

Extrait du Géologie et géo-tourisme

<http://jfmoyen.free.fr>

# Comment mesurer une productivité en recherche ?

- Academic life -

Date de mise en ligne : lundi 12 octobre 2009

## **Description :**

Parlons des choses qui fâchent.

---

Géologie et géo-tourisme

---

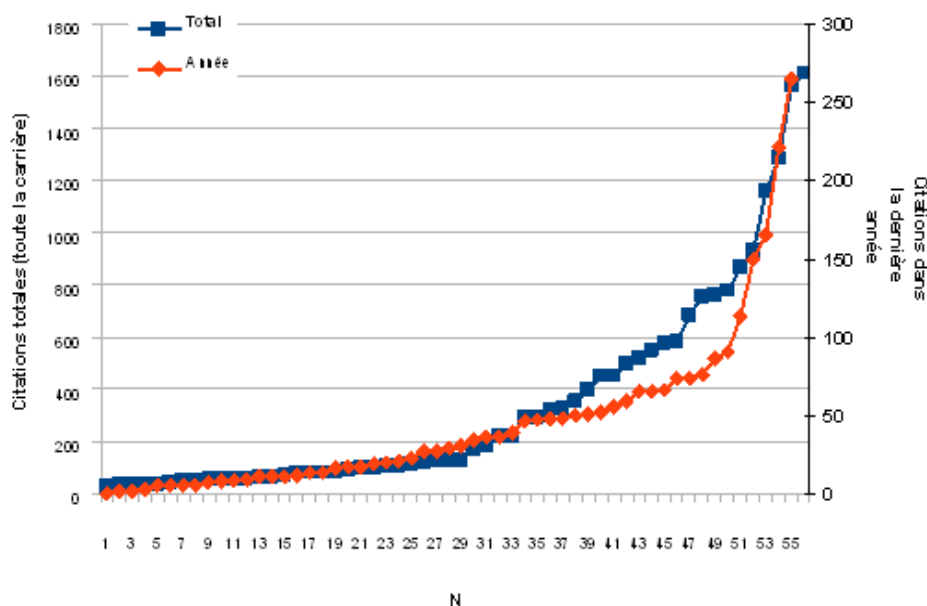
Je retrouve dans mes fichiers un tableau instructif, qui donne des informations bibliométriques pour une UMR de géologie de taille moyenne (55 personnes). Les données datent de quelques années et je ne vous dirait pas de quelle UMR il s'agit ; sachez simplement qu'on y trouve (comme souvent en géologie) des personnes enseignement supérieur, des personnes CNRS et des statuts divers (IRD, physiciens...).

Pour toutes ces personnes, j'ai leur âge ; leur nombre de citation total au cours de leur carrière, tel que connu par ISI Web of Science (c'est à dire en fait depuis, je crois, 1986) ; leur nombre de citations dans la dernière année ; et le facteur "h [1]". Dans tout les cas je ne fais pas la différence entre papiers "premier auteur" et autres.

On peut utiliser ces données comme un "proxy" de l'activité de recherche des gens, ou du moins de leur reconnaissance au sein de leur communauté. Je ne rentrerais pas dans le débat de savoir dans quelle mesure les publications sont un reflet de la qualité en recherche des gens ; je me borne à utiliser ces données, qui ont le mérite d'être quantitatives, mesurables et reproductibles, ce qui est déjà pas mal. Si on me pousse dans mes retranchements, j'ajouterais que à tout prendre, ce n'est pas une plus mauvaise mesure que un "comité d'expert" qui décide souverainement que "Untel est très bon et Telautre un peu moins".

