



HAL
open science

Contribution à l'évaluation stratégique de la recherche par l'étude des temps de premières citations

Tam Dang Nguyen, Pascal Prinderre, Christophe Sauzet, Hervé Rostaing,
Jacky Kister

► **To cite this version:**

Tam Dang Nguyen, Pascal Prinderre, Christophe Sauzet, Hervé Rostaing, Jacky Kister. Contribution à l'évaluation stratégique de la recherche par l'étude des temps de premières citations. Les systèmes d'informations élaborées, Société Française de la Bibliométrie Appliquée, Jun 2005, Ile Rousse, France. hal-01579917

HAL Id: hal-01579917

<https://hal.science/hal-01579917>

Submitted on 1 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

***CONTRIBUTION A L'EVALUATION STRATEGIQUE DE LA
RECHERCHE PAR L'ETUDE DES TEMPS DE PREMIERES CITATIONS***

Tam Dang Nguyen (*), Pascal Prinderre (), Christophe Sauzet (**), Hervé Rostaing (***),
Jacky Kister (*).**
pascal.prinderre@pharmacie.univ-mrs.fr, rostaing@crrm.u-3mrs.fr, jacky.kister@univ.u-3mrs.fr

(*) : Université Paul Cézanne (Aix-Marseille III), UMR CNRS 6171 « Systèmes Chimiques Complexes », Faculté des Sciences de St Jérôme 13397 Marseille cedex 20.

(**) : Laboratoire de pharmacie galénique Industrielle, Université de la méditerranée,
27 Bd J. Moulin 13385 Marseille Cedex 5.

(***) : Université Paul Cézanne (Aix-Marseille III), GERSIC / CRRM,
Faculté des Sciences de St Jérôme 13397 Marseille cedex 20.

Mots clés : évaluation stratégique de la recherche, différenciation et caractérisation des types de recherche, citations index, première citation, loi de hasard.

1. Introduction

Au moment des importants changements dans le paysage de la recherche académique et privée, au moment de la création de grands réseaux de recherche européens, au moment des concentrations industrielles, des fermetures de centres de recherche en Europe, au moment de la création de centres d'expertises nationaux et européens et de Très Grandes Unités de recherche (TGU) à implantation supra-régionale ; réfléchir sur l'évaluation semble être une évidence voire une nécessité.

Evaluer simultanément ou séparément les acteurs de la recherche, les laboratoires ou les institutions est actuellement un sujet politiquement sensible. La superposition des moyens de financement de la recherche, la recherche de crédits incitatifs (ANR), européens ou industriels obligent à réfléchir à une gestion des connaissances créées et à une évaluation des réalisations scientifiques différente ou complémentaire des évaluations classiques du C.N.U. ou du Comité National du CNRS.

Intégrer la notion de retour sur investissement, de transfert de technologie et de réponse à la demande sociétale en travaillant sur contrats d'objectifs oblige à mettre en place des systèmes d'évaluation de la recherche dite d'innovation ou technologique au regard de celle dite fondamentale. Faute de ces outils ou indicateurs spécifiquement dédiés, l'orientation des unités de recherche et des acteurs de recherche vers des recherches à retour rapide pour la société restera des cas d'espèces que l'évaluateur traitera d'une façon marginale.

2. Objectif

L'objectif principal de cette étude était de rechercher la durée moyenne ou médiane que mettait un article de recherche pour être cité par un autre auteur. C'est pourquoi, nous avons relevé les données temporelles correspondant à la durée entre la publication de l'article et la publication du premier article le citant. Sachant que certains articles n'ont pas encore reçu de citation, nous avons donc constitué une base de données temporelles incluant des données censurées.

A partir de ce moment, nous avons recherché un modèle mathématique qui correspond à la distribution de ces données en fonction de la quantité d'articles de l'ensemble. En utilisant la loi de hasard, (en particulier la fonction de hasard), il est possible de traiter ces données graphiquement. Très vite nous nous sommes aperçus que ces ensembles semblent suivre une distribution Log-Normal à partir de laquelle il est possible de relever 2 coefficients suffisamment significatifs pour caractériser la distribution.

Le coefficient « α » qui montre l'immédiateté de l'apparition de la première citation : plus il est petit et plus la probabilité instantanée d'apparition de cette citation est forte, ce qui signifie que si « α » est petit la probabilité d'apparition de première citation d'un article encore non-cité est très forte dans un futur proche.

Le coefficient « β » qui indique le temps médian (pour 50% des articles) de première citation de la base de données étudiées.

