L'évaluation bibliométrique des chercheurs : même pas juste… même pas fausse !

Cet article reprend avec quelques compléments celui publié en mars 2009 dans "Reflets de la physique" n°13, pp. 23-24. Merci à la Société Française de Physique pour son autorisation.

Utiliser les indices bibliométriques pour l'évaluation des chercheurs résulte d'une extrapolation injustifiée entre des cas évidents où ces indices ont un sens vers les cas où ils sont réellement utilisés en pratique. Les vérifications les plus élémentaires n'ont pas été faites pour comprendre l'influence de méthodes de calcul arbitraires sur les résultats obtenus. Dans l'état actuel des techniques, leur utilisation relève plus de la pseudo-science, comme l'astrologie, que d'une démarche scientifique.

On dit que Wolfgang Pauli, un des géniaux fondateurs de la mécanique quantique, furieux contre un article de physique sans aucun intérêt, s'était écrié « ce n'est pas juste et, pire, ce n'est même pas faux ! ». Il est vrai que ce qui n'est ni juste ni faux ne peut être scientifique ; vérification et réfutation sont au cœur du fonctionnement même des sciences de la Nature. La remarque de Pauli s'appliquerait fort bien à de nombreuses applications de la bibliométrie, promues par ceux dont la croyance semble être que, du moment qu'on manipule des chiffres, on raisonne scientifiquement.

L'évaluation bibliométrique des chercheurs n'est effectivement « même pas fausse » : oui, si l'on compare un chercheur reconnu internationalement à un doux farfelu qui n'a jamais été cité que par lui-même, les indices du premier sont bien supérieurs à ceux du second ; personne ne le conteste. Si le but de l'exercice était de reconnaître des chercheurs exceptionnels des chercheurs médiocres, nul doute qu'on pourrait recourir à la bibliométrie pour retrouver .. ce que chacun sait déjà. Mais supposons que l'on veuille réellement obtenir une information utile, comme par exemple classer les chercheurs au sein d'un groupe homogène, disons les chercheurs d'un bon laboratoire. On constate alors immédiatement de nettes fluctuations de leurs indices (H, G, etc.. peu importe), et avec une certaine surprise: des valeurs très différentes peuvent être attribuées à des chercheurs dont la qualité de production scientifique est perçue comme très similaire par la communauté scientifique. Pourquoi ?

Plusieurs raisons expliquent pourquoi les méthodes bibliométriques donnent une vue simpliste des contributions scientifiques individuelles. Elles sont sensibles à la qualité, certes, mais ce « signal » est noyé dans le « bruit » créé par une forte dépendance en fonction d'autres variables. Prenons par exemple un indice bibliométrique H, fonction d'une variable X dont nous supposons qu'elle soit assimilable à la qualité du travail scientifique, de la variable Y qui est le style du chercheur (travaille-t-il plutôt seul ou en équipe constituée, est-ce plutôt un pionnier ou quelqu'un qui préfère des domaines déjà relativement à la mode, proche des applications ou non, etc.), de la variable Z qui est son style de publication (est-il plutôt tourné vers les courtes lettres ou les articles de fond, voire les ouvrages ? est-il attiré par les revues dites de prestige, genre Nature ou Science, même si elles sont moins utilisées dans son domaine ?), et enfin W (appartient-il à une école de recherche très reconnue depuis des années, ou a-t-il choisi un petit domaine émergent, etc.). Cette liste de variables n'est évidemment pas limitative, on pourrait par exemple y ajouter le sens de la communication du chercheur et son goût pour les congrès, qui ne sont pas

toujours liés à sa créativité. N'importe quel scientifique sait que, si l'on recherche la mise en évidence de la variable X, les nombreuses autres variables vont se comporter comme un « bruit statistique». La seule façon d'avoir accès à la variable recherchée est d'éliminer ce bruit par des moyennes. Si elles sont effectuées sur de très gros échantillons d'individus, statistiquement les variables Y, Z, W prendront un peu toutes les valeurs possibles, et leur influence disparaîtra, laissant apparaître celle de X. C'est ce qui permet à la bibliométrie d'obtenir des chiffres pertinents pour, par exemple, une évaluation comparative de la production nationale dans un grand domaine de recherche. En revanche, utiliser H pour connaitre la variable X au niveau individuel est tout simplement une erreur de raisonnement qu'on ne pardonnerait dans aucun laboratoire de recherche sérieux.