## Activité globale

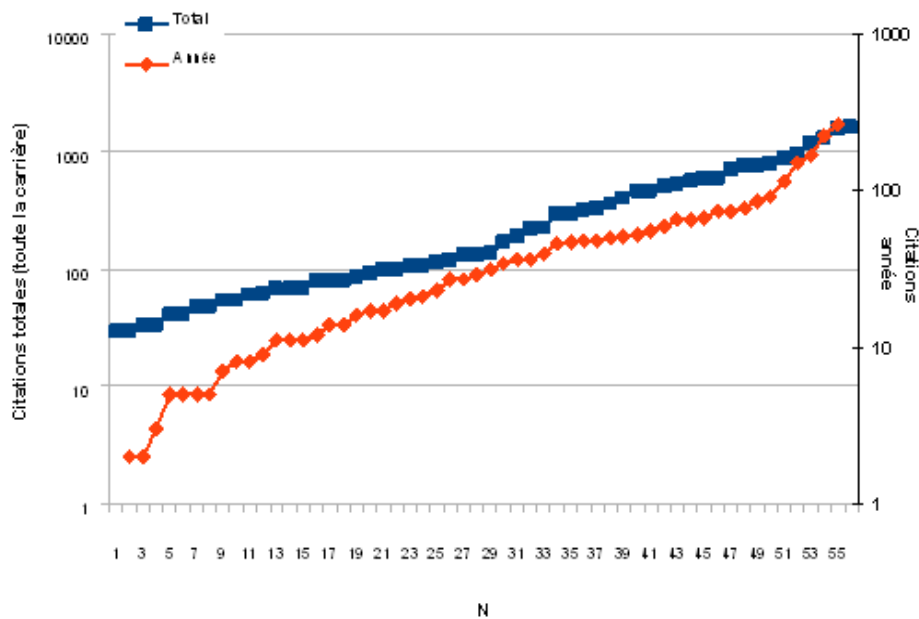
Les premiers diagrammes qu'on peut faire représentent, simplement, le nombre de citations des gens, sous forme d'histogramme ou de courbe de fréquence cumulée. On voit tout de suite sur le diagramme ci-dessous (en bleu, citations totales ; en rouge, en un an) qu'il y a une assez bonne corrélation entre citations totales et citations dans la dernière année, ce qui suggère que les gens sont assez réguliers dans leur carrière ; on voit aussi que la distribution n'est pas du tout gaussienne (dans un diagramme comme celui-ci, une distribution gaussienne ferait une sigmoïde avec un plateau à chaque bout).



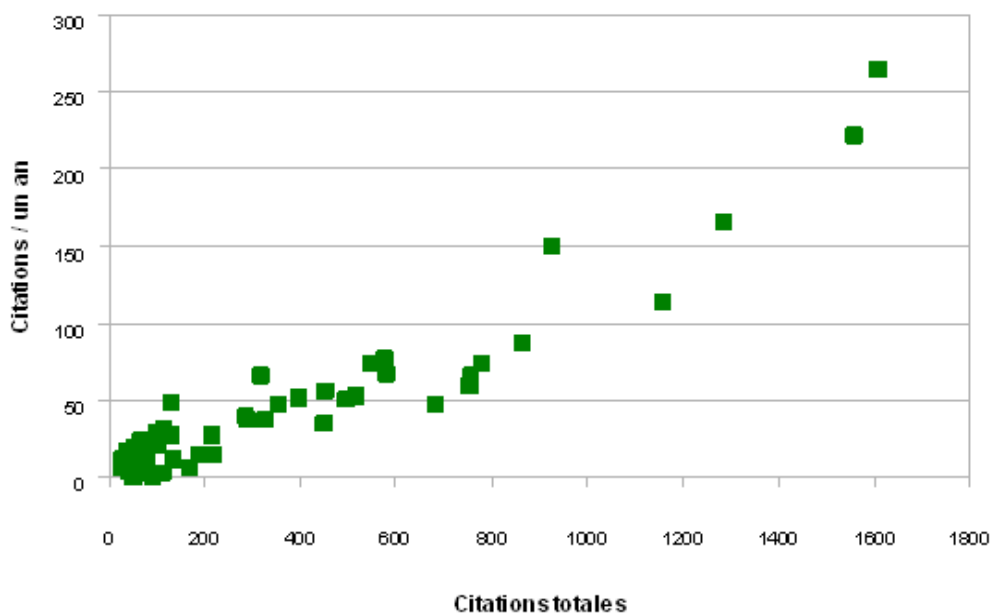
En fait, la distribution est plutôt log-normale, comme on le voit si on met les citations en échelle log : on trouve une

