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]

25 août 2009

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 la version française de Billaut, Bouyssou et Vincke (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 UMR7024 & 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, e-mail : pvincke@ulb.ac.be

Résumé

Nous proposons dans cet article une analyse critique du classement mondial des universités (« *Academic Ranking of World Universities* »), publié chaque année par l'Institut de l'Enseignement Supérieur de l'Université Jiao Tong de Shangaï (*Institute of Higher Education of the Jiao Tong University in Shanghai*), et plus connu sous le nom de « classement de Shangaï ». Après avoir rappelé comment le classement est construit, nous discuterons tout d'abord de la pertinence des critères utilisés. Nous analyserons ensuite la méthode d'agrégation proposée. Notre analyse se fondera sur les outils et les 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 globalement, l'exercice souffre d'un manque d'attention flagrant aux problèmes fondamentaux lié à la structuration du problème. Par conséquent, notre opinion est que le classement de Shangaï, malgré la grande couverture médiatique qu'il reçoit chaque année, n'est pas un outil pertinent pour juger de la « qualité » des institutions académiques, pour guider le choix des étudiants ou des familles, ni pour promouvoir des réformes du système d'enseignement supérieur. Nous indiquons le type de travail qui pourrait être conduit et qui offrirait de bonnes alternatives à ce classement.

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

Table des matières

1 Introduction **1**

2 Comment fonctionne le classement de Shangaï ? **3**
2.1 Qui sont les auteurs du classement ? . . . 3
2.2 Quels sont leurs objectifs ? . . . 3
2.3 Comment les universités ont-elles été sélectionnées ? . . . 4
2.4 Les critères utilisés . . . 4
2.4.1 Qualité de l'enseignement . . . 4
2.4.2 Qualité de l'institution . . . 5
2.4.3 Production scientifique . . . 5
2.4.4 Productivité . . . 6
2.5 Source des données . . . 6
2.6 Normalisation et agrégation . . . 6
2.7 Les résultats en 2008 . . . 6

3 Une analyse critique des critères **10**
3.1 Critères relatifs aux prix Nobel et médailles Fields . . . 10
3.2 Les chercheurs les plus cités . . . 12
3.3 Articles dans *Nature* et *Science* . . . 15
3.4 Articles indexés par Thomson Scientific . . . 15
3.5 Productivité . . . 17
3.6 Un nombre variable de critères . . . 17
3.7 Une synthèse sur les critères . . . 18
3.8 Le temps et la taille . . . 19
3.8.1 Effets du temps . . . 19
3.8.2 Effets de la taille . . . 20

4 Le classement de Shangaï au prisme de l'aide multicritère à la décision **22**
4.1 Une introduction rhétorique . . . 22
4.2 La technique d'agrégation est mauvaise . . . 23
4.3 La technique d'agrégation utilisée n'a pas de sens . . . 29
4.4 Quelques questions négligées liées à la structuration . . . 29
4.4.1 Qu'est-ce qu'une « université »? . . . 30
4.4.2 Qu'est-ce qu'une « bonne » université ? . . . 31
4.4.3 Quel est l'objectif du modèle ? . . . 33
4.4.4 Quelques bonnes pratiques d'évaluation . . . 34

5 Que faire ? 36
5.1 Que penser du classement de Shangaï ? 36
5.2 Quelques pistes 37
5.3 Pourquoi ne pas proposer votre propre classement ? 39
5.4 Le rôle de l'Europe 41

Références bibliographiques 41

1 Introduction

Une chose étrange s'est produite en 2003. Un groupe de personnes appartenant à l'institut d'enseignement supérieur (« *Institute of Higher Education* ») de l'Université Jiao Tong de Shangaï ont publié sur leur site Internet leur premier Classement Académique International des Universités (« *Academic Ranking of World Universities* » (Shanghai Jiao Tong University, Institute of Higher Education, 2003–09)), plus connu sous le nom de « *classement de Shangaï* » [1]. Ce classement consiste en une liste ordonnée de 500 universités du monde entier. Une description du classement peut être trouvée dans Liu et Cheng (2005) et l'histoire de sa création est détaillée dans Liu (2009).

Ce classement a fait quasi-instantanément l'objet d'une extraordinaire couverture médiatique. Non seulement les décideurs politiques ont utilisé les résultats de ce classement pour promouvoir des réformes de l'enseignement supérieur, mais de nombreuses institutions académiques ont commencé à utiliser leur positions dans ce classement pour leur communication. Apparemment, ce classement est un véritable succès.

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, Cheng et Liu, 2005, van Raan, 2005b). Depuis, les attaques ont été nombreuses et vigoureuses, à la fois dans la littérature académique (Buela-Casal, Gutiérez-Martínez, Bermúdez-Sánchez et Vadillo-Muñoz, 2007, Dill et Soo, 2005, Gingras, 2008, Ioannidis, Patsopoulos, Kavvoura, Tatsioni, Evangelou, Kouri, Contapoulos Ioannidis et Liberopoulos, 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). Au regard de telles attaques, 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. En fait, c'est le contraire qui s'est produit. Chaque année, une nouvelle version du classement est publiée et chaque année, la

1. Depuis 2007, les auteurs du classement de Shangaï ont également réalisé un classement des institutions en distinguant cinq champs disciplinaires (*Natural Sciences and Mathematics, Engineering / Technology and Computer Sciences, Life and Agriculture Sciences, Clinical Medicine and Pharmacy, and Social Sciences*), voir `http://www.arwu.org/ARWU-FIELD2008.htm`. Dans la mesure où la méthodologie utilisée pour ces classements par discipline est similaire à celle utilisée pour le classement global, nous ne l'analyserons pas dans ce document.

couverture médiatique sur le classement semble augmenter. De plus, des projets visant à transformer le système d'enseignement supérieur font très souvent référence au classement de Shangaï. Par exemple, l'actuelle Ministre de l'Enseignement Supérieur et de la Recherche s'est vue confiée par le Premier Ministre et le Président de la République, la mission consistant en « l'amélioration du rang de nos 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].

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 chercheurs opérationnels ayant travaillé dans le domaine des modèles d'évaluation et de décision en présence de plusieurs critères (Bouyssou, Marchant, Pirlot, Perny, Tsoukiàs et Vincke, 2000, Bouyssou, Marchant, Pirlot, Tsoukiàs et Vincke, 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).

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 paraître puérile et dénuée de tout intérêt. Nous ne sommes pas en désaccord avec eux. Néanmoins, ces lecteurs doivent se rappeler qu'il peut exister dans le paysage politique des décideurs paresseux qui pourraient vouloir utiliser les résultats d'un classement, uniquement sous prétexte qu'il existe. Plus grave encore, il pourrait y avoir des décideurs politiques souhaitant utiliser ces résultats afin de promouvoir leur propre vision d'une meilleure organisation du système d'enseignement supérieur. De plus, 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 classement est pris très au sérieux par certains. Nous le prenons également très au sérieux, et essayons de poser dessus un regard scientifique rigoureux. Notre point de vue fait l'objet de cet article. Celui-ci est organisé de la façon suivante. La section 2, décrit comment le classement de Shangaï est bâti. La section 3 est consacrée à l'étude des critères qui sont utilisés. La section 4 présente notre regard de spécialistes d'« aide multicritère à la décision » sur le classement de Shangaï. 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, dernier accès le 19 juillet 2009.

