Comitè Conjunt per a l'Avaluació Quantitativa de la Recerca

Estadístiques de Citacions

Un informe de la International Mathematical Union (IMU) en cooperació amb l' International Council of Industrial and Applied Mathematics (ICIAM) i l' Institute of Mathematical Statistics (IMS)

Versió corregida

6/12/08

Robert Adler, John Ewing ***(President)****, Peter Taylor*

6/11/2008

Sumari

Aquest és un informe sobre l'ús i el mal ús de les citacions en l'avaluació de la recerca científica. La idea que l'avaluació de la recerca cal que es dugui a terme fent servir mètodes "simples i objectius" preval d'una forma creixent avui en dia. Els mètodes "simples i objectius" són majoritàriament identificats amb mètodes *bibliomètrics*, és a dir, basats en dades de citacions i les estadístiques que se'n deriven. Existeix la creença que les estadístiques de citacions són intrínsecament més exactes perquè substitueixen judicis complexos per simples xifres, i per tant superen la possible subjectivitat de les revisions d'experts. Però aquesta creença és infundada.

- Refiar-se de les estadístiques no resulta més exacte si les estadístiques no són utilitzades adequadament. En realitat, les estadístiques poden portar a error si s'apliquen malament o no s'entenen amb exactitud. Gran part de la bibliometria moderna sembla que confia en l'experiència i en la intuïció sobre la interpretació i la validesa de les estadístiques de citacions.

- Tot i que les xifres poden semblar "objectives", la seva objectivitat pot ser il·lusòria. El significat d'una citació pot ser fins i tot més subjectiu que la revisió d'un expert. Degut al fet que aquesta subjectivitat és menys òbvia en les citacions, és molt probable que aquells que les utilitzen no siguin conscients de les limitacions d'aquest mètode.

- El fet de basar-se únicament en dades de citacions produeix, en el millor dels casos, una interpretació incompleta i sovint superficial de la recerca – una interpretació que només és vàlida si es reforça amb altres criteris. *Les xifres no són superiors a criteris sòlids.*

Fer servir les citacions per avaluar la recerca significa en última instància fer servir citacions per classificar objectes: revistes, articles, gent, programes i disciplines. Les eines estadístiques utilitzades per classificar aquests objectes són sovint mal interpretades i mal utilitzades.

- En el cas de les revistes, el factor d'impacte és l'eina usada majoritàriament per classificar-les, i es tracta simplement d'una mitjana feta a partir de la distribució de citacions d'una col·lecció d'articles dins la revista. La mitjana només capta una petita part de la informació sobre aquesta distribució i es tracta d'una estadística més aviat rudimentària. A més a més, hi ha molts factors que indueixen a confusió quan es jutja les revistes per les seves citacions, i qualsevol comparació de revistes requereix prudència quan es fa basant-se en els factors d'impacte. Fer servir només els factors d'impacte per jutjar una publicació és com basar-se en el pes d'una persona per jutjar com és la seva salut.

- Pel que fa als articles, enlloc de confiar en el recompte real de citacions per comparar articles individuals, la gent freqüentment fa servir el factor d'impacte de les revistes on l'article ha aparegut. Pensen que factors d'impacte més elevats signifiquen recomptes més elevats de citacions de cada article, però molt sovint aquest no és el cas! Es tracta d'un mal ús omnipresent que cal que sigui qüestionat en tot moment allà on es produeixi.

- Pel que fa als científics individuals, donat que els llistats complets de citacions poden ser difícils de comparar, hi ha hagut intents per trobar estadístiques simples que continguin tota la complexitat de les citacions d'un científic dins una sola xifra. El més notable d'aquests intents és l'*h-index*, que sembla que està guanyant popularitat. Però fins i tot una inspecció superficial de *l'h-index* i les seves variants mostra que són temptatives ingènues a l'hora de comprendre la complexitat d'un llistat de citacions: capten poca informació sobre la distribució de les citacions d'un científic i perden informació crucial per a l'avaluació de la recerca.

La validesa d'estadístiques com el factor d'impacte i l'*h-inex* no està ni ben entesa ni ben estudiada. La connexió d'aquestes estadístiques amb la qualitat en la recerca de vegades està establerta sobre la base de l'"experiència"; la justificació per confiar en elles és que es poden aconseguir fàcilment. Els pocs estudis que s'han fet sobre aquestes estadístiques s'han limitat a mostrar alguna correlació amb altres maneres de mesurar la qualitat més que no pas a determinar com es pot extreure informació útil de les citacions.

Nosaltres no descartem les estadístiques de citacions com a eina per avaluar la qualitat en la recerca: les citacions i les estadístiques poden proporcionar alguna informació de valor, i, a més, reconeixem que l'avaluació ha de ser pràctica. Per aquest motiu les fàcilment calculables estadístiques de citacions gairebé segur que hauran de ser part del procés. Però les citacions només proporcionen una visió limitada i incompleta de la qualitat en la recerca, i les estadístiques realitzades amb dades de citacions de vegades són mal enteses i mal emprades. La recerca és massa important perquè el seu valor sigui mesurat únicament per mitjà d'un instrument imprecís.

Esperem que aquells que estan relacionats amb l'avaluació llegiran tant el comentari com els detalls d'aquest informe per tal d'entendre no només les limitacions de les estadístiques de citacions sinó també la millor manera d'utilitzar-les. Si establim nivells elevats a l'hora de dirigir la ciència, probablement hauríem d'exigir els mateixos nivells elevats per avaluar la seva qualitat.

IMU/ICIAM/IMS Comitè Conjunt per a l'Avaluació Quantitativa de la Recerca

Robert Adler, *Technion–Israel Institute of Technology*
John Ewing (President), *American Mathematical Society*

Peter Taylor, *University of Melbourne*

Del responsable del comitè

La campanya per aconseguir més transparència i responsabilitat en el món acadèmic ha creat una "cultura de xifres" en la que les institucions i les persones creuen que les decisions justes es poden assolir a través de l'avaluació algorísmica de dades estadístiques; incapaços de mesurar la qualitat (que és l'objectiu final), els qui prenen decisions substitueixen la qualitat per xifres que puguin mesurar. Aquesta tendència requereix el comentari d'aquells qui de manera professional "tracten amb xifres": els matemàtics i els estadístics.

Introducció

La recerca científica és important. La recerca és subjacent a gran part del progrés en el nostre món modern i proporciona l'esperança de que podem resoldre alguns dels aparentment problemes intractables a què s'ha d'enfrontar l'ésser humà, des del medi ambient fins a l'expansió de la nostra població. Per aquest motiu, governs i institucions de tot el món concedeixen un suport financer considerable a la recerca científica. Naturalment volen saber que els seus diners són invertits sàviament i per això sol·liciten l'avaluació de la recerca que estan finançant per tal de tenir informació a l'hora de prendre decisions sobre futures inversions.

Tot plegat no és nou: s'ha estat avaluant la recerca durant molts anys. El que és nou, no obstant, és el fet de pensar que l'avaluació ha de ser "simple i objectiva" i que això es pot aconseguir confiant fonamentalment en les estadístiques derivades de dades de citacions més que no pas en una varietat de mètodes, un dels quals l'opinió dels mateixos científics. El primer paràgraf d'un informe recent exposa cruament aquesta visió:

> El govern té la intenció de reemplaçar l'actual model per determinar la qualitat de la recerca a les universitats (Exercici de l'Avaluació sobre la Recerca del Regne Unit (RAE)) un cop hagi finalitzat el proper cicle l'any 2008. Els indicadors, més que no pas les revisions d'experts, seran la base del nou sistema i la bibliometria (l'ús del còmput d'articles i de les seves cites) serà un índex de qualitat bàsic en aquest sistema. [Evidence Report 2007, p. 3]

Els que defensen aquesta simple objectivitat creuen que la recerca és massa important per confiar en judicis subjectius. Creuen que les estadístiques de citacions aporten claredat al procés de d'avaluació i eliminen les ambigüitats inherents a altres formes d'avaluació. Creuen que estadístiques curosament seleccionades són independents i estan lliures de parcialitat. Sobretot, creuen que aquestes estadístiques ens permeten comparar totes les parts del procés de recerca (publicacions, articles, investigadors, programes i fins i tot disciplines senceres) de manera simple i efectiva, sense haver d'utilitzar la subjectiva revisió d'experts.

Però aquesta fe en l'exactitud, independència i eficàcia de les estadístiques és fora de lloc.

- En primer lloc, l'exactitud d'aquestes estadístiques és il·lusòria. Tothom sap que les estadístiques poden mentir si no són usades adequadament. El mal ús d'estadístiques de citacions és estès i flagrant. Tot i els nombrosos intents d'advertir en contra d'aquest mal ús (per exemple, el mal ús del factor d'impacte), els governs, les institucions i fins i tot els mateixos científics continuen arribant a conclusions injustificades o fins i tot falses a partir de la mala aplicació de les estadístiques de citacions.

- En segon lloc, el fet de basar-se únicament en dades basades en citacions substitueix un tipus de judici per un altre: enlloc de la subjectiva revisió d'un expert tenim la interpretació subjectiva del que significa una citació. Els qui promouen la confiança exclusiva en indicadors basats en citacions assumeixen implícitament que cada citació significa el mateix respecte el treball citat: el seu impacte. Aquesta assumpció no està provada i probablement és incorrecta.