## Comment mesurer une productivité en recherche ?

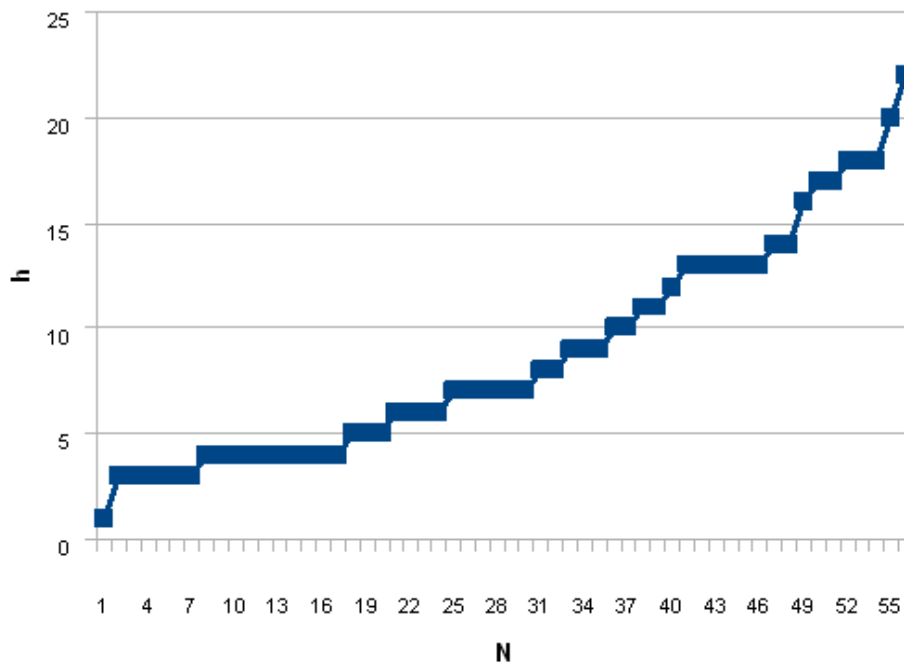
assez jolie droite, moins nette pour la courbe annuelle que pour la courbe cumulative (mais sur un an on va sentir l'effet d'un gros papier qui vient de sortir, ou de débutants pas encore connus).



On peut vérifier la corrélation annuel vs. total : elle est assez bonne, mais si on regarde bien on voit quelques points isolés sur la gauche, qui définissent peut être une courbe plus pentée : quelques jeunes qui sont beaucoup cités en "instantané", mais n'ont pas eu une carrière assez longue pour accumuler des citations. Dans l'ensemble, la bonne corrélation suggère que les gens sont plutôt réguliers sur leur carrière.



Enfin, on peut faire le même diagramme, avec l'indice h : on a une relation à peu près linéaire, ce qu'on attend puisque h varie comme la racine carrée du nombre de citations, plus ou moins.

