VQR: la bibliometria fai-da-te dell'ANVUR

2012-01-24 08:14:14 By Giuseppe De Nicolao

Lo scopo dei due articoli che seguono è analizzare alcuni potenziali errori metodologici in cui i GEV (Gruppi di Esperti della Valutazione) potrebbero incorrere nella definizione dei criteri da utilizzare nella valutazione dei prodotti della ricerca nell'ambito dell'esercizio di Valutazione della Qualità della Ricerca 2004-2010 (VQR). Come riportato nel bando del VQR, entro il 31 gennaio 2012, i GEV dovranno definire i criteri a cui si atterranno nella successiva valutazione dei prodotti della ricerca.

Il compito di definire i criteri per ogni Area Disciplinare è assai complesso e decisivo ai fini dello svolgimento del successivo processo di valutazione. La decisione da parte del Consiglio Direttivo dell'ANVUR di affidare ai GEV non solo i criteri di valutazione dei prodotti, ma anche lo sviluppo di strumenti bibliometrici e la stesura di classifiche di riviste appare del tutto singolare e senza paragoni a livello internazionale. Infatti, non solo i GEV non appaiono qualificati ad affrontare il compito, ma il tempo messo a loro disposizione (meno di due mesi) sarebbe stato del tutto insufficiente anche per dei comitati di esperti ben più agguerriti.

Anche la scelta di fare svolgere i lavori dei GEV in un clima di chiusura rispetto alla comunità scientifica non trova riscontro negli esercizi di valutazione condotti a livello internazionale. In ogni caso, le indiscrezioni trapelate destano preoccupazione, in quanto sembra profilarsi l'adozione di criteri indifendibili sul piano scientifico o addirittura errati, che non trovano alcun riscontro nella letteratura bibliometrica internazionale. In particolare, ci sono tre aspetti che meritano un'attenta riflessione:

a) La non affidabilità delle classifiche di riviste messe a punto da società scientifiche e gruppi disciplinari rispetto al ricorso alle analisi citazionali.

b) L'inconsistenza metodologica di una valutazione a più stadi degli articoli, la quale attribuisce una

classe di merito provvisoria in base alla classificazione della rivista per poi correggerla verso l'alto o verso il basso con le citazioni normalizzate in base alle citazioni degli articoli apparsi sulla stessa rivista nello stesso anno. Quando sono disponibili le citazioni dei singoli articoli, il ricorso alle classifiche delle riviste risulta superfluo e persino dannoso.

c) L'erroneità delle classifiche delle riviste basate sulla media dei loro ranks relativi a un ventaglio di indicatori bibliometrici. La classificazione basata sulla media aritmetica dei ranks è un noto errore, ampiamente stigmatizzato da decenni di letteratura scientifica.

Lo scopo dei nostri due articoli è presentare l'evidenza scientifica che rende improponibili queste tre scelte di metodo, nella speranza che i GEV ne tengano conto ed evitino di compromettere l'affidabilità dell'esercizio di valutazione tramite il ricorso a criteri ed algoritmi privi di validità scientifica.

Il primo articolo di Alberto Baccini riprende il tema dell'uso delle classifiche delle riviste e spiega l'inconsistenza della valutazione a due stadi.

Il secondo articolo di Giuseppe De Nicolao, formulato sotto forma di breve racconto, tratta il terzo punto, spiegando in modo accessibile perchè la media dei ranks da decenni non trova spazio nel trattamento scientifico dei dati.

Misurare nani e giganti

di Alberto Baccini

Il disegno del VQR italiano pone problemi inediti in riferimento alle procedure di valutazione adottate a livello internazionale. Come nel RAE/REF britannico, e nel precedente esercizio di valutazione italiano (CIVR), l'unità elementare su cui viene esercitata la valutazione è il singolo prodotto di ricerca (articolo su rivista, capitolo di libro, libro etc.); come nel RAE/REF e nel precedente CIVR si prevede che ogni prodotto di ricerca venga attribuito ad una classe di merito. Nel REF britannico è previsto l'uso esclusivo della peer review per l'attribuzione dei prodotti nelle classi di merito.[1] Ciò che differenzia il VQR dalle altre esperienze è che l'attribuzione alla classe di merito può avvenire in tre diversi modi:

1. peer review del prodotto di ricerca;
2. uso automatico di qualche indicatore bibliometrico;
3. un mix tra peer review e bibliometria.

Sui pericoli connessi all'uso della peer review nella comunità accademica italiana e sui possibili rimedi

siamo già intervenuti, indicando i rimedi possibili per ridurre le distorsioni. Qui vogliamo puntare l'attenzione sui problemi che potrebbero derivare dalla costruzione di algoritmi per l'attribuzione automatica dei prodotti alle classi di merito basandosi su dati bibliometrici.

Per rappresentare il problema può essere utile ricorrere ad una analogia. Ci sono quattro gruppi di individui caratterizzati da statura media diversa, per esempio maschietti di quattro classi consecutive di scuola primaria (prima-quarta); è nota la statura media dei bambini per fascia di età; siccome in ogni classe ci sono bambini di una sola fascia di età, la statura media per ogni fascia di età è un riferimento accettabile. Problema: misurare la statura di ogni bambino di ogni classe.

