“Luci e ombre della VQR”

Paolo Rossi

Riassunto

La Valutazione della Qualità della Ricerca (VQR) è il tentativo più sistematico finora effettuato di quantificare l’impegno e i risultati del sistema universitario e della ricerca italiana. A fronte di risultati di grande interesse (soprattutto statistico e quantitativo) occorre tuttavia rilevare alcuni limiti metodologici che hanno reso difficile una corretta interpretazione del significato qualitativo degli indicatori, mentre d’altro canto l’eccessiva (e spesso ingiustificata) enfasi su alcuni parametri di natura bibliometrica rischia di condizionare impropriamente le scelte scientifiche e culturali di un’intera generazione di studiosi, anche per la correlazione che si è voluta immediatamente stabilire tra i risultati della VQR e i finanziamenti al sistema universitario. Si propone una discussione approfondita di queste criticità, anche al fine di individuare meccanismi correttivi volti ad eliminare gli effetti potenzialmente distorsivi di un utilizzo acritico degli esiti della valutazione.

Parole chiave: *Università, Valutazione, VQR, Bibliometria, Sampling Bias.*

La VQR (Valutazione della Qualità della Ricerca) nasce sulla scia e sulla base della precedente esperienza della VTR (Valutazione Triennale della Ricerca), che era stata condotta dal CIVR in riferimento al triennio 2001-2003. Appare importante tenere a mente questo collegamento, che ha un corollario non banale nell’interpretazione dell’acronimo, inizialmente inteso come sigla di una “Valutazione Quadriennale della Ricerca (2004-2007)” che lo stesso CIVR era stato inizialmente autorizzato a svolgere ma che non fu mai portata a compimento in quanto era ormai avviato il (lento) processo che portò alla creazione dell’ANVUR e al conseguente smantellamento del CIVR.

Occorre sottolineare che, accanto ad alcuni elementi di continuità, facilmente rintracciabili nella sequenza dei documenti dell’epoca, nel passaggio dalla VTR alla VQR furono introdotte anche alcune importanti innovazioni, tali da rendere molto più ampio il processo valutativo, ma anche molto più complessa la lettura e l’interpretazione dei risultati.

La VTR, nell’ambito dei propri dichiarati limiti quantitativi, era caratterizzata da una forte coerenza interna, in quanto privilegiava, sulla base di una ristretta selezione affidata alle singole istituzioni fino al livello dipartimentale, l’analisi dei prodotti della ricerca giudicati più rappresentativi del complesso delle attività di ciascuna istituzione: di fatto si trattava di scegliere, in un arco di tempo triennale, un numero di pubblicazioni mediamente prossimo al 25% del numero di docenti presenti nel dipartimento. Ciascuno di questi prodotti di ricerca veniva poi affidato a revisori che gli attribuivano un voto, e la media dei voti rappresentava il “punteggio” del dipartimento e quindi la sua posizione nel ranking nazionale.

La VTR tendeva quindi a misurare, per quanto rozzamente, la “qualità” della ricerca dipartimentale, intesa come maggiore o minor capacità collettiva di raggiungere livelli di “eccellenza” scientifica, in misura largamente indipendente dal giudizio di merito sulla produttività e la qualità del lavoro dei singoli docenti.

Che si trattasse di una misura “rozza” può essere messo in evidenza dall’osservazione che un semplice esercizio condotto in pochi giorni dal biologo Cesareni dell’Università di Roma Tor Vergata permise di dimostrare che gli stessi risultati della VTR potevano essere ottenuti, con ottima correlazione, mediante un’analisi statistica dei dati sulle pubblicazioni raccolti dal motore di ricerca associato a Google Scholar.

L’obiettivo dichiarato della VQR era quello di superare i limiti della VTR grazie a un sistema di campionamento molto più ampio e soggetto a regole che permettessero di cogliere anche aspetti della ricerca accademica che erano evidentemente sfuggiti all’analisi precedente, tra cui in primo luogo la rilevanza quantitativa del fenomeno dell’“inattività” scientifica di una parte dei docenti. Questo obiettivo era già fortemente presente nel progetto del CIVR, che inizialmente prevedeva una “penalizzazione” per l’inattività ancor più forte di quella che fu poi effettivamente implementata. Nel passaggio della gestione all’ANVUR questo specifico obiettivo rimase centrale e condizionò largamente anche gli esiti finali del processo valutativo.

Un inevitabile corollario di tale scelta fu la quantificazione dell’obbligo individuale di sottomissione di prodotti di ricerca per la valutazione, nella misura di tre pubblicazioni per docente scelte (di solito dal docente stesso) su un arco di tempo di sette anni (2004-2010).

