibilité d'offrir à ses étudiants excellents un environnement particulièrement stimulant. Mais cela a très peu de liens avec l'aptitude qu'aurait une institution à délivrer *actuellement* un enseignement d'un niveau excellent à ses étudiants.

On peut aussi se demander dans quelle mesure les prix Nobel et les médailles Fields attribués depuis très longtemps (avant la seconde guerre mondiale et même parfois avant la première guerre mondiale) ont à voir avec la qualité actuelle des institutions. Même si le mécanisme d'actualisation adopté par les auteurs du classement tend à réduire l'impact de ces « vieux » prix et de ces « vieilles » médailles, l'impact existe. Soulignons ici que le mécanisme d'actualisation adopté est complètement arbitraire (pourquoi utiliser un schéma linéaire et non pas exponentiel ?).

En intégrant des prix et des médailles attribuées il y a longtemps, on introduit un biais en faveur des pays ayant connu très peu de changements politiques (de tels changements conduisent souvent à des reconfigurations du paysage académique) et des institutions ayant été créées depuis longtemps et ayant conservé la même appellation depuis leur création. Ceci a conduit parfois à des situations absurdes. Par exemple, les deux universités *Free University of Berlin* et *Humboldt University* créées à Berlin après la partition de l'Allemagne et, donc, de l'Université de Berlin, se sont querellées pour savoir laquelle devait recevoir l'attribution du prix Nobel d'Albert Einstein (voir Enserink, 2007). Selon le choix (arbitraire) d'affectation de ce prix à l'une ou à l'autre de ces deux institutions, leurs positions dans le classement étaient très différentes. Pour ce qui concerne la France, le site officiel des prix Nobel (`http://nobelprize.org/nobel_prizes/`) révèle plusieurs faits intéressants. De nombreux récipiendaires sont considérés comme appartenir à une institution qui n'a jamais existé (par exemple, l'« université Sorbonne », pour ce qui concerne Henri Moissan, Chimie 1906, Gabriel Lippmann, Physique 1908, Marie Curie, Chimie 1911, Charles Richet, Médecine 1913, ou Jean Perrin, Physique 1925). De nombreux autres appartiennent à une institution qui a existé mais n'existe plus, par exemple l'université de Nancy (Victor Grignard, Chimie 1912), l'université de Toulouse (Paul Sabatier, Chimie 1912), l'université de Grenoble (Louis Néel, Physique 1970) ou l'université de Paris (Jean Dausset, Médecine 1980) [13]. Une analyse détaillée de ces affiliations problématiques dans le cas de la

13. Mentionnons ici un exemple supplémentaire de ces problèmes. Le dernier lauréat français du prix Nobel de Physique, Albert Fert, est Professeur à l'Université de Paris Sud. Il travaille

France est donnée dans Billaut et al. (2009).

Tout ceci montre que l'attribution correcte des prix nécessite une connaissance fine du paysage institutionnel de chaque pays et la prise de multiples « micro-décisions ». Dans la mesure où les auteurs du classement ne fournissent aucune information sur ces micro-décisions, le moins que l'on puisse dire, c'est que ces deux critères fournissent, au mieux, des ordres de grandeur très approximatifs.

Enfin, comme un arbitre nous l'a signalé, ces deux critères sont fondés sur des récompenses qui ne couvrent pas, loin s'en faut, l'ensemble des distinctions scientifiques de très grand prestige. Des récompenses telles que la « *A. M. Turing Award* » [14] dans le domaine de l'Informatique ou la « *Bruce Gold Medal* » [15] dans le domaine de l'Astronomie, sont deux exemples de très prestigieuses récompenses qui sont ignorées par les auteurs du classement.

En résumé, les critères ALU et AWA ont un lien très lâche avec ce qu'ils sont censés mesurer. De plus, leur évaluation fait apparaître des paramètres arbitraires et soulève des problèmes sérieux au niveau du comptage. L'évaluation de ces critères est clairement entachée d'incertitude, d'imprécision et de mauvaise détermination (sur ces notions, voir Bouyssou, 1989, Roy, 1988).

3.2 Les chercheurs les plus cités

Comme mentionné par van Raan (2005a), le fait le plus saisissant est ici le fait que ce critère soit entièrement dépendant des choix faits par *Thomson Scientific*. Le découpage de la science en 21 domaines proposé par *Thomson Scientific* est-il pertinent ? Au vu du tableau 3, il semble que le choix de ces 21 domaines favorise nettement la médecine et la biologie. C'est très certainement une option raisonnable pour une entreprise commerciale comme *Thomson Scientific*, puisque ces domaines génèrent de nombreux papiers sur des sujets sensibles et médiatiques. Que ce choix soit approprié pour évaluer des universités mériterait d'être discuté et les auteurs du classement restent silencieux sur ce point.

D'autre part, comme souligné par van Raan (2005a), ces 21 catégories n'ont pas la même taille. En comptant le nombre de journaux dans chaque catégorie (notons toutefois que les journaux peuvent avoir des tailles variables en termes de nombre d'articles publiés ou en nombre de pages), on constate de grandes différences. *Space Science* contient 57 journaux, *Immunology* 120, ..., *Plant & Animal Science* 887,

dans une unité mixte de recherche (CNRS et Université de Paris Sud). Il semble qu'à cause de cela, la moitié seulement de son prix Nobel aille à son université (Fert, 2007), ce qui est très surprenant.

14. Attribuée chaque année depuis 1966 par la *Association for Computing Machinery*, voir `http://awards.acm.org/homepage.cfm?awd=140`.

15. Attribuée chaque année depuis 1898 par la *Astronomical Society of the Pacific*, voir `http://www.phys-astro.sonoma.edu/bruceMedalists/`.

Agricultural Sciences	*Materials Science*
Engineering	*Plant & Animal Science*
Neuroscience	*Computer Science*
Biology & Biochemistry	*Mathematics*
Geosciences	*Psychology / Psychiatry*
Pharmacology	*Ecology / Environment*
Chemistry	*Microbiology*
Immunology	*Social Sciences, General*
Physics	*Economics & Business*
Clinical Medicine	*Molecular Biology & Genetics*
Space Sciences	

TABLEAU 3 – Les 21 catégories utilisées par *Thomson Scientific* (source : `http://www.isihighlycited.com/isi_copy/Comm_newse04.htm`)

.

Engineering 977, *Social Science General* 1299, et *Clinical Medicine* 1305 [16].

Ce critère utilise clairement les comptages de citations réalisés par *Thomson Scientific*. Les spécialistes de bibliométrie ont souvent souligné que ces comptages manquaient de précision en raison de « pertes » de citations (dues à des appellations erronées, des numéros de pages inexacts, des problèmes d'homonymes, etc.). Van Raan (2005a) évalue la perte moyenne de citations à 7%, celle-ci pouvant monter jusqu'à 30% dans certains domaines. Étrangement, dans les réponses de Liu et al. (2005) à ces critiques, les auteurs du classement ne reconnaissent pas que ce critère utilise des comptages de citations.

Enfin, notons que *Thomson Scientific* utilise une période de 20 ans pour déterminer les noms des chercheurs les plus cités par catégorie. Dans la plupart des catégories, les personnes dans les listes ne sont donc pas particulièrement jeunes et sont très susceptibles d'avoir changé d'institution, parfois plusieurs fois, dans leur carrière.

En résumé, si on combine la dépendance de ce critère au découpage de la Science en 21 domaines par *Thomson Scientific*, l'utilisation d'une période de 20 ans et les difficultés inhérentes au comptage, on peut constater que ce critère est lié de façon extrêmement approximative à la capacité d'une institution à produire actuellement une recherche de fort impact.

16. Source : `http://www.isihighlycited.com/isi_copy/Comm_newse04.htm`.

3.3 Articles dans *Nature* et *Science*

Au delà de la question du seul choix de ces deux revues (il existe des revues ayant des « facteurs d'impact » plus élevés), le fait probablement le plus surprenant avec ce critère est le mode de pondération des auteurs lorsque les articles sont co-signés (rappelons que c'est la règle générale en « sciences dures »). Avec 100% pour l'affiliation de l'auteur correspondant, 50% pour le premier auteur s'il n'est pas l'auteur correspondant, 25% pour l'affiliation de l'auteur suivant et 10% pour les affiliations des autres auteurs, on voit rapidement que les articles publiés dans *Nature* et *Science* n'ont pas tous le même poids. Un article signé par de nombreux auteurs aura un poids plus important qu'un article signé par une seule personne (il est donc de l'intérêt d'une institution que tout article publié dans *Nature* et *Science* soit co-signé par le plus possible de membres de l'institution). Ceci nous semble contre-intuitif, voire paradoxal.

3.4 Articles indexés par *Thomson Scientific*

Comme cela est souligné dans van Raan (2005a), les auteurs du classement font entièrement confiance pour l'évaluation de ce critère aux bases de données de *Thomson Scientific*. Cela soulève de nombreux problèmes.

Tout d'abord, comme nous l'avons vu plus haut à propos des critères ALU et AWA, l'attribution correcte des articles aux institutions est loin d'être une tâche aisée. Les auteurs du classement règlent ce problème en affirmant que « l'attribution s'effectue selon l'affiliation indiquée par les auteurs de l'article » [17] (Liu et Cheng, 2005, p. 134). Ceci n'est pas satisfaisant.

Il est bien connu que les auteurs ne portent pas toujours une très grande attention à la façon dont leur affiliation est indiquée quand ils publient un article. Le problème devient particulièrement critique quand il s'agit d'articles publiés par des centres hospitaliers universitaires (ils ont généralement une dénomination propre qui n'est pas celle de l'université, et ont une adresse distincte, voir van Raan, 2005a, Vincke, 2009). Le même phénomène se produit si l'université a un nom officiel qui n'est pas en anglais. Certains auteurs utiliseront le nom officiel (ce qui peut conduire à des problèmes s'il contient des signes diacritiques ou s'il s'agit d'un sigle), d'autres essaieront de le traduire en anglais, ajoutant à la confusion. Un exemple célèbre est la difficulté qu'il peut y avoir à distinguer l'*Université Libre de Bruxelles* de la *Vrije Universiteit Brussel*. Les deux sont situées à Bruxelles et ont le même code postal. Les deux ont le même nom en anglais *Free University of Brussels*. Ainsi, ce premier problème ajoute une grande imprécision à l'évaluation du critère PUB. Affecter à chaque auteur une affiliation correcte est une tâche

17. Notre traduction de « *institutions or research organizations affiliated to a university are treated according to their own expression in the author affiliation of an article* »

difficile nécessitant une connaissance détaillée des spécificités institutionnelles de chaque pays [18].