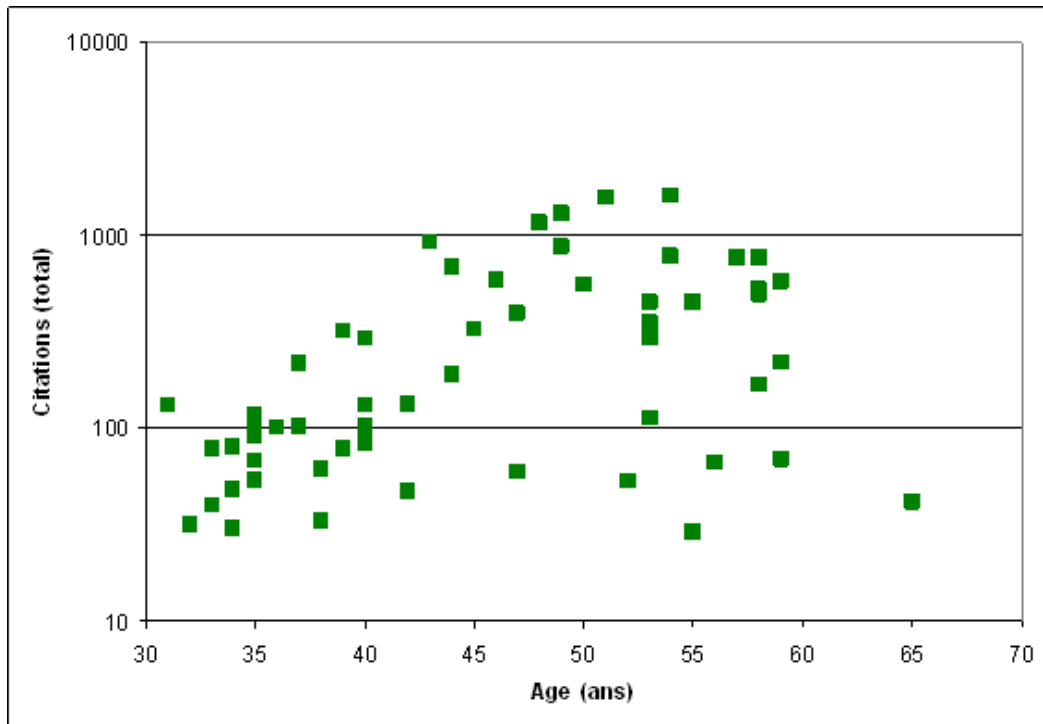


Une implication de ces diagrammes, c'est que 5 personnes dans l'UMR ont entre elles un tiers du total des citations récoltées par le labo. A l'inverse, les 20 personnes du bas de la liste, plus du tiers de l'UMR, ont moins de 5% du total des citations, et accumulent collectivement autant de références que la 5e personne la mieux classé du labo à elle seule ! C'est vrai quelle que soit la valeur qu'on regarde (cumulée ou instantanée). La dispersion des données est immense : en instantané on varie entre 2 [2] à plus de 200, deux ordres de grandeur. De même en cumulé, on varie de 29 à plus de 1600. Les indices h varient de 1 à 22, ce qui est à peu près cohérent et du même ordre.

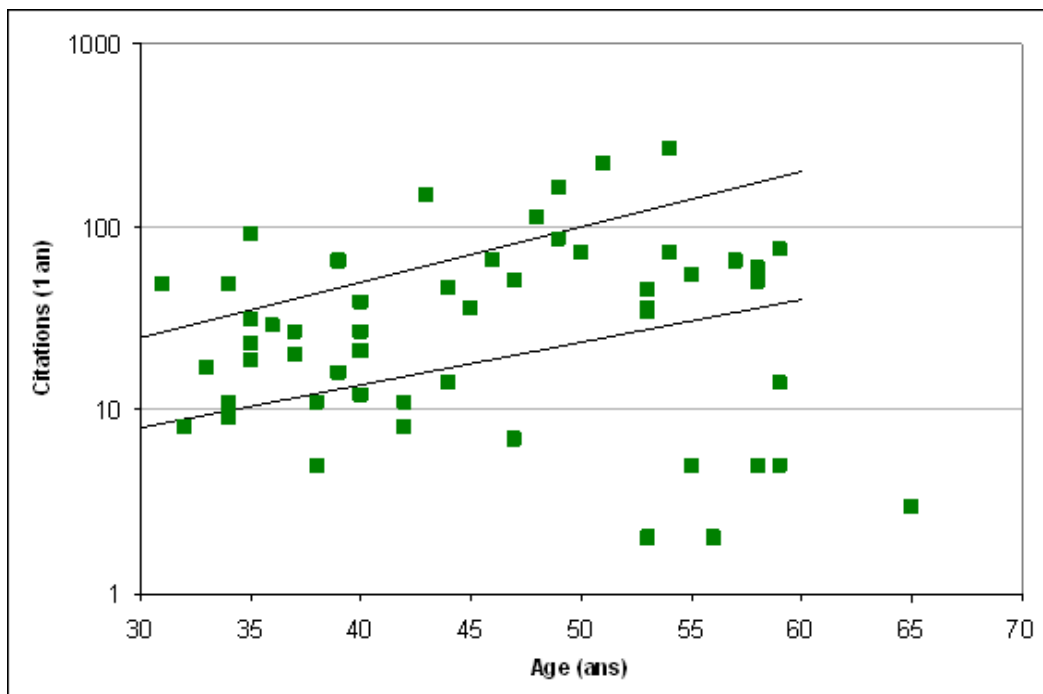
## Corrélation avec l'âge

Il faut aussi se demander si ces différences ne sont pas simplement liés à l'âge et au stade dans la carrière : un jeune, même doué, n'a mécaniquement pas le temps d'avoir amassé des citations. On peut donc faire les deux graphiques suivants, nombre de citations annuelles / cumulées en fonction de l'âge. Je fais le diagramme directement en échelle log, vu que les données couvrent deux ordres de grandeur. De façon assez spectaculaire, on a une sorte de courbe en cloche qui culmine vers 50 - 55 ans avant de redescendre ; je n'ai guère d'explications, peut-être un changement de politique de recrutement à un moment dans le passé, pour prendre des gens plus orientés recherche ?