- En tercer lloc, mentre que les estadístiques són valuoses per entendre el món on vivim, només en proporcionen una comprensió parcial. En el món modern, de vegades és moda afirmar que les mesures numèriques són superiors a les altres formes d'interpretació. Els qui promouen l'ús de les estadístiques de dades per *reemplaçar* una interpretació més completa estan recolzant aquesta creença implícitament. No només cal usar l'estadística *correctament*: també l'hem d'utilitzar *sàviament.*

No estem en contra de l'esforç d'avaluar la recerca, sinó més aviat de l'exigència que aquesta avaluació hagi de basar-se de manera predominant en "simples i objectius" indicadors de citacions (una exigència que sovint s'associa al fet de referir-se a xifres fàcils de calcular que prioritzen publicacions, persones o programes). **La recerca sovint té múltiples objectius, tant a llarg com a curt termini, i per tant és raonable que el seu valor sigui jutjat a partir de criteris múltiples**. Els matemàtics saben que hi ha moltes coses, reals i abstractes, que no poden ser ordenades simplement, en el sentit que es puguin comparar les unes amb les altres paral·lelament. La comparació sovint requereix una anàlisi més complicada, que sovint deixa indecís sobre quin partit prendre. La resposta correcta a "Quin és millor?" de vegades és "depèn!"

La petició d'utilitzar múltiples mètodes d'avaluació de la qualitat en la recerca ha estat feta amb anterioritat (per exemple [Martin 1996] o [Carey – Cowling - Taylor 2007]). Les publicacions poden ser jutjades de moltes maneres, no només a partir de les citacions. Proves d'estima com invitacions, pertinença a comitès editorials i premis sovint són també proves de qualitat. En algunes disciplines i en alguns països el finançament de beques pot jugar-hi un paper. I la revisió d'experts (el judici dels companys científics) és un component important de l'avaluació (no hauríem de descartar la revisió d'experts només perquè de vegades estigui afectada per la parcialitat quan no descartem les estadístiques de cites tot i que de vegades estan afectades pel mal ús que se'n fa. Aquesta és una petita mostra de les múltiples maneres en que l'avaluació pot ser duta a terme. Hi ha molts camins que porten a una bona avaluació, i la seva importància relativa varia entre les disciplines. Tot i això, les estadístiques "objectives" basades en citacions es converteixen repetidament en el mètode preferit d'avaluació. L'atractiu d'un procés simple i de xifres simples (preferentment una sola xifra) sembla que supera el seny i el sentit comú.

Aquest informe està escrit per científics matemàtics per redreçar el mal ús de les estadístiques en l'avaluació de la recerca científica. Evidentment, aquest mal ús de vegades també afecta la disciplina de matemàtiques, i aquest és un dels motius pels quals escrivim aquest informe. L'especial cultura de citacions de les matemàtiques, amb còmputs baixos de citacions de revistes, articles i autors, fa que sigui una disciplina especialment vulnerable a l'abús de les estadístiques de citacions. Creiem, per tant, que *tots* els científics, així com el públic en general, haurien de preocupar-se de fer servir mètodes científics sòlids a l'hora d'avaluar la recerca.

Alguns membres de la comunitat científica prescindirien totalment de les estadístiques de citacions per una reacció extrema a l'abús del passat, però això significaria descartar una eina valuosa: les estadístiques basades en citacions *poden* tenir un paper en l'avaluació de la recerca sempre i quan siguin utilitzades adequadament, interpretades amb prudència i només constitueixin una part del procés. Les citacions aporten informació sobre revistes, articles i persones. No volem amagar aquesta informació: volem fer-la més entenedora.

Aquest és el propòsit d'aquest informe. Les tres primeres seccions assenyalen les maneres en què les citacions poden ser utilitzades (o mal utilitzades) per valorar revistes, articles i científics. La

següent secció tracta dels diferents significats de les citacions i les conseqüents limitacions de les estadístiques basades en les citacions. L'última secció aconsella sobre com aconseguir un ús savi de les estadístiques i insta a moderar l'ús de les estadístiques de cites amb altres judicis, tot i que això faci les avaluacions més complicades.

"Tot hauria de ser fet de la manera més simple possible, però no més simple" va dir Albert Einstein[1]. Aquest consell d'un dels científics més preeminents del món és especialment apte en l'avaluació de la recerca científica.

La classificació de les revistes: El factor d'impacte[2]

El factor d'impacte va ser creat a la dècada dels seixanta del segle passat com una manera de mesurar la vàlua de revistes a partir del càlcul de la mitjana de citacions per article durant un període específic de temps [Garfield 2005]. La mitjana és calculada a partir de les dades recollides del *Thomson Scientific* (anteriorment anomenat "Institut per a la Informació Científica" –Institute for Scientific Information-), que publica el *Journal Citation Reports*. *Thomson Scientific* extreu referències de més de 9.000 revistes, afegint informació sobre cada article i les seves referències a la seva base de dades cada any [THOMSON: SELECTION]. A través d'aquesta informació es pot comptar quantes vegades un article en concret és citat per articles posteriors que han estat publicats a la col·lecció de revistes indexades (remarquem el fet que *Thomson Scientific* indexa menys de la meitat de les revistes matemàtiques que apareixen a *Mathematical Reviews* i *Zentralblat*, les dues majors revistes de ressenyes en matemàtiques[3])

Per una revista en particular en un any concret, el factor d'impacte es calcula fent la mitjana de les citacions a articles d'aquesta revista dels dos anys anteriors dins la totalitat d'articles publicats durant aquell any (dins la col·lecció específica de revistes indexades per *Thomson Scientific*). Si el factor d'impacte d'una revista és de 1.5 l'any 2007, significa que, de mitjana, els seus articles publicats durant el 2005 i el 2006 han estat citats 1.5 vegades en la col·lecció de totes les revistes indexades publicades l'any 2007.

El mateix *Thomson Scientific* fa servir el factor d'impacte com un dels factors per determinar quines revistes són indexades [THOMSON: SELECTION]. D'altra banda, *Thomson* promou l'ús del factor d'impacte de manera més general per comparar revistes.

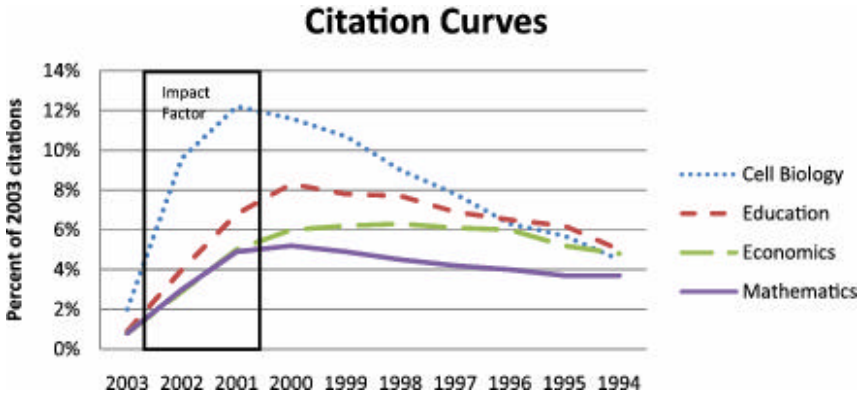
> "Com a eina per a l'administració de les col·leccions de revistes a les biblioteques, el factor d'impacte dóna a l'administrador de la biblioteca informació sobre revistes que ja són dins la col·lecció o sobre altres revistes que s'està considerant adquirir. Aquestes dades també poden combinar-se amb dades sobre el cost i la circulació per prendre decisions racionals en la compra de revistes".[THOMSON: IMPACT FACTOR]

Molts autors han assenyalat que no s'hauria de jutjar la vàlua acadèmica d'una revista fent servir només informació sobre citacions, afirmació que els autors d'aquest informe subscriuen fermament. A banda d'aquest comentari general, el factor d'impacte ha estat criticat també per altre motius (veure [Seglen 1997], [Amin - Mabe 2000], [Monastersky 2005], [Ewing 2006], [Adler 2007], i [Hall 2007]).

(i) La identificació del factor d'impacte amb una mitjana no és del tot correcta. Degut al fet que moltes revistes publiquen textos no substancials com cartes o editorials, que gairebé mai són

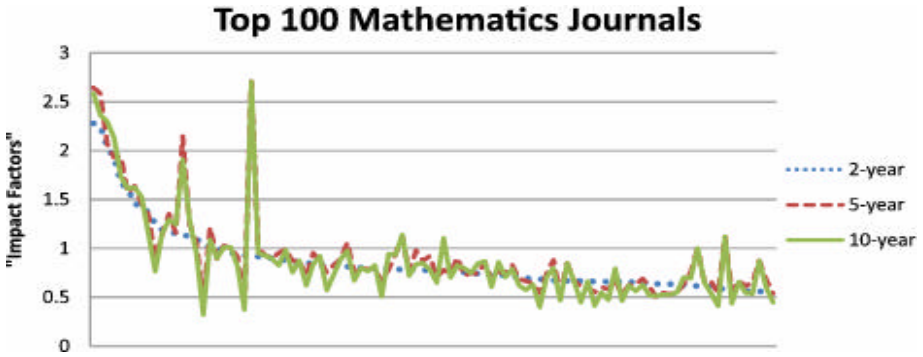
citats, aquests textos no són comptats dins el denominador del factor d'impacte. Però d'altra banda, tot i que no és freqüent, aquests textos de vegades sí que són citats, i aquestes citacions *són* comptades en el numerador. Per tant aquest factor d'impacte no és ben bé la mitjana de citacions per article. Quan les revistes publiquen un gran nombre d'aquests textos "no substancials", aquesta desviació pot ser significativa. En moltes àrees, entre elles la de les matemàtiques, aquesta desviació és mínima.

