L'évaluation de la qualité des publications en économie

Clément Bosquet
Sciences Po
Aix-Marseille School of Economics, CNRS et EHESS
clement.bosquet@sciences-po.org
Pierre-Philippe Combes
ppcombes@univmed.fr
Aix-Marseille School of Economics, CNRS et EHESS

Résumé
Nous synthétisons ici une série de travaux veillant à évaluer la qualité de la recherche en économie en France. Nous prenons soin de présenter les enjeux de ces évaluations bibliométriques et de leurs choix méthodologiques ainsi que les prolongements possibles utiles à des fins de politique publique. L'accent est principalement mis ici sur l'évaluation de la qualité des revues scientifiques puis des institutions de recherche (centres et universités), cette dernière pouvant être réalisée à partir des évaluations moyennes des articles des revues ou bien à l'aide d'indicateurs de citations individuelles dont nous comparons les avantages respectifs. Nous résumons également les résultats d'études utilisant les nouvelles données de citations Google Scholar qui permettent d'élargir le type de support traditionnellement considéré (les revues) aux livres et documents de travail, à la fois du côté des cités et des citants.

Mots clés : évaluation de la recherche, production scientifique, qualité des publications

Abstract: This article summarizes some studies that evaluate the quality of research in economics in France. We pay particular attention to the methodological choices of these bibliometric assessments. We conclude by underlying a number of possible extensions useful for public policy. Most of the article is devoted to the evaluation of the quality of scientific journals and of research institutions (research centers and universities). Publication measures can be obtained from average scores of published articles in scientific journals or from individual citation indexes of each publication. As regards the latter, we present some results using Google Scholar citations, which allow us to consider a larger variety of publications (journal articles but also books and working papers) as regards both cited and citing items.

Introduction

Cette intervention n'a pas pour objet de revenir sur les bonnes ou mauvaises raisons de l'existence d'évaluation de la production scientifique. Cependant, si évaluation il doit y avoir, il semble légitime de se demander s'il existe des façons de la mener plus pertinentes que d'autres. Nous résumons ici une série de travaux évaluant l'impact de certains choix méthodologiques relatifs aux critères d'évaluation des publications académiques en économie. Au delà de la quantité publiée, la difficulté principale est relative à l'évaluation de la qualité des travaux de recherche.

Depuis plus de dix ans, les économistes proposent des évaluations quantitatives systématiques de la production scientifique.

Deux grandes questions émergent : quels supports (revues, ouvrages, etc.) sont considérés ? Comment est mesurée la qualité de la production ? La réponse la plus fréquente que l'on peut par exemple trouver dans Combes et Linnemer (2001) ou Combes et Linnemer (2003) est la suivante. L'activité scientifique est mesurée à l'aide des publications dans les revues à comité de lecture et la qualité moyenne de la revue est attribuée à chacun des articles considérés.

Il découle de ce type de choix une question subsidiaire. De quelle façon mesure-t-on la qualité des revues ? Deux approches s'opposent pour répondre à cette question. D'une part, il est possible d'utiliser des mesures strictes de la qualité des revues, issues d'une bibliométrie purement quantitative. D'autre part, il est possible d'utiliser des mesures issues de l'appréciation par les pairs, en général plus subjectives et multicritères. Cette opposition des deux méthodes ne doit pas laisser penser qu'elles sont totalement exclusives. Par exemple, la catégorisation des revues en économie et en gestion de la section 37 du CNRS

