

La valutazione della ricerca fra scienza e feticismo dei numeri

Antonio Banfi e Giuseppe De Nicolao

Bibliometria: l'arte di misurare libri. Una strana scienza che nell'etimo nasconde le sue origini. Infatti, fu a fini biblioteconomici che nel 1927 P.L.K. Gross e E.M. Gross elaborarono un *seminal paper* nel quale si suggeriva il ricorso ad analisi quantitative per rendere più efficiente la gestione delle biblioteche di chimica: in una fase di sviluppo della ricerca e di moltiplicazione dei prodotti scientifici, occorre allocare correttamente le risorse, sulla base di valutazioni «oggettive», per garantire che le biblioteche disponessero almeno delle riviste di maggior rilevanza. In tempi relativamente rapidi, l'analisi bibliometrica abbandonò l'ambito della biblioteconomia per divenire uno degli strumenti a disposizione per l'analisi scientometrica: la misurazione della «scienza» prodotta da strutture di ricerca e singoli ricercatori. A metà degli anni '50, Eugene Garfield inventò il fattore di impatto (IF), un indice numerico fondato sulle citazioni, utilizzabile per classificare le riviste e – *a fortiori* – per fornire una stima dell'impatto dei contributi pubblicati da una determinata sede editoriale. Il ragionamento su cui si fonda l'utilizzo dell'IF (e più in generale l'analisi citazionale) è il seguente: se si assume che gli scienziati citino maggiormente i contributi che hanno un impatto rilevante sugli studi, si può supporre che alla maggiore importanza scientifica di tali pubblicazioni corrisponda un elevato livello qualitativo. Una rivista che pubblica articoli caratterizzati da molte citazioni dovrebbe quindi essere considerata come una sede editoriale di prestigio, che accetta contributi di rilievo e seleziona accuratamente i prodotti che le vengono sottoposti; non solo, i ricercatori competeranno fra loro per pubblicare in sedi editoriali di questo tipo, in un virtuoso meccanismo *self-sustaining*. Inoltre, l'analisi citazionale avrebbe il vantaggio di essere «oggettiva», sfuggendo alla soggettività propria del giudizio espresso dai pari (*peer review*). L'opera di Garfield fu pionieristica non solo per l'invenzione dell'IF, che fino a tempi recenti ha avuto un enorme successo, ma anche perché egli stesso fondò l'Institute for Scientific Information, un fornitore commerciale di dati bibliografici e citazionali: vide così la luce la prima grande banca dati commerciale per l'analisi bibliometrica, poi acquisita dalla Thompson Reuters e oggi denominata Web of Knowledge. Con Garfield nasce dunque l'industria delle informazioni bibliometriche, che oggi vede sul mercato un altro grande attore, Elsevier, proprietario della base dati Scopus. La fortuna della bibliometria ai fini della valutazione crebbe negli anni '80, grazie alle politiche dei governi conservatori del Regno Unito. L'Università, si disse, deve disseminare formazione e assicurare ricerca, specialmente nei campi di maggior interesse nazionale, ossia ingegneria e area biomedica. Poiché l'Università è finanziata con il denaro pubblico tratto dalla fiscalità generale, sono necessari esercizi di valutazione che restituiscano dati numerici facilmente interpretabili dal decisore pubblico e dall'amministratore della singola sede universitaria, in modo da incrementare la competizione e consentire un'allocazione corretta dei fondi. In realtà, il crescente uso dell'analisi bibliometrica ne fece ben presto emergere i limiti.