(ii) La causa d'establir un període de dos anys per definir el factor d'impacte és fer que l'estadística sigui actual [Garfield 2005]. Per a alguns camps, com les ciències biomèdiques, això és apropiat perquè la majoria dels articles publicats reben la majoria de citacions poc després de ser publicats. En altres camps, com les matemàtiques, la majoria de les citacions es donen més enllà del citat període de dos anys. Examinant una col·lecció de més de 3 milions de citacions recents en revistes matemàtiques (de la base de dades *Math Reviews Citation*) es veu que aproximadament el 90% de les citacions a uns revista cauen for a d'aquest marge de dos anys. Conseqüentment, el factor d'impacte es basa en un simple 10% de l'activitat citadora i deixa escapar la vasta majoria de citacions [4].



El gràfic mostra l'edat de les citacions d'articles publicats el 2003 que cobreixen 4 camps diferents (Biologia Cel·lular, Educació, Economia i Matemàtiques). Les citacions a articles publicats el 2001-2002 són les que contribueixen al factor d'impacte; totes les altres citacions són irrellevants per al factor d'impacte. Dades del *Thomsom Scientific.*

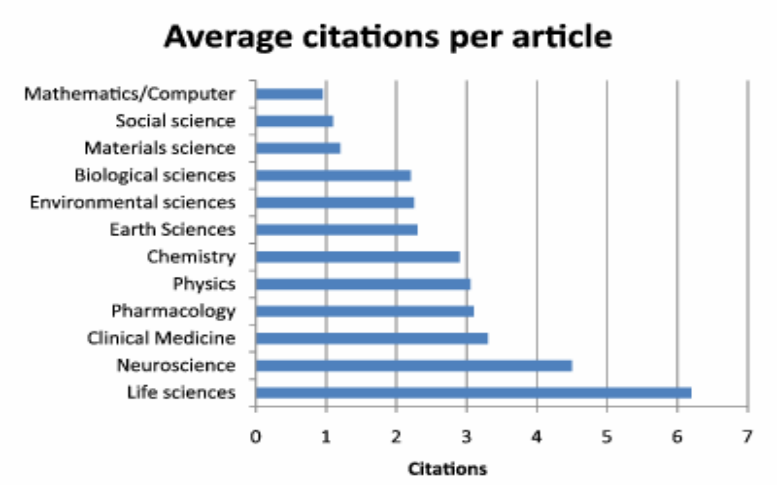
Aquest interval de dos anys significa que el factor d'impacte és erroni? Pel que fa a les revistes matemàtiques l'evidència és equívoca. *Thomson Scientific* calcula factors d'impacte de cinc anys, el qual mostra afinitat amb el factor d'impacte habitual de dos anys [Garfield 1998]. Si fem servir la base de dades de citacions *Math Reviews*, podem calcular "factors d'impacte" (és a dir, mitjanes de citacions per article) per a una col·lecció de 100 de les revistes matemàtiques més citades fent servir períodes de 2, 5 i 10 anys. El gràfic que apareix a continuació mostra que els factors d'impacte de 5 i 10 anys generalment segueixen el camí del factor d'impacte de 2 anys.



"Factors d'impacte" calculats amb dades de períodes de 2, 5 i 10 anys de 100 revistes matemàtiques. Informació extreta de la base de dades de citacions *Math Reviews*.

El pic que dispara el factor d'impacte correspon a una revista que no va publicar articles durant part d'aquest període; els pics més petits corresponen a revistes que publiquen un nombre relativament petit d'articles cada any, i el gràfic simplement reflecteix la variació normal del factor d'impacte d'aquestes revistes. És clar que el fet de canviar el nombre d'anys tinguts en compte per calcular el factor d'impacte canvia la classificació de les revistes, però aquests canvis són generalment modestos excepte en les revistes petites, en les quals el factor d'impacte també varia quan es canvia l'any d'origen de les dades (llegiu més avall).

(iii) El factor d'impacte varia considerablement entre les diferents disciplines [Amin-Mabe 2000]. Part d'aquesta diferència prové de l'observació (ii): si en algunes disciplines es donen moltes citacions fora del marge de dos anys, els factors d'impacte de les seves revistes serà molt més baix. D'altra banda, part de la diferència és simplement que la cultura de citacions varia segons la disciplina, i els científics citen articles a diferents ritmes i per motius diferents (donarem més detalls sobre aquesta última observació perquè el significat de les citacions és extremadament important). De tot això es dedueix que és impossible comparar de manera significativa dues revistes fent servir els factors d'impacte.



Mitjanes de citacions per article en les diferents disciplines, mostrant que les pràctiques de citacions difereixen remarcablement en cada disciplina. Dades de Thomson Scientific [Amin-Mabe 2000].

(iv) El factor d'impacte pot variar considerablement d'un any per l'altre, i la variació tendeix a ser més gran en les revistes més petites [Amin-Mabe 2000]. En revistes que publiquen menys de 50 articles, per exemple, el canvi mitjà en el factor d'impacte de 2002 a 2003 va ser gairebé del 50%. Això era d'esperar, evidentment, perquè les mostres de les revistes petites són petites també. Tot i això, sovint es compara revistes per un any en concret, sense tenir en compte l'alta variació de les revistes petites.

(v) Les revistes que publiquen articles en idiomes diferents de l'anglès segurament rebran menys citacions perquè gran part de la comunitat científica no els pot llegir o no els llegeix. I el tipus de revista, més que no pas la qualitat exclusivament, pot influenciar el factor d'impacte. Revistes que publiquen ressenyes, per exemple, sovint rebran moltes més citacions que revistes que no ho fan, i per tant tindran factors d'impacte més elevats (de vegades, de forma substancial) [Amin-Mabe 2000].

(vi) La crítica més important al factor d'impacte és que el seu significat no s'entén correctament. Quan fem servir el factor d'impacte per comparar dues revistes, no hi ha un model *a priori* que defineixi què vol dir ser "millor". L'únic model deriva del mateix factor d'impacte: un major factor d'impacte significa que la revista és millor. En el paradigma estadístic clàssic es defineix un model, es formula una hipòtesi (de no-diferència) i llavors es fa una estadística, que depenent dels valors que tingui permet acceptar o rebutjar la hipòtesi. Derivar informació (i possiblement un model) de les dades mateixes és una aproximació legítima a l'anàlisi estadística, però en aquest cas no està clar que la informació hagi estat derivada. De quina manera mesura la qualitat el factor d'impacte? Es tracta de la millor estadística per mesurar la qualitat? Què és el que mesura *exactament*? (L'argumentació sobre el significat de les citacions és rellevant aquí). Se sap remarcablement poc sobre cap model de qualitat per a revistes o de quina manera s'hauria de relacionar amb el factor d'impacte.

Les sis crítiques anteriors al factor d'impacte són totes vàlides, però només signifiquen que el factor d'impacte és una dada rudimentària, no inútil. Per exemple, el factor d'impacte pot ser utilitzat com a punt de partida per classificar les revistes en grups usant inicialment el factor d'impacte per definir els grups i llavors utilitzar altres criteris per tornar a definir la classificació i verificar que els grups fan sentit. Però utilitzar el factor d'impacte per avaluar revistes requereix prudència: no pot ser utilitzat per comparar revistes de diferents disciplines, per exemple, i cal examinar detingudament de quin tipus de revista es tracta abans de fer servir el factor d'impacte per classificar-la. També caldria tenir en compte les variacions anuals, especialment pel que fa a les revistes petites, i entendre que les petites diferències poden ser simplement fenòmens aleatoris. També és important reconèixer que el factor d'impacte pot no reflectir de manera exacta tot l'àmbit de l'activitat de citació en algunes disciplines, tant perquè no totes les revistes estan indexades com perquè el període tingut en compte és massa curt. Altres estadístiques basades en períodes de temps més llargs i en més revistes poden ser millors indicadors de qualitat. Finalment, les citacions només són una manera de jutjar les revistes, que hauria de ser complementada amb altra informació (aquest és el missatge central d'aquest informe).

Totes aquestes precaucions són similars a les que s'haurien de fer en qualsevol classificació basada en estadístiques. Classificar cegament les revistes d'acord als factors d'impacte en un any determinat és fer un mal ús de l'estadística. A favor d'això cal afegir que *Thomson Scientific* està d'acord amb aquesta afirmació i (sense molt èmfasi) adverteix sobre aquesta qüestió a aquells que utilitzen els factors d'impacte.

> "*Thomson Scientific* no es basa únicament en el factor d'impacte per avaluar la utilitat d'una revista, i ningú no ho hauria de fer. El factor d'impacte no hauria de ser utilitzat sense una profunda atenció als molts fenòmens que influencien els índexs de citacions, com per exemple el nombre de referències citades en l'article mitjà. El factor d'impacte s'hauria d'utilitzar sota una revisió d'expert ben fonamentada" [THOMSON: IMPACT FACTOR]

Malauradament, aquest consell sovint és ignorat.

Classificació d'articles

El factor d'impacte, així com altres estadístiques similars basades en les citacions, pot ser mal utilitzat en la classificació de revistes, però hi ha un mal ús encara més fonamental i insidiós: l'ús del factor d'impacte per comparar articles individuals, investigadors, programes o fins i tot disciplines. Es tracta d'un problema creixent que s'estén per moltes nacions i moltes disciplines, empitjorat per les recents avaluacions nacionals de la recerca.