Ovviamente una procedura di revisione di una tale mole di pubblicazioni (poco meno di 200 mila) sul modello della peer review era inimmaginabile in termini di tempi e di costi, e quindi apparve quasi inevitabile il passaggio a modalità di valutazione di tipo bibliometrico, almeno per tutte quelle aree scientifiche che fossero dotate di adeguati repertori di pubblicazioni e dei relativi indici citazionali.

A prima vista questa scelta apparirebbe giustificata proprio dai risultati di Cesareni, che sembravano indicare nella bibliometria più "spinta" uno strumento in fondo non inadeguato per una stima quantitativa dell'impatto scientifico di un'istituzione di ricerca. Vedremo tuttavia che, in combinazione con i vincoli sopra indicati per la sottomissione delle pubblicazioni, tale scelta si espone a un rischio noto in statistica con il nome di sampling bias.

Si deve comunque riconoscere che l'enorme sforzo di raccolta e di elaborazione di dati associato alla VQR ha prodotto importantissimi elementi di conoscenza del sistema italiano della ricerca, anche se quasi paradossalmente si tratta soprattutto di una conoscenza di tipo quantitativo e statistico, mentre resta sospeso il giudizio sull'attendibilità delle stime VQR sulla qualità comparata della produzione scientifica di istituzioni che è difficile confrontare al netto delle differenze profonde che le caratterizzano.

Un dato particolarmente significativo, anche perché come si è visto la sua determinazione ha condizionato pesantemente la costruzione concettuale degli indicatori della VQR, riguarda la misura del grado di inattività dei soggetti operanti nel sistema della ricerca, e in particolare di quella universitaria.

Il risultato complessivo per la percentuale dei prodotti conferiti in rapporto ai prodotti attesi è superiore al 95%, un valore decisamente superiore ad alcune pessimistiche previsioni che erano circolate prima dell'esercizio VQR. Specializzando l'analisi alle singole aree disciplinari è abbastanza immediato notare che per la maggior parte delle discipline la percentuale di "attivi" è decisamente superiore alla media, e raggiunge valori del 97-98%, mentre il valore scende (forse prevedibilmente) a poco più del 90% in alcune aree, come Medicina e Giurisprudenza, caratterizzate da una forte presenza di attività professionali extra accademiche.

Un'analisi complementare alla precedente può essere effettuata disaggregando i dati in relazione alle singole istituzioni accademiche, ma anche in questo caso non si riscontra una grande variabilità: gli Atenei con più di 500 docenti hanno valori dei conferimenti che vanno da un minimo dell'89% (Messina) a un massimo del 99% (Milano Bicocca), con la maggior parte dei casi compresa tra il 92% e il 97% e con una distribuzione geografica abbastanza casuale e non suscettibile di plausibili interpretazioni socioeconomiche, ma probabilmente riconducibile soltanto alla maggiore o minor attenzione con cui le singole istituzioni si sono presentate all'appuntamento valutativo.

Esamineremo più avanti alcune implicazioni di questi risultati sulla significatività complessiva della VQR, ma non c'è comunque dubbio che si tratta del più significativo dato quantitativo che emerge dall'intero processo, e tale dato sembra sintetizzabile nell'affermazione che nel "*contesto del sistema italiano della ricerca il fenomeno dell'assenza di produzione scientifica individuale è da considerarsi nel complesso marginale"*.

Assai più ardua è l'interpretazione dei ranking risultanti dalle somme dei punteggi attribuiti ai prodotti sottoposti alla valutazione, sia quando il confronto è effettuato tra dipartimenti non sempre perfettamente omologhi per composizione disciplinare, sia quando si comparano Atenei di ben diversa dimensione, storia, contesto territoriale e vocazione disciplinare. A fronte di risultati largamente attesi emergono da un lato eccellenze "anomale" derivanti piuttosto dalle peculiarità degli algoritmi utilizzati che non da una reale e riconosciuta superiorità scientifica. Non dimentichiamo che il sistema della ricerca ha da tempo ormai immemorabile i propri, per quanto informali, criteri di valutazione e autovalutazione, e nella maggior parte dei casi è perfettamente in grado di riconoscere, anche senza far uso di specifici indicatori quantitativi, la validità o l'opinabilità di una gerarchia di merito scientifico, tra istituzioni così come tra individui. A tale proposito ci piace rinviare il lettore al bel saggio di S. Graffi "Considerazioni sulla grandezza e decadenza dei concorsi universitari in Italia" (Quaderni di Storia 71 - 2010).