Ensuite, il est bien connu que la couverture des bases de données de *Thomson Scientific* est loin d'être parfaite (Adam, 2002). La nouvelle base de données SCOPUS créée par *Elsevier* a une couverture très différente, même si à l'évidence, l'intersection entre les deux bases de données n'est pas vide. Compter en utilisant *Thomson Scientific* au lieu de SCOPUS est un choix tout à fait légitime, à condition que l'impact de ce choix sur les résultats du classement soit analysé systématiquement, ce qui n'est pas fait par les auteurs du classement.

Il est également bien connu que la couverture des bases de données de citations a un biais important en faveur des publications en langue anglaise (voir van Leeuwen et al., 2001, van Raan, 2005a, pour une analyse de l'impact de ce biais sur l'évaluation des universités allemandes). D'autre part, il y a des disciplines (on peut penser au Droit) pour lesquelles des publications dans une langue qui n'est pas celle du pays n'a que peu de sens. De plus, il y a de grands pans de la science qui n'utilisent pas les articles publiés dans des journaux indexés pour diffuser leur recherche. Dans les SHS, par exemple, les livres restent le moyen de communication principal, tandis qu'en Informatique et en Ingénierie, ce sont les actes de congrès qui dominent. Les auteurs du classement ont tenté de corriger ce biais en défaveur des SHS en multipliant par 2 tous les papiers indexés dans *Social Science Citation Index*. Cela va sûrement dans la bonne direction. Mais il est également clair que ce coefficient est arbitraire et qu'une analyse de la robustesse des résultats à la valeur de ce coefficient devrait être menée.

Enfin, on peut aussi se demander pourquoi les auteurs du classement ont choisi de compter le nombre de papiers indexés, plutôt que d'essayer de mesurer l'impact de ces papiers. Une recherche rapide dans les bases de données de *Thomson Scientific* révèle que la plupart des articles indexés ne sont jamais cités et qu'un petit nombre d'entre eux concentrent toutes les citations, ceci étant vrai quelque soit le facteur d'impact du journal. La littérature en bibliométrie a très largement détaillé l'importance de la prise en compte de l'« impact » des recherches pour bâtir des indices pertinents de « qualité » (voir, par exemple, les travaux de Moed et al., 1995, Moed, 2006, van Raan, 1996, 2006, du *Center for Science and Technology Studies* de l'Université de Leiden, voir aussi `http://www.cwts.nl/ranking/LeidenRankingWebsite.html`).

En conclusion, le critère PUB soulève des problèmes importants et implique de nombreux choix arbitraires.

18. Clairement, ces problèmes affectent également les critères HiCi et N&S

3.5 Productivité

Le critère PY consiste dans le score total des 5 premiers indicateurs divisé par le nombre d'équivalents temps plein du corps académique. Ce critère est « ignoré » quand ce nombre ne peut pas être obtenu. Deux éléments doivent être soulignés ici.

Tout d'abord, ce critère est clairement impacté par tous les éléments d'imprécision, d'incertitude et de mauvaise détermination ainsi que par les paramètres arbitraires affectant les cinq premiers critères. De plus, les auteurs du classement utilisent des sources variées pour trouver l'information sur la taille du corps académique. Nous n'avons aucune raison de penser que ces sources sont plus fiables que celles utilisées pour les cinq premiers critères. De plus, cette variété des sources est problématique puisque la notion de « membre du corps académique » n'est pas définie précisément et qu'elle peut être interprétée de plusieurs façons (par exemple, doit-on compter les professeurs invités, les professeurs émérites, les chercheurs des EPST, les enseignants non chercheurs, les ingénieurs de recherche, etc. ?).

Ensuite, la façon dont les auteurs du classement tiennent compte du score total des cinq premiers indicateurs n'est pas totalement claire. S'agit-il des cinq scores normalisés ? Les scores sont-ils pondérés (nous pensons que c'est le cas) ? Avec quels poids (nous supposons que ce sont les mêmes poids que pour la somme globale, normalisés pour que leur somme fasse un) ?

3.6 Un nombre variable de critères

Les institutions sont évaluées dans le classement de Shangaï par six critères... mais pas toutes les institutions. En fait, il y a plusieurs cas possibles :

- les institutions non spécialisées en SHS et pour lesquelles le nombre de personnes équivalent temps plein a pu être obtenu sont évaluées sur les 6 critères : ALU, AWA, HICI, N&S, PUB et PY.
- les institutions non spécialisées en SHS et pour lesquelles le nombre de personnes équivalent temps plein n'a pas pu être obtenu sont évaluées sur 5 critères : ALU, AWA, HICI, N&S et PUB.
- les institutions spécialisées en SHS et pour lesquelles le nombre de personnes équivalent temps plein a pu être obtenu sont évaluées sur 5 critères : ALU, AWA, HICI, PUB et PY.
- les institutions spécialisées en SHS et pour lesquelles le nombre de personnes équivalent temps plein n'a pas pu être obtenu sont évaluées sur 4 critères : ALU, AWA, HICI et PUB.

Cela soulève bien des questions. Tout d'abord, en aide multicritère à la décision, on considère toujours que les objets à comparer sont évalués sur la *même* famille de critères. D'autre part, le meilleur moyen de « neutraliser » un critère n'est pas

complètement évident. Ensuite, les auteurs du classement ne rendent pas publique la liste des institutions qu'ils ont considéré comme étant spécialisées en SHS. Ils ne donnent pas non plus la liste des institutions pour lesquelles la taille du corps académique a pu être obtenue. En résumé, non seulement la famille des critères varie selon les institutions, mais il est impossible de savoir quelle famille de critères est utilisée pour évaluer une institution. Ceci est pour le moins inhabituel.

3.7 Une synthèse sur les critères

Nous avons vu que tous les critères utilisés par les auteurs du classement entretiennent des liens assez lâches avec ce qu'ils sont censés mesurer. Leur évaluation utilise des paramètres arbitraires et fait intervenir des micro-décisions qui ne sont pas documentées. Au regard de la figure 1, il semble clair que tous ces éléments impactent fortement la robustesse du classement. Malheureusement, dans la mesure où les auteurs du classement ne fournissent pas les données brutes (ce qui n'est pas franchement en accord avec les motivations académiques affichées), il est impossible d'analyser la robustesse du classement final relativement à ces éléments.

Nous avons vu précédemment que les auteurs du classement prétendent que leur travail « utilise des critères objectifs sélectionnés de façon objective », qu'il est « fondé sur des données internationalement comparables et que tout le monde peut vérifier », et qu'« il ne fait pas appel à des mesures subjectives ».

Il semble clair maintenant que les critères ont été choisis principalement sur la base de la disponibilité sur Internet des informations permettant de les renseigner, que chacun d'entre eux est lié de façon très approximative avec ce qu'il est censé mesurer et que leur évaluation fait intervenir des paramètres arbitraires et des micro-décisions non documentées. L'impact de ces éléments sur le résultat final n'est pas examiné. Les données initiales utilisées ne sont pas rendues publiques et donc ne peuvent pas être vérifiées.

Nous profitons de l'occasion pour rappeler au lecteur qu'il existe une vaste littérature sur le problème consistant à structurer des objectifs, à associer des critères aux objectifs, à juger de l'adéquation et de la cohérence d'une famille de critères. Cette littérature a deux sources principales. La première, issue de la littérature en psychologie (voir, par exemple, Ebel et Frisbie, 1991, Kerlinger et Lee, 1999, Kline, 2000), s'est concentrée sur la question de la *validité* et de la *fiabilité*. La seconde, issue de l'aide multicritère à la décision (voir, par exemple, Bouyssou, 1990, Keeney, 1992, Keeney et Raiffa, 1976, Keeney et al., 1999, Roy, 1996, Roy et Bouyssou, 1993, von Winterfeldt et Edwards, 1986), s'est intéressée à la structuration des objectifs et à la construction des critères pour mesurer l'atteinte de ces objectifs. Il semble que cette littérature ait été largement ignorée par les auteurs du classement.

3.8 Remarques finales sur les critères

Nous voudrions conclure cette analyse des critères utilisés dans le classement de Shangaï par quelques remarques additionnelles.

3.8.1 Effets du temps

Les auteurs ont choisi de publier leur classement chaque année. C'est probablement un bon choix si ce qui est visé principalement est une vaste couverture médiatique. Toutefois, compte-tenu de la durée de la plupart des programmes de recherche, nous ne pouvons pas trouver de justification sérieuse à une telle périodicité. Comme le souligne Gingras (2008), l'aptitude d'une université à produire d'excellents résultats de recherche ne change pas énormément d'une année à l'autre. La plupart des projets de recherche se déroulent sur plusieurs années et les contrats des établissements et des unités de recherche se font sur un horizon de 4 ans. Par conséquent, les variations entre une édition du classement et la suivante reflètent davantage des fluctuations aléatoires que de réels changements de fond (la recherche empirique en Psychologie a largement montré que les décideurs ont du mal à réaliser que des changements dus à des phénomènes aléatoires sont plus nombreux que les changements dus à des changements structurels, voir, par exemple, Bazerman, 1990, Dawes, 1988, Hogarth, 1987, Poulton, 1994, Russo et Schoemaker, 1989).

Le deuxième point lié au temps est relatif au choix adéquat d'une période de référence qui permettrait de mesurer la performance académique d'une institution. C'est une question difficile. Elle est résolue par les auteurs du classement d'une façon plutôt étrange. Ils ont mélangé dans le modèle plusieurs périodes de temps différentes : un siècle pour les critères ALU et AWA, 20 ans pour le critère HiCi, 5 ans pour le critère N&S et 1 an pour le critère PUB. Ces choix ne sont pas justifiés par les auteurs du classement. Comme le souligne van Raan (2005a), la « performance académique » peut avoir deux significations très différentes : d'une part le prestige d'une institution fondé sur ses performances passées et d'autre part sa capacité actuelle à attirer d'excellents chercheurs. Il est important de ne pas confondre ou mélanger ces deux aspects.