En certa manera, no és un fenomen nou. Sovint es demana als científics que facin judicis sobre historials de publicacions i se senten coses com ara "publica en bones revistes" o "la majoria dels seus articles estan a revistes de baix nivell". Aquesta avaluació pot ser sensata: la qualitat de les revistes en les quals un científic publica generalment (o constantment) és un dels molts factors que es poden utilitzar per avaluar la recerca global d'un científic. El factor d'impacte, en canvi, ha incrementat la tendència a atribuir les propietats d'una revista en particular a cada article que conté (i a cada autor).

Thomson Scientific promou aquesta pràctica de manera implícita:

> "Potser l'ús més important i recent de l'impacte es troba en el procés d'avaluació acadèmica. El factor d'impacte pot ser utilitzat per obtenir una aproximació al prestigi de les revistes en les quals els individus han estat publicats." [THOMSON: IMPACT FACTOR]

Aquí hi ha alguns exemples de les maneres en què la gent ha interpretat aquest consell, reportades per matemàtics de tot el món:

> Exemple 1: La meva universitat ha introduït recentment una nova classificació de revistes fent servir les publicacions del *Science Citation Index Core*. Les revistes han estat dividides en tres grups basant-se exclusivament en el factor d'impacte. Hi ha 30 revistes en el primer grup, que no en conté cap de matemàtiques. La segona llista conté 667 revistes, de les quals 21 són de matemàtiques. El fet de publicar en la primera llista permet triplicar l'ajuda per a la recerca per part de la universitat; el fet de publicar en la segona llista permet duplicar l'ajut. Haver publicat en la llista principal atorga 15 punts; haver publicat en qualsevol de les revistes contemplades per *Thomson Scientific* atorga 10 punts. Aconseguir una promoció requereix un determinat nombre mínim de punts.
>
> Exemple 2: Al meu país, el personal acadèmic de la universitat amb posicions permanents és avaluat cada 6 anys. La clau per obtenir reconeixement acadèmic és tenir avaluacions

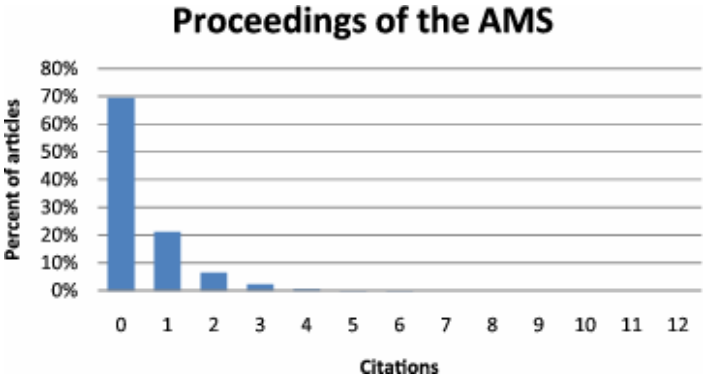
positives successivament. A banda del currículum vitae, el factor d'avaluació més important consisteix a classificar cinc articles publicats. Últimament se'ls dóna 3 punts si apareixen a revistes del primer terç de la llista del *Thomson Scientific*, 2 punts si apareixen al segon terç i un punt si apareixen a l'últim terç (les tres llistes es creen segons el factor d'impacte).

Exemple 3: Al nostre departament, cada membre del personal acadèmic és avaluat a través d'una fórmula relacionada amb el nombre d'articles "equivalents a un únic autor", multiplicat pel factor d'impacte de la revista on han aparegut. Les promocions i les contractacions són fetes parcialment basant-se en aquesta fórmula.

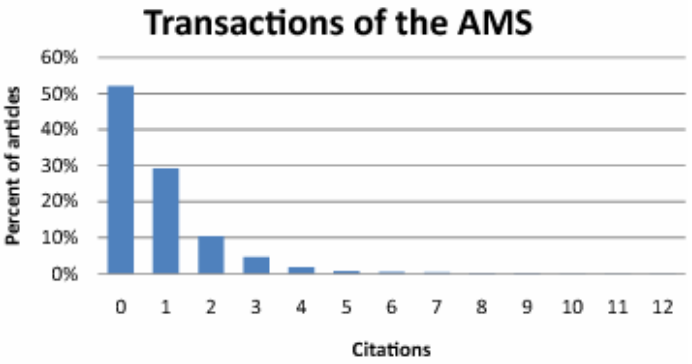
En aquests exemples, així com en altres que ens han estat reportats, es mostra com el factor d'impacte està sent utilitzat explícitament o implícitament per comparar articles individuals i també els seus autors: si el factor d'impacte de la revista A és més elevat que el de la revista B, un article publicat a la revista A ha de ser superior a un article publicat a B, i segurament un autor A serà superior a un autor B. En alguns casos, aquest raonament també es fa servir per classificar departaments o fins i tot disciplines senceres.

És ben sabut que la distribució de còmputs de citacions d'articles individuals en una revista és altament enganyós, aproximat a l'anomenada llei potencial ([Seglen 1996], [Garfield 1987]). Això té conseqüències que es poden plasmar en exemples concrets.

La distribució dels articles en la revista *Proceedings of the American Mathematical Society* en el període 2000-2004 es pot veure a sota. Aquesta revista publica articles curts, generalment tenen una llargada inferior a 10 pàgines. Durant aquest període, va publicar 2.381 articles (unes 15.000 pàgines). Consultant els articles del 2005 en la base de dades *Math Reviews*, el còmput mitjà de citacions per article (és a dir, el factor d'impacte) és 0,434.



La revista *Transactions of the AMS* publica articles més llargs que generalment són més substancials, tant pel que fa a l'abast com pel que fa al contingut. En el mateix període de temps, aquesta revista va publicar 1.165 articles (més de 25.000 pàgines), amb còmputs de citacions que van de 0 a 12. La mitjana de citacions per article va ser 0,846 – gairebé el doble que el de *Proceedings of the American Mathematical Society.*



Ara considerem dos matemàtics, un que publiqui un article a *Proceedings of the American Mathematical Society* i l'altre a *Transactions of the AMS.* Segons algunes de les pràctiques institucionals mencionades anteriorment, el segon seria considerat superior al primer, degut al fet d'haver publicat un article en una revista amb un factor d'impacte més elevat –de fet, el doble!. És aquesta avaluació vàlida? Són els articles de *Transaction of the AMS* el doble de bons que els de *Proceedings of the American Mathematical Society*?

Quan afirmem que un article de *Transactions* és millor (basant-nos en el càlcul de les citacions) que un de *Proceedings*, no ens hem de preguntar res sobre mitjanes sinó més aviat sobre probabilitats: quina és la probabilitat que ens estiguem equivocant? Quina és la probabilitat que un article seleccionat aleatòriament de la revista *Proceedings* tingui al menys les mateixes citacions que un article seleccionat aleatòriament de la revista *Transactions*?

Es tracta d'un càlcul elemental, i la resposta és 62%. Això vol dir que estem equivocats el 62% del temps, i que un article seleccionat aleatòriament de *Proceedings* serà tan bo (o millor) que un article seleccionat aleatòriament de *Transactions* (tot i que el factor d'impacte de la primera és la meitat que el de la segona!). Ens equivoquem més sovint que no encertem. *La majoria* de la gent troba aquest fet sorprenent, però és una conseqüència de la distribució altament enganyosa i de l'estret marge de temps contemplat per calcular el factor d'impacte (aquest és el motiu de l'alt nombre d'articles no citats[5]). Mostra la vàlua del pensament estadístic precís més que no l'observació intuïtiva.

Aquest és el típic comportament pel que fa a revistes, i no hi ha res d'especial en la tria d'aquestes dues publicacions (per exemple, *Journal of the AMS* té un factor d'impacte 2,63 en el mateix període –sis vegades el de *Proceedings*-), però un article seleccionat aleatòriament de *Proceedings* és al menys igual de bo que un article de *Journal*, pel que fa a les citacions, el 32% de les vegades)

Per tant, **tot i que que és incorrecte dir que el factor d'impacte no dóna informació sobre un determinat article en una revista, sí que és cert que la informació que dóna és sorprenentment imprecisa i que pot ser dramàticament desorientadora**.

De tot això es dedueix que els tipus de càlculs duts a terme en els tres exemples anteriors –fer servir el factor d'impacte com a representació dels còmputs reals de citacions de cada article- té poc fonament racional. Realitzar afirmacions que són incorrectes més de la meitat de les vegades (o un terç de les vegades) segurament no és una bona manera de dur a terme una avaluació.

Un cop ens adonem que no té sentit substituir el factor d'impacte pel còmput de cites de cada article concret, deduïm que no fa sentit utilitzar el factor d'impacte per avaluar els autors d'aquests articles, els programes en els quals treballen o (amb menys motiu) les disciplines que representen. El factor d'impacte i les mitjanes aritmètiques en general són massa simples per fer que aquest tipus de comparacions siguin significatives, si no es disposa d'informació complementària.

És evident que classificar científics no és el mateix que classificar els seus articles, però si es vol classificar els articles d'algú fent servir només les citacions per mesurar la qualitat d'un article determinat, cal començar per comptar les citacions d'aquest article. El factor d'impacte de la revista no és un substitut fidedigne.

Classificació dels científics

Mentre que el factor d'impacte ha estat l'estadística basada en citacions més coneguda, hi ha altres estadístiques més recents que ara són promogudes activament. Aquí tenim una petita mostra de tres d'aquestes estadístiques creades per classificar individus.

h-index: L'*h-index* d'un científic és l'*n* més elevat pel qual ha publicat *n* articles, cada un amb un mínim d'*n* citacions.