Infatti, l'equazione fra qualità e impatto è tutt'altro che scontata, specialmente in certi ambiti disciplinari. Inoltre vi è il rischio di confondere la valutazione aggregata con la valutazione individuale: i dati bibliometrici sono grezzi e non consentono, da soli, una valutazione accurata dei singoli ricercatori, mentre possono essere sfruttati – *cum grano salis* – per valutare *gruppi* di scienziati (dipartimenti, atenei, centri di ricerca). La barriera fra valutazione aggregata e valutazione individuale non deve essere infranta, pena conseguenze deleterie: ossia l'incentivazione di comportamenti opportunistici, quando non abusivi. Per fare un esempio, si correrebbe il rischio di scoraggiare ricerche difficili e complesse, solo perché producono citazioni in tempi più lunghi e con maggiore fatica; oppure ricerche di nicchia, particolarmente innovative o multidisciplinari, che per un tempo prolungato potranno non essere prese in considerazione dalle comunità scientifiche: lasciamo al lettore di immaginare le conseguenze di tutto ciò in ambito biomedico. D'altro canto si favoriranno comportamenti volti a massimizzare l'utile dei ricercatori in termini di citazioni, con

esiti al limite della frode: il declino dell'IF, ad esempio, è dovuto anche al proliferare di riviste e consorterie accademiche in grado di orchestrare politiche editoriali che, benché di scarsa qualità, si dimostrano capaci di generare un elevato numero di citazioni.

L'affermazione dell'analisi bibliometrica si legava dunque ad alcuni assunti di carattere politico ed economico: favorire una competizione sana fra ricercatori e premiare i migliori (individui, dipartimenti, atenei) attraverso la concessione di maggiori fondi per la ricerca e eventualmente di meccanismi premiali per gli individui. Il ricorso a sistemi di valutazione bibliometrica si lega dunque a politiche pubbliche di incentivazione e disincentivazione, il cui fine è massimizzare il rendimento dei fondi allocati. Da questo punto di vista, la bibliometria sembra offrire indubbi vantaggi: i costi sono limitati (per lo più dovuti all'acquisto dell'accesso ai dati da parte delle banche dati) e gli esiti verificabili e trasparenti. Ma un uso non accorto di tali strumenti finisce per generare esiti contrari alle aspettative. Non è interesse del decisore pubblico, né del *taxpayer*, che i fondi vengano assorbiti da scienziati che si trasformano in meri collezionisti di citazioni. Tantomeno è interesse pubblico che il progresso scientifico sia orientato in modo determinante dall'adozione di strumenti di valutazione che sfavoriscono alcune linee di ricerca solo perché meno produttive dal punto di vista citazionale. Fisica, matematica, la quasi totalità delle scienze umane, ma anche alcuni ambiti della ricerca biomedica sono scarsamente produttivi o comunque meno produttivi di altri dal punto di vista citazionale: un dato di fatto che nulla ha a che vedere con la qualità della produzione scientifica. Non solo: vi sono ambiti di ricerca che, per le loro proprie abitudini e oggetti di ricerca, o anche per il numero esiguo di ricercatori coinvolti, restituiscono un numero di citazioni complessivo talmente basso da mettere in dubbio la rilevanza statistica di un'analisi citazionale.

In un noto articolo, il premio Nobel per la chimica Ernst ha scritto: «as an ultimate plea, the personal wish of the author remains to send all bibliometrics and its diligent servants to the darkest omnivoric black hole that is known in the entire universe, in order to liberate academia forever from this pestilence» (*Chimia* 2010.64 p. 90). Anche se i toni non sono sempre tanto esasperati, di tutti i problemi cui si è fin qui accennato vi è nella comunità scientifica internazionale diffusa consapevolezza. In un certo senso si potrebbe dire che dopo decenni di esperienza, la comunità scientifica internazionale è almeno in parte vaccinata dai rischi che comporta un uso dissennato degli indici citazionali e lo stesso si potrebbe dire del regolatore. Ad esempio il sistema di valutazione britannico, il più antico e più noto di tutti (RAE/REF), si basa ormai in modo pressoché esclusivo sulla revisione dei pari. Il che, tuttavia, crea ulteriori problemi.