2 Comment fonctionne le classement de Shangaï ?

Cette section décrit la façon dont le classement de Shangaï est construit, d'après ce qui en est expliqué dans Shanghai Jiao Tong University, Institute of Higher Education (2003–09) et Liu et Cheng (2005). Nous nous concentrons sur la dernière édition du classement publié en 2008 [3], malgré le fait que la méthodologie ait quelque peu évolué dans le temps.

2.1 Qui sont les auteurs du classement ?

L'institut d'enseignement supérieur de l'Université Jiao Tong de Shangaï est un petit groupe de personnes dirigé par le Professeur Nian Cai Liu. Les auteurs du classement reconnaissent qu'ils n'ont aucune connaissance particulière en bibliométrie (Nian Cai Liu est un chimiste spécialisé dans les polymères). L'analyse qui suit semble montrer qu'ils n'ont aucune connaissance particulière en aide multicritère à la décision, ni dans le développement de systèmes d'évaluation.

Les auteurs du classement insistent sur le fait qu'ils n'ont reçu 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), par exemple.

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 [4].

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. Notons que 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

3. Voir `http://www.arwu.org/rank2008/ARWU2008Methodology(EN).htm`, dernier accès le 19 juillet 2009.

4. Notons que la Chine est très friande de classements d'écoles et d'universités. Le classement des universités chinoises proposé par la *Concodia University Alumni Association* (CUAA), voir `http://www.ambafrance-cn.org/IMG/pdf/Classement_CUAA_2009.pdf` (dernier accès le 19 juillet 2009.) n'est pas construit sur les mêmes bases que le classement de Shangaï et sert davantage de référence en Chine pour classer leurs institutions.

d'obtenir des données comparables sur le plan international, ils ont décidé de classer les universités sur la seule base de leurs performances académiques (Shanghai Jiao Tong University, Institute of Higher Education, 2003–09).

2.3 Comment les universités ont-elles été sélectionnées ?

Les auteurs du classement indiquent 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 rangées de 1 à 100. Les suivantes sont classées par groupes de 50 (jusqu'au rang 201) et ensuite par groupes de 100.

2.4 Les critères utilisés

Les auteurs du classement utilisent six critères regroupés en quatre domaines. Ils sont présentés ci-dessous.

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 [5]. Ceci définit le premier critère, noté ALU.

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

2.4.2 Qualité de l'institution

Ce domaine comporte deux critères. Le premier compte le nombre de membres du « personnel 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, ..., 10% pour les lauréats entre 1911 et 1920. 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 chercheurs les plus cités, dans chacun des 21 domaines de la Science, selon Thomson Scientific. Ces chercheurs les plus cités, dans chacun des 21 domaines, consistent en une liste d'environ 250 personnes qui ont reçu le plus grand nombre de citations dans un domaine, d'après les informations contenues dans les bases de données de Thomson Scientific (source `http://hcr3.isiknowledge.com/popup.cgi?name=hccom`, dernier accès le 19 juillet 2009.). Dans certaines catégories, Thomson Scientific liste plus de 250 personnes, nous supposons que c'est en raison d'ex aequo. 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 personnel académique de l'institution pendant les cinq dernières années. Cela soulève le problème du comptage 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, 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 sur une période d'un an. Puisqu'il est bien connu que ces bases de données ne couvrent pas de façon satisfaisante les sciences humaines et sociales, un coefficient de 2 est attribué à chaque publication indexée dans l'indice de citation des sciences sociales (*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) de l'institution (Liu et Cheng, 2005). Ce critère est « ignoré » quand le nombre de personnes équivalent temps plein n'a pas pu être obtenu. Ceci définit le critère PY.

2.5 Source des données

Hormis pour le nombre de personnes équivalent temps plein 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 données brutes utilisées par les auteurs du classement ne sont pas communiquées.

2.6 Normalisation et agrégation

Chacun des six critères est « mesuré » 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 (donc un pourcentage). 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 prétendent que des ajustements sont réalisés, quand l'analyse révèle des « anomalies statistiques ». La 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 nouveau score « global » entre 0 et 100. Les scores finaux sont à nouveau normalisés, de sorte que la meilleure institution reçoive le score 100 et les autres un pourcentage. Ce score final normalisé est utilisé pour classer les institutions.

2.7 Les résultats en 2008

Le tableau 1 présente la liste des 20 meilleurs universités mondiales d'après l'édition 2008 du classement de Shangaï. Le tableau 2 présente la liste des 20

meilleures universités européennes.

Un regard rapide au tableau 1 montre une domination flagrante des universités américaines dans ce classement : seulement 3 parmi les 20 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. Des pays relativement petits comme la Suisse, la Norvège, la Suède ou les Pays-Bas, semblent mieux tirer leur épingle du jeu dans le classement que des pays plus grands 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.

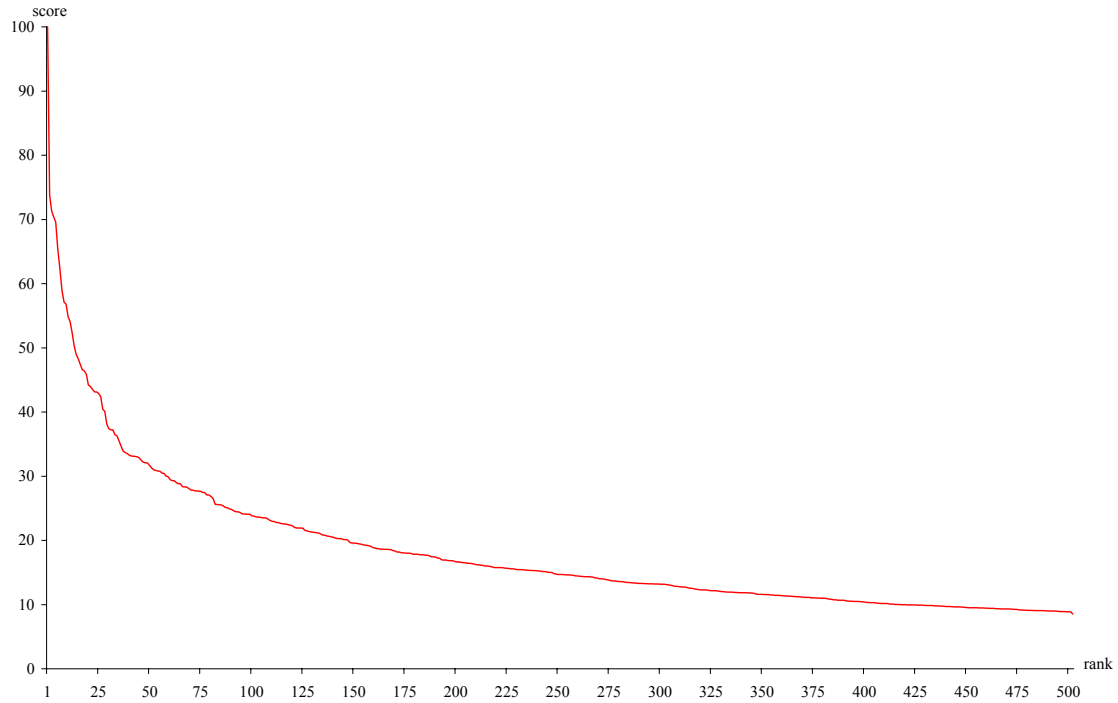


FIGURE 1 – Distribution des scores normalisés des 500 universités du classement de Shangaï (2008). Source : Shanghai Jiao Tong University, Institute of Higher Education (2003–09).