Aquesta és l'estadística més popular mencionada aquí. Va ser proposada per J.E. Hirsch [Hirsch 2006] per tal de mesurar "la producció científica d'un investigador" centrada en el límit superior de la distribució de cites d'una persona. L'objectiu era substituir els recomptes de publicacions i la distribució de cites per una sola xifra.

m-index: L'*m-index* d'un científic és l'*h-index* dividit pel nombre d'anys transcorreguts des del seu primer article.

Això també va ser proposat per Hirsch en l'article mencionat. La intenció era compensar els investigadors joves per no haver tingut temps de publicar articles o aconseguir cites.

g-index: El *g-index* d'un científic és l'*n* més elevat pel qual els *n* articles més citats tenen un total d'almenys n_2 citations.

Aquest índex va ser proposat per Leo Egghe in 2006 [Egghe 2006]. L'*h-index* no té en compte el fet que alguns articles a la part de dalt d'*n* poden tenir recomptes extraordinàriament elevats de citacions. El *g-index* vol compensar aquest fet.

Hi ha més índexs (molts més) que inclouen variants d'aquests tres que tenen en compte l'edat dels articles o el nombre d'autors ([Batista - Campiteli - Kinouchi - Martinez 2005], [Batista - Campiteli - Kinouchi 2006], [Sidiropouls-Katsaros-Manolopoulos 2006]).

En l'article on definia l'*h-index*, Hirsch va escriure que ell proposava aquest sistema com "un índex fàcil de calcular, que dóna una estimació de la importància, la significació i l'amplitud de l'impacte de l'acumulació de contribucions a la recerca d'un científic." [Hirsch 2005, p. 5]. Continuava tot afegint que "aquest índex pot proporcionar un criteri útil per comparar individus que competeixen per al mateix recurs quan un dels criteris importants d'avaluació siguin els assoliments científics".

Cap d'aquestes afirmacions està basada en proves convincents. Per recolzar la seva reivindicació que l'*h-index* mesura la importància i la significança de la recerca acumulativa d'un científic, Hirsch analitza *l'h-index* d'un conjunt de guanyadors del premi Nobel (i, separadament, membres de l'Acadèmia Nacional). Hirsch demostra que les persones incloses en aquests grups generalment tenen un *h-index* elevat. Es pot concloure que és probable que un científic tingui un *h-index* elevat si és un premi Nobel. Però sense informació addicional no sabem quina és la probabilitat que algú arribi a ser un premi Nobel o un membre de l'Acadèmia Nacional pel fet que tingui un *h-index* elevat. Aquest és el tipus d'informació que ens falta per establir la validesa de *l'h-index.*

Al seu article, Hirsch també afirma que es pot fer servir l'*h-index* per comparar dos científics:

> "Sostinc que dos individus amb *h* similars són comparables pel que fa al seu impacte científic en general, fins i tot si el seu nombre total d'articles i de citacions és molt diferent. En canvi, el fet que dos individus (de la mateixa edat científica) tinguin un nombre similar d'articles publicats i de citacions però un valor *h* molt diferent significa amb tota probabilitat que el que té *l'h* més elevat és el científic més dotat." [Hirsch 2005, p. 1]

Sembla que aquestes assercions es poden refutar amb el sentit comú. (Penseu en dos científics, cadascun amb 10 articles amb 10 citacions, però un d'ells amb 90 articles més amb 9 citacions cada un; o suposeu que un d'ells té exactament 10 articles de 10 citacions i l'altre 10 articles de 100 citacions cada un. Algú podria considerar-los equivalents?)[6]

Hirsch lloa les virtuts de l'*h-index* i reivindica que "l'*h* és preferible a altres criteris d'una xifra utilitzats habitualment per avaluar la producció científica d'un investigador..."[Hirsch 2005, p. 1], però ni defineix "preferible" ni explica perquè és necessari trobar "criteris d'una xifra"

Tot i que hi ha hagut certa crítica a aquest mètode, l'anàlisi seriosa ha estat escassa. Gran part d'aquesta anàlisi consisteix a mostrar "validesa convergent", és a dir, el fet que l'*h-index* es correlaciona bé amb altres estadístiques de publicacions o de cites, com ara el nombre d'articles publicats o el nombre total de citacions. Aquesta correlació no té res de remarcable perquè totes aquestes variables són funcions del mateix fenomen bàsic: les publicacions. En un article notable sobre *l'h-index* [Lehmann - Jackson - Lautrup 2006] els autors duen a terme una anàlisi més curosa i demostren que *l'h-index* (de fet, *l'm*-index) no és tan "bo" com el fet de simplement considerar el nombre de citacions per article. Fins i tot aquí, no obstant, els autors no defineixen adequadament què significa el terme "bo". Quan s'aplica el paradigma clàssic estadístic [Lehmann - Jackson - Lautrup 2006], es comprova que l'*h-index* és menys fiable que altres mesures.

Un conjunt de variants de *l'h-index* han estat inventades per comparar la qualitat d'investigadors no només dins una disciplina sinó també entre diverses disciplines ([Batista - Campiteli - Kinouchi 2006], [Molinari - Molinari 2008]). Altres afirmen que l'*h-index* pot ser utilitzat per comparar instituts i departaments [Kinney 2007]. Són intents sovint exageradament ingenus per reduir un complex historial de citacions a una sola xifra: efectivament, l'avantatge principal d'aquests nous índexs sobre els simples histogrames de còmputs de citacions és el fet que els índexs descarten gairebé tots els detalls dels historials de citacions, i això fa que sigui possible comparar dos científics qualssevol. De tota manera, fins i tot exemples senzills mostren que la informació descartada és necessària per entendre un historial de recerca. Sens dubte **la comprensió hauria de**

ser l'objectiu en l'avaluació de la recerca, no simplement garantir que dos individus són comparables.

En alguns casos, els cossos d'avaluació nacionals recullen l'*h-index* o alguna de les seves variants com a part de la informació que tenen en compte. Aquest és un mal ús de les dades. Malauradament, tenir una sola xifra per classificar cada científic és una idea seductora – idea que es pot estendre més àmpliament a un públic que sovint no comprèn l'ús adequat del raonament estadístic en entorns molt més simples.

El significat de les citacions

Aquells que promouen estadístiques de citacions com a mesura predominant per a la qualitat en la recerca no responen a la pregunta principal: què signifiquen les citacions? Apleguen grans quantitats de dades sobre còmputs de citacions, processen les dades per treure'n estadístiques, i llavors asseguren que el procés d'avaluació resultant és "objectiu". Però és la *interpretació* de les estadístiques el que condueix cap a l'avaluació, i la interpretació es basa en el *significat* de les citacions, que és considerablement subjectiu.

A la literatura que recolza aquesta aproximació, és sorprenentment difícil trobar afirmacions clares sobre el significat de les citacions.

> "El concepte que hi ha darrere de la indexació de cites és essencialment simple. Si reconeixem que el valor de la informació es determina per aquells que la utilitzen, quina millor manera de mesurar la qualitat d'un treball que calculant l'impacte que té sobre la comunitat en general. La població més àmplia possible dins la comunitat erudita (per exemple qualsevol que utilitza o cita el material de la font) determina la influència o impacte de la idea i de qui l'ha originada en el nostre corpus de coneixement." [THOMSON: HISTORY]
>
> "Tot i que quantificar la qualitat de científics individuals és difícil, la visió general és que és millor publicar molt que publicar poc i que el còmput de citacions d'un article (en comparació amb els hàbits de citacions dins el camp) és una mesura útil de la seva qualitat." [Lehman - Jackson - Lautrup 2006, p. 1003]
>
> "Les citacions sovint reflecteixen el valor d'una revista i l'ús que se'n fa..." [Garfield 1972, p. 535]
>
> "Quan un físic o un investigador biomèdic cita un article d'una revista, està indicant que l'article citat l'ha influenciat en una o altra manera." [Garfield 1987, p. 7]
>
> "Les citacions són un reconeixement de deute intel·lectual." [THOMSON: FIFTY YEARS]

Els termes rellevants són "qualitat", "vàlua", "influència" i "deute intel·lectual". El terme "impacte" ha esdevingut la paraula genèrica per assignar significat a les citacions – terme que va aparèixer primer en un article curt escrit el 1955 per Eugene Garfield per promoure la idea de crear un índex de cites. Garfield va escriure:

> "Per tant, en el cas d'un article molt significatiu, l'índex de cites té un valor quantitatiu, perquè pot ajudar l'historiador a mesurar la influència de l'article – és a dir, el seu *factor d'impacte*". [Garfield 1955, p. 3]

Queda clar que aquí, com a tot arreu, es pretén que el terme "factor d'impacte" suggereixi que l'article que cita ha "estat construït sobre" el treball citat – és a dir, que les citacions són el mecanisme a través del qual la recerca avança.

Hi ha força literatura sobre el significat real de les citacions que suggereix que les citacions són més complicades del que aquestes afirmacions vagues ens porten a pensar. Per exemple, en el seu article de 1983 sobre l'avaluació de la recerca, Martin i Irvine van escriure:

> "Un fet subjacent a tots aquests problemes amb l'ús de citacions com a mesura de qualitat és la nostra ignorància de les raons per les quals els autors citen uns treballs en particular i no uns altres. Els problemes descrits més amunt [...]. Una anàlisi simple de les citacions pressuposa un model altament racional de cerca de referències, en el qual les citacions serveixen essencialment per reflectir l'apreciació científica de treballs anteriors d'alta qualitat o importància, i els citadors potencials tenen tots les mateixes possibilitats de citar treballs concrets..." [Martin- Irvine 1983, p. 69]

En el seu article de 1988 sobre el significat de les citacions [Cozzens 1989], Cozzens afirma que les citacions són el resultat de dos sistemes subjacents en la gestió de les publicacions científiques: el sistema de "recompensa" i el sistema "retòric". El primer tipus associa en la majoria dels casos el significat d'una citació al reconeixement per part de l'article que cita d'un "deute intel·lectual" amb l'article citat. El segon, en canvi, té un significat bastant diferent: la referència a un article anterior que explica algun resultat que potser no és en absolut un resultat de l'autor citat. Aquestes citacions retòriques són simplement una manera de dur a terme converses científiques, i no d'establir deutes intel·lectuals. Naturalment, en alguns casos una citació pot tenir tots dos significats.