a opté officiellement pour la seconde solution tout en précisant « La section n'a cependant pas ignoré les indicateurs bibliométriques dans son analyse champ par champ, revue par revue. »[1] Néanmoins, dès que l'on s'éloigne de critères purement bibliométriques, il est tout de suite très difficile de justifier tel ou tel choix et de parvenir à des choix faisant l'unanimité. Le choix d'un indicateur bibliométrique fait rarement l'unanimité mais il présente l'avantage d'être complètement transparent quant à sa méthode de calcul.
Qu'elle soit faite par les pairs ou par la bibliométrie, l'évaluation moyenne d'une revue n'a, à l'origine, pas vocation à être unique ou à juger de la pertinence d'une ou de quelques publications données mais de refléter la qualité moyenne d'un ensemble évidemment bien plus large comme l'intégralité des articles de la revue, et ce souvent sur plusieurs années. Afin de mener une entreprise d'évaluation de publications particulières, s'appuyer sur la qualité moyenne des revues peut mal évaluer, en positif comme en négatif, le véritable impact de ces publications. La variation intra-revue de l'impact de ses articles est en effet tout à fait considérable. Cependant, il existe désormais des approches plus directes qui permettent à la fois d'étendre le support de publications considérées (à des ouvrages par exemple) et d'individualiser la mesure de l'impact scientifique. Sont alors utilisés directement des indices de citation du chercheur ou de ses publications, sans passer par l'impact moyen de la revue dans lequel elles sont publiées. Google Scholar, par exemple, ne considère non seulement plus uniquement les revues à comité de lecture mais tout type de supports « académiques » présents sur internet incluant les ouvrages et les documents de travail. Il permet aussi d'attribuer à chaque support non pas la qualité moyenne des supports équivalents mais son propre nombre de citations. Ainsi, deux articles publiés dans la même revue mais avec des nombres de citations différents peuvent être considérés comme d'impact différent.
Le reste de cet article est organisé de la manière suivante. Dans la section 2, nous revenons sur les méthodes d'estimation bibliométriques de la qualité des revues, avec une illustration sur l'économie. La section 3 montre brièvement les utilisations qu'il peut en être fait pour évaluer les institutions et présente une approche directe alternative. La section 4 présente des exemples de travaux de prolongement que permettent ces processus d'évaluation.

Evaluation des revues

Il existe de nombreux classements des revues scientifiques en économie issus d'évaluations quantitatives. Les écarts d'impact entre revues obtenus sont souvent tout aussi importants pour l'évaluation des institutions que la hiérarchie des revues elle-même. Ainsi peut-on considérer qu'il vaut mieux être troisième avec 80% de la note du premier que deuxième avec 70% de celle-ci. Un consensus relativement large, même si on peut le discuter, s'étend au delà des frontières de la science économique : les citations reçues par les publications de la revue constituent une mesure objective de son impact. Dès lors, le principal reproche que l'on peut faire aux classements existants est de considérer un nombre trop restreint de revues. Ainsi, pour l'économie, Laband et Piette (1994) en considèrent 130, Kalaitzidakis, Mamuneas et Stengos (2003), 159, Kodrzycki et Yu (2006), 181, et Ritzberger (2008), 271. Or, il existe plus de 1200 revues référencées dans EconLit, la base de données bibliographique la plus utilisée en économie. La contribution importante de Combes et Linnemer (2010) consiste à donner un classement et un score à l'ensemble des revues EconLit en utilisant une méthodologie en trois étapes décrite ci-après.
Comme le font la plupart des approches précédemment citées, Combes et Linnemer (2010) considèrent que la note d'une revue est une moyenne pondérée d'indices de citations. Le choix de ces indices, et leur poids dans la moyenne, est relativement ad hoc, mais transparent, cherchant à capter des dimensions différentes de l'impact d'une revue. Il peut facilement être modifié si d'autres postulats souhaitent être effectués. La contrainte importante provient du fait que la source principale de citations existant est le Journal of Citation Reports (JCR) de Thomson-Reuters qui ne contenait, en 2007 date de l'étude, que 304 revues en économie. Pour les 898 autres revues recensées dans EconLit, aucun comptage de citations n'est effectué. Combes et Linnemer (2010) proposent d'estimer sur les 304 revues recensées dans le JCR un modèle économétrique reliant leur impact à certaines de leurs caractéristiques observables. Celles-ci étant disponibles non seulement pour ces revues mais également pour les 898 autres, une fois le modèle estimé, il peut être réutilisé pour prédire l'impact des revues pour lesquelles il ne peut être calculé directement.
L'indice d'impact choisi est une moyenne pondérée de *2x3=6* indices. Les trois indices de citations sont constitués de deux indicateurs d'impact fondés sur les citations JCR des articles (tels que proposés par