De plus, quand un chercheur rédige un article et qu'il y inclut des références, ce n'est pas un acte destiné à la bibliométrie : le but premier des citations n'est pas de dresser une sorte de palmarès, mais de donner au lecteur des informations qui lui sont utiles pour lire l'article en question. C'est donc un processus relatif, fortement contextuel. Par commodité, on peut par exemple citer un article de revue plutôt que les sources originales, pour gagner de la place. Parfois, on cite un article qui permet de raccourcir sa propre rédaction, et on choisira alors le texte juste pour une question de similarité. On peut même citer un article qu'on considère comme faux dans le but d'en corriger les erreurs ! Comme il s'agit de faciliter la répétition des expériences dans d'autres laboratoires, on privilégiera dans les citations les articles qui décrivent des méthodes ou des appareillages. Pour les idées scientifiques plus abstraites, en revanche, c'est généralement des articles dérivés qu'on cite, pas le grand article original et fondateur. C'est donc un emploi très dérivé des citations, pour ne pas dire un contresens sur leur fonction réelle, que de les prendre comme élément de base pour l'évaluation de la qualité scientifique. Pire encore, par un effet pervers maintes fois signalé, cela risque d'entraîner des changements dans la manière dont les citations se feront à l'avenir, aux dépends de la qualité de la rédaction scientifique et donc de la recherche elle-même.

La base qui est utilisée pour une évaluation bibliométrique en « sciences dures » est celle du WOS de l'ISI. Première remarque : les ouvrages scientifiques ne sont pas pris en compte dans le calcul du facteur H que cette base donne en deux clics ! Premier paradoxe, chacun s'accordant à penser qu'une des meilleures façons pour un chercheur de « laisser une trace dans un domaine » est précisément de publier un ou des ouvrages de référence. Deuxième remarque : les indices G, H, etc.., classiquement utilisés pour classer les individus sont aussi fondamentalement biaisés que le classement de Shanghai[1]. Dans le calcul de ces indices, la contribution d'un auteur est la même s'il est seul signataire ou s'il a dix co-auteurs ! Il pourrait paraître élémentaire de calculer des indices G', H', établis sur la base de nombre de citations divisé par le nombre d'auteurs, ce qui découlerait de la logique la plus élémentaire, mais personne ne le fait sur les données de l'ISI[2]. Le biais est évident : si trois amis décident de tout publier ensemble pendant toute leur carrière, chacun de leurs indices H fera un bond vers le haut. Troisième faiblesse : tout est centré sur le court terme. Dans beaucoup de

[1] Il est trivialement additif, ne tenant aucun compte de la taille des établissements

[2] Il ne s'agit pas de dire que cette façon de faire serait bonne, mais juste qu'elle serait moins fausse que la méthode habituelle. Les indices G', H' ne seraient pas plus pertinents que G,H,.. ; ce qui est intéressant est la variabilité, le bouleversement des classements individuels qui en résulterait.

domaines, de petites percées techniques entraînent une bouffée de publications, vite oubliées, mais nombreuses. Ainsi les indices sont très orientés vers les sujets à la mode, même s'ils disparaissent vite. Il n'y a aucune raison particulière à cette faiblesse, et on pourrait facilement imaginer d'effectuer des calculs plus adaptés où l'on prenne en compte surtout les publications qui ont une influence à long terme ; mais là encore personne ne le fait.[3]. Quatrième faiblesse, bassement technique peut-être, mais très réelle dans certains cas : la base ISI elle-même est inhomogène, ayant fluctué au gré des habitudes de travail des opérateurs de saisie qui l'ont alimentée au cours des décennies. Cela induit tout une série de corrections qui sont nécessaires, qu'il serait trop long de discuter ici, mais qui demandent un travail de spécialiste. Seul un travail précis permettrait de reconstruire des vrais indices, mais bien évidemment dans la pratique personne ne s'en donne la peine : il est tellement plus simple de faire un classement avec des chiffres faux obtenus en trois clics d'ordinateur !