3.8.2 Effets de la taille

Cinq critères parmi les six utilisés par les auteurs du classement sont obtenus par des comptages (prix et médailles, chercheurs les plus cités, articles dans *Nature* et *Science*, articles indexés par *Thomson Scientific*). Il n'est donc pas surprenant que la valeur de ces critères soit fortement corrélée à la taille de l'institution. Comme le montrent Zitt et Filliatreau (2006), l'utilisation de critères liés à la taille de l'institution implique que « *big is made beautiful* ». Le fait que les critères

soient fortement corrélés n'est donc pas une surprise et, au contraire des auteurs du classement, il ne faut pas voir dans cette corrélation un signe de pertinence.

3.8.3 Pouvoir discriminants des critères

Puisque les critères utilisés par les auteurs du classement sont relatifs à l'« excellence académique », on doit s'attendre à ce qu'ils soient peu discriminants en dehors de la tête du classement. Une simple analyse statistique révèle que c'est le cas. Le tableau 4 donne la valeur moyenne des cinq premiers critères utilisés par le classement en fonction du rang des institutions (chiffres fondés sur le classement 2009). Il est clair que les critères ALU et AWA ne contribuent quasiment en rien au classement des institutions qui ne sont pas dans les 100 premières. De plus, les critères HiCi et N&S ont de très faibles valeurs en bas de la liste. Ainsi, pour ces institutions, le critère PUB explique quasiment tout. On est assez loin d'un véritable classement multicritère.

4 Le classement de Shangaï au prisme de l'aide multicritère à la décision

Dans la section précédente, nous avons proposé une analyse critique des critères utilisés par les auteurs du classement. Nous abordons maintenant les questions liées à la méthodologie utilisée par les auteurs du classement pour agréger ces critères.

4.1 Une introduction rhétorique

Supposons que vous donniez un cours de master sur l'aide multicritère à la décision. L'évaluation des étudiants consiste en un travail personnel : les étudiants doivent proposer et justifier une technique d'aide multicritère à la décision pour aborder un problème particulier. Le sujet donné cette année consiste à mettre au point une technique qui permettrait de classer les pays en fonction de leur « richesse ». Considérons les trois mémoires d'étudiants suivants.

Le premier étudiant a proposé une technique assez complexe qui présente la particularité suivante. Le fait que le pays a soit classé avant le pays b ne dépend pas uniquement des données collectées sur les pays a et b, mais aussi de ce qui se passe pour un troisième pays c. Même s'il est possible d'imaginer des situations où un tel phénomène pourrait se justifier (voir Luce et Raiffa, 1957, Sen, 1993), nous pensons que vous serez d'accord pour dire que cet étudiant « *a tout faux* ». Il est en effet difficilement concevable que le fait que a soit plus riche que b dépende de la richesse de c !

Rang	ALU	AWA	HiCi	N&S	PUB
1–100	25,56 (11%)	26,32 (16%)	35,76 (1%,16,15)	31,61 (0%,14,95)	54,00 (11,94)
101–200	8,25 (50%)	5,25 (69%)	18,02 (1%, 6,25)	16,73 (0%, 5,19)	40,83 (8,01)
201–302	5,24 (63%)	1,19 (91%)	10,58 (17%, 5,92)	12,05 (0%, 3,77)	35,64 (7,29)
303–401	3,08 (81%)	1,59 (91%)	8,42 (22%, 5,33)	7,99 (2%, 3,31)	28,90 (6,63)
402–501	0,83 (94%)	0,29 (97%)	5,24 (39%, 4,59)	6,32 (5%, 3,17)	26,99 (6,03)

TABLEAU 4 – Valeurs moyennes pour les critères en fonction des rangs (source : classement 2009). Pour les critères ALU et AWA, nous indiquons entre parenthèses le pourcentage d'institutions ayant un score nul. Pour les critères HiCi et N&S, nous indiquons entre parenthèses le pourcentage d'institutions ayant un score nul ainsi que l'écart-type du score. Pour le critère PUB, nous indiquons entre parenthèses l'écart-type du score. L'irrégularité apparente des rangs est due aux ex-aequo.

Le second étudiant propose une technique simple fonctionnant de la façon suivante. Pour chaque pays, il a collecté le PIB (Produit Intérieur Brut) et le PIBph (PIB par habitant). Il suggère de classer les pays en utilisant une somme pondérée du PIB et du PIBph pour chaque pays. Là aussi, nous pensons que vous serez d'accord pour dire que cet étudiant « *a tout faux* »: soit on veut mesurer la richesse totale d'un pays et on utilisera le PIB, soit on veut mesurer la richesse moyenne de ses habitants et on utilisera le PIBph. Mais une combinaison linéaire du PIB et du PIBph n'a pas de sens : le PIB mesure la *production*, le PIBph la *productivité*. Prendre α fois la production plus $(1-\alpha)$ fois la productivité est une opération qui est légitime sur le plan arithmétique mais qui n'a pas de sens, excepté bien entendu si α vaut 1 ou 0. Le lecteur qui ne serait pas entièrement convaincu par cet argument est invité à tester le résultat d'une telle opération en utilisant les informations sur le PIB et le PIBph des pays qui sont facilement disponibles sur Internet.

Considérons maintenant le mémoire du troisième étudiant qui a proposé un modèle complexe, mais qui :

- ne s'est pas interrogé sur la pertinence de la tâche,
- n'a pas réfléchi à ce qu'était « la richesse » et comment elle devait être mesurée,
- n'a pas réfléchi aux impacts potentiels de son travail,
- n'a utilisé que des informations facilement disponibles sur Internet, sans se poser de question sur leur pertinence et leur précision,
- a mélangé ces informations en utilisant des paramètres arbitraires, sans se poser la question de l'influence de ces paramètres sur les résultats.

Clairement, vous serez d'accord pour dire que cet étudiant « *a tout faux* ». Il est passé à côté de la difficulté du sujet, en le réduisant à un vulgaire exercice de « cuisine ».

Nous sommes désolés de dire que les auteurs du classement ne nous semblent pas être dans une situation meilleure que ces trois étudiants. Nous expliquons ci-après pourquoi nous pensons qu'ils ont, dans leur travail, réuni tout ce que nous avons trouvé inacceptable dans le travail de nos trois étudiants.

4.2 La technique d'agrégation est déficiente

Dans tout cours d'introduction au multicritère on enseigne le fait que si on agrège plusieurs critères en utilisant une somme pondérée, les poids utilisés *ne doivent pas* être interprétés comme s'ils reflétaient l'« importance » des critères. Cela peut sembler étrange à première vue mais est en fait très intuitif. Les poids (ou encore *constantes d'échelle* comme ils sont appelés en aide multicritère à la décision) sont étroitement liés à la *normalisation* des critères. Si la normalisation change, les poids doivent aussi changer. En effet, on peut choisir de mesurer un

critère en mètres ou en kilomètres. Si on utilise les mêmes poids pour ce critère dans les deux cas, on obtiendra évidemment des *résultats absurdes*.

Ceci a deux conséquences essentielles. Tout d'abord, les poids dans une somme pondérée ne peuvent pas être fixés sur la base d'une vague notion d'« importance » des critères. La comparaison des poids ne reflète pas une comparaison de leur importance. Si le poids d'un critère mesuré en mètres est 0,3, alors pour que ce critère garde la même « importance », ce poids doit être multiplié par 1 000 si on décide de le mesurer en kilomètres. Dès lors, la comparaison de ce poids avec les poids des autres critères ne reflète en aucune manière son importance relative. Ceci a des conséquences importantes sur la façon adéquate de définir les poids dans une somme pondérée (voir Bouyssou et al., 2006, Keeney et Raiffa, 1976). Quoiqu'il en soit, cela n'a aucun sens de demander à quelqu'un de fixer des poids sans référence claire à la normalisation choisie (comme les auteurs du classement le font sur leur site, voir `http://www.arwu.org/rank/2004/Questionnaire.htm`). Cela soulève également le problème sur la façon dont les auteurs du classement ont choisi leur jeu de poids. Ils n'expliquent rien à ce sujet. Tout porte à croire que les poids ont été fixés arbitrairement. La seule explication qui nous semble rationnelle est que, dans la première version du classement, les auteurs avaient utilisé uniquement cinq critères avec des poids égaux. Si l'utilisation de poids égaux peut se justifier dans certains cas (voir Einhorn et Hogarth, 1975), nous n'avons aucune raison de croire que ces raisons s'appliquent ici.

Ce qui précède a une conséquence plus importante encore. Si l'on change la normalisation des critères, on doit *absolument* changer les poids pour refléter cette modification. Si cela n'est pas fait, on court le risque d'obtenir des résultats qui n'ont aucun sens. Dans la mesure où tous les ans, les auteurs du classement normalisent les critères en donnant le score de 100 à la meilleure institution pour chaque critère, et dans la mesure où tous les ans, le score non normalisé de la meilleure institution change, les poids devraient changer tous les ans de manière à refléter cette nouvelle normalisation. Mais les auteurs du classement ne changent pas les poids pour refléter ces changements de normalisation [19]. À cause du changement de normalisation, ceci revient donc à utiliser chaque année des poids différents.