## Comment mesurer une productivité en recherche ?



Cet effet mis à part, on voit une sorte d'éventail : aucun jeune n'a un très fort nombre de citations, par contre chez les moins jeunes on trouve tout les cas, de presque rien à des nombres énormes. En gros, chacun évolue sur une trajectoire (pas forcément linéaire d'ailleurs, voir plus bas) qui peut être plus ou moins pentue. La dérivée de cette trajectoire, c'est le nombre de citations par an, qu'on peut grapher aussi :



Il y a une corrélation globalement positive (ce qui est logique, plus on a écrit de publis dans sa carrière et plus il y a de chance qu'une d'entre elles soit citée !), mais même comme ça, il y a des différences énormes, d'un ordre de grandeur, pour un groupe d'âge donné. On retrouve encore plus nettement le groupe des "anciens", dont l'activité de recherche n'est pas évidente sur ces graphes (et qui n'ont jamais du avoir de publis très citées, sinon ils

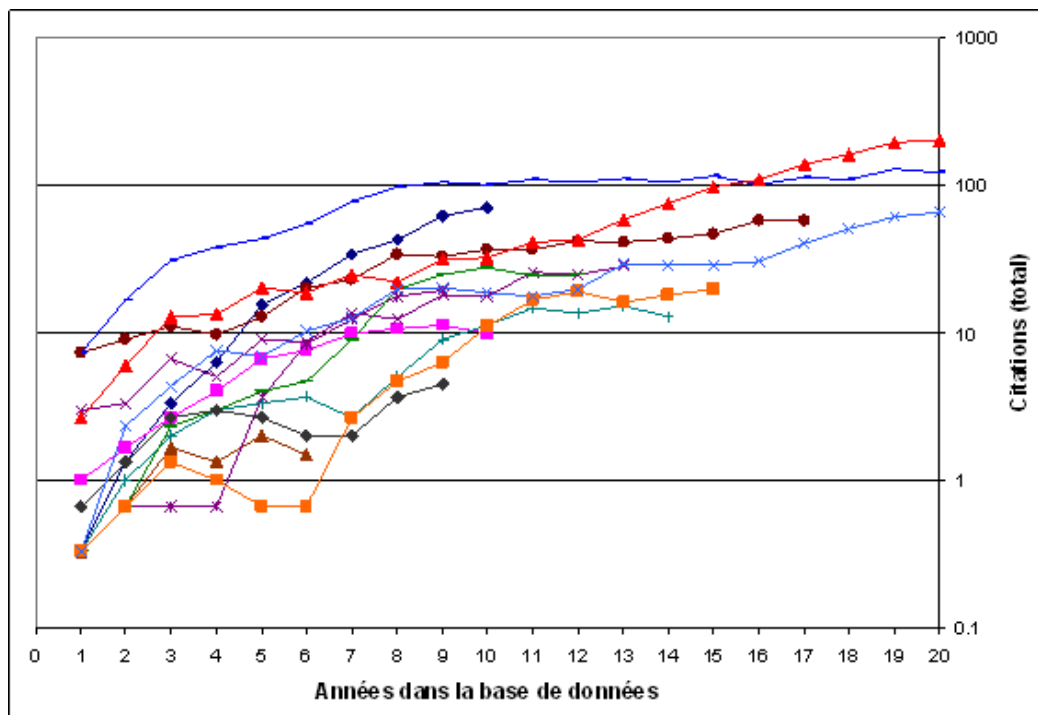
## Comment mesurer une productivité en recherche ?

continueraient à accumuler des citations de vieux papiers) ; il s'agit sans doute du même effet de génération, de recrutement à une époque où les publications n'étaient pas aussi importantes que maintenant, etc [3]. La situation est plus inquiétante chez les "jeunes" (les MC, en gros) : on y trouve des gens avec un taux de citation annuel comparable à celui des cadors de 50 ans, mais aussi des gens qui semblent bien partis pour rester toute leur vie sur une trajectoire très basse.

En gros et au pif, je délimiterais trois zones dans ce diagramme : la zone "haute" qui est celle des bons ou très bon, avec une reconnaissance internationale (ou en passe de l'acquérir, pour les plus jeunes) ; celle du milieu, les chercheurs "corrects", qui font bien leur boulot sans être des leaders internationaux dans ce qu'ils font ; celle du bas, qui est celle des gens dont l'activité de recherche est, disons pudiquement, faible. C'est pour ce dernier groupe qu'on peut se poser des questions, et chercher à voir quelle est leur contribution au fonctionnement de l'Université ; j'en parlerais dans un prochain billet. Notez que ça représente quand même presque le tiers de l'UMR &mdash; un tiers des "chercheurs" et "enseignants-chercheurs" de ce labo ne fait pas, ou si peu, de recherche.... Et ça inclut des jeunes, sensés être en pleine productivité ; il y a des fois des recrutements étonnants.