Cozzens observa que la majoria de les citacions són retòriques. Això ho confirma l'experiència de la majoria de matemàtics (a la base de dades de citacions *Math Reviews*, per exemple, gairebé el 30% dels més de 3 milions de citacions citen llibres i *no* articles de recerca en revistes). Per què és important aquest fet? Perquè al contrari que les citacions de "recompensa", que acostumen a referir-se a articles fonamentals, la tria de quin article citar retòricament depèn de molts factors: el prestigi de l'autor citat (l'efecte-halo), la relació dels autors que citen amb els autors citats, la disponibilitat de la revista (potser les revistes d'accés lliure són més susceptibles de ser citades?), la conveniència de fer referències a diversos resultats d'un sol article, etc. Pocs d'aquests factors estan directament relacionats amb la "qualitat" de l'article citat.

Fins i tot en els casos en què les citacions són de "recompensa", poden reflectir diversitat de motius, com són "actualitat, reconeixement negatiu, informació operativa, persuasió, reconeixement positiu, alerta als lectors i consens social". [Brooks 1996] En la majoria dels casos, les citacions van responen a més d'un d'aquests motius. Alguns resultats notables poden patir l'efecte de l'obliteració quan són immediatament incorporats en els treballs d'altres, que a partir de llavors serveixen de base per a futures citacions. Altres citacions no són recompenses per

investigacions destacades, sinó més aviat advertències sobre resultats o raonaments erronis. Aquest informe presenta molts exemples d'aquestes citacions "d'advertència".

La sociologia de les citacions és un tema complex, de fet és un tema que queda més enllà d'aquest informe, però fins i tot aquesta argumentació superficial mostra que el significat de les citacions no és simple i que les estadístiques basades en citacions no són ni de lluny tan "objectives" com els seus partidaris asseguren.

Algú podria argumentar que el significat de les citacions és irrellevant perquè les estadístiques basades en citacions estan altament correlacionades amb altres mesures de qualitat de la recerca (com la revisió d'experts). Per exemple, l'informe *Evidence* mencionat anteriorment afirma que les estadístiques de cites poden (i haurien de) substituir altres formes d'avaluació a causa de la següent correlació:

> "Les proves han demostrat que les tècniques bibliomètriques poden crear indicadors de qualitat en la recerca que són congruents amb la percepció dels investigadors." [Evidence Report 2007, p. 9]

Sembla que la conclusió hagi de ser que les estadístiques basades en citacions, independentment del seu significat precís, haurien de substituir altres mètodes d'avaluació, perquè sovint hi coincideixen. A part de la circularitat d'aquest argument, la fal·làcia d'aquest raonament és fàcil de veure.

Utilitzar les estadístiques sàviament

L'entusiasta excés de confiança en les estadístiques per avaluar recerca no és un fenomen nou ni isolat. És descrit de forma eloqüent l'any 2001 en el conegut llibre *Damned lies and statistics*, escrit pel sociòleg Joel Best:

> "Hi ha cultures en les quals la gent creu que determinats objectes tenen poders màgics; els antropòlegs anomenen aquests objectes fetitxes. En la nostra societat, les estadístiques són una mena de fetitxe. Tendim a considerar les estadístiques com si fossin màgiques, com si fossin més que simples xifres. Les tractem com a representacions poderoses de la veritat; actuem com si destil·lessin la complexitat i la confusió de la realitat en simples fets. Fem servir estadístiques per convertir problemes socials complicats en estimacions, percentatges i índexs més fàcils d'entendre. Les estadístiques dirigeixen les nostres preocupacions; ens mostren de què ens hauríem de preocupar i quant ens hauríem de preocupar. En certa manera, el problema estadístic es converteix en estadística i, perquè tractem les estadístiques com a vertaderes i incontrovertibles, adquireixen una mena de control màgic i fetitxista sobre com veiem els problemes socials. Pensem en les estadístiques com en fets que descobrim, no com en xifres que creem." [Best 2001, p 160]

Aquesta creença mística en la màgia de les estadístiques de citacions es pot trobar en tota la documentació per a l'exercici de l'avaluació de la recerca, tant a nivell nacional com a nivell institucional. També es troba en els treballs dels qui promouen l'*h-index* i les seves variants.

Aquesta actitud també és evident en els intents recents de millorar el factor d'impacte, tot utilitzant algoritmes matemàtics més sofisticats, com els algoritmes *PageRank,* per analitzar les citacions ([Bergstrom 2007], [Stringer - Sales - Pardo - Nunes 2008]). Els defensors d'aquest canvi reivindiquen la seva eficàcia de manera injustificada pel que fa a l'anàlisi i difícil d'avaluar. Degut a que estan basades en càlculs més complicats, les suposicions (sovint amagades) que hi ha al darrere no són fàcils de discernir per a la majoria de la gent[7]. La intenció és que tractem les xifres amb reverència – com a veritats més que no pas com a creacions.

La recerca no és la primera activitat finançada públicament que és sotmesa a examen, i durant les últimes dècades s'ha intentat dur a terme actuacions quantitatives d'avaluació de temes que van des dels sistemes educatius (escoles) fins a la sanitat (hospitals i fins i tot cirurgians individuals). En alguns casos, els estadístics han intervingut per aconsellar els avaluadors sobre quantificacions raonables i sobre l'ús correcte de les estadístiques. **Si cal consultar doctors quan es practica la medecina, segurament caldrà consultar els estadístics (i atendre els seus consells) quan es practica l'estadística**. Es poden trobar dos exemples excel·lents en [Bird 2005] i [Goldstein - Spiegelhalter 1996]. Tot i que els dos tracten l'actuació de l'avaluació en altres àmbits que el de la recerca (el primer en el seguiment del sector públic i el segon en sanitat i educació) cadascun d'ells aporta perspicàcia sobre l'ús sensat de les estadístiques per avaluar la recerca.

L'article de Goldstein i Spiegelhalter, en concret, tracta l'ús dels quadres comparatius (*rànquings*) basats en simples xifres (per exemple qualificacions d'estudiants o resultats mèdics), i és especialment rellevant per a l'avaluació de la recerca mitjançant la classificació de revistes, articles o autors basada en estadístiques de cites. En el seu article, els autors estableixen un marc de tres parts per a qualsevol actuació d'avaluació:

Dades:

> "Cap mena de tripijoc estadístic pot vèncer les insuficiències bàsiques relatives tant a la *propietat* com a la *integritat* de les dades recollides." [Goldstein - Spiegelhalter 1996, p. 389]

Aquesta és una observació important per a l'avaluació basada en citacions. El factor d'impacte, per exemple, està basat en un subconjunt de dades que només inclou les revistes seleccionades per *Thomson Scientific* (Cal remarcar que el factor d'impacte és el criteri de selecció més important). Hi ha qui ha qüestionat la integritat d'aquestes dades [Rossner - VanEpps - Hill 2007]; altres assenyalen que altres grups de dades podrien ser més complets [Meho - Yang 2007]; diversos grups han proposat fer servir el *Google Scholar* per implementar estadístiques basades en citacions, com per exemple l'*h-index*, però les dades contingudes en el *Google Scholar* sovint no són exactes (ja que dades com els noms dels autors són extretes automàticament d'enviaments a través de la web). Les estadístiques de citacions de científics individuals de vegades són difícils d'obtenir perquè els autors no estan identificats d'una sola manera i en alguns entorns i certs països això pot ser un impediment enorme per recopilar dades exactes sobre citacions. L'específica compilació de dades que es fa servir per a l'anàlisi de les citacions és freqüentment passada per alt, i per tant és probable que s'extreguin conclusions imperfectes d'estadístiques basades en dades imperfectes.

Anàlisi i presentació de les estadístiques:

> "Posarem especial atenció a l'especificació d'un *model* estadístic apropiat, a la importància crucial de la *incertesa* en la presentació de tots els resultats, a les tècniques d'*ajust* de resultats per als factors confusionaris i finalment al grau de confiança que es pot posar en els *rànquings* explícits." [Goldstein - Spiegelhalter 1996, p. 390]

Com hem escrit prèviament, a la majoria de casos en què les estadístiques de citacions són utilitzades per classificar articles, persones i programes, no s'especifica cap model amb antelació. Enlloc d'això, les pròpies dades suggereixen un model, que sovint és vague. Sembla que un procés circular classifiqui els objectes més amunt perquè estan classificats més amunt (en la base de dades). Freqüentment es para poca atenció a la *incertesa* de *qualsevol* d'aquests rànquings, i s'analitza poc com aquesta incertesa (per exemple, variacions anuals en el factor d'impacte) afectaria els rànquings. Finalment, els factors confusionaris (per exemple, la disciplina en particular, el tipus d'articles que publica una revista, si un científic particular és experimentalista o teòric) són ignorats amb freqüència, especialment en els rànquings fets en actuacions nacionals d'avaluació.