Strategia 1: si misura l'altezza di ogni bambino. Si può così calcolare il rapporto tra l'altezza di ogni bambino e quella media dei bambini della stessa età; e per ogni classe si può calcolare l'altezza media e lo scarto rispetto alla media di riferimento.

Strategia 2: ad ogni bambino di ogni classe viene attribuita la statura media dei bambini della fascia di età di appartenenza.

Strategia 3: E' una strategia a due stadi. Nel primo si attribuisce ad ogni bambino la statura media della fascia di età di appartenenza. Nel secondo si adottano alcune "correzioni": si misura anzitutto il bambino; se la sua statura è superiore alla statura media dei bambini di un anno più grandi, la sua altezza è quella media della classe superiore. Se la sua altezza è inferiore all'altezza del suo vicino di banco, allora la sua altezza è quella media dei bambini della classe inferiore. [La regola non prevede cosa accade se le due condizioni si verificano contemporaneamente, cioè se è più alto della media dei bambini più grandi, ma più piccolo del gigante suo vicino di banco].

Ci sono pochi dubbi che la Strategia 1 sia la più corretta per misurare la statura dei bambini.

La strategia 2 è approssimativa; per ogni bambino non dà informazioni affidabili, ma si può sperare che in media non distorca troppo. Ha l'indubbio vantaggio di evitare le procedure di misurazione di ogni singolo bambino. Quindi potrebbe essere adottata quando non sia possibile misurare ogni bambino o quando si voglia risparmiare sui costi di misurazione.

La strategia 3 introduce ulteriori distorsioni rispetto alla strategia 2, e soprattutto richiede comunque di misurare ogni bambino. Tra le tre è sicuramente quella che si allontana di più dal senso comune, quella meno efficiente e più costosa. Essa stabilisce che anche di fronte all'evidenza empirica, il gigante della classe dei più piccoli non può essere più alto della media dei bambini che hanno un anno di più. Stabilisce inoltre che l'altezza di ogni bambino dipende dall'altezza del vicino di banco. Per cui chi ha la sfortuna di avere come vicino di banco un bambino particolarmente alto, deve essere accorciato. I danni prodotti nella misurazione dalla strategia 3 sono ridotti se, come avviene per le stature, la distribuzione statistica è gaussiana ed i giganti, così come i nani, sono molto rari. Se i giganti ed i nani fossero molto diffusi nella popolazione, le situazioni paradossali si moltiplicherebbero.

Ma che c'entra tutto questo con il VQR?

Misurare l'altezza di gruppi di bambini non è poi così diverso dal misurare l'ìmpatto di un prodotto scientifico. L'ANVUR per classificare i prodotti di ricerca potrebbe decidere di ricorrere alla

Strategia 2: ogni articolo viene assegnato alla classe di merito della rivista in cui è pubblicato.

Questa strategia è semplice e poco costosa. Ha l'indubbio vantaggio di poter essere usata anche quando non si abbiano informazioni relative all'impatto specifico del singolo prodotto di ricerca, ma solo sulla qualità o "rilevanza" o utilità della rivista. Concediamo che l'ANVUR, nel caso decida di utilizzare le classifiche delle riviste, lo faccia a ragion veduta, abbia cioè ben valutato che valga la pena incorrere nei ben documentati effetti dell'adozione di queste procedure sui comportamenti della comunità scientifica (modifica dei temi di ricerca, modifica delle strategia di pubblicazione, penalizzazione delle ricerche interdisciplinari, curiosity driven, di nicchia, di interesse nazionale etc.).

Il problema diventa così eminentemente tecnico: chi stila le classifiche delle riviste e con quali tecniche. I modi per costruire le classifiche sono essenzialmente due: si ricorre a strumenti bibliometrici, o a consultazioni di esperti.[2]

Il ricorso a classifiche bibliometriche delle riviste, come quelle contenute nel Journal Citation Reports o SCIMAGO, è una pratica diffusa nell'accademia anche se pressoché inedita in un esercizio nazionale di valutazione. Già si è detto dell'abbandono di questa pratica in Australia. Può essere utile ricordare che nel prossimo REF britannico molti panel "will make use of citation data, where it is available, as an indicator of the academic impact of the outputs, to inform its assessment of output quality" (PANEL A: medicina e biologia), ma che nessun panel userà classifiche di riviste per la valutazione dei singoli prodotti di ricerca. Alcuni panel hanno addirittura reso noto che non intendono ricevere alcuna informazione bibliometrica riguardante le riviste dall'Agenzia.[3] In effetti non ha molto senso usare classifiche bibliometriche per classificare gli articoli, perché quando quelle sono disponibili, sono disponibili anche dati migliori (citazioni dei singoli articoli), e quindi non si capisce quale sia il vantaggio di usarle.