Illustrons ce qui peut se passer avec un exemple simple utilisant deux critères. Considérons les données du tableau 5. Dans ce tableau, on a 8 institutions a, b, c, d, e, f, g et h qui sont évaluées sur deux critères notés g_1 et g_2. Ces critères sont normalisés pour donner le score 100 à la meilleure institution (ici, h pour les deux critères, h pour « Harvard »). Ceci définit les deux critères normalisés g_1^n et g_2^n. Par exemple, nous avons $g_2^n(f) = 22 = (110 \times 100)/500$. Agrégeons ces deux critères par une somme pondérée utilisant des poids égaux. Cela donne la colonne « Score » de tableau 5 (il n'est pas nécessaire ici de normaliser à nouveau le score

19. Keeney (1992, p. 147) appelle ceci « *the most common critical error* ».

global puisque le score de h est déjà égal à 100). Si nous utilisons ce score global pour classer les institutions, nous obtenons le classement suivant ($a \succ b$ signifie que a est préférée à b) :

$$h \succ a \succ b \succ c \succ d \succ e \succ f \succ g.$$

Institutions	g_1	g_2	g_1^n	g_2^n	Score	Rang
h	2 000	500	100,00	100,00	100,0	1
a	160	435	8,00	87,00	47,5	2
b	400	370	20,00	74,00	47,0	3
c	640	305	32,00	61,00	46,5	4
d	880	240	44,00	48,00	46,0	5
e	1 120	175	56,00	35,00	45,5	6
f	1 360	110	68,00	22,00	45,0	7
g	1 600	45	80,00	9,00	44,5	8

Tableau 5 – Somme pondérée : exemple avec des poids identiques

Considérons maintenant une situation similaire dans laquelle tout reste inchangé, excepté la performance de h sur le critère g_2 qui passe de 500 à 700. Cela conduit aux données du tableau 6. Les deux critères sont à nouveau normalisés de sorte à donner 100 à la meilleure institution sur chaque critère (à nouveau ici, h pour les deux critères). Mais parce que le score de h sur g_2 a changé, cela a un impact sur *tous* les scores normalisés donnés par g_2^n. Si on décide d'agréger les deux critères normalisés en utilisant les mêmes poids que précédemment, on obtient le classement suivant :

$$h \succ g \succ f \succ e \succ d \succ c \succ b \succ a.$$

On observe alors que la modification du score de h sur g_2 a inversé le classement pour *toutes* les autres institutions ! Les raisons de ce paradoxe sont claires. Puisque le score de h sur g_2 a changé, la normalisation du critère g_2^n a également changé. La normalisation ayant changé, les poids associés aux critères devraient changer si l'on souhaite être cohérent.

Il est clair que l'on peut rendre cet exemple encore plus frappant. Supposons qu'à partir des données du tableau 5, on considère une situation où tout reste inchangé, sauf la performance de h qui augmente sur g_2 en passant de 500 à 700 et les performances de a s'améliorent à la fois sur g_1 (passant de 160 à 165) et sur g_2 (passant de 435 à 450). Le lecteur vérifiera sans peine que la position de a dans cette nouvelle configuration, au lieu de s'améliorer comme on pouvait légitimement s'y attendre, se détériore (a est maintenant placée en dernière position).

Institutions	g_1	g_2	g_1^n	g_2^n	Score	Rang
h	2 000	**700**	100,00	100,00	100,00	1
a	160	435	8,00	62,14	35,07	8
b	400	370	20,00	52,86	36,43	7
c	640	305	32,00	43,57	37,79	6
d	880	240	44,00	34,29	39,14	5
e	1 120	175	56,00	25,00	40,50	4
f	1 360	110	68,00	15,71	41,86	3
g	1 600	45	80,00	6,43	43,21	2

TABLEAU 6 – Somme pondérée avec des poids identiques : h s'améliore sur g_2

Puisque les auteurs du classement sont tombés dans ces pièges élémentaires liés à la normalisation, on peut légitimement s'interroger sur la validité de leurs résultats. En aucun cas, on ne peut conclure que « s'améliorer sur un critère permet de monter dans le classement ». Un argument qui pourrait venir à la rescousse des résultats du classement de Shangaï, consisterait à dire que les données sont telles que quelque soient les poids, les résultats sont toujours les mêmes. Au regard de la figure 1, il est clair qu'un tel argument ne s'applique pas ici.

Notons que le fait de ne pas changer les poids quand l'échelle change a un autre effet étrange. Si une institution b est faible sur un critère, de sorte qu'une institution concurrente a est classée juste devant elle, l'intérêt de l'institution b est que la meilleure institution sur ce critère augmente ses performances. En effet, si les poids restent inchangés, cette amélioration diminuera mécaniquement l'écart entre a et b, ce qui permettra éventuellement à b de passer devant a. Donc si une institution est faible sur un critère, son intérêt est que son écart avec la meilleure institution pour ce critère augmente !

Nous concluons par une dernière remarque sur la technique d'agrégation utilisée. Même si les auteurs du classement n'étaient pas tombés dans les pièges expliqués précédemment (il est possible de remédier simplement au problème de la normalisation évoqué plus haut), il faut souligner que la somme pondérée reste une méthode très peu attractive pour agréger des critères. La presque totalité de Bouyssou et al. (2000) est consacrée à des exemples qui expliquent pourquoi. Rappelons seulement ici le problème central de l'existence d'*alternatives efficaces non supportées*. Une alternative b (une institution dans notre cas) est dite dominée s'il existe une autre alternative qui a des évaluations au moins aussi bonnes sur tous les critères et strictement meilleures sur au moins un critère. Une alternative est dite efficace si elle n'est pas dominée. Clairement, toute alternative non dominée apparaît comme un candidat possible pour la tête de classement. Toute alternative non dominée devrait potentiellement pouvoir être classée en première position avec

un choix approprié des paramètres du modèle d'agrégation.

Malgré cela, en utilisant une somme pondérée des critères, il peut exister des alternatives non dominées qui ne pourront *jamais* être classées en première position, quelle que soit la valeur des poids. Le tableau 7 donne un exemple (extrait de Bouyssou et al., 2000) d'une telle situation, en utilisant deux critères à maximiser. Observons que les trois alternatives a, b et c sont non dominées (l'alternative d est clairement dominée par toutes les autres). Intuitivement, l'alternative c est un bon candidat pour être classée en première position : elle a une bonne performance sur les deux critères, tandis que a (b, respectivement) est excellente sur le critère 1 (2, respectivement) mais mauvaise sur l'autre critère. Toutefois, si on agrège les deux critères en effectuant une somme pondérée des critères, il n'est pas possible de trouver des poids qui permettront à l'alternative c d'être classée première. En effet, supposons qu'il existe de tels poids α et $(1-\alpha)$ qui le permettent. Classer c avant a implique que $11\alpha+11(1-\alpha) > 5\alpha+19(1-\alpha)$, c'est-à-dire, $\alpha > 8/15 \approx 0{,}53$. Classer c avant b implique $11\alpha+11(1-\alpha) > 20\alpha+4(1-\alpha)$, c'est-à-dire, $\alpha < 7/16 \approx 0{,}44$. La figure 2 illustre cette impossibilité, due au fait que c est dominé par une combinaison linéaire de a et b.

	g_1	g_2
a	5	19
b	20	4
c	11	11
d	3	3

Tableau 7 – Somme pondérée : alternatives efficaces non supportées

Des travaux récents en aide multicritère à la décision ont permis de mettre au point des techniques d'agrégation qui ne présentent pas de tels défauts (Belton et Stewart, 2001, Bouyssou et al., 2000).

4.3 L'agrégation réalisée n'a pas de sens

Les critères ALU, AWA, HiCi, N&S et PUB sont des critères de *comptage*. Sous réserve des remarques faites à la section 3, ils sont globalement relatifs à l'aptitude qu'a une institution à produire une grande quantité de bons papiers et de bons chercheurs. Leur agrégation est sémantiquement cohérente. En revanche, le critère Py est de nature totalement différente. Si les cinq premiers critères représentent la « production », le dernier représente la « productivité ». Mais le bon sens et une analyse économique élémentaire suggèrent fortement que faire une somme pondérée de la production et de la productivité, bien que ce soit possible

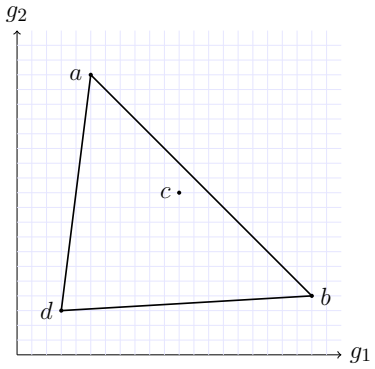

FIGURE 2 – Alternatives efficaces non supportées

sur un pur plan arithmétique, conduit à un indice « vide de sens ». Le seul argument auquel on pourrait penser en faveur de cette mesure, est que le poids du dernier critère est assez faible (même si, comme nous l'avons évoqué précédemment, les poids d'une somme pondérée doivent être interprétés avec beaucoup de précaution). Néanmoins, le fait de mélanger une mesure de la production avec une mesure de la productivité nous semble très problématique et signale un manque de réflexion sur ce qui doit guider la construction d'un indice pertinent. Les projets des auteurs du classement, comme annoncé dans Liu et al. (2005, p. 108), de construire un classement avec un poids de 50% pour le critère PY, indiquent clairement que ce problème sémantique n'a pas été complètement compris. Cela entame sérieusement la confiance qu'il est possible d'avoir dans les résultats à venir du classement, si toutefois le lecteur gardait encore une telle confiance.

4.4 Questions négligées liées à la structuration

Quand on met au point un classement, la bonne pratique suggère (Bouyssou et al., 2000, JRC/OECD, 2008) que la réflexion doit commencer par quelques questions simples mais fondamentales :

1. Quelle est la définition des objets évalués ?
2. Quel est le but du modèle ? Qui l'utilisera ?
3. Comment structurer les objectifs ?
4. Comment définir une « famille cohérente de critères »?
5. Comment prendre en compte l'incertitude, l'imprécision, et la mauvaise détermination ?

Pour les trois dernières questions, nous avons vu à la section 3 que le travail des auteurs du classement peut faire l'objet de critiques sévères. C'est particulièrement

vrai pour la dernière question : dans la mesure où les données initiales ne sont pas rendues publiques par les auteurs du classement et où les micro-décisions qui ont conduit à ces données ne sont pas documentées, il est totalement impossible d'analyser la robustesse du classement proposé. L'analyse partielle conduite dans Saisana et D'Hombres (2008) montre que cette robustesse est très faible.

Concentrons-nous maintenant sur les deux premières questions, en gardant à l'esprit quelques bonnes pratiques pour la construction d'un modèle d'évaluation.