Se da un canto si può dire superata la rozza contrapposizione fra soggettività della *peer review* e oggettività degli indici bibliometrici, poiché si è riconosciuto che le citazioni riservate a un articolo altro non sono se una forma mediata di revisione dei pari, rimane il problema del costo enorme che comporta condurre attraverso la *peer review* un intero esercizio di valutazione nazionale. Anche conferendo una limitata quota di prodotti scientifici ai valutatori, resta il rischio che il costo complessivo dell'operazione superi i benefici attesi. Il dibattito scientometrico si muove dunque fra questi due scogli: valutazioni bibliometriche a basso costo e scarsamente precise, specie se fondate su numeri non sufficientemente grandi, e valutazioni dei pari più accurate ma molto dispendiose. Di qui l'adozione di soluzioni intermedie, quali la *informed peer review*, ossia una revisione dei pari integrata da dati ulteriori, in modo da renderla più celere e meno *time-consuming*: un altro esempio è dato dalle classifiche di riviste, variamente elaborate, che consentano di trasferire in modo più o meno automatico la valutazione del contenitore (la rivista) sul contenuto (l'articolo).

L'elaborazione di sistemi di valutazione è dunque un compito delicato. Il dibattito scientometrico è vivace e l'analisi bibliometrica è in continua evoluzione. Tutto ciò indica che in questa materia ancor più che per

altre occorre cautela nell'elaborazione delle regole e vigilanza *ex post* sul loro funzionamento, sia da parte di chi redige le regole stesse, che da parte delle comunità scientifiche. L'Italia è da questo punto di vista un Paese immaturo. L'adozione di strumenti di valutazione, pur auspicata e necessaria, è avvenuta senza sufficiente consapevolezza sia da parte del legislatore, che da parte di valutatori e valutati, con conseguenze nocive per il sistema dell'università e della ricerca.

L'impreparazione culturale dell'accademia italiana ha favorito una sorta di feticismo dei numeri. Chi maneggia in modo accorto i metodi quantitativi è consapevole del loro valore strumentale, dei possibili vantaggi rispetto alle valutazioni dei pari, ma anche dei loro limiti e pericoli. Invece, è prevalsa l'ingenuità di neofiti convinti che nei numeri risiedesse una sorta di potere magico tale da assicurare esiti positivi ad *una qualsiasi* valutazione, *purché quantitativa*. Di qui una sistematica sottovalutazione dell'importanza delle conoscenze necessarie a raccogliere, maneggiare e interpretare i dati. In un contesto ai limiti della superstizione, la valutazione numerica è stata ritenuta un toccasana nei confronti di un'accademia dipinta come irrecuperabile. Uno degli esiti principali del feticismo dei numeri è l'idea che, grazie alla sua presunta oggettività, una valutazione numerica di tipo automatico sia da ritenersi preferibile alle valutazioni dei pari. Che questa oggettività sia solo apparente è evidente: valutare le ricerche in base alle citazioni implica una decisione soggettiva che privilegia ricerche ad alto impatto rispetto a ricerche meno popolari, ma forse altrettanto o più importanti. Si dirà tuttavia che avere delle regole oggettive evita ogni arbitrio e che, errore per errore, è meglio avere regole inequivoche. Ai profani sfugge però la natura sistematica degli errori associati ai criteri bibliometrici. Una valutazione umana, condotta con l'eventuale ausilio di informazioni bibliometriche, è per sua natura imprecisa: diversi esperti daranno diverse valutazioni dello stesso lavoro scientifico o dello stesso ricercatore. Tuttavia, gli errori commessi non avranno natura sistematica: un revisore potrà sopravvalutare l'importanza delle citazioni mentre un altro potrà sottovalutarla, un revisore premierà argomenti di moda mentre un altro sarà più attento a valorizzare ricerche di frontiera. Nel caso di strumenti bibliometrici automatici, invece, gli errori sono *sistematici*. Se si contano le citazioni, intere aree di indagine risulteranno sistematicamente penalizzate con il risultato di orientare le ricerche dell'intera comunità nazionale verso tematiche *mainstream*. Questi pericoli sono noti alla comunità scientifica internazionale e ciò spiega il largo consenso nel respingere l'uso di metodi bibliometrici automatici per la valutazione di singoli articoli e di singoli ricercatori: le distorsioni sistematiche renderebbero inutilizzabili i risultati e incentiverebbero comportamenti opportunistici su scala nazionale.