Cerchiamo dunque di capire quali possono essere le principali cause di errore sistematico presenti nella procedura VQR, con l'obiettivo e l'auspicio di un loro progressivo superamento.

Nell'opinione di chi scrive, la "madre di tutti gli errori" è rintracciabile nel non aver dato corso a quanto previsto dalla Legge 1 del 2009, che istituiva l'ANPRePS (Anagrafe Nazionale dei Professori e dei Ricercatori e delle Pubblicazioni Scientifiche). Questo repertorio nazionale dei metadati dei prodotti della ricerca avrebbe dovuto rendersi disponibile nei termini previsti dalla legge, e possibilmente avrebbe dovuto costituirsi come repository delle pubblicazioni, sia pure ad accesso limitato per i prodotti non open access, (ma, in prospettiva di medio periodo, ad accesso aperto per tutte le pubblicazioni realizzate con finanziamenti pubblici, come del resto previsto da una successiva norma). Se ciò fosse avvenuto, non soltanto

l'intera procedura di valutazione sarebbe stata molto più semplice sotto il profilo organizzativo e operativo, ma soprattutto avrebbe potuto essere condotta in una forma atta a evitare il sampling bias cui facevamo cenno poc'anzi.

Questo spiacevole fenomeno, ben noto in statistica, si verifica quando per effettuare una determinata analisi si utilizza un campione non casuale ma in qualche modo preselezionato. Un esempio famoso è quello di una famosa ditta di cibi in scatola che prima di lanciare sul mercato un nuovo prodotto (fagioli stufati con costolette di maiale) fece un ampio sondaggio in un quartiere di Londra usualmente considerato un buon campione in quanto la popolazione era socialmente molto composita e a fronte di un 90% di risposte negative rinunciò alla produzione senza rendersi conto che la comunità di quel quartiere, per quanto eterogenea, era prevalentemente di origine ebraica.

Nella VQR avviene una forte preselezione, in quanto, con maggiore o minore abilità, tutti i partecipanti tentano di sottoporre i propri migliori prodotti. Ma da un punto di vista statistico non è certo la stessa cosa avere soltanto tre buoni prodotti o scegliere i migliori tra decine di buoni prodotti.

Un altro modo di vedere lo stesso problema consiste nel partire dal presupposto, certamente realistico, dell'impossibilità di ottenere dai contemporanei, con un qualunque strumento, sia esso la bibliometria o la peer review, un giudizio realmente oggettivo sulla qualità di un singolo prodotto della ricerca. Pertanto l'unica speranza di poter convertire il risultato di una valutazione quantitativa della ricerca in un giudizio qualitativo attendibile risiede nella legge dei grandi numeri, che ci assicura che la media di un campione (sufficientemente ampio) tende a convergere al valore atteso se il campione è effettivamente casuale.

Il "salto" dalla quantità alla qualità non è quindi concettualmente inconcepibile, ma richiederebbe una ben diversa metodologia, e in particolare dovrebbe fondarsi, questa volta davvero à la Cesareni, su una ricognizione dell'intero repertorio dei prodotti della ricerca nazionale, quale soltanto l'ANPRePS potrebbe consentire. Tale ricognizione potrebbe essere anche largamente automatica, perché sarebbero i grandi numeri, e non l'accuratezza del giudizio individuale, ad assicurarci la cancellazione delle maggiori fluttuazioni e a far convergere il risultato verso il valore "atteso" (nel senso statistico della parola).

Ovviamente in questa impostazione non c'è, e non ci deve essere, alcuno spazio per l'attribuzione di un peso valutativo alla presenza di inattivi. Abbiamo già visto che si tratta di un fenomeno largamente marginale, i cui effetti sulla valutazione rischiano tuttavia di essere distorsivi quando, per motivi dovuti alla peculiare composizione disciplinare o a specifiche circostanze storiche (e geografiche), si dovessero riscontrare in casi particolari dei picchi di inattività.

Nella valutazione complessiva di un'istituzione ciò che conta è quello che l'istituzione stessa è riuscita a produrre, non ciò che alcuni suoi membri non hanno fatto. Una penalizzazione dell'inattività non penalizza in alcun modo gli inattivi, che per definizione non hanno bisogno di risorse premiali, ma può danneggiare anche fortemente "attivi" eccellenti che si trovino inseriti in un contesto meno vivace della media, ma che potrebbe trarre comunque beneficio, o comunque compensazione, dal loro contributo, se valutato senza l'attribuzione di un handicap in partenza.