Nous voudrions conclure cette brève présentation du classement de Shangaï par trois affirmations extraites de Liu et Cheng (2005). Le classement de Shangaï

- « *utilise des critères objectifs sélectionnés avec soin* »,
- « *est fondé sur des données internationales comparables que tout le monde peut vérifier* »,

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 (2008). Source : Shanghai Jiao Tong University, Institute of Higher Education (2003–09).

Rang	Institution	Pays	ALU	AWA	HiCi	N&S	PUB	Py	Score
4	Cambridge	UK	90.3	91.5	53.6	56.0	64.1	65.0	70.4
10	Oxford	UK	59.0	57.9	48.4	52.0	66.0	45.7	56.8
22	U Coll London	UK	31.2	32.2	38.6	44.3	65.8	35.4	44.0
24	Swiss Fed Inst Tech - Zurich	Suisse	35.9	36.3	36.1	38.1	53.6	56.0	43.1
27	Imperial Coll	UK	18.6	37.4	39.9	38.2	61.8	39.4	42.4
40	U Manchester	UK	24.4	18.9	28.2	28.3	60.5	30.4	33.6
42	U Paris 06	France	36.6	23.6	23.1	27.3	58.2	21.3	33.1
45	U Copenhagen	Danemark	27.4	24.2	26.3	25.4	54.5	33.4	33.0
47	U Utrecht	Pays-Bas	27.4	20.9	28.2	28.8	53.3	26.0	32.4
49	U Paris 11	France	33.3	46.2	14.6	20.4	47.0	23.1	32.1
51	Karolinska Inst Stockholm	Suède	27.4	27.3	31.8	18.3	50.1	25.7	31.6
53	U Zurich	Suisse	11.2	26.8	24.7	27.5	50.2	32.4	31.0
55	U Edinburgh	UK	20.2	16.7	26.3	32.3	49.7	30.0	30.8
55	U Munich	Allemagne	33.1	22.9	16.3	25.6	52.7	31.8	30.8
57	TU Munich	Allemagne	41.1	23.6	25.3	18.9	44.8	30.6	30.5
61	U Bristol	UK	9.7	17.9	28.2	28.1	47.8	33.5	29.5
64	U Oslo	Norvège	23.1	33.4	17.9	17.0	46.7	29.8	29.0
67	U Heidelberg	Allemagne	17.7	27.2	17.9	20.4	49.2	29.3	28.4
68	U Helsinki	Finlande	16.8	17.9	21.9	20.8	53.8	30.1	28.3
70	Moscow State U	Russie	49.1	34.2	0.0	8.3	53.2	33.4	28.1

TABLEAU 2 – Les 20 meilleures universités européennes dans le classement de Shangaï (2008). Source : Shanghai Jiao Tong University, Institute of Higher Education (2003–09).

– et « *ne comporte aucune mesure subjective* » [6].

3 Une analyse critique des critères

Nous commençons notre analyse du classement de Shangaï par un examen détaillé des six critères utilisés.

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*. Ce n'est peut-être pas un problème majeur pour les médailles Fields qui ne récompensent que des candidats de moins de 40 ans. En revanche, c'est une difficulté sérieuse pour les prix Nobel. En effet, un examen minutieux de la liste de ces prix révèle 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 conduisit sa recherche alors qu'il était employé par l'office suisse des brevets à Zurich. Il reçut le prix Nobel longtemps après, lorsqu'il était affilié à l'université de Berlin (nous reviendrons sur l'exemple d'Albert Einstein plus loin). 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 a très peu à voir 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 au passage que le mécanisme d'actualisation adopté est

6. 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* »

complètement arbitraire (pourquoi utiliser un schéma linéaire et non pas exponentiel ?).

Les options retenues par les auteurs du classement sur ces deux critères induisent d'autres biais. Un biais en faveur des pays ayant connu très peu de changements radicaux sur le plan politique depuis 1901. Un biais en faveur des institutions ayant été créées depuis longtemps et ayant conservé la même appellation depuis leur création. Ce n'est pas le cas partout, et notamment en Europe (il suffit de penser aux guerres et aux changements politiques radicaux intervenus dans nos pays depuis 1901). D'ailleurs, en Europe, il y a parfois des situations absurdes. Par exemple, les deux universités *Free University of Berlin* et *Humboldt University* créées à Berlin après la partition de l'Allemagne se sont querellées pour savoir laquelle devait recevoir l'attribution du prix Nobel d'Albert Einstein (voir Enserink (2007) sur ce cas stupéfiant). 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.

Malheureusement, l'Allemagne n'est pas un cas isolé. Le tableau 3 donne la liste de quelques lauréats français des prix Nobel avec leur affiliation, telle qu'elle est mentionnée sur le site officiel (`http://nobelprize.org/nobel_prizes/`). Un examen attentif de ce tableau révèle plusieurs faits intéressants. Tout d'abord, il y a dans cette liste des institutions qui n'ont *jamais* existé (ou alors peut-être au Moyen-Âge). Par exemple, l'« Université Sorbonne », que ce soit en français ou en anglais, n'a jamais existé en tant que telle et n'existe pas actuellement (même si un bâtiment du Quartier Latin à Paris porte ce nom). Un autre problème se cache dans les affiliations multiples. À quelle institution doit-on rattacher le prix Nobel de Louis de Broglie ? Soit à l'Université Sorbonne (qui n'existe pas comme nous venons de le dire) ou à l'Institut Henri Poincaré, qui n'est pas non plus une université. Il y a aussi de nombreux cas où le prix Nobel est affecté à la bonne institution, mais cette institution n'existe plus (Université de Strasbourg, Université de Toulouse, Université de Nancy, ...)[7]. Un autre cas intriguant est celui du prix attribué à George Charpak qui a deux affiliations ESPC et CERN. Le CERN n'étant pas une université, le prix est-il attribué dans sa totalité à l'ESPC ou bien seulement en partie ? (notons au passage que l'ESPC est le nouveau nom de l'*École municipale de physique et de chimie industrielle*). Tout ceci montre que la correcte attribution

7. Remarquons au passage une certaine similitude entre ces anciennes appellations et les appellations des nouveaux Pôles de Recherche et d'Enseignement Supérieur. Gageons qu'il s'agit d'un hasard et que nos décideurs politiques ne font pas volontairement coller la structuration et l'appellation des établissements d'enseignement supérieur français aux dénominations utilisées sur ce site, dans l'objectif de se voir attribuer tous ces prix. Comme nous l'indiquerons plus loin, l'Université de Strasbourg existe de nouveau depuis 2009. À ce sujet, rappelons que Strasbourg était en territoire annexé par l'Allemagne en 1909 lorsque le prix Nobel de Karl Braun a été attribué.

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.

Nous voudrions conclure ici par une nouvelle désolante. Le dernier lauréat français du prix Nobel de Physique, Albert Fert, est Professeur à l'Université de Paris Sud. Il travaille 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). Nous ne comprenons vraiment pas pourquoi.

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 entachée d'incertitude, d'imprécision et de mauvaise détermination (sur ces notions, voir Bouyssou, 1989, Roy, 1988).