L'adozione da parte di ANVUR (Agenzia nazionale di valutazione) di strumenti bibliometrici artigianali non può essere giustificata invocando l'esperienza internazionale. Basti considerare le metodologie di valutazione nell'ambito della VQR, l'esercizio di valutazione della ricerca che dovrebbe quantificare la qualità della ricerca prodotta da dipartimenti e atenei. In tale ambito, è necessario valutare un largo numero di pubblicazioni. L'approccio più semplice, ma anche più costoso, è il ricorso integrale alla revisione dei pari: la strada utilizzata nel Regno Unito. Non sorprende che prima ancora dell'ANVUR sia stata l'agenzia di valutazione inglese HEFCE a valutare la possibilità di una valutazione bibliometrica automatica. Dopo aver condotto uno studio pilota, la conclusione è stata inequivocabile: «Bibliometrics are not sufficiently robust at this stage to be used formulaically or to replace expert review in the Research Excellence Framework». Anche l'Australia aveva sperimentato un metodo automatico basato sulla classificazione delle riviste in classi di merito. La preparazione delle classifiche aveva suscitato aspre controversie e, dopo averle utilizzate in un esercizio di valutazione, nel maggio 2011 il governo australiano le ha ritirate ammettendo che «There is clear and consistent evidence that the rankings were being deployed inappropriately [...] in ways that could produce harmful outcomes». Queste informazioni erano a disposizione dell'ANVUR nel momento in cui si accingeva a definire le procedure della VQR. È pertanto, ancor più significativa la decisione dell'ANVUR di predisporre un sistema di valutazione ibrido in cui le

scienze dure adottavano gli indicatori bibliometrici (rigettati come inadeguati dell'HEFCE) mentre le scienze umane e sociali, pur adottando ufficialmente la revisione dei pari, attribuivano un ruolo ambiguo, e talora dirimente, a classifiche di riviste da predisporre in tempi del tutto incongrui. La VQR italiana si è così venuta a configurare come un vero e proprio *monstrum* in cui non solo alcune aree scientifiche erano valutate su base prevalentemente bibliometrica ed altre in base alla revisione dei pari, ma le regole bibliometriche differivano da un'area scientifica all'altra. Con un'innovazione senza basi scientifiche e precedenti internazionali, veniva dato mandato a Gruppi di Esperti della Valutazione, designati dal direttivo dell'Agenzia, di costruirsi i propri criteri. A dispetto del nome, tali gruppi non avevano pressoché nessuna competenza in materia, cosicché sono state inventate dal nulla una decina di schemi bibliometrici più amatoriali che scientifici. Criteri così poco scientifici finiscono per favorire o danneggiare le discipline in modo casuale. Nella migliore delle ipotesi, ciò renderà i risultati della valutazione inutilizzabili mentre, nella peggiore delle ipotesi, essi produrranno risultati arbitrari e inadatti a essere utilizzati a scopo premiale. Lo stesso Ministro Profumo, messo di fronte all'evidente inadeguatezza della VQR, si è trovato costretto a dichiarare pubblicamente che "questa VQR è una sperimentazione", evidenziando, seppur implicitamente, lo spreco di denaro pubblico che essa comporta.