Outre d'être un guide de médiocre qualité, cette utilisation superficielle de la bibliométrie risque d'avoir des effets pervers plus graves, allant bien au-delà d'une évaluation biaisée. C'est en effet la qualité même de la communication scientifique entre chercheurs qui est en jeu, donc à terme la qualité de la recherche. Il est évident que, si les critères de sélection et de promotion des chercheurs privilégient l'impact bibliométrique instantané, ces derniers seront tentés de modifier leur stratégie de publication ; ils y feront entrer des considérations qui n'ont plus rien de scientifique. Comme la communication des résultats est au cœur même du fonctionnement de la recherche, cela va progressivement influencer son style en la poussant dans une direction opposée de celle de la qualité et du travail de fond. Nous observons déjà le phénomène : par exemple, pour augmenter leur « impact factor », certains journaux pratiquent ouvertement une politique de citations biaisée, en faisant pression sur les auteurs ; il suffit de faire savoir que « si votre manuscrit contient au moins 4 références à des articles de notre journal, il sera publié plus facilement », et le tour est joué. La généralisation de telles pratiques ne favorise assurément pas une recherche de qualité[4] !

Pour finir, ces indices d'évaluation individuelle :

1. n'ont jamais été testés rationnellement, pour les corréler avec d'autres évaluations ; ce sont des méthodes qui se présentent elles-mêmes comme des méthodes d'évaluation de la recherche, mais qui paradoxalement ne sont pas passées par l'évaluation scientifique de leur fiabilité et leur pertinence. Même les vérifications les plus élémentaires comme celles que nous avons signalées plus haut n'ont pas été faites, afin de comprendre l'influence des méthodes de calcul choisies sur les résultats obtenus.

[3] Dans le cas du calcul de l'impact factor des journaux, c'est même exactement le contraire qu'on fait, puisqu'on ne prend en compte les citations que sur deux ans ; pour quelques domaines de recherche techniques, ceci peut éventuellement avoir un sens, mais certainement pas pour le cœur des grandes disciplines scientifiques.

[4] Pour revenir à la remarque de Pauli et aux débuts de la mécanique quantique, on est dans un cas où le processus de mesure risque de perturber fortement la quantité mesurée, mais ici la raison n'est pas due aux lois fondamentales de la nature ! Il s'agit d'un comportement humain : des erreurs dans l'évaluation d'une activité réagissent négativement sur l'activité elle-même.

2. personne ne semble avoir pris le temps d'essayer honnêtement de les améliorer pour essayer d'obtenir des chiffres qui soient plus pertinents. Porter le débat vers la recherche intelligente de la qualité réelle de la recherche ne semble pas d'actualité.
3. ne sont « même pas faux », car certes ils contiennent un peu d'information sur les individus, mais la plupart du temps cette information est triviale et déjà connue. Lorsqu'ils sont utilisés sur une population homogène de chercheurs (dossiers au Comité National par exemple), ils déterminent plus un style de travail du chercheur (porté au travail en équipe, ou plutôt éclaireur dans son domaine, etc..), alors que tous ces styles sont nécessaires pour une recherche équilibrée et efficace. Les utiliser pour l'évaluation des chercheurs est donc une faute de raisonnement, une espèce d'extrapolation injustifiée entre des cas triviaux où les indices ont un sens vers les cas réels où ils seront utilisés.
4. leur succès vient évidemment de la fausse facilité qu'ils procurent : avoir un chiffre en quelques clics est évidemment bien commode.
5. la foi dans ces indices est devenue une espèce de croyance, qui échappe au rationnel. La comparaison qui vient à l'esprit est l'astrologie ou la numérologie, qui elles aussi se parent de vertus du scientifique mais ne se sont jamais passées avec succès sous les fourches caudines d'une véritable évaluation scientifique.