L'uso di classifiche delle riviste stilate attraverso consultazione di esperti diventa più interessante quando non esistono altre e migliori informazioni bibliometriche (e si ritenga di poter procedere con una elevata approssimazione). Per esempio, in Francia il panel di valutazione delle discipline economiche utilizza la classifica delle riviste messa a punto dal CNRS. La procedura è del tutto ragionevole: molti ricercatori delle aree economiche scrivono su riviste che non sono coperte nei database internazionali. Non ci sono quindi informazioni sull'impatto dei loro lavori. Una lista di riviste economiche giudicate di qualità dalla comunità scientifica permette di attribuire il bollino di qualità agli articoli che vi sono pubblicati. Sono ormai molti anni che viene svolta una consultazione pubblica, descritta qui, che ha prodotto diverse versioni della classifica.

In Italia, ROARS lo ha già discusso, non esiste una lista delle "riviste scientifiche", simile a quelle usate per esempio dalle agenzie di valutazione di Australia, Francia e Norvegia, e che possa essere utilizzata dall'ANVUR per distinguere il lavoro scientifico da quello non scientifico. Le uniche liste di riviste sono quelle messe a punto dalle società disciplinari per alcune aree delle scienze umane e sociali, con i problemi di affidablità che abbiamo già discusso qui. Allo stato attuale non sembra quindi che la

procedura sia applicabile facilmente.

Vogliamo sperare che i GEV non siano chiamati a stilare classifiche delle riviste. Si tratterebbe di una procedura inedita, almeno a conoscenza di chi scrive, nelle procedure di valutazione internazionali. Inoltre i GEV dovrebbero fare in un paio di mesi quello che altrove ha richiesto anni di lavoro a gruppi ben più numerosi (le liste australiane per esempio hanno richiesto due anni di lavoro). Si porrebbe inoltre un problema di credibilità dei risultati finali.

Un recente paper di Serenko e Dohan passa in rassegna 23 lavori dedicati al tema della consistenza tra classifiche stilate da esperti e classifiche bibliometriche concludendo che i risultati delle classifiche stilate da esperti "merely reflect their present research preferences rather than an objective assessment of each journal's quality" e che quindi "the final ranking closely corresponds to the research profile of the group of respondents". Far stilare classifiche così delicate a gruppi molto ristretti di ricercatori, come nel caso dei GEV, potrebbe condizionare i risultati finali dell'esercizio di valutazione complessivo.[4]

Strategia 3. UNA PROCEDURA A DUE STADI PER LA CLASSIFICAZIONE DEI PRODOTTI DI RICERCA

Si dice che alcuni GEV stiano procedendo all'adozione della Strategia 3. La procedura di classificazione dei prodotti consiste di due stadi: nel primo si distribuiscono le riviste in 4 classi di merito e si attribuisce ad ogni prodotto la classe di merito della rivista. Nel secondo stadio si corregge la classe di merito di una posizione verso l'alto se le citazioni ricevute dal singolo prodotto di ricerca sono superiori alle citazioni medie delle riviste della classe superiore; si corregge la classe di merito di una posizione verso il basso se le citazioni ricevute dal singolo prodotto di ricerca sono inferiori alle citazioni medie ricevute dagli articoli usciti nella stessa rivista nello stesso anno. [La Strategia 3 non prevede cosa accade se le due condizioni si verificano contemporaneamente, cioè se l'articolo è più citato della media degli articoli pubblicati dalle riviste di fascia superiore, ma meno citato rispetto alle citazioni medie ricevute dagli articoli usciti nella stessa rivista nell'anno in cui è stato pubblicato].

Indici citazionali che calcolano la media delle citazioni ricevute da una rivista (Impact factor, citazioni per articolo pubblicato, ma non l'h-index) sono molto sensibili ai valori estremi. Le distribuzioni con cui si ha a che fare in bibliometria non sono gaussiane, ma a code pesanti. Questo significa che è molto frequente la presenza di valori estremi (vicini di banco giganti) che modificano sensibilmente la media. Tra parentesi, è per questo che sarebbe opportuno usare con estrema cautela i valori medi in bibliometria. E questa è la ragione per cui non ha alcun senso confrontare le citazioni ricevute da un articolo con quelle medie ricevute dagli articoli della stessa rivista nello stesso anno. Se in quell'anno qualcuno piazza su quella rivista l'articolo più importante della carriera, o magari il peggiore della carriera, quello che raccoglie una valanga di citazioni negative, i valori medi annuali della rivista saranno molto elevati, e gli articoli che ricevono un numero di citazioni nella norma saranno penalizzati nella valutazione.

Nella tabella 1 c'è un esempio di ciò che potrebbe accadere se l'ANVUR adottasse la strategia 3.