[1] Catégorisation des revues en Economie et en Gestion, Section 37 (Economie/Gestion), du Comité National de la Recherche Scientifique, Version 3.01, octobre 2011.

deux projets Internet : RedJasper et EigenFactor[2]) et d'un indice-H de citations Google Scholar.[3] De façon cruciale, chacun de ces indices est introduit en deux variantes dans la moyenne générale : Sa valeur brute et sa valeur normalisée par la valeur de l'indice prise au niveau mondial par chacun des domaines représentés dans la revue. Ce deuxième élément corrige ainsi totalement pour les différences entre domaines de recherche en termes de citations reçues qui existent au niveau mondial.[4] Cela tend à prendre en compte la critique principale des indices de citation, à savoir qu'ils sont nécessairement plus élevés pour les domaines surreprésentés dans la profession. A l'inverse, comme l'on peut aussi argumenter que cette surreprésentation est justement un aspect de l'impact scientifique, les deux aspects sont considérés dans l'indice total du fait que la moyenne des deux est calculée.

Les variables observables utilisées pour prédire l'indice d'impact de la revue sont de deux natures. Il s'agit d'une part d'indices fondés sur les citations Google Scholar des revues qui sont disponibles pour l'ensemble des 1202 revues. La deuxième famille est constituée de l'impact moyen des auteurs de la revue, mesuré via les indices RedJasper et Eigenfactor. Autrement dit, on utilise le fait que si une revue non référencée dans le JCR a des auteurs publiant dans des revues référencées par le JCR, l'impact de cette revue est vraisemblablement lié à l'impact des autres revues dans lesquels ses auteurs publient. Cette hypothèse n'est pas forcément valide a priori, mais il s'avère qu'a posteriori la qualité de prédiction du modèle est très élevée, puisque 98% de la variance de l'indice est expliquée par ces variables. Il ne reste plus qu'à prédire les indices des 898 autres revues à l'aide du modèle estimé.

Les inconvénients de cette procédure sont de deux types. Le premier est commun à l'ensemble des méthodes pondérant plusieurs indices pour construire un score : les indices de citation choisis, et leur poids dans le calcul de la moyenne (choisis ici égalitaires), sont arbitraires. Le deuxième inconvénient tient à l'acceptation de l'hypothèse de validité du modèle estimé pour les revues référencées par le JCR pour les revues qui ne le sont pas. Notons que pour quelques revues qui depuis ont intégré la base JCR, comme le *Journal of the European Economic Association*, le modèle s'est avéré tout à fait pertinent.

Le premier avantage de cette approche est tout d'abord que les critères sont « objectifs ». Dans la mesure où il s'agit de moyennes pondérées d'indices de citations, il est possible de discuter facilement du choix des indices comme du poids qui leur est donné. Il est également possible d'extraire des sous-ensembles de revues par domaine pour obtenir des classements « intra-domaine ». Un classement cardinal continu étant obtenu, on peut appliquer ensuite une fonction choisie pour obtenir un degré de sélectivité plus ou moins fort, choix encore laissé à l'analyste. Finalement, et encore une fois, le support couvert est considérablement élargi (plus de 1200 revues, dont de nouvelles revues créés récemment) par comparaison avec les autres exercices du même type.