Franck Laloë (1) et Rémy Mosseri (2)

Remerciements : l'un des auteurs (F.L.) remercie Y. Gingras, de l'Université de Montréal, pour d'utiles discussions ainsi que pour des données bibliométriques concernant l'article EPR. On consultera avec profit les articles de Y. Gingras, en particulier « La fièvre de l'évaluation de la recherche ; du mauvais usage des faux indicateurs »
http://www.cirst.uqam.ca/Portals/0/docs/note_rech/2008_05.pdf

(1) laloe@lkb.ens , Laboratoire Kastler-Brossel, Ecole Normale Supérieure, CNRS et UPMC, 24 rue lhomond, 75005 Paris
(2) Laboratoire de Physique Théorique de la Matière Condensée, UMT CNRS 7600, UPMC, 4 Place Jussieu, 75252, paris Cedex 05

Recette

Si vous êtes un bon chercheur, et si vous voulez un meilleur indice H, voici quelques conseils :

1. Travaillez dans un groupe d'au moins 5 ou 6 personnes, si possible plus, dont toutes les publications sont systématiquement communes ; cela doit permettre au moins de doubler votre indice, voire mieux. De plus, ce regroupement d'efforts permet la mise en commun de « moyens » (matériels et humains, post-docs par exemple) qui peut même parfois augmenter encore votre productivité réelle, ce qui n'est pas à négliger. Inutile de préciser que, plus ces collègues seront brillants, plus vous en profiterez ; choisissez les donc bien !

2. Favorisez les grands domaines ; on constate une corrélation entre le taux de citations et la taille du domaine scientifique, due au fait que les articles des petits domaines citent beaucoup d'articles plus généraux et pas l'inverse. Evitez les travaux en marge du courant général de votre domaine, même si vous êtes génial : il faudra 10 ans pour que vos travaux soient vraiment reconnus, et alors ce seront les travaux dérivés des vôtres qui seront cités. Bref, ne vous laissez par trop obnubiler par l'intérêt scientifique de vos recherches : prendre des risques pour tenter une percée scientifique est rarement payant avant des décennies !

3. Ne perdez surtout pas de temps à publier des ouvrages, quel que soit leur impact international ; ce sont des pertes sèches pour les indicateurs.

4. Surtout, n'accordez pas trop d'importance à la mission fondamentale des chercheurs, la production de connaissance, en particulier quand vous rédigez vos articles ; c'est la communication qui passe d'abord.

Quelques cas d'école

Il est instructif d'étudier le taux de citation de quelques articles particulièrement célèbres. Ce sont bien évidemment des « points singuliers », dont on ne peut tirer directement des conséquences pour l'immense majorité des articles scientifiques. Mais, si on garde cette réserve à l'esprit, leur histoire bibliométrique nous apporte un enseignement utile.

1. A tout seigneur tout honneur, prenons l'article célèbre de EPR (Einstein-Podolsky-Rosen) de 1935 : « *Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?* » (*Phys. Rev.*, vol. 47, 1935, p. 777-780). On dit parfois que c'est l'article de physique le plus cité de tous les articles de physique. Il joue un rôle important dans le développement des idées quantiques, car en un sens il clôt la célèbre série de débats entre Einstein et Bohr, qui a permis d'établir la solidité et la cohérence du formalisme de la mécanique quantique. Presque simultanément, l'article a reçu un important écho dans la grande presse, puisque le New York Times y a consacré un long article le mois même de sa parution dans Physical Review.

