Scuola POLITECNICA

Conoscere la VQR

2017

a cura del Gruppo Monitoraggio e Valutazione
della Scuola Politecnica*

(*) Antonio Boccalatte, Saverio Giulini, Paolo La Barbera, Paolo Molfino, Federico Scarpa
Coordinato dal Preside, Prof. Aristide F. Massardo

L'Università è un'istituzione autonoma che produce e trasmette criticamente la cultura mediante la ricerca e l'insegnamento. Per essere aperta alle necessità del mondo contemporaneo deve avere, nel suo sforzo di ricerca e d'insegnamento, indipendenza morale e scientifica nei confronti di ogni potere politico ed economico.

(Magna Charta Universitatum, Bologna)

UNIVERSITÀ DEGLI STUDI DI GENOVA

Premessa

Al termine della presentazione da parte del gruppo di Monitoraggio e Valutazione del lavoro di monitoraggio delle prestazioni delle strutture della Scuola Politecnica nel Gennaio 2017 si era deciso di proseguire l'attività anche (pag. 115 del documento):

"…….. nell'ambito dello Stato della Ricerca, verrà sviluppata una approfondita analisi dei risultati delle due valutazioni VQR 2004-2010 e 2011-2014 per poter meglio comprendere gli aspetti positivi e negativi della nostra struttura in questo particolare esercizio. In questo compito il gruppo di Monitoraggio e Valutazione sarà affiancato dai nostri esperti attivi nei GEV delle aree 08 e 09, i professori Colombini, Franco, Solari (area 08), Bottaro, Perego e Martinoia (area 09)".

Non appena resi pubblici i risultati della seconda VQR abbiamo iniziato a riunirci e a discutere il da farsi e abbiamo deciso, dopo ampia discussione, di operare in due direzioni:

1. un'analisi delle criticità principali della procedura voluta da Anvur per la valutazione delle prestazioni scientifiche delle Università italiane con lo scopo di evidenziare da un lato gli aspetti maggiormente critici (raccolti dall'ampia letteratura) e dall'altro di suggerire correzioni che riteniamo doverose od essenziali per un migliore risultato.
2. Un'analisi di dettaglio dei risultati delle due VQR 2004-2010 e 2011-2014, che consenta, basandosi sui soli dati pubblici evitando in tal modo le problematiche dell'ottenimento dei dati dai docenti dei SSD sottonumerali, di analizzare in dettaglio le perfomance dei singoli SSD, dei Dipartimenti e confrontare le loro evoluzioni negli anni, anche in relazione ai risultati riportati nel report del Gennaio 2017.

Questa prima pubblicazione riguarda quanto riportato al punto 1, mentre una seconda pubblicazione si occupa dell'analisi dei nostri risultati, anche nella speranza che non sia troppo tardi per opportune azioni di coordinamento ed indirizzo da parte dei SSD, dei Dipartimenti e dell'Ateneo.

Come Preside della Scuola va il mio ringraziamento a tutto il Gruppo di Monitoraggio ed a colleghi dei GEV, ma uno personalissimo va a ***Federico Scarpa, Saverio Giulini*** da un lato e a ***Marco Colombini e Sergio Martinoia*** dall'altro. Leggendo il documento e partecipando alle riunioni e agli scambi di mail ho potuto capire molte cose che prima credevo ***leggende metropolitane*** o che ero portato a ***liquidare con eccessiva semplicità***.

Spero vivamente che la lettura del documento da parte di tutti i docenti della Scuola, ma forse anche di altre strutture dell'Ateneo, possa aiutare a meglio comprendere in quale "***imbuto senza ritorno***" siamo finiti.

Aristide Fausto Massardo

Premessa degli autori

A valle dell'esercizio di valutazione VQR 2011-2014, ed in particolare dopo la pubblicazione da parte di ANVUR della lista dei cosiddetti Dipartimenti "eccellenti", è apparso utile, oltre che doveroso secondo quanto deciso nel Gennaio 2017 dal Gruppo di Monitoraggio e Valutazione della Scuola come detto in Premessa, estendere un breve documento idoneo a chiarire alcuni aspetti importanti, anche tecnici, necessari per comprendere i meccanismi che hanno condotto a tale risultato.