Numero di citazioni ricevute	Articoli 2008	Articoli 2009	Strategia 2 Classificazione 2008 e 2009	Strategia 3 Classificazione 2008	Strategia 3 Classificazione 2009	STRATEGIA 1 Impatto normalizzato
0	3	3	A	B	B	0,0
1	2	3	A	B	B	0,8
2	2	4	A	B	B	1,5
3	0	0	A	B	B	2,3
4	1	2	A	B	B	3,1
5	0	1	A	B	B	3,8
6	2	0	A	B	B	4,6
7	0	1	A	B	A	5,4
8	1	1	A	B	A	6,1
9	0	1	A	B	A	6,9
10	1	0	A	B	A	7,7
11	1	1	A	B	A	8,4
12	0	1	A	B	A	9,2
13	1	0	A	A	A	10,0
15	1	0	A	A	A	11,5
16	1	0	A	A	A	12,3
18	0	1	A	A	A	13,8
19	1	0	A	A	A	14,6
35	1	0	A	A	A	26,9
47	0	1	A	A	A	36,1
87	1	0	A	A	A	66,8
Totale	*19*	*20*				
IF CAT WOS: 1,302						

Tabella 1.Dati citazionali 2008-2009 degli articoli pubblicati da ACM Computing Surveys e confronto tra diverse strategie di classificazione. L'impatto normalizzato è calcolato rispetto all'IF medio 1.302 della categoria “Computer science, theory and methods”

La rivista è ACM Computing Surveys. Secondo l’edizione 2010 del Journal of Citation Report, con un Impact Factor pari a 8,0 è la migliore rivista (anche in termini di 5-Year Impact Factor e Article Influence) della categoria “Computer science, theory and methods”.[5] Per ogni anno è riportata la frequenza degli articoli per numero di citazioni, il totale annuale delle citazioni e il numero medio di citazioni per articolo nei due anni. Vediamo il risultato cui porta l’applicazione della strategia 3. Ci sono due articoli, uno per anno che hanno ricevuto 8 citazioni (proprio il valore dell’IF della rivista). Siccome la rivista su cui sono pubblicati è la stessa, il GEV li attribuisce alla stessa fascia di merito iniziale, quella più elevata.

- L’articolo pubblicato nel 2008 ha un gigante per vicino di banco: in quell’anno c’è un articolo che ha ricevuto ben 87 citazioni; il numero medio di citazioni per articolo di quell’anno è 12,4; 8 è minore di 12,4 e quindi l’articolo del 2008 è retrocesso nella classe di merito inferiore.
- L’articolo del 2009 ha un vicino di banco alto "solo" 47 citazioni; la media delle citazioni nell’anno è 6,8; 8 è maggiore di 6,8 e quindi l’articolo resta nella classe di merito iniziale.

Più in generale, possiamo vedere che nel 2008 la presenza del vicino gigante (87 citazioni) contribuirebbe al declassamento di ben 11 articoli su 19 (58%), tutti quelli con meno di 12,4 citazioni; nel 2009, anche se c'è un vicino un po' meno gigante (47 citazioni), si verifica addirittura il declassamento di 13 articoli su 20 (65%) Risultati, per così dire, idiosincratici; che rendono del tutto

incredibile il risultato aggregato finale.

E allora cosa potrebbe fare l'ANVUR?

Come potrebbe l'ANVUR attribuire i prodotti di ricerca ad una fascia di merito, dati i vincoli di tempo, costo e assetto istituzionale in cui si trova ad operare? Potrebbe adottare con un mix semplice di bibliometria e peer review.

Potrebbe applicare la strategia 1 (misurare direttamente l'impatto della ricerca) per tutti quei prodotti per i quali è noto il numero di citazioni ricevute.

In alcune aree disciplinari i GEV avranno a disposizione, per tutti i prodotti o per una quota significativa di essi, il numero di citazioni ricevute dal singolo prodotto di ricerca. Questo numero, tratto da una banca dati solida (Scopus e WOS), è l'indicatore convenzionale più utilizzato in letteratura. Come scrivono Rafols et al. (2012)

> Bibliometric measures based on citations to publications provide an internal measure of the impact of the contribution, and hence a proxy of scientific performance. The number of citations per publication (or 'citation impact') is neither an indicator of quality nor importance. Instead, it is a reflection of one form of influence (influence on one's scientific peers) that a publication may exert, which can be used in evaluations provided certain caveats are met

Per ognuno di questi prodotti, i GEV hanno anche una informazioni indiretta di qualità: il prodotto risponde agli standard prevalenti nella disciplina poiché ha superato una peer review per accedere alla pubblicazione su una rivista. Si troveranno dunque nella situazione informativa ideale sognata da ogni bibliometrico.

Non c'è dunque alcuna ragione perché in questi casi si ricorra a dati diversi dal numero di citazioni. Ovviamente il numero di citazioni in sé non è immediatamente significativo nell'attribuzione ad una classe di merito, poiché il comportamento citazionale cambia a seconda delle discipline ed è quindi necessario normalizzare il numero di citazioni rispetto a qualche misura citazionale di riferimento (se la valutazione riguarda una singola disciplina il problema della normalizzazione non si pone). Si tratta del problema forse più dibattuto nella letteratura bibliometrica, per il quale, come scrivono Rafols et al. 2012

> the most extensively adopted practice is to normalise by the discipline to which is assigned the journal in which the article is published.

Le citazioni ricevute da ogni articolo potrebbero essere normalizzate in riferimento al numero medio di citazioni ricevute dagli articoli pubblicati nelle riviste appartenenti alla stessa categoria disciplinare della rivista che lo contiene, nel periodo di riferimento. Una volta calcolato questo indicatore, si potrebbe ordinare in senso decrescente gli articoli, e attribuire ogni articolo ad una delle quattro categorie di merito. Il 25% degli articoli con i valori più elevati finirebbe nella categoria di merito più alta; il 25% successivo in seconda categoria e così via. Oppure applicare le soglie previste nel bando (20%, 20%, 10%, 50%).