A titre d'exemple de différences possibles entre des classements purement bibliométriques des revues et les classements issus des pairs, même lorsque ceux-ci se réclament au moins partiellement d'une approche en termes d'impact bibliométrique, nous proposons une succincte comparaison des classements des revues selon Combes et Linnemer (2010) et selon le comité de la section 37 du CNRS.[5] L'analyse prend pour donné le nombre de revues appartenant à chacune des 5 catégories considérées dans le classement du Comité CNRS. Elle alloue ensuite les revues entre ces catégories à partir du classement Combes et Linnemer (2010). Autrement dit, sachant qu'il y a 6 revues appartenant à la plus haute catégorie par exemple, les 6 revues les mieux classées par Combes et Linnemer (2010) sont mises dans cette catégorie et l'on recommence avec les catégories suivantes. Ainsi, chaque revue appartenant à l'intersection d'EconLit et du classement du Comité CNRS (qui considère beaucoup moins de revues en économie[6]) appartient donc à deux catégories, celle selon le classement du Comité CNRS et celle selon le classement Combes et Linnemer (2010). Le Tableau 1 est une matrice de transition qui donne pour chaque catégorie du classement Combes et Linnemer (2010) (lignes) le nombre de revues dans chacune des catégories du classement de la section 37 du CNRS (colonnes), et vice et versa. Ainsi, parmi les 6 revues classées dans la catégorie la plus élevée par le Comité CNRS, 5 l'auraient aussi été en utilisant la hiérarchie issue de l'indice Combes et Linnemer (2010) mais une revue n'aurait été que dans la catégorie suivante. A l'inverse, une revue classée dans la deuxième catégorie par le CNRS aurait dû se trouver dans la première. De façon générale, on s'aperçoit que les deux hiérarchies des revues sont très loin de se correspondre. Si le terme diagonal de la matrice est toujours le plus élevé, signifiant que pour un certain nombre de revues celle-ci

[2] Il s'agit d'indices récursifs valorisant davantage les citations provenant elles-mêmes des revues les plus citées.
[3] Une revue obtient un indice-H de valeur *h* s'il a *h* articles ayant reçus au moins *h* citations chacun.
[4] A cette fin, l'économie présente le grand avantage d'avoir un système normalisé de classification des domaines, utilisé par toutes les revues.
[5] Version 3.01, octobre 2011.
[6] Il considère à l'inverse des revues de gestion non prises en compte ici.

ont été placées par le Comité CNRS au niveau où le classement Combes et Linnemer (2010) les aurait aussi situées, les termes sur et sous diagonaux sont aussi très importants. Un grand nombre de revues sont surclassées selon le classement du Comité CNRS par rapport à la hiérarchie Combes et Linnemer (2010) alors qu'un grand nombre d'autres sont déclassées. Au delà des différences dans le nombre de revues considérées (bien plus faible dans le classement du Comité CNRS) et du choix ad hoc du nombre de classes et de revues par classe alors que les indices Combes et Linnemer (2010) sont continus, l'avis des pairs semble s'éloigner relativement fortement d'une hiérarchie fondée uniquement sur des indices de citation. Il est difficile d'interpréter ce résultat mais on peut lui donner deux grandes familles d'explications. La première, que le regard des pairs prend en compte des critères plus larges que celui des citations, ce qui conduit à des différences. La seconde que les pairs distordent, consciemment ou pas, la réalité de l'impact des revues. A ce stade, il est difficile de trancher entre ces deux explications.

Tableau 1 – Matrice de transition entre classements

		Classement CNRS				
		1*	1	2	3	4
Clst. CL	1*	5	1	0	0	0
	1	1	43	19	7	0
	2	0	23	50	31	5
	3	0	3	36	77	37
	4	0	0	4	38	42

Evaluation des institutions

Evaluer la quantité et la qualité des publications des chercheurs est une étape préalable indispensable à l'évaluation de leurs institutions d'appartenance. Une fois disponibles des indices de qualité des revues, ceux-ci peuvent être utilisés pour définir des indices de quantité et qualité des publications des institutions, ce que nous présentons dans une première sous-section. Dans une deuxième, nous proposons une stratégie alternative fondée sur les citations reçues par chaque publication, et non sur les indices moyens d'impact des revues.

Evaluation à partir de la qualité moyenne des revues

Dans Bosquet, Combes et Linnemer (2010), nous calculons un ensemble d'indices de la production scientifique en économie pour l'ensemble des universités et centres de recherche français (105 centres regroupés dans 76 universités), en 1998 et 2008. Cela nous permet de réaliser des études de sensibilité quant aux choix méthodologiques possibles.

Pour effectuer ce travail, notre méthodologie a été la suivante. Nous avons d'abord recensé l'ensemble des articles publiés par tous les enseignants-chercheurs (Universités) et chercheurs (CNRS, INRA, Ecoles) en économie français (soit 2832 personnes équivalent temps plein) dans les 1202 revues EconLit. Ensuite, nous avons calculé des indices prenant ou pas en compte le nombre d'auteurs par article, le domaine (Codes JEL), la qualité de la revue (comme expliquée dans la section précédente) et la longueur relative des articles, et ce pour différentes périodes de temps : 5 dernières années, 40 dernières années, avec un facteur de décompte dans le temps, par année depuis la thèse pour chaque chercheur, etc.