L'impostazione della valutazione collettiva che abbiamo qui suggerito avrebbe anche il vantaggio di rimuovere due ordini di problemi che si sono presentati con forza nel corso della VQR. Scomparirebbe l'esigenza di attribuire una "proprietà" individuale ai lavori in collaborazione, che comparirebbero comunque in maniera anonima, eventualmente con un "peso" legato al numero dei collaboratori per non sottostimare il maggior impegno complessivo insito in un lavoro collettivo. E l'anonimato del giudizio aggregato potrebbe anche attenuare fortemente, se non far scomparire del tutto, la preoccupazione avvertita da molti di una "riconversione" del giudizio collettivo in un giudizio individuale, peraltro attribuito con metodologie completamente inappropriate per un giudizio di tal genere. Il "voto" sui singoli prodotti, qualunque fosse l'algoritmo (possibilmente trasparente) con cui viene attribuito, non dovrebbe essere ricavabile in alcun modo dai risultati aggregati, nemmeno per i diretti interessati, dato il significato largamente stocastico di ciascun singolo voto.

Non ci nascondiamo che una procedura del genere sopra descritto continuerebbe a manifestare qualche criticità di tipo metodologico nelle aree cosiddette "non bibliometriche", non esistendo il parametro (numero di citazioni), per quanto opinabile, sul quale calcolare le medie. Riteniamo però che, una volta depurata dalla preoccupazione che la valutazione possa trasformarsi in un giudizio individuale, la definizione di fasce di merito potrebbe essere oggetto di un consenso più largo di quello attuale, e comunque anche nel limite in cui si decidesse, per una particolare area o settore, che non è possibile individuare parametri "oggettivi" di merito, il mero conteggio di tutti i prodotti della ricerca di un'istituzione sarebbe un primo indicatore delle sue potenzialità scientifiche, e anche in questo caso i grandi numeri potrebbero in parte compensare l'eccessiva schematicità dell'approccio.

Il sampling bias non è comunque purtroppo l'unica seria limitazione della metodologia adottata per la

VQR. Un altro grave ostacolo per una corretta interpretazione dei risultati consiste nella pretesa di formulare un ranking (tra dipartimenti della stessa area o tra istituzioni), ossia una graduatoria che preveda l'attribuzione a ciascun soggetto di una posizione caratterizzata da un numero intero progressivo. Il modo più facile per "smontare" questa modalità di presentazione dei risultati di una valutazione consiste nella semplice osservazione che il ranking riduce a una distanza unitaria intervalli di valori potenzialmente molto diversi tra loro. In una "graduatoria" potrebbe esserci una distanza trascurabile tra il primo e il quinto classificato (per di più ulteriormente ridotta quando si voglia tener conto, come si dovrebbe, dell'errore statistico insito nelle procedure di misurazione adottate), mentre la distanza tra il quinto e il sesto potrebbe essere anche macroscopica. La logica del ranking, come si vede, può stravolgere i reali valori in gioco, che implicherebbero un giudizio sostanzialmente equivalente per i primi cinque "classificati" e uno scarto importante dei primi rispetto al gruppo dei successivi.

Bisognerebbe quindi passare sempre e comunque a una logica di rating, che sulla base della distribuzione empirica dei risultati individui dei cluster (non più di quattro o cinque per categoria) all'interno dei quali non ha realmente senso operare ulteriori distinzioni. La valutazione della ricerca scientifica non è una gara sportiva che debba concludersi con un podio e un medagliere, ma una complessa procedura finalizzata in primo luogo a una miglior comprensione delle potenzialità e delle criticità del sistema, per favorire uno sviluppo per quanto possibile armonico, e non una stratificazione gerarchica che per vari motivi di sociologia della ricerca ma anche di equilibrio generale tra i territori non risulterebbe in ultima analisi vantaggiosa né per la ricerca né per il Paese.

Un altro limite tecnico riscontrabile nell'analisi dei dati della VQR deriva dal non aver tenuto nel debito conto l'effetto che le differenti dimensioni delle strutture esaminate producono inevitabilmente sulla larghezza delle distribuzioni dei risultati a causa delle fluttuazioni statistiche. Si tratta del cosiddetto "effetto imbuto", per cui la distribuzione dei valori intorno al valor medio si restringe progressivamente al crescere delle dimensioni della struttura, fino al limite per cui le strutture più grandi tendono ad assestarsi intorno ai valori medi per il semplice motivo che sono esse stesse, in larga misura, a definirli. Il confronto tra realtà di peso molto differente (sia inter-ateneo che intra-ateneo) può diventare significativo soltanto mettendo in atto adeguate procedure correttive quali la riduzione, per quanto possibile, a distribuzioni normalizzate (operazione peraltro non banale quando si tenga conto della sostanziale incomparabilità tra le procedure di valutazione di tipo bibliometrico e quelle basate sulla peer review). Rendere almeno parzialmente accettabile un tale confronto ci sembra preliminare a qualunque operazione che abbia poi obiettivi di premialità. Non è certo un caso il fatto che in molte discipline sportive la competizione si effettui soltanto all'interno di categorie relativamente omogenee.