Se non si è troppo raffinati, ci si potrebbe anche rifare alle categorie già presenti nei database commerciali che l'ANVUR utilizzerà (WOS e Scopus), ed alle statistiche aggregate presenti nel Journal of Citation Report o in Scimago, come si fa di norma nelle ricerche bibliometriche. Nel nostro esempio si potrebbe prendere l'IF medio aggregato delle 97 riviste della categoria "Computer science, theory and methods" che è 1,302; ciò significa che in media ognuno dei 10.933 articoli pubblicati tra 2008 e 2009 sulle 97 riviste della categorie ha ricevuto 1,3 citazioni (per un totale di 14.231 citazioni). La tabella 1 riporta il calcolo dei valori normalizzati. Avere ricevuto 8 citazioni non è poi così male. E soprattutto non è difficile notare che con questo sistema due articoli usciti sulla stessa rivista e che hanno ottenuto lo stesso numero di citazioni sono bibliometricamente indistinguibili tra loro.[6]

Per tutti quei prodotti per cui la strategia 1 non sia applicabile (articoli su riviste non indicizzate, libri, altri prodotti) i GEV potrebbero ricorrere alla peer review, utilizzando opportuni accorgimenti per renderla impermeabile alle manipolazioni da parte di cricche disciplinari. Solo nel caso in cui si ritenga che la peer review sia davvero troppo facilmente manipolabile, si potrebbe adottare con cautela la strategia 2, ricorrendo però a classifiche di riviste messe a punto fuori dai confini nazionali. Quello di cui non c'è davvero bisogno è l'adozione della Strategia 3, e più in generale la creazione di bibliometria e classifiche fai-da-te.

---

[1] Nel REF britannico è previsto l'uso esclusivo della peer review per l'attribuzione dei prodotti nelle classi di merito; a tale scopo i panel di valutatori decidono se avere o meno informazioni bibliometriche sul prodotto di ricerca sottoposto a valutazione. I revisori devono tenere conto, dando loro peso diverso, della qualità, dell'impatto accademico (diffusione) e dell'impatto socio-economico del singolo prodotto di ricerca. Per l'impatto accademico i panel decidono se avvalersi o meno di qualche indicatore bibliometrico. La peer review effettuata nel corso del REF non è una duplicazione del processo di revisione che un articolo ha già superato quando è stato pubblicato su una rivista, poiché l'insieme informativo su cui deve basarsi il giudizio del revisore è ben più ampio.

[2] Con questa espressione comprendo anche le classifiche costruite sottoponendo questionari a tutti i ricercatori di una certa disciplina.

[3] La logica dietro a questo ragionamento non è antibibliometrica; solo a favore della strategia 1. Se c'è informazione bibliometrica sulle riviste, c'è informazione bibliometrica anche sugli articoli pubblicati in quelle riviste. Si usi quella.

[4] Rafols et. al. 2012 hanno svolto un complesso esercizio in cui hanno costruito il ranking di alcune istituzioni di ricerca usando tre metodi diversi: dati bibliometrici Web of Science; classificazioni

bibliometriche delle riviste (IF), e una classificazione delle riviste messa a punto da una società scientifica. I risultati finali cambiano fortemente a seconda del metodo utilizzato.

[5] Per semplicità lavoriamo con un solo valore di IF, quello dell'ultimo anno disponibile nel Journal of Citation Report. Il valore dell'IF delle riviste è rilasciato da Thomson Reuters annualmente.

[6] Per la precisione in questo caso andrebbe fatto notare che l'articolo del 2008 ha lo stesso numero di citazioni di quello pubblicato nel 2009, ma ha avuto un anno di tempo in più per essere citato. Ma ai fini di un esercizio aggregato di valutazione questo punto potrebbe essere considerato trascurabile.

La classifica di Nonna Papera

ovvero

Perchè non si possono usare le medie dei ranks per classificare le riviste

di Giuseppe De Nicolao

1. La genesi del "Gedeon Score"

Il prof. Gedeone P. rientrò nel suo ufficio e richiuse la porta quasi sbattendola. Non sopportava essere contraddetto e, ancor peggio, fare la figura dell'incompetente. Il presidente del Nucleo di Valutazione, Paolo L., aveva la mania di spaccare il capello in quattro. Quando Gedeone aveva preso la parola dicendo che il primo passo per la valutazione dei prodotti della ricerca era classificare le riviste di ogni settore in 3 o 4 livellli di qualità, Paolo L. aveva replicato citando la vicenda della valutazione australiana, il cosiddetto ERA (Excellence of Research in Australia). L'ERA 2010 aveva appunto fatto uso di una classificazione delle riviste, suscitando però infinite dispute e controversie nella comunità accademica. Nel maggio 2011, lo stesso Ministro Carr, davanti ad una commissione del Senato Australiano, non solo aveva dichiarato che la prossima edizione dell'ERA avrebbe rinunciato alla classificazione, ma aveva dovuto ammettere:

> There is clear and consistent evidence that the rankings were being deployed inappropriately ... in ways that could produce harmful outcomes