Interpretació i impacte:

> "Les comparacions tractades en aquest article són de gran interès general, i es tracta d'una àrea on una atenció meticulosa a les limitacions és alhora essencial i susceptible de ser ignorada. El fet que resultats ajustats puguin ser d'alguna manera mesures vàlides de "qualitat" institucional és una qüestió, però els analistes també haurien de ser conscients de l'efecte potencial dels resultats pel que fa a futurs canvis de comportament per part de les institucions i els individus buscant millorar el seu "rànquing" subsegüent." [Goldstein - Spiegelhalter 1996, p. 390]

L'avaluació de la recerca és *també* una qüestió d'interès general. Per a un científic individual, una avaluació pot tenir efectes profunds i a llarg termini sobre la seva carrera; per a un departament, pot canviar les perspectives d'èxit per al futur; per a les disciplines, un conjunt d'avaluacions pot marcar la diferència entre la prosperitat o l'esllanguiment. En una tasca tan important, no cal dir que s'hauria d'entendre tant la validesa com les limitacions de les eines utilitzades per dur-la a terme. Fins a quin punt les citacions mesuren la qualitat en la recerca? Sembla que els còmputs de citacions estiguin correlacionats amb la qualitat, i existeix una noció intuïtiva segons la qual els articles d'alta qualitat són altament citats. Però tal i com s'ha explicat més amunt, alguns articles, especialment en algunes disciplines, són citats per altres motius que la qualitat, i per tant els articles altament citats no tenen perquè ser d'alta qualitat. Cal una millor comprensió de la interpretació exacta dels rànquings basats en estadístiques. De més a més, si les estadístiques de citacions juguen un paper principal en l'avaluació de la recerca, és evident que els autors i fins i tot els editors trobaran maneres de manipular el sistema a favor seu [Macdonald - Kam 2007]. Les implicacions d'aquest fenomen a llarg termini no són clares ni han estat estudiades.

Val la pena llegir avui l'article de Goldstein i Spiegelhalter perquè deixa clar que l'excés de confiança en simples estadístiques en l'avaluació de la recerca no és un problema aïllat. Els governs, les institucions i els individus han lluitat amb problemes similars en el passat en altres contextos, i han trobat maneres de comprendre millor les eines estadístiques i augmentar-les amb altres mitjans d'avaluació. Goldstein i Spiegelhalter acaben el seu article amb una declaració d'esperança:

"Finalment, tot i que hem estat crítics en general amb moltes temptatives actuals per proporcionar judicis sobre les institucions, no desitgem donar la impressió que creiem que totes aquestes comparacions són necessàriament errònies. Pensem que la comparació d'institucions i l'intent d'entendre perquè les institucions difereixen les unes de les altres és una activitat extremadament important que es portarà millor a terme amb un esperit de col·laboració que de confrontació. Potser és l'únic mètode segur per obtenir informació objectiva que pugui portar a la comprensió i finalment resultar en millores. *El problema real amb els procediments simplistes que hem criticat és que distreuen l'atenció i els recursos d'aquest gran objectiu."* [Goldstein - Spiegelhalter 1996, p. 406]

Seria difícil trobar una frase més encertada per expressar els objectius que haurien de ser compartits per tothom involucrat en l'avaluació de la recerca.

Comitè conjunt IMU/ICIAM/IM sobre l'Avaluació Quantitativa de la Recerca

Robert Adler, *Technion–Israel Institute of Technology*
John Ewing (President), *American Mathematical Society*
Peter Taylor, *University of Melbourne*

Referències

Adler, Robert. 2007. The impact of impact factors. *IMS Bulletin*, Vol. 36, No. 5, p. 4.
http://bulletin.imstat.org/pdf/36/5

Amin, M.; Mabe, M. 2000. Impact factor: use and abuse. *Perspectives in Publishing*, No. 1, October, pp. 1-6.
http://www.elsevier.com/framework_editors/pdfs/Perspectives1.pdf

Batista, Pablo Diniz; Campiteli,Monica Guimaraes; Kinouchi, Osame; Martinez, Alexandre Souto. 2005. Universal behavior of a research productivity index. arXiv: physics, v1, pp. 1-5.

arXiv:physics/0510142v1

Batista, Pablo Diniz; Campiteli,Monica Guimaraes; Kinouchi, Osame; . 2006. Is it possible to compare researchers with different scientific interests?. *Scientometrics*, Vol 68, No 1, pp. 179-189.
http://dx.doi.org/10.1007/s11192-006-0090-4

Bergstrom, Carl. Eigenfactor: measuring the value and presitige of scholarly journals. College & Research Libraries News, Vol 68, No. 5, May 2007
http://www.ala.org/ala/acrl/acrlpubs/crlnews/backissues2007/may07/eigenfactor.cfm
(See also **http://www.eigenfactor.org/methods.pdf** .)

Best, Joel. 2001. Damned lies and statistics: untangling the numbers from the media, politicians, and activists. University of California Press, Berkeley.

Bird, Sheila; et al. 2005. Performance indicators: good, bad, and ugly; Report of a working party on performance monitoring in the public services. *J.R.Statist. Soc.* A (2005), 168, Part 1, pp. 1-27.
http://dx.doi.org/10.1111/j.1467-985X.2004.00333.x

Brooks, Terrence. 1986. Evidence of complex citer motivations. *Journal of the American Society for Information Science*, Vol 37, No. 1, pp. 34-36, 1986.
http://dx.doi.org/10.1002/asi.4630370106

Carey, Alan L.; Cowling, Michael G.; Taylor, Peter G. 2007. Assessing research in the mathematical sciences. *Gazette of the Australian Math Society*, A.L. Carey, Vol. 34, No. 2, May, pp. 84-89.
http://www.austms.org.au/Publ/Gazette/2007/May07/084CommsCarey.pdf

Cozzens, Susan E. 1989. What do citations count? The rhetoric-first model. *Scientometrics*, Vol 15, Nos 5-6, (1989), pp. 437-447.
http://dx.doi.org/10.1007/BF02017064

Egghe, Leo. 2006. Theory and practice of the g-index. *Scientometrics*, vol. 69, No 1, pp. 131-152.

http://dx.doi.org/10.1007/s11192-006-0144-7

Evidence Report. 2007. The use of bibliometrics to measure research quality in the UK higher education system. (A report produced for the Research Policy Committee of Universities, UK, by Evidence Ltd., a
June 2008 Citation Statistics IMU-ICIAM-IMS 21
company specializing in research performance analysis and interpretation. Evidence Ltd. has "strategic alliance" with Thomson Scientific.)
http://bookshop.universitiesuk.ac.uk/downloads/bibliometrics.pdf

Ewing, John. 2006. Measuring journals. *Notices of the AMS*, vol. 53, no. 9, pp. 1049-1053.
http://www.ams.org/notices/200609/comm-ewing.pdf

Garfield, Eugene. 1955. Citation indexes for science: A new dimension in documentation through association of ideas." *Science*, 122(3159), p.108-11, July 1955.
http://garfield.library.upenn.edu/papers/science1955.pdf

_____. 1972. Citation analysis as a tool in journal evaluation. *Science*, 178 (4060), pp. 471-479, 1972.
http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf

_____. 1987. Why are the impacts of the leading medical journals so similar and yet so different? *Current Comments* #2, p. 3, January 12, 1987.
http://www.garfield.library.upenn.edu/essays/v10p007y1987.pdf

_____. 1998. Long-term vs. short-term journal impact (part II). *The Scientist* 12(14):12-3 (July 6, 1998).
http://garfield.library.upenn.edu/commentaries/tsv12(14)p12y19980706.pdf

_____. 2005. Agony and the ecstasy—the history and meaning of the journal impact factor. Presented at the *International Congress on Peer Review and Bibliomedical Publication*, Chicago, September 16, 2005.
http://garfield.library.upenn.edu/papers/jifchicago2005.pdf

Goldstein, Harvey; Spiegelhalter, David J. 1996. League tables and their limitations: Statistical issues in comparisons of institutional performance. *J. R. Statist. Soc. A*, 159, No. 3. (1996), pp 385-443.
http://links.jstor.org/sici?sici=0964-1998%281996%29159%3A3%3C385%3ALTATLS%3E2.0.CO%3B2-5
http://dx.doi.org/10.2307/2983325

Hall, Peter. 2007. Measuring research performance in the mathematical sciences in Australian universities. *The Australian Mathematical Society Gazette*, Vol. 34, No. 1, pp. 26-30.
http://www.austms.org.au/Publ/Gazette/2007/Mar07/26HallMeasuring.pdf

Hirsch, J. E. 2006. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA*, Vol. 102, No. 46, pp. 16569–16573.

http://dx.doi.org/10.1073/pnas.0507655102

Kinney, A. L. 2007. National scientific facilities and their science impact on nonbiomedical research. *Proc Natl Acad Sci USA*, Vol. 104, No. 46, pp. 17943-17947.
http://dx.doi.org/10.1073/pnas.0704416104

Lehmann, Sune; Jackson, Andrew D.; Lautrup, Benny E. 2006. Measures for measures, *Nature*, Vol 444, No. 21, pp. 1003-1004.
http://www.nature.com/nature/journal/v444/n7122/full/4441003a.html

June 2008 Citation Statistics IMU-ICIAM-IMS 22
Macdonald, Stuart; Kam, Jacqueline. 2007. Aardvark et al.: quality journals and gamesmanship in management studies. *Journal of Information Science*, Vol. 33, pp. 702-717.
http://dx.doi.org/10.1177/0165551507077419

Martin, Ben R. 1996. The use of multiple indicators in the assessment of basic research, *Scientometrics*, Vol 36, No. 3 (1996), pp. 343-362.
http://dx.doi.org/10.1007 /BF02129599