4.4.1 Qu'est-ce qu'une « université »?

La question peut paraître curieuse pour un lecteur venant des États-Unis ou du Royaume-Uni. Toutefois, pour un lecteur venant d'un pays d'Europe continentale, cette question est parfois délicate. Prenons l'exemple de la France, qui est un exemple particulièrement complexe. En France co-existent :

- des universités publiques (appelées habituellement *Universités*). Il faut noter ici que l'histoire de ces universités est longue et mouvementée. Après 1968, la plupart d'entre elles ont été découpées en plus petites universités. De plus, il existe des universités de création récentes qui sont de taille limitée et n'offrent pas de formation dans tous les domaines de la Science et/ou des programmes de la licence au doctorat dans tous ces domaines. Enfin, quand on s'intéresse au système français, il faut garder à l'esprit que ces universités attirent rarement les meilleurs étudiants. Ceux-ci choisissent d'entrer dans le système des « Grandes Écoles ». Notons enfin que les frais d'inscription dans ces universités sont généralement très faibles.
- des *Grandes Écoles* (principalement en Ingénierie, en Gestion et en Sciences Politiques) qui sont des institutions très particulières. Elles sont généralement de petite taille et la plupart d'entre elles ne délivrent qu'un diplôme de master. Ce sont des institutions très sélectives qui recrutent des étudiants après un concours national. Elles ont une longue histoire et des réseaux d'anciens élèves très actifs. Très peu d'entre elles sont impliquées dans des programmes doctoraux. Les frais d'inscription dans les Grandes Écoles varient beaucoup. Certaines d'entre elles sont très chères (généralement les écoles de Gestion) tandis que pour d'autres, les frais d'inscription sont comparables à ceux des universités. Enfin, dans certaines d'entre elles (les Écoles Normales Supérieures), les étudiants sont payés.
- des instituts de recherche publics et privés qui peuvent avoir des centres de recherche communs, entre eux ou avec des universités ou des Grandes Écoles. Parmi les centres de recherche publics, mentionnons : le CNRS, l'INSERM, l'INRA, l'INRIA, etc. Une part très significative de la recherche en France est effectuée dans ces instituts, même s'ils n'ont aucun étudiant et ne délivrent aucun diplôme. De plus, il y a un grand nombre de centres de recherche

privés, le plus célèbre étant l'Institut Pasteur (dont sont issus beaucoup des prix Nobel français en médecine).

Avec un paysage institutionnel aussi complexe, ce qui doit compter comme une université est loin d'être évident [20]. Il semble également que ce n'était pas évident non plus pour les auteurs du classement, qui ont inclus dans leurs classements de 2003 à 2005 le *Collège de France*, une institution qui n'a aucun étudiant et ne délivre aucun diplôme. Si une telle institution peut compter comme une institution universitaire, alors presque toutes les organisations peuvent y prétendre. La situation française est particulièrement complexe mais elle est loin d'être exceptionnelle. L'Allemagne et l'Italie ont également des importants centres de recherche publics à côté de leurs universités.

Tout système d'évaluation doit commencer par définir clairement les objets qui sont évalués. Cette définition est absente du classement de Shangaï.

4.4.2 Qu'est-ce qu'une « bonne » université ?

Les auteurs du classement sont intéressés par les « universités de classe mondiale » (« *world-class universities* »). De même qu'ils n'ont pas défini de ce qu'ils considèrent comme une université, ils n'ont pas non plus défini ce qu'est une université de classe mondiale. Toutefois, les critères qu'ils utilisent permettent implicitement de définir ce qu'ils entendent par là. La seule chose importante est l'« excellence » en matière de recherche. De plus, cette excellence est mesurée en utilisant des critères très particuliers, évalués d'une façon très particulière (voir la section 3). Pourquoi ne pas prendre en compte d'autres productions scientifiques comme les ouvrages ou les thèses de doctorat ? Pourquoi compter les articles plutôt que mesurer leur impact ? Les questions sont nombreuses.

Ce qui laisse peut-être le plus perplexe, c'est que la définition implicite d'université de classe mondiale retenue par les auteurs du classement ignore complètement les contraintes que peuvent subir les institutions. Certaines universités ont une (plus ou moins) totale liberté pour organiser leur gouvernance, pour embaucher

20. Les effets de cette complexité sur le classement sont importants. La France arrive au 6ème rang des pays ayant le plus de personnes dans la catégorie HiCi (toutes disciplines confondues), avec un total de 161 personnes. Une étude détaillée de leurs affiliations (telles que données sur le site `http://www.isihighlycited.com`) montre que 30% de ces 161 personnes sont affiliées à une université, 7% à une grande école, 41% au CNRS ou à un institut de recherche, 20% à un autre type d'organisme et 2% ne sont affiliées à rien du tout. Donc globalement, seulement 37% de ces personnes font profiter les institutions françaises de leur renommée. Pour avoir un élément de comparaison, les États-Unis qui sont au premier rang comptent 4 096 personnes dans la catégorie HiCi. Un comptage révèle que 66% sont affiliés à une université ou à une école, 11% à un institut, 20% à un autre type d'organisme et 3% ne sont affiliés à rien du tout. La tendance est donc inverse, 66% des personnes les plus citées aux États-Unis profitent aux universités américaines.

ou licencier du personnel académique ou non académique, pour décider des salaires, pour sélectionner les étudiants, pour décider des frais d'inscription. D'autres n'ont presque aucune liberté sur tous ces points (c'est globalement le cas des universités françaises). Elles ne peuvent pas sélectionner les étudiants, ne peuvent pas décider des frais d'inscription, elles ne sont pas totalement impliquées dans la sélection de leur personnel académique et licencier quelqu'un est presque impossible. Compte tenu de ces différences, doit-on simplement les ignorer, comme c'est le cas implicitement dans le classement ? Ceci n'est légitime que si l'on admet qu'il existe un unique modèle d'université de classe mondiale. Cette hypothèse nécessiterait une justification empirique approfondie qui n'est pas donnée par les auteurs du classement.

De même, les ressources (*inputs*) consommées par les institutions dans leur « processus de production scientifique » sont totalement ignorées. La seule ressource qui est explicitement prise en compte est le taille du « corps académique » de l'institution, quand il a pu être obtenu. Mais il y a de nombreuses autres ressources qui devraient être intégrées pour juger de l'efficacité du processus de production scientifique. Notons simplement ici que les frais d'inscription, les donations (le budget annuel d'Harvard était de plus de 3×10^9 USD en 2007, Harvard University, 2007, p. 38, ce qui est supérieur au PIB du Laos), la qualité du campus, les bibliothèques universitaires (la bibliothèque d'Harvard possède plus de 15×10^6 volumes, Harvard University, 2007), la liberté de chercher et de publier dans tous domaines, etc. sont aussi des ingrédients du succès très importants pour une université. En ignorant toutes ces ressources, on aboutit inévitablement une vision très étroite de l'excellence académique [21].

La définition implicite qui est utilisée par les auteurs du classement devrait être maintenant claire. Il s'agit d'une institution :

1. qui est ancienne et qui a toujours gardé son nom (de préférence un nom simple, en anglais, sans signe diacritique) à travers son histoire,
2. issue d'un pays n'ayant pas connu de changements majeurs sur le plan politique et social depuis 1901,
3. issue d'un pays où l'organisation du système d'enseignement supérieur est simple (pas de système dual universités/grandes écoles, pas de centres de recherche associés),
4. issue d'un pays où l'on parle l'anglais,
5. qui a une grande liberté dans sa gouvernance,
6. qui a une grande liberté pour embaucher ou licencier son personnel et décider des salaires,

21. Signalons ici notre désaccord avec le principe 8 énoncé dans International Ranking Expert Group (2006) : un processus de production, qu'il soit scientifique ou non, ne peut pas être analysé sans considérer de façon *explicite* les extrants (*outputs*) *et* les intrants (*inputs*).

7. qui a de grands moyens financiers.

Nous ne prétendrons pas qu'une telle définition implicite n'a pas de sens [22]. En revanche, nous voulons souligner que cette définition correspond plus ou moins à la définition de la *Ivy League* (ainsi que de ses précurseurs anglais, Oxford et Cambridge) et que, de façon peu surprenante, ces institutions sont très bien représentées en tête du classement de Shangaï.

4.4.3 Quel est l'objectif du modèle ? Qui peut l'utiliser ?

De notre point de vue, l'intérêt de procéder à un classement des universités ne va pas de soi. En effet, qui peut bénéficier d'un tel classement ?

Les étudiants et les familles qui cherchent des informations seraient certainement très intéressés par un modèle qui évaluerait les *programmes*. Nous savons tous qu'une bonne université peut être particulièrement performante dans certains domaines mais beaucoup moins dans d'autres. D'après Bourdin (2008), pour les étudiants, « les principaux facteurs intervenant dans le choix du pays d'accueil sont la langue d'enseignement, le montant des droits de scolarité et le coût de la vie dans le pays d'accueil ». De plus, il semble clair que les étudiants et les familles seront aussi intéressés par des choses aussi triviales que la qualité du logement, les équipements sportifs, la qualité de l'enseignement, la réputation du diplôme dans les entreprises, les salaires moyens après le diplôme, le dynamisme de l'association des anciens élèves, la vie sur le campus, etc. Pour un système offrant de tels détails nous renvoyons à Berghoff et Federkeil (2009) et Centre for Higher Education Development (2008).

Les recruteurs sont peu susceptibles d'être influencés par des prix Nobel délivrés il y a longtemps à des membres d'un département, s'ils veulent recruter des étudiants de niveau master venant d'un autre département. Ils seront en revanche plus intéressés par leur « employabilité ». Au-delà des critères mentionnés pour les étudiants et les familles, d'autres éléments comme leur maîtrise des langues, leur expérience internationale, leurs stages, etc., seront de première importance pour eux.

De la même façon, un classement global des universités n'est sans doute que de peu d'intérêt pour les présidents d'universités (ou les recteurs, selon les pays) qui souhaitent travailler afin d'améliorer la qualité de leur institution. Clairement,

22. Il est cependant clair qu'elle n'avantage pas les institutions situées en Europe continentale. De ce fait, l'Union Européenne pourrait avoir intérêt à se préoccuper de la question du classement des universités. Alors que nous écrivions la première version de ce texte, nous avons découvert qu'un appel d'offre européen était lancé à ce sujet. Cet appel d'offres a depuis été remporté par le CHERPA-Network, un consortium dirigé par le *Centre for Higher Education Policy Studies* de l'Université de Twente (Pays-Bas) et le *Centrum für Hochschulentwicklung* (Allemagne). Le projet durera deux ans et coûtera approximativement 1,1 millions d'euros. .

ils voudront avant tout disposer d'un outil permettant d'identifier les forces et les faiblesses de leur institution, d'identifier ses principaux concurrents et de trouver des pistes possibles d'amélioration. Sauf s'ils ont un contrat qui spécifie explicitement qu'ils doivent améliorer leur position dans le classement de Shangaï, nous ne voyons pas comment un classement global d'une institution peut constituer, pour eux, un outil de gestion utile.