Nelle orecchie di Gedeone risuonava ancora fastidiosamente la voce del collega che declamava l'epitaffio della classificazione australiana, ufficialmente sconfessata ed abbandonata. Intanto, un suono attirò il suo sguardo su un nuovo messaggio, inviato proprio da Paolo. Il collega, per infierire, gli mandava un articolo, scritto da due australiani, che riassumeva tutte le ragioni che sconsigliavano il ricorso alla classificazione delle riviste nella valutazione delle strutture di ricerca ed anche il link ad un articolo in italiano apparso su un blog dal nome strano, "Roars". Si innervosì ancora di più. Era convinto che tutte quelle obiezioni tecniche non fossero altro che pretesti. Quando si tratta di dare i voti e premiare chi eccelle, c'è sempre qualcuno che rema contro.

Eppure non poteva essere così complicato. Si rincuorò subito: avrebbe mostrato ai colleghi che era facile costruire una classificazione oggettiva. Andò sulla pagina del Journal Citation Report dell'ISI Web of Knowledge. Per ogni rivista, erano elencati diversi indicatori bibliometrici, alcuni noti, altri più misteriosi. Quelli candidabili per stilare una classifica delle riviste erano:

- Impact Factor
- 5-Year Impact Factor
- Eigenfactor Score
- Article Influence

Scegliendone uno, era immediato ottenere la classifica con un colpo di mouse. La scelta più semplice sarebbe stata prendere uno dei due Impact Factors, ma rinunciò subito, ricordandosi che sulla mailing list di facoltà era già circolato quell'articolo allarmista sulla "Top ten in journal impact factor manipulation". Inoltre, la settimana prima, un collega informatico gli aveva raccontato che le manipolazioni dell'Impact Factor erano talmente notorie che le tecniche per il loro riconoscimento automatico erano oggetto di articoli di ricerca. Gedeone non si sentiva abbastanza preparato per difendere la scelta di uno degli altri due indici.

Ebbe un'idea che gli parve il classico "uovo di Colombo": indubbiamente, tutti e quattro gli indicatori sono affetti da qualche errore - nessuno è perfetto - ma la loro media aritmetica compenserà gli errori tra di loro, fornendo un compromesso accettabile da tutti. All'improvviso entusiasmo seguì un'altrettanto rapida disillusione: sarebbe stato come sommare le mele con le pere. Non era possibile mediare numeri del tutto eterogenei, anche come ordine di grandezza.

Doveva esserci una soluzione semplice. Ebbe una seconda illuminazione: per avere grandezze omogenee, avrebbe costruito quattro classifiche, una per indicatore, e attribuito ad ogni rivista quattro punteggi, coincidenti con la posizione nelle diverse classifiche. Per esempio, il JEE (Journal of Excellent Engineering) nella sua categoria ISI otteneva i seguenti piazzamenti:

- 1° - Impact Factor
- 3° - 5-Year Impact Factor
- 6° - Eigenfactor Score
- 10° - Article Influence

Per attribuire un voto oggettivo al JEE bastava un ultimo passaggio: calcolare la media aritmetica delle quattro posizioni in classifica. Ecco fatto: con orgoglio digitò

"Gedeon Score" of JEE = (1 + 3 +6 +10)/4 = 5

Quanto più basso il punteggio tanto migliore la rivista: se una rivista fosse stata in prima posizione in tutte e quattro le classifiche, avrebbe avuto uno score pari a (1+1+1+1)/4 = 1, il minimo possibile. Confrontando il “Gedeon Score” delle riviste, sarebbe stato immediato costruire una classifica delle riviste per una qualsiasi categoria disciplinare. A quel punto, la strada era in discesa: il primo quartile (vale a dire il 25% delle riviste con lo score più basso) avrebbe identificato le riviste di livello A, il secondo quartile quelle di livello B e così via.

Era fiero di se stesso, ma c’era ancora un piccolo problema. Per alcune riviste, non erano disponibili tutti e quattro gli indicatori. Sorrise con sufficienza: se c’erano solo due indicatori avrebbe calcolato la media di due indicatori, invece di quattro. Finalmente una procedura semplice, chiara, basata su dati oggettivi.

Preparò in fretta e furia un documento che spiegava il calcolo del Gedeon Score e stava per spedirlo alla mailing list di facoltà quando si accorse di un errore di formattazione. Assorbito dal problema, non si era accorto che si era fatto tardi e per di più era anche venerdì. Meglio andare a casa. Avrebbe terminato il lavoro lunedì mattina.

2. Il manuale di Nonna Papera

Nel weekend lo attendeva un compito ingrato, a lungo rimandato, ma non più procrastinabile. Aveva promesso alla moglie di fare ordine in cantina, facendo sparire un po’ di ciarpame ed anche un paio di scatoloni di libri e fumetti risalenti alla sua adolescenza. Prima di buttare via tutto, volle aprire e vedere cosa c’era dentro. Sfogliò con nostalgia alcuni numeri di Tex che emanavano un tipico odore di cantina. Mentre passava da un “satanasso” pronunciato da Kit Carson ad un intrigo di Mefisto, notò un intruso in mezzo a Tex, Zagor e Alan Ford: un Manuale di Nonna Papera che non ricordava di aver mai posseduto. Incuriosito, sfogliò alcune pagine. In mezzo alle ricette di cucina, trovò una storiella che attirò la sua attenzione. Si intitolava “La classifica della nonna” ed era una specie di esercizio aritmetico.