Senza la pretesa di esaurire con questo il tema della "misurabilità" della qualità, che è oggetto di un ampio dibattito anche a livello internazionale, vorremmo segnalare almeno un altro elemento di critica e di conseguente preoccupazione: l'uso sistematico di indicatori compositi, nei quali la soggettività di chi definisce l'indicatore ha sempre e comunque un peso preponderante, in quanto non esiste e non può esistere alcun criterio oggettivo che consenta di sommare tra loro quantità incommensurabili.

Le considerazioni svolte fino a questo punto si potrebbero considerare puramente "accademiche" se i risultati della VQR fossero usati soltanto come elementi conoscitivi della realtà del sistema della ricerca. Purtroppo invece assistiamo alla pretesa, da parte di chi ha il compito di governare il sistema, di un utilizzo diretto di tali risultati ai fini della determinazione di alcuni importanti parametri adottati per la ripartizione del finanziamento pubblico al sistema universitario.

Appare comunque criticabile l'idea stessa di una valutazione che si traduca immediatamente in una formula di finanziamento, senza l'indispensabile mediazione che deriverebbe da una definizione, tutta politica, di indirizzi e di obiettivi strategici, che dovrebbero tener conto anche dell'esigenza di contenere gli squilibri territoriali e di garantire opportunità di sviluppo culturale a un'ampia fascia della cittadinanza, esigenza che passa anche attraverso la creazione e il mantenimento di una larga rete di formazione e di ricerca, senza la pretesa che tutti i nodi di tale rete si collochino agli stessi livelli di "eccellenza".

Non bisogna poi dimenticare mai il rischio di condizionamento culturale che una valutazione fortemente basata su criteri legati alla bibliometria può determinare sulle comunità scientifiche, in quanto questo tipo di criteri privilegia inevitabilmente tutte le ricerche mainstream penalizzando le ricerche di nicchia e quelle particolarmente originali, e spesso scoraggiando anche le iniziative più interdisciplinari.

Se la misurazione dell'eccellenza è così fortemente suscettibile di critiche di metodo e di merito, come abbiamo più sopra cercato di argomentare, allora è addirittura pericoloso per il sistema ancorare a tale misurazione i meccanismi della "premialità" che poi in concreto, poiché le risorse sono complessivamente limitate e spesso decrescenti, si traducono soprattutto in meccanismi di penalizzazione, con effetti poten-

zialmente devastanti su realtà che in molti casi andrebbero invece aiutate a svilupparsi positivamente.

La crescita della cultura della valutazione è certamente un fatto positivo per il sistema universitario e per il mondo della ricerca, ma una valutazione i cui esiti non siano riconosciuti e accettati dalla comunità dei valutati rischia facilmente di produrre effetti di rigetto (di cui si sono già intravvisti alcuni segnali) e rappresenta quindi un autentico pericolo proprio ai fini dell'obiettivo fondamentale, che dovrebbe essere quello di una maggior consapevolezza della responsabilità sociale di chi opera autonomamente grazie a risorse messegli a disposizione della collettività, e quindi dell'obbligo di trasparenza e di rendicontazione del proprio operato che tale ruolo, tutto sommato privilegiato, dovrebbe imporgli.

PAOLO ROSSI

(Bologna 1952) Ordinario di Fisica Teorica all'Università di Pisa dal 2000. Attivo nella ricerca dal 1976, con cinque anni di esperienza all'estero (MIT, CERN) e oltre 100 articoli pubblicati in riviste internazionali, è attualmente impegnato nel campo delle applicazioni di modelli fisico-matematici a fenomeni biologici e sociali. È stato Direttore del Dipartimento di Fisica (2003-2010), Preside della Facoltà di Scienze MFN (2010-2012), e ora è Consigliere d'Amministrazione. Dal 2007 fa parte del Consiglio Universitario Nazionale, nel cui ambito coordina la Commissione di Ricerca.
Indirizzo: Dipartimento di Fisica, Largo Pontecorvo 3, 56127 Pisa
e-mail: rossi@df.unipi.it
home page: http://www.df.unipi.it/~rossi