3.2 Les chercheurs les plus cités

Ici, le fait le plus saisissant est la dependence complète des auteurs du classement par rapport aux choix faits par Thomson Scientific. Le découpage de la science en 21 domaines proposé par Thomson Scientific est-il pertinent ? Au regard du tableau 4, 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 gênèrent de nombreux papiers sur des sujets sensibles et médiatiques. Que ce choix soit approprié pour évaluer des universités mériterait d'être discuté et, malheureusement, 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 regardant le nombre de journaux dans chaque catégorie (mais rappelons-nous aussi 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, *Engineering* 977, *Social Science General* 1299, et *Clinical Medicine* 1305 [8].

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. Les couplages entre articles ont des vides (dus à des appellations erronées, des numéros de pages faux, des problèmes d'homonymes,

8. Source : `http://www.isihighlycited.com/isi_copy/Comm_newse04.htm#ENGINEERING`, dernier accès le 19 juillet 2009.

Institution	Nom	Prix	Date
Sorbonne University	Henri Moissan	Chimie	1906
Sorbonne University	Gabriel Lippmann	Physique	1908
Sorbonne University	Marie Curie	Chimie	1911
Sorbonne University	Charles Richet	Médecine	1913
Sorbonne University	Jean Perrin	Physique	1925
Sorbonne University & Institut Henri Poincaré	Louis de Broglie	Physique	1929
Strasbourg University	Karl Braun	Physique	1909
École municipale de physique et de chimie industrielle	Pierre Curie	Physique	1903
Nancy University	Victor Grignard	Chimie	1912
Toulouse University	Paul Sabatier	Chimie	1912
University of Grenoble	Louis Néel	Physique	1970
Université de Paris	Jean Dausset	Médecine	1980
Université Louis Pasteur & Collège de France	Jean-Marie Lehn	Chimie	1987
ESPC & CERN	Georges Charpak	Physique	1992
Collège de France	Pierre-Gilles de Gennes	Physique	1991
Collège de France & École Normale Supérieure	Claude Cohen-Tannoudji	Physique	1997

TABLEAU 3 – Quelques lauréats français du prix Nobel (source : `http://nobelprize.org/nobel_prizes/`, dernier accès le 19 juillet 2009.)

21 catégories utilisées par Thomson Scientific	
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 4 – Les 21 catégories utilisées par Thomson Scientific (source : `http://www.isihighlycited.com/isi_copy/Comm_newse04.htm#ENGINEERING`, dernier accès le 19 juillet 2009.)

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 leur système utilise des comptages de citations.

Ici aussi, le problème de la bonne affectation des personnes aux institutions n'est pas trivial. Par exemple, dans la discipline « Computer Science », la France compte huit personnes « *highly cited* ». Sans citer de nom (disponibles sur le site), les institutions récompensées sont l'« INRIA-Futurs », l'« Université Bordeaux 1 », l'« Institut de Recherche en Informatique de Toulouse », l'« INRIA », l'« Institut de Mathématiques Luminy », l'« École Polytechnique », le « CNRS » et l'« École Normale Supérieure ». A priori, les seules institutions qui bénéficient de ces personnalités pour leur classement sont donc l'Université de Bordeaux 1, l'École Polytechnique et l'École Normale Supérieure, les autres institutions étant des laboratoires ou des organisme de recherche. Ces personnes sont, à quelques exceptions près, des directeurs de recherche au sein d'organismes de recherche, leur renommée ne profite donc à aucune université française. Toujours dans cette discipline, ces 8 noms sont à mettre en regard des 253 noms d'américains, qui, quant à eux, sont pour la plupart affiliés à des universités bien identifiées.

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 pas particulièrement jeunes et sont très susceptibles d'avoir changé d'institution, parfois plusieurs fois, dans leur carrière.

En résumé, si on combine la dependence 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 une recherche de fort impact.

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 (celui qui soumet l'article), 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). Nous tenons à souligner que ceci est largement contre-intuitif, voire paradoxal. Soulignons également que les problèmes d'affiliation indiqués précédemment sont également présents pour ce critère.

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 un nombre important de problèmes.

Tout d'abord, comme nous l'avons vu plus haut, l'attribution 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 » [9] (Liu et al., 2005). Cela conduit à de nombreux problèmes. Tout d'abord, 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

9. 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* »

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 de 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 [10]. Affecter à chaque auteur une affiliation correcte est une tâche difficile nécessitant une connaissance détaillée des spécificités institutionnelles de chaque pays.

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 soient analysés systématiquement. Ce n'est pas ce qui est fait dans le classement de Shangaï.

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, Moed, Tijssen, Visser et van Rann, 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 (pensons au Droit par exemple) 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 sciences humaines et sociales, 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 sciences humaines et sociales 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 sensibilité des résultats à la valeur de ce coefficient doit 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

10. Soulignons que même les experts en bibliométrie peuvent ignorer certains de ces problèmes. Van Raan (2005a) mentionne qu'un article dont l'affiliation est « CNRS, Illkirsh » doit être attribué à l'*Université Louis Pasteur*. Même si cela est certainement exact, ce n'est pas systématiquement correct puisque cela peut signifier que l'article est signé par un chercheur du CNRS travaillant dans une unité mixte de recherche avec l'Université Louis Pasteur. Notons également que l'Université Louis Pasteur est maintenant fusionnée avec deux autres universités pour constituer l'Université de Strasbourg, actuellement la plus grande université française.

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.

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

3.5 Productivité

Le critère PY est le score total des 5 premiers indicateurs divisé par le nombre d'équivalents temps plein du personnel 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 n'indiquent pas les sources qu'ils ont utilisées pour trouver l'information sur le nombre de personnes équivalent temps plein. Nous n'avons aucune raison de penser que ces sources sont plus fiables que celles qu'ils ont utilisées pour évaluer les cinq premiers critères. Si les informations sont fournies par les institutions elles-mêmes, il y a un risque potentiellement sérieux, puisque la notion de « membre du personnel 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 enseignants non chercheurs, les ingénieurs de recherche, les chercheurs purs, 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 sciences sociales 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 sciences sociales 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 sciences sociales 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 sciences sociales 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 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 Sciences humaines et sociales. Ils ne donnent pas non plus la liste des institutions pour lesquelles le critère Py pouvait être obtenu. En résumé, non seulement la famille des critères varie selon les institutions, mais de plus il est impossible de savoir quelle famille de critères est utilisée pour évaluer une institution. C'est extrêmement surprenant et 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 (un comportement qui ne semble pas franchement en accord avec les motivations académiques qu'ils annoncent), 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 très grande 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 (Ebel et Frisbie, 1991, Ghiselli, 1981, Green, Tull et Albaum, 1988, Kerlinger et Lee, 1999, Kline, 2000, Nunally, 1967, Popham, 1981), s'est concentrée sur la question de la validité et de la fiabilité. La seconde, issue de l'aide multicritère à la décision (Bouyssou, 1990, Fryback et Keeney, 1983, Keeney, 1981, 1988a,b, 1992, Keeney et McDaniel, 1999, Keeney et Raiffa, 1976, Keeney, Hammond et Raiffa, 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 ces domaines aient été largement ignorés par les auteurs du classement.