3. Etude graphique

Un moyen simple et rapide pour étudier une base de données temporelles présentant des données censurées est l'étude graphique. Celle-ci a pour avantage principal de repérer les individus

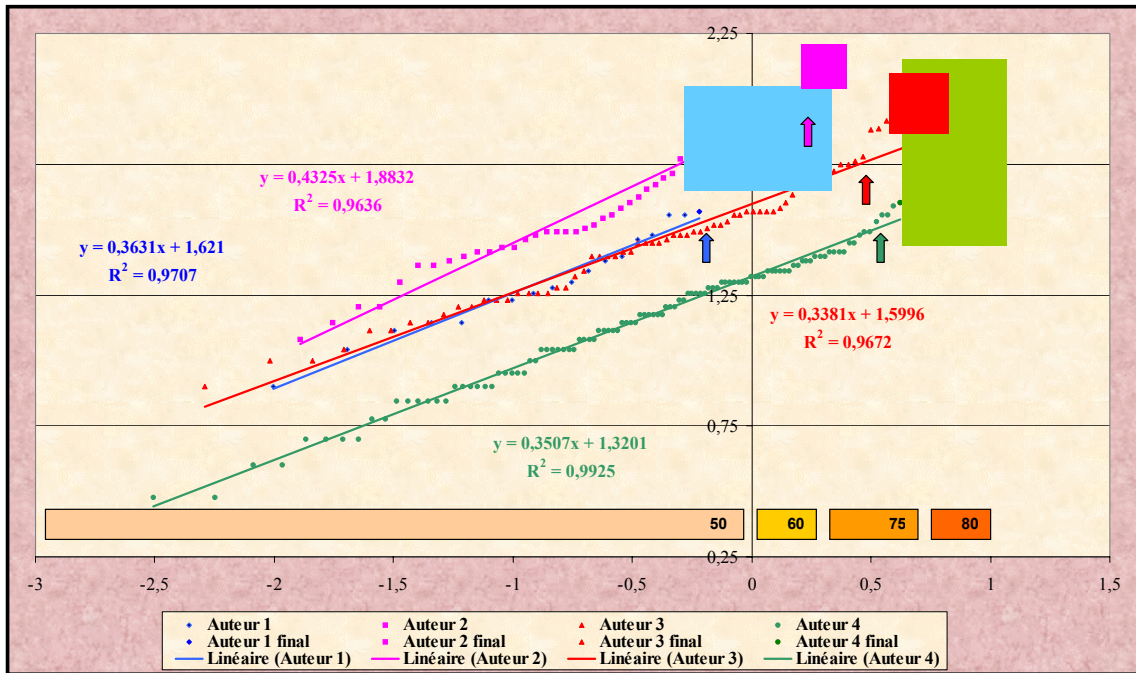
qui ne suivent pas la même distribution que les autres, grâce à une régression linéaire. (Cependant pour rechercher la loi de hasard correspondant à la distribution, il est nécessaire de reporter les données sur différents « papier de hasard » qui correspondent respectivement à une fonction donnée). Dans notre cas, la distribution correspondant le mieux est la fonction log-normal. De plus pour des raisons pratiques, nous allons utiliser le logiciel Excel pour tracer ces graphiques.

Nous avons donc collecté les données concernant 4 auteurs afin de les soumettre à notre étude :

- Individus : article de recherche.
- Base de données : ISI Web of Sciences.
- Événement : première citation (sauf auto citation).
- Donnée temporelle (T) : temps entre la publication et la première citation (mois).
- Début de l'étude : janvier 1981.
- Fin de l'étude : décembre 2004.
- Loi de hasard Log-Normal : $\log(T_H) = \alpha \Phi^{-1}(1 - \text{Exp}(-H)) + \beta$. **Φ^{-1} : Loi normal inverse ; H : hasard cumulé ; T_H : Donnée Temporelle).**

- Le hasard cumulé se calcule par rapport au rang inverse attribué à chaque individu en éliminant ceux présentant des données censurées.
- Le graphique représente donc le temps de première citation en fonction (en ordonnée) de la densité des individus (en abscisse) avec une modification par la loi Log Normal.

➤ **Représentation graphique et régression linéaire:**



➤ **Constatations :**

La loi Log-Normal semble correspondre à ces 4 ensembles de données même si nous pouvons observer une augmentation de la pente en fin de distribution.

Concernant l'Auteur 2, il semblerait qu'il existe 2 sous-ensembles à cause d'une modification importante de la pente au milieu de la distribution bien que le coefficient de régression soit suffisamment voisin de 1 (0,9636) pour considérer que l'ensemble des individus suivent la même distribution (une étude plus détaillée sera nécessaire afin de distinguer les 2 sous-ensembles).

➤ **Théorie sur la fonction de hasard Log-Normal :**

Le graphique représente les distributions respectives des 4 ensembles de données temporelles qui correspondent à la fonction de hasard Log-Normal. A partir des régressions linéaires, nous pouvons relever 2 coefficients qui permettent de caractériser ces ensembles.

En effet, la fonction de hasard a pour principale qualité d'indiquer la probabilité instantanée d'apparition de l'événement (première citation) en fonction du temps, c'est à dire qu'il est possible de savoir la proportion d'individus qui subira l'événement dans un futur très proche à un instant donnée. Cette proportion dépend de la valeur de la pente « α » :

1. si $\alpha > 0,8$: la probabilité instantanée est croissante.
2. si $\alpha \neq 0,5$: la probabilité instantanée est constante.
3. si $\alpha < 0,2$: la probabilité instantanée est décroissante.