Quel est l'écho bibliométrique de cet article ? L'article a rapidement été cité par N. Bohr, qui a publié une réponse dans le même journal et avec le même titre ; quelques échanges et controverses occasionnelles ont eu lieu assez rapidement (Kemble par exemple), mais on peut dire qu'en gros la communauté scientifique n'a pas cité cet article à l'époque. Un sentiment général vague prévalait que « Bohr avait su répondre » et que ces questions étaient trop délicates pour la majorité des physiciens. Ainsi, si l'impact factor des revues avait existé à l'époque, l'article EPR n'y aurait joué aucun rôle.

Son histogramme de citations est cependant intéressant. Jusqu'au milieu des années 60, l'ISI ne donne que des citations sporadiques. Cela ne veut pas dire qu'il n'ait pas été cité, mais probablement que les citations étaient, soit dans des livres (comme ceux de M. Jammer), soit dans des journaux de philosophie ou de philosophie des sciences – dans les deux cas non couverts pas l'ISI. Mais, à partir de 1970, les physiciens redécouvrent l'article, le taux de citation donné par le SCI commence à monter, dépasse 10 par an, pour atteindre 50 autour de 1990. En 2000, il avoisinait 200, et ne cesse de croître depuis !

Bien évidemment, ce démarrage dans les années 1970 correspond à la parution des articles de John Bell, qui sont le direct prolongement du travail de EPR. Il aura donc fallu plus de 30 ans pour qu'un article aussi important laisse sa signature bibliométrique !

2. L'article de John Bell « On the Einstein-Podolsky-Rosen paradox » de 1964 (Physics, vol. 1, pp. 195-200, 1964), lui, a commencé à être cité de façon significative (une dizaine de fois par an) dix ans après sa parution. C'est dans les années 1980 qu'il a commencé à véritablement décoller, avec une croissance devenue rapide depuis 2000. En 2004, il recueillait environ 200 citations annuelles dans les journaux scientifiques répertoriés dans l'ISI.

3. L'article d'Alfred Kastler où il a proposé la méthode du pompage optique (J. Phys. Rad. Vol. 11, page 255, 1950) a une histoire bibliométrique très différente. Contrairement aux deux précédents, il n'a pas subi un long temps de latence avant de recueillir des citations. La raison en est claire :

l'auteur avait fondé un laboratoire expérimental avec Jean Brossel, de sorte que la mise en évidence du pompage optique a donné lieu à des citations dès les années qui ont suivi. Mais, curieusement, ce taux est resté modeste, ce qui est très étonnant pour un article d'une telle cette importance : il faut en effet attendre 20 ans pour qu'il soit cité plus de 10 fois en un an ! En 2008, il a été cité 9 fois.

Le chiffre le plus frappant est le nombre total de citations recueillies par cet article: environ 400 fois seulement en tout sur une période couvrant plus de 50 ans, alors qu'il s'agit d'un travail qui a fondé un grand domaine de recherche. Quel paradoxe, alors que des milliers d'articles ont été publiés pour relater des travaux directement basés sur l'utilisation du pompage optique ! Aux travaux de nature plutôt fondamentale s'ajoutent tous ceux qui utilisent le pompage optique pour ses applications (en particulier horloges atomiques et magnétomètres). On peut même voir une filiation entre l'article fondateur de Kastler et les expériences récentes sur les gaz ultra-froids (condensation de Bose-Einstein) : il mentionne explicitement la possibilité de refroidir un gaz atomique par irradiation lumineuse. Force est donc de constater que, parmi cette foule d'articles où l'on trouve en bonne place les mots « pompage optique », tout au plus quelques pourcents ont cité le travail original. Visiblement des milliers d'auteurs, en toute bonne foi, ont tout simplement estimé inutile de se référer à un article aussi connu, estimant que les lecteurs le connaissaient déjà – voilà qui illustre bien ce que nous disions plus haut du caractère contextuel des citations, et du fait qu'elles ne sont pas un palmarès.