3.8 Le temps et la taille

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

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 une période de 3 ou 4 ans. Les contrats des établissements et des unités de recherche se font sur un horizon de 4 ans. La Commission du Titre d'Ingénieurs délivre l'habilitation à délivrer le titre d'ingénieur pour une période allant jusqu'à 6 ans. Par conséquent, les changements de rang 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 Bazerman, 1990, Dawes, 1988, Hogarth, 1987, Kahneman, Slovic et Tversky, 1981, 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. Une réponse implicite est donnée 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 [11]. Il y a peut-être une justification derrière ces choix, mais elle n'est pas donnée par les auteurs du classement. Comme le souligne van Raan (2005a), la « performance académique » peut avoir deux significations très différentes : le prestige d'une institution fondé sur ses performances passées et sa capacité actuelle à attirer d'excellents chercheurs. Il est importante 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 N&S, articles indexés par Thomson Scientific). Il n'est donc pas surprenant que la valeur de ces critères soient fortement corrélée à la taille de l'institution. Comme cela est montré avec force par Zitt et Filliatreau (2006), l'utilisation de tels critères relatifs à la taille de l'institution est le signe que « *big is made beautiful* ». Le fait que les critères soient fortement corrélés n'est donc pas une surprise. Au contraire des auteurs du classement, il ne faut pas voir dans cette forte corrélation des critères utilisés un signe de pertinence.

De plus, 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 entre les institutions dès que l'on quitte la tête du classement. Une simple analyse statistique révèle que c'est en effet le cas. Le tableau 5 donne la valeur moyenne des cinq premiers critères utilisés par le classement en fonction du rang des institutions (chiffres fondés sur le classement 2008).

Il est clair que les critères ALU et AWA ne contribuent quasiment en rien au classement des institutions qui ne sont pas dans le top 100. 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.

11. Soulignons ici que cette période de temps nous semble extrêmement courte. Le calcul des facteurs d'impact des journaux par Thomson Scientific utilise une période de 2 ans, voir `http://admin-apps.isiknowledge.com.gate6.inist.fr/JCR/help/h_impfact.htm#impact_factor`, dernier accès le 19 juillet 2009 et des facteurs d'impact sur des périodes de 5 ans sont également calculés.

Rank	ALU	AWA	HiCi	N&S	PUB
0–97	26.32 (10%)	26.15 (17%)	35.57 (1%,16.2)	31.96 (0%,15.1)	54.56 (12.4)
101–200	8.03 (52%)	5.18 (69%)	18.02 (1%, 6.4)	16.87 (0%, 5.3)	41.62 (8.0)
201–302	5.01 (65%)	1.08 (92%)	10.84 (16%, 5.9)	12.24 (1%, 3.6)	35.52 (7.1)
303–401	3.05 (82%)	1.55 (91%)	8.04 (23%, 5.3)	8.06 (1%, 3.6)	29.13 (6.8)
402–503	0.71 (94%)	0.45 (96%)	5.25 (37%, 4.3)	6.50 (7%, 3.0)	26.70 (5.8)

TABLEAU 5 – Valeurs moyennes pour les critères en fonction des rangs (source : classement 2008). 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.

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. Cette analyse est dans la même lignée que les observations de van Raan (2005a) et de Ioannidis et al. (2007). Nous nous concentrons maintenant sur les questions inhérentes 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 !

Le second étudiant propose une technique simple fonctionnant de la façon suivante. Pour chaque pays, il a trouvé le PIB (Produit Intérieur Brut) et le PIBph (PIB par habitant). L'étudiant suggère alors 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 utilise le PIB, soit on veut mesurer la richesse moyenne de ses habitants et on utilise le PIBph. Mais une mesure qui combine les deux en prenant une combinaison linéaire n'a pas de sens : le premier critère mesure la *production*, le second la *productivité*. Prendre α fois la production plus $(1-\alpha)$ fois la productivité est une opération légitime sur le plan arithmétique, mais c'est une opération qui n'a pas de sens, excepté bien entendu si α vaut 1 ou 0. Le lecteur qui n'est pas totalement convaincu de 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 la copie 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 mauvaise

Une des premières choses enseignées systématiquement dans n'importe quel cours de base en aide multicritère à la décision est la suivante. Si vous agrégez 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 *facteurs 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, vous pouvez choisir de mesurer un critère en mètres, mais vous pouvez aussi décider le mesurer en kilomètres. Si vous utilisez les mêmes poids pour ce critère dans les deux cas, vous obtiendrez 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 seule base d'une vague notion d'« importance » des critères. La comparaison des poids ne reflète pas une comparaison de leur importance. En effet, 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 1000 si vous décidez 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 web). Cela soulève également le problème sur la façon dont les auteurs du classement ont choisi leur jeu de poids. Ils n'expliquent rien sur ce point. Tout porte à croire que les poids ont été fixés arbitrairement. La seule explication qui nous semble rationnelle est que, dans les premières versions du classement, les auteurs du classement avaient utilisé uniquement cinq critères avec des poids égaux. Même 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 [12]. À cause du changement de normalisation, ceci revient donc à utiliser chaque année des poids différents.

Illustrons ce qui peut se passer avec un simple exemple utilisant deux critères. Considérons les données du tableau 6. 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' comme 'Harward'). 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 6 (il n'est pas nécessaire ici de normaliser à nouveau le score 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) :

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

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 7. 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

12. Keeney (1992, p. 147) appelle ceci « l'erreur critique la plus habituelle » (« *the most common critical error* »).

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 6 – Somme pondérée : exemple avec des poids identiques

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. Parce que la normalisation a changé, les poids associés à ce critère devraient changer si l'on souhaite être cohérent : au lieu d'utiliser des poids égaux à 0.5, il faudrait utiliser d'autres poids pour refléter les changements de normalisation.

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

Ce « paradoxe » peut être renforcé en utilisant un autre exemple. Reprenons les données indiquées dans le tableau 6. Supposons que rien n'ait changé sauf que h augmente sa performance sur g_2 et que a, qui est en deuxième position, augmente ses performances à la fois sur g_1 et sur g_2. Intuitivement, on est en droit de s'attendre à ce que, puisque h et a occupaient les deux premières positions et qu'elles seules se sont améliorées, elles restent en tête de classement. Cette intuition

est trompeuse. Les détails sont fournis au tableau 8. Le classement final devient :

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

À nouveau, une situation contre-intuitive s'est produite : l'institution a, qui était en deuxième position et qui était la seule avec h à avoir progressé, est maintenant classée dernière !

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	**165**	**450**	8.25	64.29	36.27	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 8 – Somme pondérée avec des poids identiques : h s'améliore sur g_2 et a s'améliore sur g_1 et g_2

Notons que le fait de ne pas changer les poids quand l'échelle change a un effet très étrange, en plus de celui mentionné précédemment. Si une institution b est faible sur un critère, de sorte qu' une institution concurrente a est 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, l'amélioration du meilleur 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 !

Puisque les auteurs du classement sont tombés dans ces pièges élémentaires, on peut tout à fait conclure que leurs résultats n'ont aucune valeur. Même si l'exemple précédent était bien choisi, on comprend que chaque année, pour peu que Harward modifie ses performances, la suite du classement peut être complètement bouleversée, même si rien d'autre ne change. Qu'une institution modifie ses scores à la marge sur tel ou tel critère ne fait qu'ajouter à la confusion ; en aucun cas, on ne peut conclure que « s'améliorer sur un critère permet de monter dans le classement ». Inversement, nous laissons le soin au lecteur de construire un exemple où une institution dont les performances se dégraderaient se retrouverait mieux classée qu'une institution dont les performances seraient inchangées (voire se seraient améliorées légèrement). 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 [13].