Martin, Ben R., Irvine, John. 1983. Assessing basic research. *Research Policy*, Vol 12 (1983), pp. 61-90.
http://dx.doi.org/10.1016/0048-7333(83)90005-7

Meho, Lokman; Yang, Kiduk. 2007. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, Vol 58, No 13, pp. 2105-2125.
http://dx.doi.org/10.1002/asi.20677

Molinari, J. F., Molinari, A. 2008. A new methodology for ranking scientific institutions. To appear in *Scientometrics*.
http://imechanica.org/files/paper.pdf

Monastersky, R. 2005. The number that's devouring science. *Chronicle Higher Ed.* Vol. 52, No. 8.
http://chronicle.com/free/v52/i08/08a01201.htm

Rossner, Mike; Van Epps, Heather; Hill, Emma. 2007. Show me the data. *Journal of Cell Biology*, Vol 179, No 6, December 17, pp. 1091-1092.
http://dx.doi.org/10.1083/jcb.200711140

Seglen, P. O. 1997. Why the impact factor for journals should not be used for evaluating research; *BMJ*, 314:497 (15 February).
http://www.bmj.com/cgi/content/full/314/7079/497

Sidiropoulos, Antonis; Katsaros, Dimitrios; Manolopoulos, Yannis. 2006. Generalized h-index for disclosing latent facts in citation networks. V1, arXiv:cs.
arXiv:cs/0607066v1 [cs.DL]

Stringer MJ, Sales-Pardo M, Nunes Amaral LA (2008) Effectiveness of journal ranking schemes as a tool for locating information. PLoS ONE 3(2): e1683
http://dx.doi.org/10.1371/journal.pone.0001683

THOMSON: JOURNAL CITATION REPORTS. 2007. (Thomson Scientific website)
http://scientific.thomson.com/products/jcr/

THOMSON: SELECTION. 2007. (Thomson Scientific website)
http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/

THOMSON: IMPACT FACTOR (Thomson Scientific website)
http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/

June 2008 Citation Statistics IMU-ICIAM-IMS 23
THOMSON: HISTORY (Thomson Scientific website)
http://scientific.thomson.com/free/essays/citationindexing/history/

THOMSON: FIFTY YEARS (Thomson Scientific website)
http://scientific.thomson.com/free/essays/citationindexing/50y-citationindexing/

Notes

1 Aquesta citació va ser atribuida a Einstein al *Reader's Digest*. Oct. 1977. Sembla ser una derivació de la seva citació real: "No es pot negar que l'objectiu suprem de tota teoria és fer els elements bàsics irreductibles tan simples i d'un nombre tan limitat com sigui possible sense haver de renunciar a la suficient representació d'una sola dada d'experiència", de "On the Method of Theoretical Physics" The Herbert Spencer Lecture, pronunciada a Oxford (10 juny 1933); també publicada a *Philosophy of Science*, Vol. 1, No. 2 (abril 1934), pp. 163-169.

2 Tot i que ens concentrem en el factor d'impacte de *Thomson Scientific* en aquesta secció, observem que *Thomson* promou l'ús de dues altres estadístiques. A més a més, estadístiques similars basades en còmputs de mitjanes de citacions en revistes poden ser derivades d'altres bases de dades, tals com Scopus, Spires, Google Scholar, i (per a les matemàtiques) la base de dades de citacions *Math Reviews*. Aquesta última consisteix en citacions d'unes 400 revistes matemàtiques des de l'any 2000 fins a l'actualitat, identificades com a ítems que han estat llistats al Math Reviews des de 1940; inclou més de 3 milions de citacions.

3 *Thomson Scientific* indica (març 2008) que indexa revistes de les següents categories:

Matemàtiques (217)
Matemàtiques aplicades (177)
Matemàtiques interdisciplinàries (76)
Física, matemàtica (44)
Probabilitat i estadística (96)

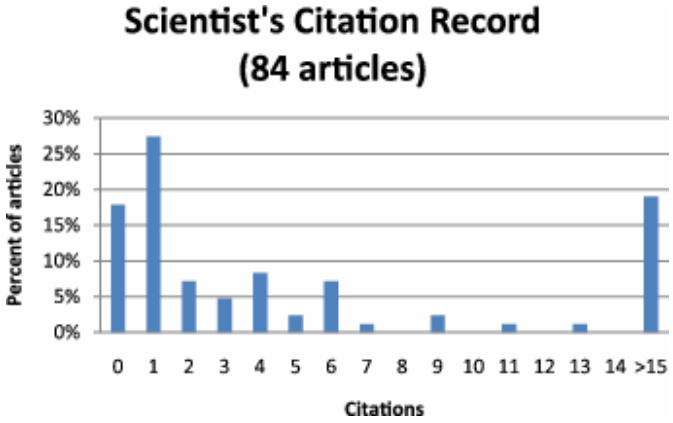
Les categories se superposen, i el nombre total de revistes és aproximadament 400. En canvi, *Mathematical Reviews* inclou ítems de més de 1.200 revistes cada any, i considera més de 800 revistes com a "centrals" (en el sentit que cada ítem en la revista és inclòs al Math Reviews). Zentralblatt cobreix un nombre similar de revistes matemàtiques.

4 La base de dades de citacions *Mathematical Reviews* inclou (març 2008) més de 3 milions de referències d'aproximadament 400 revistes publicades des del 2000 fins al present. Les referències es corresponen a ítems de la base de dades de MR i engloben moltes dècades. En contrast amb el Science Citation Index, s'hi inclouen citacions a revistes i a llibres. És un fet curiós que aproximadament el 50% de les citacions es refereixen a ítems apareguts a la dècada anterior; 25% citen articles apareguts a la dècada prèvia a aquesta; 12,5 % citen articles de la dècada anterior a aquella; etc. Aquest tipus de comportament depèn de cada disciplina, és clar.

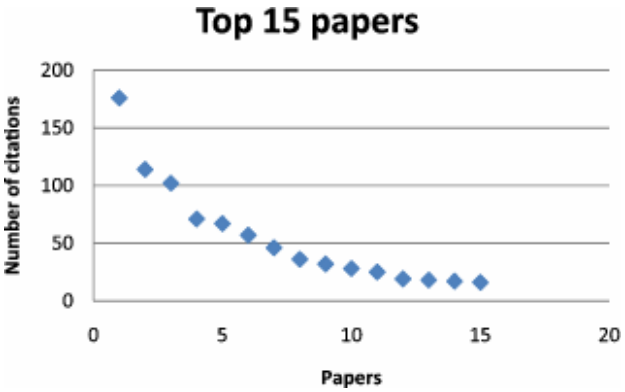
5 La mala distribució combinada amb el curt període (utilitzant només les revistes d'un any com a font de citacions i cinc anys com a objectiu) fa que un gran nombre d'articles tingui molt poques o cap citació. Això fa intuïtivament obvi el fet que articles escollits de manera aleatòria sovint siguin equivalents.

El fet que molts articles no tinguin citacions (o en tinguin molt poques) també és una conseqüència del llarg període de fer citacions en matemàtiques – els articles sovint triguen molts anys a acumular citacions. Si triem períodes de temps més llargs per les revistes font de les citacions i pels articles citats, els còmputs de citacions creixen de manera substancial i esdevé més fàcil distingir les revistes segons el comportament de les citacions. Aquesta és l'aproximació utilitzada a [Stringer- et al. 2008] per analitzar citacions, tot mostrant que, per períodes de temps suficientment llargs, la distribució dels còmputs de citacions semblen ser *log-normal*. Això proporciona un mecanisme per comparar les distribucions, i certament és més sofisticat que el fet d'utilitzar factors d'impacte. De tota manera, un altre cop considera únicament les citacions.

6 Per il·lustrar quanta informació es perd quan només es fa servir l'*h-índex*, aquí tenim un exemple de la vida real d'un distingit matemàtic cap a la meitat de la seva carrera que ha publicat 84 articles de recerca. La distribució de les citacions té el següent aspecte:



Observeu que lleugerament per sota del 20% de les publicacions tenen 15 o més citacions. La distribució de citacions reals per aquests 15 articles és:



El l'anàlisi de Hirsch, de tota manera, tota aquesta informació no és tinguda en compte. Només es té en consideració el fet que l'*h-index* és 15, que significa que els 15 primers articles tenen 15 o més citacions.

[7]L'algoritme a [Bergstrom 2007] fa servir un algoritme *pagerank* per donar un pes a cada citació, i llavors calcula un "factor d'impacte" utilitzant les mitjanes de "pes" de les citacions. Els algoritmes *pagerank* tenen mèrit perquè tenen en compte "el valor" de les citacions. D'altra banda, la seva complexitat pot ser perillosa perquè els resultats finals són més difícils d'entendre. En aquest cas, totes les "autocitacions" són descartades (és a dir, les citacions d'articles d'una revista J a artides publicats en J durant els 5 anys anteriors són descartats). Aquestes no són "autocitacions" en el sentit normal del mot, i un cop d'ull a la base de dades de citacions Math Reviews suggereix que això descarta aproximadament un terç de totes les citacions.

L'algoritme a [Stringer-et al. 2008] és interessant, en part perquè intenta tractar els diferents períodes de temps per a les citacions i la comparació d'articles seleccionats de manera aleatòria en una revista amb els d'una altra publicació. De nou, la complexitat dels algoritmes dificulta a la majoria de la gent l'avaluació dels seus resultats. Una hipòtesi notable apareix a l'article a la pàgina 2: "La nostra primera assumpció és que els articles publicats a la revista J tenen una distribució normal de "qualitat"..." Això sembla contradir l'experiència comú.