Ainsi une brusque augmentation de la pente en fin de distribution indique une forte diminution de cette probabilité. Bien que les données soient insuffisantes en quantité, il est fort probable que ces individus ne suivent pas la même distribution que le reste de l'ensemble.

Le coefficient « β » permet quant à lui d'estimer avec précision le temps que mettent 50% des individus ayant subi l'évènement (temps médian).

➤ **Interprétation :**

Une vision rapide des distributions montre que nous pouvons distinguer 3 profils différents pour les 4 ensembles étudiés.

Le premier profil est attribué à l'auteur 2 qui possède à la fois les plus forts coefficients α (0,4325) et β (1,8832) :

1. α est suffisamment proche de la valeur 0,5 pour dire que la probabilité instantanée de première citation est pratiquement constante dans le temps.
2. β indique que le temps médian de première citation est de 76,4 mois.

Le deuxième profil concerne les auteurs 1 et 3 qui possèdent un coefficient α plus faible (0,3631 et 0,3381) tout comme le coefficient β (1,621 et 1,5996) :

1. α est légèrement en dessous de la valeur 0,5 mais suffisamment éloigné pour dire que la probabilité instantanée de première citation de ces 2 ensembles est légèrement croissant en fonction du temps. Ce qui signifie qu'un article encore non cité aura une probabilité de plus en plus forte d'être cité au fur et à mesure que le temps passe. De plus, il existe une augmentation de pente en fin de distribution qui est beaucoup plus évidente de cette probabilité. Ce changement montre clairement que les articles qui se situent dans cette zone auront très peu de chance d'être cités. Ce qui veut dire que ces articles ne sont plus vraiment d'actualité.
2. β indique que le temps médian de citation est de 41,8 mois et 39,8 mois, ce qui représente pratiquement la moitié par rapport au premier profil. Cela montre un retour (de première citation) pour ces 2 ensembles qui est beaucoup plus rapide.

Le troisième profil (Auteur 4) possède non seulement un faible coefficient α (0,3507) mais surtout un très faible coefficient β (1,3201). En réalité, ce profil se distingue du précédent seulement par le coefficient β qui montre que le retour (de première citation) est beaucoup plus rapide puisque le temps médian de première citation est seulement de 20,9 mois. Ce faible coefficient est expliqué par une forte utilisation d'articles courts, près de 40 % (les autres ensembles en possèdent moins de 5%), à parution rapide « short notes » qui favorise forcément un retour plus rapide des citations. En effet, lors de la collecte de données temporelles, ce type d'articles n'a pas été éliminé. Dès lors, la perturbation provient non seulement des individus de type « short notes » mais aussi des données qui peuvent provenir de citations faites par ce type d'articles. Pour connaître le vrai profil de cet ensemble, il serait nécessaire d'éliminer cette perturbation. Par contre, il serait très certainement possible de quantifier le gain au niveau du temps de première citation par l'utilisation de ce type d'articles par une comparaison des données α et β issus d'études avec ou sans short notes, citées par des short notes

4. Conclusion

A partir des articles scientifiques, il est possible d'appréhender le type de recherche effectué par une communauté scientifique ou un individu de cette communauté. Bien que ce travail préliminaire demande confirmation sur un plus grand nombre d'auteurs et une généralisation aux journaux scientifiques, l'étude graphique devrait permettre de distinguer la recherche dite fondamentale par rapport à la recherche dite d'innovation ou technologique à partir des profils de courbes de distributions obtenus et des coefficients α et β de la droite de régression. Ils devraient permettre de constater un retour plus ou moins rapide de l'investissement scientifique en fonction de leurs valeurs et d'être utilisés comme outils d'évaluation de la recherche.

île Rousse 2005
Journée sur les systèmes d'information élaborée

La comparaison des journaux (grands journaux de classe A) et des journaux plus spécifiques à des communautés réduites devrait aussi confirmer une notion de cible de « publications » adaptée à des communautés différentes et complémentaires (technologie et innovation / fondamentale).

Séparer ces communautés et ces supports de publications peut apparaître complètement artificiel. C'est souvent le cas de l'évaluation traditionnelle qui néglige les brevets, les rapports techniques, les actes de congrès, les articles de vulgarisation au profit des taux de citations et des facteurs d'impact sans avoir « souvent » une image réelle des cibles de ces supports (type de documents et type de journaux) pour un domaine scientifique.

Cette évaluation se complique encore avec le caractère transdisciplinaire d'une recherche, ou avec l'arrivée d'une discipline qui n'existe pas encore comme une vraie communauté. Le réflexe premier est de parler alors de « dispersion » ou de « mauvaise perception » pour ce qui est le plus souvent du manque d'expertise de l'évaluation et du manque de pertinence du système d'évaluation. Ce travail est une contribution à l'évaluation stratégique de la recherche.