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 lié au lien entre poids et normalisation évoqué plus haut), il faut souligner que la somme pondérée des critères 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 poids des critères.

Malgré cela, en utilisant une somme pondérée des critères, il reste des alternatives non dominées qui ne pourront *jamais* être classées en première position, quelle que soit la valeur des poids. Le tableau 9 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 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.

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).

13. Pour s'en convaincre, le lecteur pourra tester un classement avec les données du tableau 6 en utilisant les poids 0.6 et 0.4.

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

TABLEAU 9 – Somme pondérée : alternatives efficaces non supportées

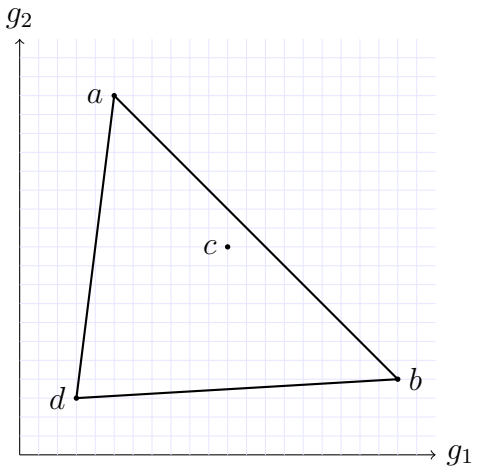


FIGURE 2 – Alternatives efficaces non supportées

4.3 La technique d'agrégation utilisée n'a pas de sens

Les critères ALU, AWA, HiCi, N&S et PUB sont des critères de *comptage*. Il est clair, sous réserve des remarques faites à la section 3, qu'ils sont globalement relatifs à l'aptitude qu'a une institution à produire une grande quantité de bons papiers et de bons chercheurs. Ils mesurent le potentiel de recherche d'une institution. Ceci est sémantiquement cohérent. En revanche, le critère Py est totalement différent. 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 (une fois de plus) ce soit possible 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 indique un manque de réflexion sur ce qui doit guider la construction d'un indice. 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 Quelques 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 quelque peu mouvementée. Après 1968, la plupart d'entre elles ont été découpées en plus petites universités (ceci explique pourquoi le nom de la plupart des universités françaises comporte un numéro). De plus, il y a des universités créées récemment en France qui sont petites et qui n'offrent pas de formation dans tous les domaines de la Science et/ou qui n'offrent pas des programmes de la licence au doctorat dans tous ces domaines. Enfin, quand on regarde le 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 d'ingénieurs, de gestion et de sciences politiques) 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 étudiants sont payés (les Écoles Normales Supérieures).
- 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 (spécialisé en recherche biomédicale), l'INRA (spécialisé en sciences de l'agriculture) et l'INRIA (spécialisée en informatique). Une part très significative de la recherche en France est effectuée dans de telles institutions, même si elles 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 sans doute l'Institut Pasteur (beaucoup des prix Nobel français en médecine y sont liés).

Avec un paysage institutionnel aussi complexe, ce qui doit compter comme une université est loin d'être évident. 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 le peuvent aussi. 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 devrait au moins commencer par définir clairement les objets qui sont évalués. Cette définition ne figure pas dans le 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* »). Mais, 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 la définition implicite d'université de classe mondiale, retenue par les auteurs du classement, qui ignore complètement les contraintes que peuvent subir les institutions.

Certaines universités ont une (plus ou moins) totale liberté pour organiser leur gouvernance, pour embaucher 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 (ce qui 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. Sachant ces différences dans les contraintes des institutions, doit-on simplement les ignorer, comme c'est le cas implicitement dans le classement ? Ce n'est le cas que si l'on admet qu'il existe un unique modèle d'université de classe mondiale. Cette hypothèse nécessiterait une justification

approfondie qui n'est pas donnée par les auteurs du classement.

De la même façon, les ressources (*inputs*, intrants) 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 « personnel 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'Harward était de plus de $\$3.10^9$ 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'Harward possède plus de 15.10^6 volumes, Harvard University, 2007), la liberté de chercher et de publier dans tous les domaines d'intérêt, etc. sont aussi des ingrédients du succès très importants pour une université. En ignorant toutes ces ressources, on a inévitablement une vision très étroite de l'excellence académique [14].

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. issu d'un pays n'ayant pas connu de changements majeurs sur le plan politique et social depuis 1901,
3. issu 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. issu 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,
7. qui a de grands moyens financiers.

Nous ne prétendrons pas qu'une telle définition implicite n'a pas de sens. 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.

14. 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*).

4.4.3 Quel est l'objectif du modèle ?

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 (bien que nous réalisons que chaque famille a tendance à considérer son enfant comme un futur prix Nobel) 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, la puissance 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 des étudiants sont peu susceptibles d'être impressionnés par quelques prix Nobel délivrés il y a longtemps à des membres d'un certain département d'une institution, 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, les présidents d'universités 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 (aussi surprenant que cela semble, cela a pu arriver), nous ne voyons pas comment un classement global d'une institution peut constituer un outil de gestion utile. Que cela puisse servir comme argument de communication est clair, mais est sans grand rapport avec la science.

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 aucunement 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. À moins que les auteurs du classement ne soient en mesure de fournir des preuves empiriques montrant que la productivité scientifique d'une institution est lié à sa taille et que les institutions de taille moyenne ne peuvent pas produire de recherche de qualité, ce qu'ils ne font pas, nous ne comprenons pas pourquoi tout cela peut intéresser les décideurs politiques, sauf, bien entendu, pour soutenir leurs objectifs stratégiques. Mais là aussi, cela est sans rapport avec la science.

4.4.4 Quelques bonnes pratiques d'évaluation

Comme expliqué dans Bouyssou et al. (2000) et Bouyssou et al. (2006), il y a quelques bonnes pratiques qu'il faut suivre quand on construit 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 de bonne conduite élémentaire a été oubliée par les auteurs de ce classement.

La seconde règle que nous souhaitons mentionner 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. C'est l'enseignement de base de tout cours introductif 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, Dörner, 1996, Hatchuel et Molet, 1986, Mintzberg, 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. Les seuls conseils de prudence sont « tout exercice de classement est sujet à controverse, et aucun classement n'est absolument objectif » et que « Les personnes doivent être prudentes avec tout classement et ne doivent pas être totalement confiantes en un classement, y compris le '*Academic Ranking of World Universities*'. Au lieu de cela, les personnes doivent utiliser les classements comme une sorte de référence et lire la méthodologie de classement avec attention avant de lire les listes fournies par le classement » (Shanghai Jiao Tong University, Institute of Higher Education, 2003–

09) [15]. 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 dirigez une université et que vous voulez augmenter votre position dans le classement. C'est, en fait, 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 Sciences Humaines et Sociales. 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 comme le football où quelques riches équipes sont en compétition pour attirer les meilleurs joueurs mondiaux. Nous ne sommes pas persuadés, pour dire le moins, qu'il s'agisse du meilleur moyen pour permettre aux hommes d'accroître leurs connaissances.