Enfin, les décideurs politiques devraient être principalement intéressés par un système d'évaluation qui les aiderait à juger de l'efficacité du *système d'enseignement supérieur* de leur pays. Si un pays a beaucoup d'institutions de taille moyenne, peu d'entre elles seront bien classées dans le classement de Shangaï. Mais cela ne signifie pas que le système, dans sa globalité, est inefficace. Vouloir de « grandes » institutions « visibles » dans chaque pays pourrait bien se révéler un grand gaspillage de ressources. Ici encore seules des preuves empiriques permettraient de trancher. Ces preuves ne sont pas fournies par les auteurs du classement.

4.4.4 L'oubli des bonnes pratiques

Comme détaillé dans Bouyssou et al. (2000) et Bouyssou et al. (2006), il existe des règles de « bonnes pratiques » à observer quand on veut construire un modèle d'évaluation. Nous n'en citerons que deux d'entre elles.

La première est évidente. Si on évalue une personne ou une organisation, on doit permettre à cette personne ou à cette organisation de vérifier les données qui sont utilisées pour l'évaluer. Ne pas faire ainsi conduit inévitablement à un cauchemar bureaucratique, où chacun est évalué sur la base de données qui restent « sous le manteau ». Nous avons déjà constaté que cette règle a été oubliée par les auteurs de ce classement.

La seconde est moins simple que la première mais est néanmoins cruciale. Quand un système d'évaluation est conçu, ses créateurs ne doivent pas s'attendre à ce que les personnes et les organisations qui sont évaluées réagissent passivement au système : il s'agit du message central de tout cours de base en Gestion. Les personnes et les organisations vont adapter leur comportement, consciemment ou non, en réaction au système d'évaluation. Ces effets en retour (notre traduction de *feedback effects*) sont inévitables et peuvent engendrer des effets pervers (tout ceci a été bien documenté dans la littérature en gestion, voir Berry, 1983, Boudon, 1979, Moisdon, 2005, Morel, 2002). Une bonne pratique est alors la suivante : essayer d'anticiper les effets pervers les plus évidents qui peuvent être générés par un système d'évaluation et essayer de concevoir un système dans lequel ces effets pervers indésirables seraient les moins dommageables possibles. Il ne semble pas que les auteurs du classement aient suivi ce conseil. Leurs seuls conseils de prudence sont que « tout exercice de classement est sujet à controverse, et aucun classement n'est absolument objectif » et que « on doit être prudent avec tout classement, y

compris le '*Academic Ranking of World Universities*'. On doit utiliser les classements comme une référence et lire la méthodologie de classement avec attention avant d'aborder les résultats » (ARWU, 2003–09) [23]. C'est un minimum. Mais au-delà de cela, on s'attend aussi de la part des créateurs d'un système d'évaluation, qu'ils analysent clairement les limites de ce qu'ils ont créé, afin de réduire, autant que faire se peut, ses utilisations illégitimes, et les auteurs du classement restent silencieux sur ce point.

Supposons que vous dirigiez une université et que vous voulez augmenter votre position dans le classement. C'est relativement simple. Il y a de grands domaines dans votre université qui ne contribuent aucunement à votre position dans le classement. Nous pensons au Droit et à la plupart des SHS. Supprimez ces domaines. Vous dégagerez des moyens financiers importants que vous pourrez alors utiliser pour vous « acheter » des groupes de recherche qui pourront contribuer à améliorer votre position dans le classement. Plusieurs indicateurs fournis par *Thomson Scientific* sont utiles pour cela (après tout, la liste des potentiels futurs prix Nobel en médecine n'est pas si longue que cela). Et quoiqu'il en soit, si le groupe ne reçoit pas le prix Nobel, il publiera dans des journaux qui comptent pour le classement et il contribuera à augmenter la liste des chercheurs les plus cités dans votre institution. Cela tend à promouvoir une certaine vision de la science, qui la fait ressembler à un sport professionnel où quelques riches équipes sont en compétition pour attirer les meilleurs joueurs mondiaux. Nous ne sommes pas persuadés qu'il s'agisse du meilleur moyen d'accroître les connaissances scientifiques.

Pour un gouvernement, réformer avec pour objectif de mieux figurer dans le classement de Shangaï est également simple. Prenons l'exemple du gouvernement français, puisque nous avons évoqué brièvement l'organisation complexe du système d'enseignement supérieur français. La plupart des universités françaises ont été découpées en de plus petites universités vers la fin des années 60. L'idée était alors de créer des organisations plus faciles à diriger. La vénérable *Université de Paris* a ainsi donné naissance à pas moins de 13 universités. Mais nous avons vu que ceci a un impact négatif pour le classement. Il faut donc inciter ces universités à se regrouper à nouveau. En négligeant l'impact plutôt marginal du dernier critère, un rapide calcul montre qu'en regroupant les universités parisiennes qui sont orientées principalement vers les Sciences exactes et la Médecine (il n'est évidemment d'aucun intérêt de se regrouper avec des personnes qui font des choses aussi futiles que du Droit ou des SHS), c'est-à-dire Paris 5, 6, 7 et 11 conduira (en utilisant les données du classement 2007, utiliser les données des classements ulté-

23. Notre traduction de « *Any ranking exercise is controversial, and no ranking is absolutely objective* » and that « *People should be cautious about any ranking and should not rely on any ranking either, including the 'Academic Ranking of World Universities'. Instead, people should use rankings simply as one kind of reference and read the ranking methodology carefully before looking at the ranking lists* »

rieurs ne changerait pas notre conclusion) à une institution qui sera globalement au niveau de Harvard. Bingo ! On ne dépense pas un Euro de plus, on n'augmente certainement pas non plus la production et le potentiel scientifique du pays, on a créé une organisation énorme qui sera sans doute très lourde à gérer... mais on a largement augmenté la position des institutions françaises dans le classement de Shangaï ! On peut même envisager d'aller plus loin encore : les centres de recherche publics (CNRS, INSERM, etc.), bien que généralement très efficaces, ne comptent pour rien dans le classement. On peut les supprimer et transférer les ressources et le personnel de ces centres de recherche dans l'énorme institution universitaire qui vient d'être créée. Avec tout cela, on battra sans aucun doute Harvard. Inutile de dire que toutes ces manipulations pourront avoir des effets désastreux à long terme [24].

5 Conclusion

Dans ce qui a été probablement la première analyse sérieuse du classement de Shangaï, van Raan (2005a) écrivait que « à partir des considérations précédentes, nous concluons que le classement de Shangaï ne doit pas être utilisé à des fins d'évaluation, ni même pour faire du *benchmarking* » [25] et que « le problème le plus sérieux de ces classements est qu'ils sont considérés comme des "quasi-évaluations" des universités considérées. C'est absolument inacceptable » [26]. Les conclusions de van Raan étaient principalement fondées sur des considérations bibliométriques, à propos desquelles les auteurs du classement ont été incapables de répondre de façon convaincante (Liu et al., 2005, van Raan, 2005b).

Notre propre analyse est principalement fondée sur l'aide multicritère à la décision. Ajouter ce point de vue à l'analyse bibliométrique de van Raan (2005a) conduit inévitablement à une conclusion encore plus radicale. Nous avons vu que les critères utilisés par les auteurs du classement ne sont liés que de façon très lâche à ce qu'ils sont censés mesurer, que l'évaluation de ces critères implique l'utilisation de nombreux paramètres arbitraires et de multiples micro-décisions qui ne sont pas documentées. De plus, nous avons observé que la méthode d'agrégation

24. Ce qui précède nous semble permettre une lecture intéressante des évolutions récentes du paysage académique français : fusion des trois universités de Strasbourg (à la date du 1 janvier 2009, voir `http://demain.unistra.fr/`), regroupements d'institutions sous forme de Pôles de Recherche et d'Enseignement Supérieur (PRES) ou de Réseau thématique de recherche avancée (RTRA), loi relative aux Libertés et Responsabilités des Universités (LRU)(Numéro 2007-1199 et datée du 10 août 2007).

25. Notre traduction de « *From the above considerations we conclude that the Shanghai ranking should not be used for evaluation purposes, even not for benchmarking* »

26. Notre traduction de « *The most serious problem of these rankings is that they are considered as 'quasi-evaluations' of the universities considered. This is absolutely unacceptable* »

utilisée est incroyablement naïve et qu'elle conduit à des situations paradoxales. Enfin, on a observé que les auteurs du classement n'ont prêté que très peu d'attention à des questions fondamentales liées à la structuration du problème [27] [28]. Il ne nous semble donc pas excessif de conclure que *le classement de Shangaï est un exercice qui n'a absolument aucune valeur.*

À ce stade, le lecteur pourrait légitimement espérer que nous proposions un autre classement, qui serait meilleur [29]. Nous ne le ferons pas. En effet, nous ne sommes pas convaincu qu'un tel exercice soit réellement utile et nous avons expliqué plus haut pourquoi il nous semblerait préférable de se consacrer à un classement des *programmes* ou des *systèmes nationaux d'enseignement* plutôt que des universités. Si une telle exercice devait cependant être conduit, il nécessiterait

27. Même si nous nous sommes centrés ici sur le classement de Shangaï, notre analyse semble s'appliquer assez bien au classement du *Times higher Education Supplement* (Times Higher Education Supplement, 2008), comme montré dans Vincke (2009). Pour une critique vigoureuse du classement du *US News & World Report*, nous renvoyons à Ehrenberg, 2003.

28. Dans la mesure où les auteurs du classement ont décidé d'ignorer plus ou moins complètement l'avis de van Raan (2005a), nous pensons probable qu'ils en feront de même avec le notre. Les premiers échanges que nous avons eus avec les auteurs du classement en mai 2009 au sujet d'une version anglaise préliminaire de ce texte (Billaut et al., 2009) laissent en effet à penser qu'ils ne semblent pas vouloir démarrer une discussion méthodologique sur des bases solides. De fait, la version du classement parue en novembre 2009 ne tient aucun compte de nos observations.