Nonna Papera organizza una festa per Qui, Quo, Qua ed anche per Gilberto (il nipote di Pippo, noto studente prodigio) e per Pennino, un personaggio meno noto, nipote di Paperoga. La nonna vuole approfittare dell'occasione per premiare con caramelle e cioccolatini i due ragazzi che sono più bravi a scuola e nello sport. Per questo scopo, la nonna trascrive, per tutti e cinque i ragazzi, i voti in Matematica, in Inglese ed anche il numero di canestri che hanno messo a segno nel torneo di basket scolastico (vedi Tabella 1). Gilberto è bravissimo a scuola – ha ben due “10” - ma un po’ meno nel basket. Non potendo sommare voti scolastici e canestri, il manuale di Nonna Papera consiglia di costruire tre classifiche distinte e poi attribuire ad ogni ragazzo un voto uguale alla media delle sue posizioni nelle tre classifiche di Matematica, Inglese e canestri. Al lettore è chiesto di ricavare la soluzione. I "pari merito" sono trattati nel modo più logico: se nella classifica dei canestri, Qui e Quo sono primi alla pari, si spartiscono i primi due "ranks": (1+2)/2 = 1,5 punti a testa.

	Gilberto		Qui		Quo		Qua		Pennino	
	Voto	rank	Voto	rank	Voto	rank	Voto	rank	Voto	rank
Matematica	10	***1°***	7	***2°***	6	***3°***	5	***4°***	4	***5°***
Inglese	10	***1°***	6	***3°***	7	***2°***	4	***5°***	5	***4°***
Canestri	20	***5°***	22	***1°-2°***	22	***1°-2°***	21	***3°-4°***	21	***3°-4°***
Media dei rank		***2,33***		***2,17***		***2,17***		***4,17***		***4,17***

Tabella 1. Nonna Papera deve assegnare due premi. A sorpresa, Qui e Quo sono i vincitori e Gilberto solo terzo. Il paradosso è dovuto all'uso della media dei ranks come criterio di valutazione.

Gedeone sorrise riconoscendo nel metodo proposto il suo “Gedeon Score”. Fece rapidamente i conti a mente, scoprendo che Gilberto, a dispetto dei suoi “10” in pagella, sarebbe finito in terza posizione (Gedeon Score = 2,33), lasciando caramelle e cioccolatini a Qui e Quo, che arrivavano primi e secondi a pari merito (Gedeon Score = 2,17). Un risultato paradossale e ingiusto nei confronti di Gilberto, se si considera che nessun altro otteneva voti scolastici superiori al “7”. Durante il resto del weekend, Gedeone continuò a rimuginare, cercando di capire quale fosse il punto debole del “Gedeon Score”. Non venendo a capo di nulla, domenica sera si decise: l’indomani, avrebbe chiesto un parere a Peppe, il massimo esperto di analisi dati di tutta la facoltà.

3. Prima lo boccio e poi lo inseguo con il forcone

“Posso chiederti un parere tecnico?” “Ma certo”, rispose Peppe mentre invitava Gedeone ad entrare e accomodarsi sulla sedia dal rivestimento logoro e persino strappato in un angolo. Per prevenire una possibile brutta figura, Gedeone raccontò che un collega della Ruritania gli aveva esposto un metodo semplice per classificare le riviste scientifiche. Il collega straniero intendeva proporre questo nuovo metodo all'ANVUR, l'Agenzia Nazionale di Valutazione del sistema Universitario della Ruritania, il cui acronimo, per pura coincidenza, è uguale a quello dell'agenzia italiana. Mentre ascoltava i dettagli del “Gedeon Score”, Peppe cominciò ad agitarsi sulla sedia. Alla fine esplose indignato:

"Certo che il tuo collega ne capisce proprio poco di analisi dei dati!” (a dire il vero, l’espressione fu più colorita). "Non ha senso sommare o mediare le posizioni in classifica, i cosiddetti ranks, come pure non ha senso sommare o mediare i percentile ranks (le posizioni mormalizzate sulla scala 1-100). Il motivo è semplice: tra una posizione e la successiva può esserci un distacco enorme (e Gedeone pensò ai “10” di Gilberto confrontati ai “7” e “6” di Qui e Quo) oppure piccolissimo. Fare le medie dei ranks è un po’ come sommare le mele con le pere. Fornisce risultati arbitrari. Se lo vedo fare da un mio studente, prima lo boccio e poi lo inseguo con il forcone."

Gedeone si affrettò a dire che anche lui aveva subito sospettato che il criterio fosse sbagliato. Anzi, si domandava se non ci fosse qualche riferimento bibliografico da spedire al collega straniero. Peppe partì in quarta:

"In generale, è difficile trovare lavori scientifici che discutono gli errori. Nessuna rivista accetterebbe di pubblicare un lavoro che spiega perché uno svarione è uno svarione. Chi conosce la materia, lo sa già. Punto. Al massimo, qualche libro di testo mette in guardia gli studenti dagli errori più comuni. Questa volta, però, sei fortunato perchè proprio questo svarione ha avuto un ruolo nel dibattito sugli high-stakes tests negli USA."