Pour un gouvernement, utiliser le classement de Shangaï pour effectuer des réformes est aussi simple que potentiellement dangereux. 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 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 Sciences Humaines et Sociales), c'est-à-dire Paris 5, 6, 7 et 11 (ce ne sont pas les noms officiels mais leur appellation la plus répandue), conduira (en utilisant les données du classement 2007) à une institution qui sera globalement au niveau de Harward. Bingo ! On ne dépense pas un Euro de

15. 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* »

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 ! Peut-on faire plus ? Certainement. Les centres de recherche publics (CNRS, INSERM, etc.), bien que très efficaces, ne comptent pas dans le classement. Vous pouvez certainement les supprimer et transférer les fonds et les personnes dans cette énorme organisation qui vient d'être créée. Avec tout cela, on battra sans aucun doute Harward... Inutile de dire que toutes ces manipulations pourront avoir des effets désastreux à long terme. Mais comme ce ne sera qu'à long terme... et il nous faut ici admettre que les décideurs politiques français ne semblent pas échapper à la règle voulant que les décideurs politiques ont souvent une bien étrange manière de gérer les nécessaires compromis entre les effets à court terme et les effets à long terme... Il semble même qu'en France, ce chantier catastrophe a déjà commencé. En témoignent les trois universités de Strasbourg qui ont fusionné, les regroupements sous forme de Pôles de Recherche et d'Enseignement Supérieur partout en France et la récente loi sur l'organisation et l'autonomie des universités publiques [16].

5 Que faire ?

Résumons à présent nos observations sur le classement de Shangaï et essayons de dégager quelques conclusions sur les bases de ce que nous avons trouvé, à la fois au niveau scientifique et à un niveau plus stratégique.

5.1 Que penser du classement de Shangaï ?

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 » [17] 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 » [18]. Nous sommes bien entendu entièrement d'accord. Les conclusions de van Raan étaient principalement fondées sur des considérations bibliométriques, à propos desquelles les auteurs du

16. numéro 2007-1199 et datée du 10 août 2007

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

18. 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* »

classement ont été incapables de répondre de façon convaincante (Liu et al., 2005, van Raan, 2005b).

Notre propre analyse est fondée sur un point de vue qui reflète notre intérêt pour 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. En effet, 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 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.

Dès lors, il n'est pas injuste de conclure que *le classement de Shangaï est un exercice qui n'a absolument aucune valeur*. N'importe lequel de nos étudiants qui aurait proposé une telle méthodologie dans un mémoire de master aurait certainement été recalé.

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). Nous laissons son étude détaillée au soin du lecteur intéressé (pour une critique vigoureuse du classement du *US News & World Report*, nous renvoyons à Ehrenberg, 2003).

5.2 Quelques pistes

Un point de vue optimiste serait que, après avoir lu notre rapport, les auteurs du classement décident immédiatement d'arrêter ce travail en s'excusant d'avoir créé autant de confusion dans le monde académique et que les décideurs politiques décident immédiatement d'arrêter d'utiliser des classements internationaux « bien connus » comme outil de promotion de leurs objectifs stratégiques. Toutefois, nous vivons dans le monde réel et nous craignons que cela n'arrive pas. Dans la mesure où les auteurs du classement ont décidé d'ignorer plus ou moins complètement l'avis de van Raan (2005a), nous pensons qu'ils en feront de même avec le notre [19]. Nous pensons donc qu'ils continueront encore longtemps à produire leur classement annuel. De même, nous jugeons peu vraisemblable que les décideurs politiques de nos pays abandonnent des arguments aussi simples et si bien repris par les médias. Dès lors, il nous faudra vivre dans un monde où des classements très médiocres

19. Les premiers échanges que nous avons eus avec les auteurs du classement au sujet de notre rapport, qui leur a été envoyé, laissent en effet à penser qu'ils ne s'arrêteront pas et qu'ils ne semblent pas vouloir démarrer une discussion méthodologique sur des bases solides.

seront régulièrement publiés et utilisés. Qu'est-ce qui peut être fait ? Plusieurs choses.

La première chose, et la plus facile, serait d'arrêter d'être naïfs. « Quel est le meilleur vin du monde ? », « Quelle est la meilleure voiture du monde ? », « Quelle est la ville la plus agréable en Europe ? », etc. Toutes ces questions peuvent être intéressantes si votre objectif est de vendre de nombreux exemplaires de votre journal ou de votre livre. En dehors de cela, il est évident que ces questions n'ont aucun intérêt, à moins d'être précédées par un long et difficile travail de structuration. Soyons plus précis : la « meilleure voiture du monde » est un concept vide de sens si l'on n'a pas identifié les futurs propriétaires du véhicule, structuré leurs objectifs, étudié les attributs qui pourraient servir à mesurer l'atteinte de ces objectifs, appliqué des procédures d'agrégation raisonnables de ces informations et effectué une analyse de robustesse importante. En faisant ainsi, il sera possible d'obtenir un modèle qui pourra éventuellement aider une personne à choisir une voiture, ou encore qui pourra aider un gouvernement à préparer de nouveaux standards pour l'émission de gaz non polluants. Mais sans ce travail, la question n'a aucun sens [20]. Si la question n'a aucun sens pour les voitures, doit-on s'attendre à un miracle si l'on se penche sur des objets incroyablement plus complexes comme les universités ? Certainement pas. Il n'existe pas de « meilleure université du monde » *in abstracto*, pas plus qu'il n'existe de « meilleure voiture du monde ». Une première mesure immédiate que nous suggérons est la suivante. Arrêtons de parler de tous ces classements. Ils n'ont pas de sens. Expliquons cela dans nos institutions pour que ces classements ne soient pas utilisés dans leur communication. C'est bien entendu particulièrement important pour nos lecteurs « chanceux » qui appartiennent à des institutions bien classées. Ils doivent résister à la tentation de dire ou de penser « qu'il y a bien une part de vérité derrière tout ça » et arrêter d'utiliser cette publicité gratuite offerte par ces classements.

Puisque la production de classements médiocres n'est pas prête de s'arrêter, un moyen plus pro-actif de les combattre est de produire des classements alternatifs, qui produisent des résultats complètement différents. Il n'est pas nécessaire que ces classements soient bien mieux construits que le classement de Shangaï. Leur principale utilité sera de *diluer* ses effets particulièrement dévastateurs. Un pas dans cette direction a été fait par l'*École Nationale Supérieure des Mines de Paris* (ENSMP, 2007). Comme chacun sait, il s'agit d'une Grande École française très prestigieuse. Sa taille fait qu'elle n'apparaîtra jamais en bonne position dans le classement de Shangaï. Par conséquent, l'ENSMP a décidé de produire un classement alternatif au classement de Shangaï. Il peut être expliqué très simplement

20. C'est également vrai pour les vins. Quelques « experts » ont rendue populaire l'idée de noter les vins. Ce processus leur a sans doute permis de s'enrichir, mais il en a aussi résulté un appauvrissement incroyable de la variété des vins produits en Europe

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. En effet, les performances des institutions sont fondées sur des faits qui se produisent très longtemps après que le diplôme a été obtenu (il n'est, en effet, 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, et École Nationale d'Administration). Diluer les effets du classement de Shangaï nécessitera clairement davantage de classements de ce genre. Nous invitons nos lecteurs à contribuer à ce travail de dilution.

5.3 Pourquoi ne pas proposer votre propre classement ?

Nous avons été très négatifs sur le classement de Shangaï (mais, nous le pensons, pas injustes). Le lecteur peut à présent se demander de façon légitime pourquoi nous ne proposons pas un autre classement, qui serait meilleur. Nous ne le ferons pas ici pour plusieurs raisons.