Le premier résultat que l'on peut alors observer est la relative insensibilité de la hiérarchie des universités à l'indice de publication choisi. Les autres tendances qui se dégagent de cette étude sont les suivantes. On observe une augmentation des taux de publiant dans le temps (entre 1998 et 2008), une augmentation de la part des publications des chercheurs français dans le monde, une baisse des disparités entre chercheurs/centres/universités et une courbe de productivité en cloche en fonction de l'âge. Chaque université ou centre de recherche peut également se situer dans la hiérarchie française en volume total et en production moyenne par chercheur et peut également évaluer comment sa position évolue dans le temps. Une analyse effectuée pour chaque grand domaine de recherche pris séparément permet aussi de déterminer ceux pour lesquels l'université ou le centre sont les plus visibles.[7]

Evaluation à partir des citations individuelles

Les approches fondées sur la prise en compte de la qualité des publications à partir de la qualité moyenne des revues présentent le risque de masquer des différences potentiellement importantes entre publications d'une même revue. De nouveaux outils permettent cependant désormais de mesurer l'impact individuel de chaque publication, et donc de ne plus avoir besoin d'utiliser les indices moyens d'impact des revues. Dans Bosquet et Combes (2011a) et Bosquet et Combes (2011b), nous ne calculons plus des taux de publiant ou des volume totaux et moyens d'articles publiés en prenant en compte la qualité moyenne des revues, mais

[7] Pour plus de détails, se référer directement au rapport ou à sa note de synthèse, disponibles en ligne aux liens http://www.vcharite.univ-mrs.fr/pp/combes/RapportRanking010310.pdf et http://www.vcharite.univ-mrs.fr/pp/combes/SyntheseRapportRanking.pdf, respectivement.

des indicateurs de citations totales, par article, des indices-H et des indices-G des institutions à partir de ces mêmes indicateurs calculés pour chacun de leurs membres.[8] Ces indices de citation sont issus de Google Scholar. Ici également, de nombreux tests de robustesse sont effectués selon le fait que l'on prend en compte ou pas le nombre d'auteurs par article ou l'âge de ces articles, et les évaluations sont également effectuées sur différentes périodes de temps.
Les avantages de cette nouvelle perspective par rapport à une analyse des publications sont les suivants. D'une part, le type de supports considérés pour chaque chercheur est bien plus large, puisqu'étendu à tous ceux présents sur des sites Internet académiques (articles mais aussi ouvrages, documents de travail, etc.). L'extension est également très importante du côté des citations, puisque sont considérées non pas seulement celles émanant de revues académiques mais également de tous ces mêmes supports présents sur des sites académiques. Pour l'ensemble des économistes exerçants en France en 2008, on compte par exemple 42 448 entrées GS ayant au total 265 578 citations. Les contributions interdisciplinaires sont également mieux prises en compte puisque l'on n'est plus obligé de se restreindre au champ économique que représente EconLit. Ainsi, les publications des économistes en sociologie, histoire, statistiques ou gestion sont par exemple considérées, et l'on pourrait élargir aux mathématiques pures, à la physique etc.
La conclusion principale de cette étude est qu'il n'y a pas de fortes différences dans la hiérarchie des institutions françaises entre cette approche utilisant les citations et l'approche précédemment présentée utilisant la qualité moyenne des revues, même si, au niveau d'un chercheur donné, les différences sont plus importantes. Néanmoins, les institutions dont le cœur de métier est davantage orienté « gestion » par exemple progressent légèrement dans les classements.[9]