Un esempio anche peggiore di feticismo dei numeri sono i parametri numerici utilizzati per l'abilitazione scientifica nazionale. Secondo la L. 240/2010, l'abilitazione scientifica certifica una qualificazione scientifica che, come tale, non è connotata in termini comparativi. Il D.M. 76/2012, sulla scia di precedenti documenti ANVUR, ha invece condizionato il conseguimento dell'abilitazione ad una comparazione con i professori già in ruolo: al candidato che tenta di ottenere l'abilitazione di prima fascia si chiede di «essere migliore» del 50% dei professori ordinari in servizio. Per effettuare la comparazione, si ricorre a indicatori bibliometrici come il numero di articoli scientifici, il numero di citazioni e l'indice di Hirsch. Analoghe soglie bibliometriche vengono applicate per gli aspiranti commissari. Abbiamo già ricordato che esiste un vasto consenso internazionale sull'impossibilità di utilizzare criteri bibliometrici automatici per valutare i singoli ricercatori: il decreto ministeriale inaugura invece una via italiana al reclutamento universitario anomala e foriera di problemi tecnici e procedurali. Solo l'ignoranza della natura dei dati citazionali – soggetti a definizioni arbitrarie, errori e ritardi temporali – può giustificare la decisione di usarli come criterio dirimente in ambito amministrativo. Inoltre, le soglie numeriche non sono definite in modo assoluto, ma richiedono il calcolo degli indicatori bibliometrici dei professori già in servizio per trovare la mediana, ovvero il valore che divide in due parti la popolazione, valore che gli aspiranti abilitati e commissari devono superare. Non esistendo un'anagrafe nazionale della ricerca, ANVUR ha invitato i docenti a completare i loro siti docente procedendo quindi alla rilevazione su dati inaffidabili perché caricati su base puramente volontaria. Il seguito è degno di una commedia dell'assurdo: ANVUR calcola le mediane e dopo poco le ritratta dichiarando che aveva usato una formula sbagliata. I dati su cui sono stati effettuati i calcoli vengono tenuti segreti. Per le scienze umane e sociali, le mediane richiedono la conoscenza degli elenchi delle riviste «eccellenti» ed anche di quelli delle riviste scientifiche. Gli elenchi vengono pubblicati giorni se non settimane dopo le mediane suscitando il sospetto di aggiustamenti fuori tempo massimo. Infine, quando ANVUR pubblica le liste delle riviste scientifiche si scopre che esse contengono di tutto, dai quotidiani come *il Mattino* di Padova fino all'Annuario del Liceo di Rovereto. Il caso fa ridere tutta Italia quando approda sulla prima pagina del *Corriere della Sera*, e l'eco della vicenda varca persino i confini nazionali giungendo sulle colonne di *Times Higher Education* e di *Science*. Nel frattempo, fioccano i ricorsi amministrativi degli aspiranti commissari presso il TAR del Lazio, timido antipasto di quello che potrà accadere a valle degli esiti delle abilitazioni quando saranno i candidati bocciati ad adire le vie giurisdizionali.

Esiste una spiegazione plausibile a questa commedia dell'assurdo? Certamente, chi crede nel potere taumaturgico dei numeri è più soggetto a ignorare che essi hanno un'origine, che non sono tutti ugualmente affidabili e che richiedono un'interpretazione. Veri esperti di valutazione non sarebbero incorsi negli errori che abbiamo ricordato. Purtroppo, in Italia la valutazione è stata caricata di valenze fortemente ideologiche e affidata a scienziati, che per quanto abili nel loro campo, non disponevano delle competenze necessarie ad emanciparsi da pregiudizi e feticismi. Scienziati talmente a digiuno delle più elementari nozioni di bibliometria e scientometria da costruire procedure suicide destinate a screditare se stessi, l'agenzia di valutazione nazionale e a produrre un micidiale contenzioso amministrativo.

Al momento, non sono visibili facili vie d'uscita. Nel medio periodo è auspicabile che l'ubriacatura numerologica funga da "vaccinazione" per la comunità accademica italiana aiutandola a superare i feticismi bibliometrici a favore di un più sano approccio scientifico che usa tutti i dati disponibili alla luce delle metodologie scientifiche, senza chiedere risposte che essi non possono dare. Inoltre, una limitazione dei danni esigerebbe una pronta revisione del D.M. 76/2012 per eliminare regole e parametri privi di fondamento scientifico e destinati a trasferire buona parte delle procedure di abilitazione scientifica nelle aule del tribunale amministrativo. Tutto ciò se si vuole che la valutazione della ricerca sia funzionale al miglioramento qualitativo della ricerca stessa e non si riveli l'ennesimo spreco di danaro ed energie.