29. Une autre façon de réagir au classement de Shangaï a été mise en avant par l'*École Nationale Supérieure des Mines de Paris* (ENSMP, 2007). Cette Grande École française prestigieuse a une taille telle qu'elle n'apparaîtra jamais en bonne position dans le classement de Shangaï. L'ENSMP a donc décidé de proposé un classement alternatif. Il peut être expliqué très simplement puisqu'il est fondé sur un critère unique : chaque institution reçoit comme score le nombre de ses anciens élèves dirigeant une des 500 plus grandes entreprises mondiales identifiée par le journal *Fortune* (nous renvoyons à ENSMP, 2007, pour plus de détails sur la façon dont ce nombre est calculé). Nous ne trouvons pas ce classement très attractif. Les performances des institutions sont fondées sur des faits qui se produisent très longtemps après que le diplôme a été obtenu (il n'est pas fréquent de voir un jeune diplômé devenir P-DG d'une grande entreprise). Ce nombre a donc peu de choses à voir avec la performance *actuelle* de l'institution. D'autre part, ce critère dépend évidemment des structures industrielles des pays (les institutions provenant de pays où l'industrie est très concentrée sont clairement avantagées) et les effets de réseaux auront également un impact majeur sur son évaluation (et on sait que ces effets sont de la plus grande importance pour bien comprendre le système français des Grandes Écoles). Nous ne considérons cependant pas que le classement de l'ENSMP est pire que celui de Shangaï. Au contraire, il montre clairement que la plupart des critères utilisés par les auteurs du classement sont arbitraires. Finalement, et c'est le plus intéressant, le classement de l'ENSMP donne des résultats très différents du classement de Shangaï. Il y a cinq institutions françaises parmi les dix premières institutions de ce classement (École Polytechnique, HEC, Sciences-Po Paris, École Nationale d'Administration et ENSMP). Parmi ces cinq institutions, trois ne figurent pas dans les 500 institutions classées dans le classement de Shangaï (HEC, Sciences Po Paris et École Nationale d'Administration). Il s'agit là d'une tentative intéressante pour contrer les effets du classement de Shangaï en adoptant une stratégie de *dilution* consistant à proposer des classements alternatifs produisant qui produisent des résultats complètement différents.

clairement un travail de longue haleine qui devrait être réalisé par un groupe pluridisciplinaire d'experts spécialistes en évaluation, en systèmes d'enseignement et en bibliométrie. Soulignons néanmoins que la combinaison de techniques issues de la Recherche Opérationnelle avec des outils d'évaluation bibliométriques sophistiqués semble offrir une base solide pour l'évaluation des systèmes d'enseignement supérieur. Parce que la question des techniques bibliométriques a été déjà largement étudiée (voit Moed et al., 1995, Moed, 2006, van Raan, 2005c, 2006, Zitt et al., 2005), mentionnons simplement ici quelques techniques de Recherche Opérationnelle potentiellement utiles. Nous avons montré plus haut qu'un manque de connaissances de base dans les techniques d'agrégation et leurs propriétés peut invalider une technique d'évaluation. Au delà de l'analyse multicritère, la Recherche Opérationnelle a également développé des outils sophistiqués qui peuvent aider à structurer des problèmes (Checkland, 1981, Checkland et Scholes, 1990, Eden, 1988, Eden et al., 1983, Friend et Hickling, 1987, Rosenhead, 1989) et qui peuvent se combiner avec d'autres outils d'agrégation sophistiqués (Ackermann et Belton, 2006, Bana e Costa et al., 1999, Belton et al., 1997, Montibeller et al., 2008, Phillips et Bana e Costa, 2007). Un développement parallèle concerne des méthodes conçues pour mesurer l'efficacité d'« unités de décision » (notre traduction de *decision-making units*), qui transforment des intrants (*inputs*) en extrants (*outputs*). Ces méthodes sont connues sous l'appellation de *Data Envelopment Analysis* (DEA) (Banker et al., 1984, Charnes et al., 1978, Cherchye et al., 2008, Cook et Zhu, 2008, Cooper et al., 1999). L'utilité de ces techniques pour construire des modèles d'évaluation dans l'enseignement supérieur a déjà été reconnue (Bougnol et Dulá, 2006, Johnes, 2006, Leitner et al., 2007, Turner, 2005, 2008). Nous pensons qu'une combinaison de ces deux types d'approches pourra donner une base solide à des modèles d'évaluation intéressants et bien fondés. Cela pourrait par exemple être combiné avec l'approche interactive développée par le Centre d'enseignement supérieur en Allemagne (Berghoff et Federkeil, 2009, Centre for Higher Education Development, 2008) permettant à chaque utilisateur du système de choisir les critères qu'il souhaite utiliser.

Références

F. Ackermann et V. Belton. Problem structuring without workshops? Experiments with distributed interaction in a PSM. *Journal of the Operational Research Society*, 58: 547–556, 2006.

D. Adam. Citation analysis: The counting house. *Nature*, 415(6873):726–729, 2002.

ARWU. Academic Ranking of World Universities, 2003–09. Shanghai Jiao Tong University, Institute of Higher Education, voir `http://www.arwu.org`.

C. A. Bana e Costa, L. Ensslin, É. C. Corrêa et J.-C. Vansnick. Decision Support Systems

in action: Integrated application in a multicriteria decision aid process. *European Journal of Operational Research*, 113:315–335, 1999.

R. D. Banker, A. Charnes et W. W. Cooper. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30(9):1078–1092, 1984.

M. H. Bazerman. *Judgment in managerial decision making*. Wiley, New York, 1990.

V. Belton et T. J. Stewart. *Multiple criteria decision analysis: An integrated approach*. Kluwer, Dordrecht, 2001.

V. Belton, F. Ackermann et I. Shepherd. Integrated support from problem structuring through alternative evaluation using COPE and V•I•S•A. *Journal of Multi-Criteria Decision Analysis*, 6:115–130, 1997.

S. Berghoff et G. Federkeil. The CHE approach. In D. Jacobs et C. Vermandele, editors, *Ranking universities*, pages 41–63, Brussels, 2009. Édition de l'Université de Bruxelles.

M. Berry. Une technologie invisible ? Le rôle des instruments de gestion dans l'évolution des systèmes humains. Mémoire, Centre de Recherche en Gestion. École Polytechnique, 1983. Disponible à `http://crg.polytechnique.fr/fichiers/crg/publications/pdf/2007-04-05-1133.pdf`.

J.-C. Billaut, D. Bouyssou et Ph. Vincke. Should you believe in the Shanghai ranking? An MCDM view. Cahier du LAMSADE # 283, LAMSADE, 2009. Disponible à `http://hal.archives-ouvertes.fr/hal-00388319/en/`. Une version courte de ce document est parue dans *Scientometrics* **84** (1), 237–263, 2010.

R. Boudon. *Effets pervers et ordre social*. PUF, Paris, 1979.

M.-L. Bougnol et J. H. Dulá. Validating DEA as a ranking tool: An application of DEA to assess performance in higher education. *Annals of Operations Research*, 145: 339–365, 2006.

J. Bourdin. Le défi des classements dans l'enseignement supérieur. Rapport au Sénat 442, République française, 2008. Disponible à `http://www.senat.fr/rap/r07-442/r07-442.html`.

D. Bouyssou. Modelling inaccurate determination, uncertainty, imprecision using multiple criteria. In A.G. Lockett et G. Islei, editors, *Improving Decision Making in Organisations*, LNEMS 335, pages 78–87. Springer-Verlag, Berlin, 1989.

D. Bouyssou. Building criteria: A prerequisite for MCDA. In C. A. Bana e Costa, editor, *Readings in multiple criteria decision aid*, pages 58–80. Springer-Verlag, Heidelberg, 1990.

D. Bouyssou, Th. Marchant, M. Pirlot, P. Perny, A. Tsoukiàs et Ph. Vincke. *Evaluation and decision models: A critical perspective*. Kluwer, Dordrecht, 2000.

D. Bouyssou, Th. Marchant, M. Pirlot, A. Tsoukiàs et Ph. Vincke. *Evaluation and decision models: Stepping stones for the analyst*. Springer, New York, 2006.

R. L. Brooks. Measuring university quality. *The Review of Higher Education*, 29(1): 1–21, 2005.

G. Buela-Casal, O. Gutiérez-Martínez, M. P. Bermúdez-Sánchez et O. Vadillo-Muñoz. Comparative study of international academic rankings of universities. *Scientometrics*, 71(3):349–365, 2007.

Centre for Higher Education Development. CHE ranking. Technical report, CHE, 2008. http://www.che.de.

A. Charnes, W. W. Cooper et E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2:429–444, 1978. Correction: European Journal of Operational Research, 3:339.

P. Checkland. *Systems thinking, systems practice.* Wiley, New York, 1981.

P. Checkland et J. Scholes. *Soft systems methodology in action.* Wiley, New York, 1990.

L. Cherchye, W. Moesen, N. Rogge, T. van Puyenbroeck, M. Saisana, A. Saltelli, R. Liska et S. Tarantola. Creating composite indicators with DEA and robustness analysis: the case of the technology achievement index. *Journal of Operational Research Society*, 59:239–251, 2008.

W. A. Cook et J. Zhu. *Data Envelopment Analysis: Modeling Operational Processes and Measuring Productivity.* CreateSpace, 2008.

W. W. Cooper, L. M. Seiford et K. Tone. *Data Envelopment Analysis. A comprehensive text with models, applications, references and DEA-solver software.* Kluwer, Boston, 1999.

N. Dalsheimer et D. Despréaux. Analyses des classements internationaux des établissements d'enseignement supérieur. *Éducation & formations*, 78:151–173, 2008.

R. M. Dawes. *Rational choice in an uncertain world.* Hartcourt Brace, Fort Worth, 1988.

D. Desbois. Classement de Shanghai : peut-on mesurer l'excellence académique au niveau mondial ? *La revue trimestrielle du réseau Écrin*, 67:20–26, 2007.

D. Dill et M. Soo. Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher Education*, 49:495–533, 2005.

R. L. Ebel et D. A. Frisbie. *Essentials of educational measurement.* Prentice-Hall, New York, 1991.

C. Eden. Cognitive mapping. *European Journal of Operational Research*, 36:1–13, 1988.

C. Eden, S. Jones et D. Sims. *Messing about in problems.* Pergamon Press, Oxford, 1983.

R. G. Ehrenberg. Method or madness? Inside the *USNWR* college rankings. Working paper, ILR collection, Cornell University, 2003. Disponible à http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1043&context=workingpapers.