Au delà de la mesure, des guides possibles de la politique scientifique

Dans notre démarche scientifique, nous ne considérons pas que l'évaluation des revues, chercheurs ou institutions est une fin en elle-même mais une étape préalable à des études pouvant être instructives en termes de politique de la recherche. Ainsi, dans Bosquet et Combes (2012) et Bosquet et Combes (2011c), nous nous intéressons à l'évaluation des déterminants de la production de recherche des universités. Nous tentons alors de répondre au type de questionnement suivant : quelle est le rôle pour l'activité de publication de la taille des laboratoires, de leur diversité thématique, de leur composition en termes de ratio maîtres de conférence / professeurs, de la charge d'enseignement des chercheurs à l'université, de leurs liens avec des centres de recherche étranger, etc. ? Ce type de quantification devrait être susceptible de guider la politique de recherche et d'éventuellement identifier des stratégies plus efficaces que d'autres en termes de publications ou de citations reçues. Dans un autre travail en cours, Bosquet, Combes et Garcia-Peñalosa (2012), nous cherchons à évaluer les origines des différences de publication des hommes et des femmes et les comparons à leur statut (maître de conférences versus professeur par exemple) sur le marché académique. Cela peut par exemple permettre de déceler une éventuelle discrimination de ces dernières, à niveau de publication donné. De nombreuses autres utilisations des indices de publication peuvent être envisagées.

Références

Clément Bosquet et Pierre-Philippe Combes : *Comparaison des mesures Econlit et Google Scholar de la production de recherche en économie en France en 2008*. Direction Générale de la Recherche et de l'Innovation (DGRI) du Ministère de l'Enseignement Supérieur et de la Recherche, 2011a.
Clément Bosquet et Pierre-Philippe Combes : Un panorama de la recherche française en économie comparant les approches Google Scholar et Econlit. *GREQAM Working Paper,* (2011-56), à paraître dans *Revue d'Economie Politique*, 2011b.
Clément Bosquet et Pierre-Philippe Combes : *Déterminants de la production d'articles de recherche en France*. Direction Générale de la Recherche et de l'Innovation (DGRI) du Ministère de l'Enseignement Supérieur et de la Recherche, 2011c.

[8] Un chercheur a un indice-G de valeur g s'il a g articles ayant reçus en moyenne g citations chacun, ou g^2 citations au total.
[9] Pour plus de détails, se référer directement au rapport ou à sa note de synthèse, disponibles en ligne.

Clément Bosquet et Pierre-Philippe Combes : Do large departments make academics more productive? Agglomeration and peer effects in research. *Mimeo Greqam*, 2012.
Clément Bosquet, Pierre-Philippe Combes et Cecilia Garcia-Peñalosa : Gender differences in a micro labour market : promotions amongst academic economists in France. *Mimeo Greqam*, 2012.
Clément Bosquet, Pierre-Philippe Combes et Laurent Linnemer : *La publication d'articles de recherche en économie en France en 2008. Disparités actuelles et évolutions depuis 1998*. Direction Générale de la Recherche et de l'Innovation (DGRI) du Ministère de l'Enseignement Supérieur et de la Recherche, 2010.
Pierre-Philippe Combes et Laurent Linnemer : La publication d'articles de recherche en économie en France. *Annales d'Economie et de Statistiques*, 62:5-47, Avril/Juin 2001.
Pierre-Philippe Combes et Laurent Linnemer : Where are the economists who publish ? Publication concentration and rankings in Europe based on cumulative publications. *Journal of the European Economic Association*, 1(6):1250-1308, December 2003.
Pierre-Philippe Combes et Laurent Linnemer : Inferring missing citations. A quantitative multi-criteria ranking of all journals in economics. *GREQAM Working Paper*, (2010-25), 2010.
Pantelis Kalaitzidakis, Theofanis P. Mamuneas et Thanasis Stengos : Ranking of academic journals and institutions in economics. *Journal of the European Economic Association*, 1(6):1346-1366, December 2003.
Yolanda K. Kodrzycki et Pingkang Yu : New approaches to ranking economics journals. *B.E. Journals in Economic Analysis and Policy : Contributions to Economic Analysis and Policy*, 5(1):1-42, 2006.
David N. Laband et Michael J. Piette : The relative impacts of economics journals : 1970-1990. *Journal of Economic Literature*, 32(2):640-666, June 1994.
Klaus Ritzberger : A ranking of journals in economics and related fields. *German Economic Review*, 9(4):402-430, November 2008.