Tout d'abord, ce n'est pas notre domaine de recherche. Nous ne sommes pas spécialisés dans les systèmes d'enseignement, nous ne sommes pas experts en bibliométrie. Contrairement aux auteurs du classement, nous pensons qu'un classement digne de ce nom, doit être réalisé par un groupe pluridisciplinaire d'experts spécialistes en évaluation, en systèmes d'enseignement et en bibliométrie.

Ensuite, nous ne sous-estimons pas le travail que nécessite la construction d'un tel classement. Même si nous pensons que le classement de Shangaï est de très faible qualité, le produire est une tâche énorme. Un meilleur classement nécessitera clairement davantage de travail encore.

D'autre part, nous ne sommes absolument pas convaincus qu'un tel exercice

puisse être utile à une personne sérieuse. Nous avons expliqué pourquoi nous serions tentés de classer des programmes ou des systèmes nationaux d'enseignement plutôt que des universités. Nous avons également donné quelques pistes sur ce qui, à notre sens, pourrait être fait ou être absolument prohibé. Nous espérons que ces indications seront utiles aux lecteurs qui souhaiteront se charger de cette lourde tâche.

Soulignons enfin 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'éducation supérieure, ce qui pourrait aider véritablement à améliorer la qualité de la science (la seule chose, après tout, qui nous intéresse réellement). Parce que la question des techniques bibliométriques a été déjà largement étudiée (voit Moed, De Bruin et van Leeuwen, 1995, Moed, 2006, van Raan, 2005c, 2006, Zitt, Ramanana-Rahary et Bassecoulard, 2005), nous allons nous concentrer ici sur quelques techniques de Recherche Opérationnelle potentiellement utiles.

Nous avons déjà mentionné qu'un manque de connaissances de base dans les techniques d'agrégation et leurs propriétés, un des domaines de l'aide multicritère à la décision, peut invalider une technique d'évaluation. Nous voulons signaler ici que 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, Jones et Sims, 1983, Friend et Hickling, 1987, Ostanello, 1990, Rosenhead, 1989) et qui peuvent se combiner avec d'autres outils d'agrégation sophistiqués (Ackermann et Belton, 2006, Bana e Costa, Ensslin, Corrêa et Vansnick, 1999, Belton, Ackermann et Shepherd, 1997, Montibeller, Ackermann, Belton et Ensslin, 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) dans la littérature anglo-saxonne (Banker, Charnes et Cooper, 1984, Charnes, Cooper et Rhodes, 1978, Cherchye, Moesen, Rogge, van Puyenbroeck, Saisana, Saltelli, Liska et Tarantola, 2008, Cook et Zhu, 2008, Cooper, Seiford et Tone, 1999, Norman et Stoker, 1991). L'utilité de ces techniques pour construire des modèles d'évaluation dans le système d'enseignement supérieur a déjà été reconnu (Bougnol et Dulá, 2006, Johnes, 2006, Leitner, Prikoszovits, Schaffhauser-Linzatti, Stowasser et Wagner, 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 particulièrement intéressante, développée par le Centre d'enseignement supérieur en Allemagne (Berghoff et Federkeil, 2009, Centre for Higher Education Development, 2008), et permettant à

chaque utilisateur du système de choisir les critères qu'il souhaite utiliser.

5.4 Le rôle de l'Europe

Nous pensons que l'Union européenne a une *grande* responsabilité dans la mise au point d'outils d'évaluation alternatifs. L'Europe a une histoire longue, complexe et, malheureusement, souvent violente. Elle est néanmoins riche de sa diversité politique, culturelle et linguistique. Mais nous avons vu qu'une telle diversité est un point négatif dans le classement de Shangaï. Les responsables de l'Union Européenne ont désormais un choix à faire. S'ils pensent qu'un classement particulièrement défectueux produit en Chine peut être une bonne base pour engager des réformes majeures dans les systèmes d'enseignement supérieur européens, ils sont fondés à ne rien faire du tout. S'ils pensent, au contraire, que la diversité culturelle, linguistique et politique de l'Europe a une quelconque valeur, alors une action urgente est nécessaire.

Alors que nous écrivions ce texte, nous avons découvert qu'un appel d'offre européen était lancé à ce sujet [21]. Nous voulions terminer ce rapport sur ce signal positif.

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.

C. A. Bana e Costa, L. Ensslin, É. C. Corrêa et J.-Cl. 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

21. Voir `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/08/1942&format=HTML&aged=0&language=EN&guiLanguage=en`, dernier accès le 19 juillet 2009. À l'heure où nous en écrivons la traduction en français, nous apprenons que cet appel d'offres a été remporté 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). Les autres partenaires sont le *Centre for Science and Technology Studies* de l'Université de Leiden (Pays-Bas), le *Research Division INCENTIM* de la Katholieke Universiteit Leuven (Belgique), l'Observatoire des Sciences et des Techniques Paris, et deux fondations : la *European Federation of National Engineering Associations* et la *European Foundation for Management Development*. Le projet durera deux ans et coûtera approximativement 1.1 millions d'euros. Nous souhaitons à ce consortium une grande réussite dans cette entreprise.

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/`.

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.

D. Dörner. *The logic of failure.* Perseus Books, Jackson, 1996.

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. *Orga-*

nizational 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.

D. G. Fryback et R. L. Keeney. Constructing a complex judgmental model: An index of trauma severity. *Management Science*, 29:869–883., 1983.

E. E. Ghiselli. *Measurement theory for the behavioral sciences.* W. H. Freeman, San Francisco, 1981.

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`.

P. E. Green, D. S. Tull et G. Albaum. *Research for marketing decisions.* Englewood Cliffs, 1988.

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`.

A. Hatchuel et H. Molet. Rational modelling in understanding and aiding human decision making: About two case studies. *European Journal of Operational Research*, 24:178–186, 1986.

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`.

D. Kahneman, P. Slovic et A. Tversky. *Judgement under uncertainty: Heuristics and biases*. Cambridge University Press, Cambridge, 1981.

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. Measurement scales for quantifying attributes. *Behavioral Science*, 26:29–36, 1981.

R. L. Keeney. Structuring objectives for problems of public interest. *Operations Research*, 36:396–405, 1988a.

R. L. Keeney. Building models of values. *European Journal of Operational Research*, 37(2):149–157, 1988b.

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

R. L. Keeney et T. L. McDaniel. Identifying and structuring values to guide integrated ressource planning at BC gas. *Operations Research*, 47(5):651–662, September-October 1999.

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 ap-

proach. *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. Mintzberg. *The structuring of organizations*. Prentice Hall, Englewood Cliffs, 1979.

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.

M. Norman et B. Stoker. *Data Envelopment Analysis: The Assessment of performance*. Wiley, London, 1991.

J. C. Nunally. *Psychmometric Theory*. McGraw-Hill, New York, 1967.

A. Ostanello. Action evaluation and action structuring: Different decision aid situations reviewed through two actual cases. In C. A. Bana e Costa, editor, *Readings in multiple criteria decision aid*, pages 36–57. Springer-Verlag, Berlin, 1990.

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.

W. J. Popham. *Modern educational measurement*. Prentice-Hall, New York, 1981.

E. C. Poulton. *Behavioral decision theory: A new approach*. Cambridge University Press, Cambridge, 1994.

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.

Shanghai Jiao Tong University, Institute of Higher Education. Academic ranking of world universities (ARWU), 2003–09.

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.