On peut aussi essayer de regarder, pour une personne donnée, les "trajectoires de publication" en fonction du temps : ici, j'ai extrait pour une douzaine de personnes (pas les mêmes que dans la base de donnée UMR...) le nombre de citations par an, en fonction de l'année d'ancienneté (en fait pour les gens qui ont plus de 20 ans dans le métier, on regarde en fonction du temps écoulé depuis l'an 1 de Web of Science qui est 1988).

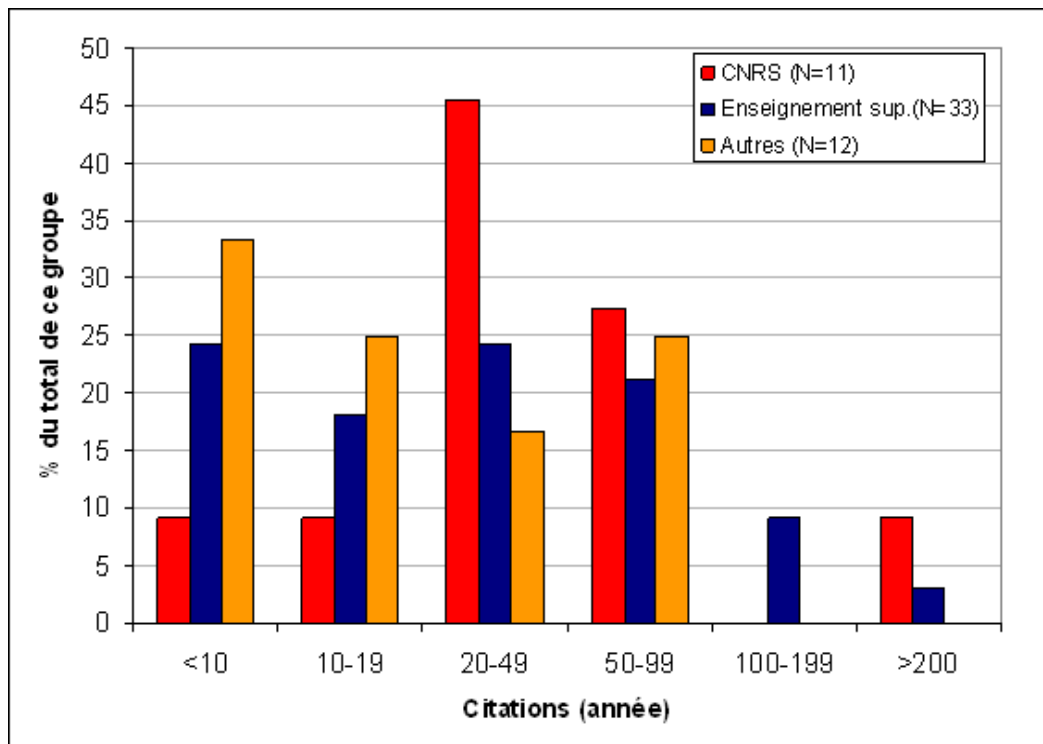
Encore une fois en échelle log, on voit que la plupart des gens arrivent en 8-10 ans à un "plateau", dont ils ne décollent plus ensuite (une exception dans l'échantillon). Le plateau varie d'un ordre de grandeur selon les gens (entre 100 et 10, en gros) ... et pour la plupart des gens, la trajectoire dans les 3-4 premières années est un bon prédicteur de la hauteur du plateau, les courbes individuelles sont pratiquement parallèles et ne se croisent que rarement.