H. J. Einhorn et R. Hogarth. Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 13:171–192, 1975.

M. Enserink. Who ranks the university rankers? *Science*, 317(5841):1026–1028, 2007.

ENSMP. Professional ranking of world universities. Technical report, École Nationale Supérieure des Mines de Paris (ENMSP), 2007. Disponible à http://www.ensmp.fr/Actualites/PR/EMP-ranking.pdf.

A. Fert. Comment le classement de Shanghaï désavantage nos universités. *Le Monde*, 2007. 27 Août.

R. Florian. Irreproducibility of the results of the Shanghai academic ranking of world universities. *Scientometrics*, 72:25–32, 2007.

J. K. Friend et A. Hickling. *Planning under pressure: The strategic choice approach.*

Pergamon Press, New York, 1987.

Y. Gingras. Du mauvais usage de faux indicateurs. *Revue d'Histoire Moderne et Contemporaine*, 5(55-4bis):67–79, 2008. Disponible à `http://www.cairn.info/load_pdf.php?ID_ARTICLE=RHMC_555_0067`.

Harvard University. Harvard university fact book, 2006–2007. Technical report, Harvard University, 2007. Disponible à `http://www.provost.harvard.edu/institutional_research/FACTBOOK_2007-08_FULL.pdf`.

HEE 2002. Ranking and league tables of higher education institutions. *Higher Education in Europe*, 27(4), 2002. Special Issue.

HEE 2005. Ranking systems and methodologies in higher education. *Higher Education in Europe*, 30(2), 2005. Special Issue.

HEE 2007. Higher education ranking and its ascending impact on higher education. *Higher Education in Europe*, 32(1), 2007. Special Issue.

HEE 2008. University rankings: Seeking prestige, raising visibility and embedding quality. *Higher Education in Europe*, 33(2-3), 2008. Special Issue.

CHERI / HEFCE. Counting what is measured or measuring what counts? league tables and their impact on higher education institutions in England. Report to HEFCE by the Centre for Higher Education Research and Information (CHERI) 2008/14, Open University, and Hobsons Research, April 2008. Disponible à `http://www.hefce.ac.uk/pubs/hefce/2008/08_14/`.

R. Hogarth. *Judgement and choice: The psychology of decision.* Wiley, New York, 1987.

International Ranking Expert Group. Berlin principles on ranking of higher education institutions. Technical report, CEPES-UNNESCO, 2006. Disponible à `http://www.che.de/downloads/Berlin_Principles_IREG_534.pdf`.

J. P. A. Ioannidis, N. A. Patsopoulos, F. K. Kavvoura, A. Tatsioni, E. Evangelou, I. Kouri, D. G. Contapoulos Ioannidis et G. Liberopoulos. International ranking systems for universities and institutions: a critical appraisal. *BioMed Central*, 5(30), 2007. doi: 10.1186/1741-7015/5/30. Disponible à `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2174504`.

J. Johnes. Measuring efficiency: A comparison of multilevel modelling and data envelopment analysis in the context of higher education. *Bulletin of Economic Research*, 58(2):75–104, 2006.

JRC/OECD. Handbook on constructing composite indicators. methodology and user guide. Technical report, JRC/OECD, OECD Publishing, 2008. ISBN 978-92-64-04345-9. Disponible à `http://www.olis.oecd.org/olis/2005doc.nsf/LinkTo/NT00002E4E/$FILE/JT00188147.PDF`.

T. Kävelmark. University ranking systems: A critique. Technical report, Irish Universities Quality Board, 2007. Disponible à `http://www.urank.se/Dokument/Torsten_Kalvemark_University_Ranking_Systems_A_Critique.pdf`.

R. L. Keeney. *Value-focused thinking. A path to creative decision making.* Harvard University Press, Cambridge, 1992.

R. L. Keeney et H. Raiffa. *Decisions with multiple objectives: Preferences and value tradeoffs.* Wiley, New York, 1976.

R. L. Keeney, J. S. Hammond et H. Raiffa. *Smart choices: A guide to making better decisions*. Harvard University Press, Boston, 1999.

F. N. Kerlinger et H. B. Lee. *Foundations of behavioral research*. Wadsworth Publishing, New York, 4 edition, 1999.

O. Kivinen et J. Hedman. World-wide university rankings: A scandinavian approach. *Scientometrics*, 74(3):391–408, 2008.

P. Kline. *Handbook of psychological testing*. Routledge, New York, 2 edition, 2000.

T. N. van Leeuwen, H. F. Moed, R. J. W. Tijssen, M. S. Visser et A. F. J. van Rann. Language biases in the coverage of the *Science Citation Index* and its consequences for international comparisons of national research performance. *Scientometrics*, 51(1): 335–346, 2001.

K.-H. Leitner, J. Prikoszovits, M. Schaffhauser-Linzatti, R. Stowasser et K. Wagner. The impact of size and specialisation on universities' department performance: A DEA analysis applied to Austrian universities. *Higher Education*, 53(4):517–538, 2007.

N. C. Liu. The story of academic ranking of world universities. *International Higher Education*, 54:2–3, 2009.

N. C. Liu et Y. Cheng. The academic ranking of world universities. *Higher Education in Europe*, 30(2):127–136, 2005.

N. C. Liu, Y. Cheng et L. Liu. Academic ranking of world universities using scientometrics: A comment to the "fatal attraction". *Scientometrics*, 64(1):101–109, 2005.

R. D. Luce et H. Raiffa. *Games and Decisions*. Wiley, New York, 1957.

S. Marginson. Global university rankings: where to from here? Technical report, Asia-Pacific Association for International Education, 2007. Communication to the Asia-Pacific Association for International Education, National University of Singapore, 7-9 March 2007. Disponible à `http://www.cshe.unimelb.edu.au/people/staff_pages/Marginson/APAIE_090307_Marginson.pdf`.

H. F. Moed, R. E. De Bruin et T. N. van Leeuwen. New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, 33:381–422, 1995.

H. M. Moed. Bibliometric rankings of world universities. Technical Report 2006-01, CWTS, Leiden University, 2006. Disponible à `http://www.cwts.nl/hm/bibl_rnk_wrld_univ_full.pdf`.

J.-C. Moisdon. Vers des modélisations apprenantes ? *Économies et Sociétés. Sciences de Gestion*, 7-8:569–582, 2005.

G. Montibeller, F. Ackermann, V. Belton et L. Ensslin. Reasoning maps for decision aiding: An integrated approach for problem structuring and multi-criteria evaluation. *Journal of the Operational Research Society*, 59:575–589, 2008.

C. Morel. *Les Décisions Absurdes*. Bibliothèque des Sciences Humaines. Gallimard, Paris, 2002.

L. D. Phillips et C. A. Bana e Costa. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. *Annals of Operations Research*, 154:51–68, 2007.

E. C. Poulton. *Behavioral decision theory: A new approach*. Cambridge University Press,

Cambridge, 1994.

A. F. J. van Raan. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36:397–420, 1996.

A. F. J. van Raan. Fatal attraction: Ranking of universities by bibliometric methods. *Scientometrics*, 62:133–145, 2005a.

A. F. J. van Raan. Reply to the comments of Liu et al. *Scientometrics*, 64(1):111–112, 2005b.

A. F. J. van Raan. Measurement of central aspects of scientific research: performance, interdisciplinarity, structure. *Measurement: Interdisciplinary Research and Perspectives*, 3(1):1–19, 2005c.

A. F. J. van Raan. Challenges in the ranking of universities. In J. Sadlak et N. C. Liu, editors, *World-Class University and Ranking: Aiming Beyond Status*, pages 81–123, Bucharest, 2006. UNESCO-CEPES. ISBN 92-9069-184-0.

M. J. Rosenhead. *Rational analysis for a problematic world*. Wiley, New York, 1989.

B. Roy. Main sources of inaccurate determination, uncertainty and imprecision in decision models. In B. Munier et M. Shakun, editors, *Compromise, Negotiation and group decision*, pages 43–67. Reidel, Dordrecht, 1988.

B. Roy. *Multicriteria methodology for decision aiding*. Kluwer, Dordrecht, 1996. Original version in French "*Méthodologie multicritère d'aide à la décision*", Economica, Paris, 1985.

B. Roy et D. Bouyssou. *Aide multicritère à la décision : méthodes et cas*. Economica, Paris, 1993.

J. E. Russo et P. J. H. Schoemaker. *Confident decision making*. Piatkus, London, 1989.

M. Saisana et B. D'Hombres. Higher education rankings: Robustness issues and critical assessment. How much confidence can we have in higher education rankings? Technical Report EUR 23487 EN 2008, IPSC, CRELL, Joint Research Centre, European Commission, 2008. Disponible à `http://composite-indicators.jrc.ec.europa.eu/Seminar_Eurostat_2008/EUR23487_Saisana_DHombres.pdf`.

A. K. Sen. Internal consistency of choice. *Econometrica*, 61:495–521, 1993.

A. Stella et D. Woodhouse. Ranking of higher education institutions. Technical report, Australian Universities Quality Agency, 2006. Disponible à `http://www.auqa.edu.au/files/publications/ranking_of_higher_education_institutions_final.pdf`.

Times Higher Education Supplement. THES ranking, 2008. `http://www.thes.co.uk/worldrankings/`.

V. T'kindt et J.-C. Billaut. *Multicriteria Scheduling*. Springer Verlag, Berlin, 2nd revised edition, 2006.

D. Turner. Benchmarking in universities: League tables revisited. *Oxford Review of Education*, 31(3):353–371, 2005.

D. Turner. World university rankings. *International Perspectives on Education and Society*, 9:27–61, 2008.

Ph. Vincke. University rankings. In D. Jacobs et C. Vermandele, editors, *Ranking universities*, pages 11–26, Brussels, 2009. Édition de l'Université de Bruxelles.

D. von Winterfeldt et W. Edwards. *Decision analysis and behavioral research*. Cambridge

University Press, Cambridge, 1986.

M. Zitt et G. Filliatreau. Big is (made) beautiful: Some comments about the Shanghai-ranking of world-class universities. In J. Sadlak et N. C. Liu, editors, *World-Class University and Ranking: Aiming Beyond Status*, pages 141–160, Bucharest, 2006. UNESCO-CEPES. ISBN 92-9069-184-0.

M. Zitt, S. Ramanana-Rahary et E. Bassecoulard. Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, 63(2):373–401, 2005.