Gli high-stakes tests sono quelle prove di esame il cui risultato ha un importanza rilevante per chi lo sostiene. Può trattarsi dell'esame per la patente o del test di ammissione all'università. Proprio nel secondo caso, si era diffusa l'abitudine di ottenere il punteggio finale del candidato calcolando la media dei suoi percentile ranks nei test parziali, per esempio di Matematica e Inglese. Agli statistici era evidente che si trattava di una procedura errata, ma non si riusciva a sradicarne l'uso.

"Per tale motivo" spiegò Peppe "nel 1993, uno statistico, Bruce Thompson, pubblicò un "position paper" con lo scopo di chiarire la questione una volta per tutte. Ecco, dovrei averlo nel mio hard disk. Adesso ti spedisco il pdf via e-mail. Non spaventarti per la qualità tipografica: è un dattiloscritto, ma la qualità tecnica è ineccepibile. Non so se vorrai girarlo al tuo collega, potrebbe rimanerci male. Infatti, Thompson ci va giù duro contro chi sostiene questi metodi."

Mentre diceva queste cose, Peppe aprì il pdf e andò a pagina 27, leggendo ad alta voce:

> The only reason for using percentile ranks is ignorance, and it is questionable whether a defense of ignorance will be viable.

"Se in Ruritania adotteranno questo metodo, verranno strapazzati per anni da tutta la letteratura bibliometrica, che citerà la loro classificazione come un tipico esempio delle cose da non fare. Senza offesa per il tuo amico ruritano, a noi due, dopo aver ascoltato le spiegazioni di Paolo, una stupidata simile non sarebbe neppure passata per l'anticamera del cervello. Quando conosci il numero di citazioni di un articolo, che bisogno hai di passare attraverso la classificazione delle riviste? Per trovare la fascia di merito dell'articolo (primo quartile, secondo quartile, eccetera), basta confrontare le citazioni con la distribuzione delle citazioni in quella categoria disciplinare, un metodo molto più semplice e rigoroso."

4. Non tutti i numeri si possono sommare

Gedeone se ne tornò in ufficio con la coda tra le gambe. Stampò l'articolo di B. Thompson. Il lungo titolo non lasciava adito a dubbi:

GRE [Graduate Record Examination] percentile ranks cannot be added or averaged: a position paper exploring the scaling characteristics of percentile ranks, and the ethical and legal culpabilities created by adding percentile ranks in making "High-Stakes" admission decisions

Il punto fondamentale dell'articolo era che i numeri si possono sommare solo quando corrispondono a misure prese con una scala graduata ad intervalli costanti. Quando i numeri indicano la posizione in una scala ordinale (primo, secondo, terzo, ...), le distanze tra una posizione e la successiva potrebbero non essere costanti e l'addizione, anche se possibile sul piano formale, è un'operazione priva di senso. Gedeone non fu particolarmente rincuorato dall'apprendere che poteva invocare qualche attenuante alla sua cantonata:

> It seems counterintuitive to many persons, even to some educated people with terminal degrees

serving on faculty at world-class universities, that some numbers simply cannot be added ... Most of us have paradigms about numbers that were unconsciously formulated, typically in the primary grades of elementary school. When we are given several numerals, we are used to presuming that we can add them up. Few of us were ever admonished that we can only add numbers when the numerals represent data derived using an equal interval measurement ruler. In fact, few of us consciously recognize that addition itself does presume equal-interval measurement.

Nell'appendice dell'articolo, Bruce Thompson aveva persino raccolto una raffica di citazioni scientifiche contrarie all'uso della somma dei ranks. Insomma, il “Gedeon Score” era null’altro che la riproposizione di uno svarione senza appello, già stigmatizzato dagli esperti. Gedeone trascinò nel cestino il file che stava per mandare alla mailing list di facoltà e vuotò il cestino. Meglio pensare ad altro. La moglie, per festeggiare il ritrovamento del suo amato Manuale di Nonna Papera, per cena avrebbe preparato la squisita focaccia del paleolitico (vedi figura).

Avvertenza. Il Prof. Gedeone P. e i suoi colleghi sono personaggi di fantasia come pure di fantasia è il Journal of Excellent Engineering. Il Manuale di Nonna Papera, la cui autrice siede nel Consiglio Direttivo dell'ANVUR, spiega come cucinare la focaccia del paleolitico, ma non contiene la "La classifica della nonna". Infine, non è l'ANVUR della Ruritania, nazione che esiste solo nei romanzi e nei film, ma è l'ANVUR italiana che potrebbe adottare il "Gedeon Score" per l'esercizio di Valutazione della Qualità della Ricerca 2004-2010. Tutte le altre informazioni riportate nell'articolo sono fedeli alla realtà.

Scarica qui l'articolo di B. Thompson che spiega le ragioni per cui ranks e percentile ranks non possono essere sommati e mediati.