## En fonction du statut

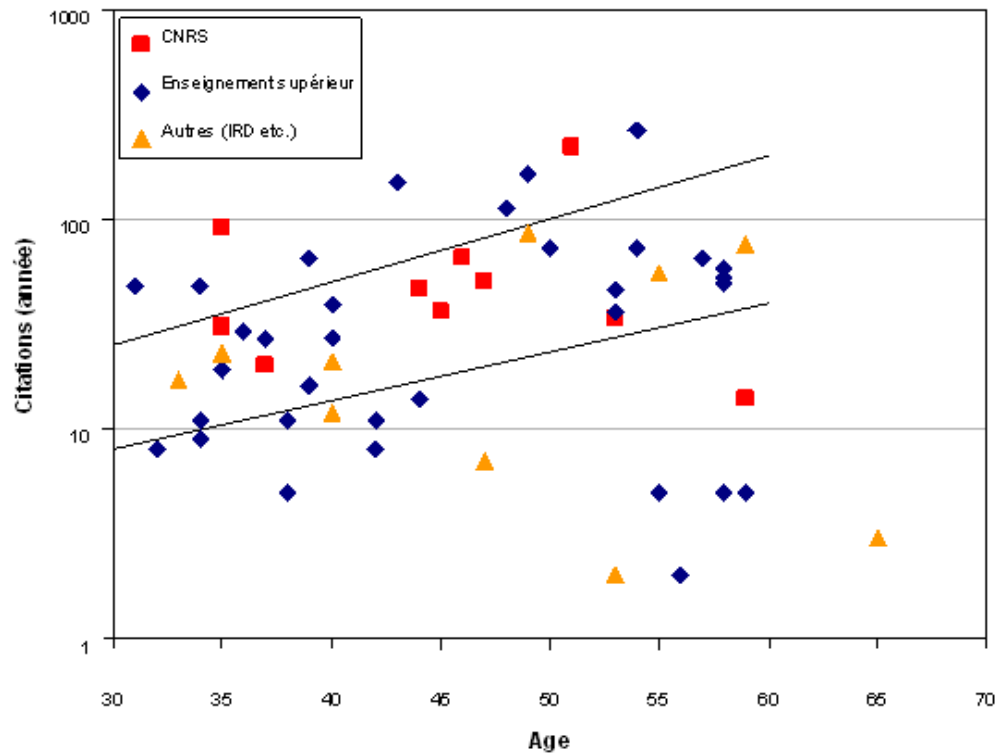
## Comment mesurer une productivité en recherche ?

Pour finir, on peut regarder l'activité des gens en fonction de leur statut : les chercheurs du CNRS n'ayant pas de charge d'enseignement, ils devraient donc avoir une recherche plus soutenue.



Eh bien... Comme vous le voyez sur la figure ci-dessus, la différence est plus subtile. Si on regarde le côté gauche du tableau, les CNRS à très faible taux de citation sont rares, comparés aux enseignants-chercheurs : c'est en partie un effet du groupe des « anciens » évoqué plus haut, qui sont presque uniquement des enseignants-chercheurs. Le pic des CNRS se situe dans le groupe 20-50 citations par an, alors que pour les EC c'est plus étalé sur la gauche. À droite par contre, on ne voit pas vraiment de différence, parmi les gens actifs la distribution des CNRS et celle des EC est très similaire.

Si on reprend la figure précédente en fonction de l'âge, on ne distingue pas non plus de distribution claire (si ce n'est que les "jeunes" CNRS sont plus actifs que les jeunes MC, et qu'il y a peu de « vieux » CNRS ; le seul exemple est comparable aux « vieux » MC) [4]. De façon un peu suprenante aussi, les statuts "particuliers" (IRD, etc.), sensés avoir des missions autres que la recherche, ne se distinguent pas non plus du paquet (alors qu'il serait justifié qu'ils soient "plus bas" !).



[1] Cet indice est défini comme étant le nombre  $h$  d'articles cités plus de  $h$  fois chacun ; si  $h$  vaut 5 par exemple, la personne concernée a 5 articles qui ont été cités chacun 5 fois ou plus. C'est donc un indice qui est plus ou moins logarithmique par rapport au nombre de citations, et qui est construit pour éviter les biais du type "a publié un gros article cité par tout le monde et plus rien depuis", ou "publie très vite plein de petits papiers que personne ne lit". L'idée est de construire un indice robuste qui reflète le statut global de quelqu'un dans sa communauté

[2] En fait il y a même deux personnes à 0 citations dans l'année

[3] Au passage, comme ces données ont 5 ans, une bonne partie de ce groupe a du partir en retraite maintenant, et la distribution ne doit plus tellement montrer cette chute sur la droite

[4] Il faudrait faire des statistiques co-variées, mais je ne sais pas faire. Si quelqu'un veut s'y coller je lui passe le fichier, anonymisé bien sûr.