Il presente libello, dopo una sintetica descrizione della VQR 2011-2014, effettua una rassegna delle principali critiche, anche non sempre consistenti, rivolte sulla letteratura specializzata, al nostrano esercizio di valutazione della ricerca, da studiosi qualificati in bibliometria, scientometria, valutazione e politiche per il finanziamento della ricerca.

Consapevoli che non è sufficiente conoscere i concetti di media e di varianza per sentirsi e agire come esperti di bibliometria, scientometria e valutazione, abbiamo tentato di limitare le critiche personali, riportando piuttosto una serie di opinioni sull'argomento.

Si tratta quindi di una *review* che non nasce, o quasi, dalle opinioni degli estensori, privi di specifica expertise nel campo della valutazione della ricerca.

Dal documento potranno essere tratte speriamo interessanti informazioni sui meccanismi della VQR e su possibili buone pratiche idonee a migliorare la prestazione dei singoli e delle strutture.

Infine, il documento vorrebbe stimolare sull'argomento quella discussione approfondita sul tema che, ad oggi, ancora manca nei dipartimenti, nella Scuola e forse nell'intero Ateneo, dove spesso si è limitata alla negazione della valutazione vista come strumento di "costrizione" della libertà dei docenti.

Vista infatti l'entità del finanziamento (1.3 miliardi di euro) destinato ai dipartimenti "vincitori" dell'ultima iniziativa premiale del MIUR (180 Dipartimenti che ricevono da 1.0 a 1.5 milioni di euro all'anno per cinque anni) sembrano esistere oggi le condizioni affinché tale discussione, e gli eventuali provvedimenti a seguire, abbia finalmente luogo. Certamente tutto ciò sarà fatto a livello di Scuola e di Dipartimenti della stessa, ma anche a livello si singoli settori scientifico disciplinari (SSD) e di macrosettori.

Genova, 15 luglio 2017 G.M.V. Politecnica

Nota bene

Sebbene il presente documento raccolga una serie di critiche anche severe alle modalità con cui gli esercizi VQR sono stati progettati prima e implementati poi, è d'obbligo sottolineare come questo momento di valutazione metta a disposizione una messe di dati preziosa per chiunque voglia approfondire lo stato della ricerca della propria struttura e mettere in cantiere accorgimenti confacenti ad un miglioramento delle prestazioni nel senso stimolato (in modo più o meno condivisibile) dall'azione di governo.

E' inutile nascondersi, e occorre prendere piena coscienza del fatto, che un risultato VQR scadente è indice sicuro della presenza, nel dipartimento o nell'aggregato considerato, di studiosi (ricercatori, professori) che si caratterizzano per una produzione estremamente limitata in quantità, qualità o entrambe. Di pari, risulta agli estensori inutile, ed anzi controproducente, ricorrere ad astensioni o ad altri atteggiamenti negazionisti.

Se da un lato è doveroso pretendere criteri di valutazione che, diversamente dagli attuali, siano caratterizzati da metodi fondati e da un utilizzo di strumenti adeguati, è opportuno precisare che qualunque metodo si fosse adottato, anche più scientifico ed efficiente, non avrebbe prodotto vistosi cambiamenti alle code della distribuzione, nel senso che sarebbero rimasti in ampia misura inalterati sia i punteggi molto alti sia quelli molto bassi. Le variazioni più consistenti avrebbero riguardato però le valutazioni intermedie e queste in alcuni casi possono provocare (ed hanno provocato) effetti di una certa rilevanza.

Intervenire in modo capillare per migliorare la performance VQR è un compito difficile e a volte anche sgradevole. Una critica seria verso metodi e procedure adottate, però, non deve trasformarsi nell'alibi per non agire.

Nelle pagine a seguire sono evidenziate alcune osservazioni critiche, anche propositive, riguardo la procedura valutativa sulla qualità della ricerca proposta dall'ANVUR.

Introduzione

Dopo il primo esercizio sperimentale VTR 2001-2003 (Miur, 2003), un esercizio basato su una valutazione di tipo *peer-review*, in pratica privo di ricadute sulle strutture accademiche, l'implementazione della successiva VQR 2004-2010, il cui progetto è stato formalizzato con DM 15 luglio 2011 (Miur,2011), introduce una brusca virata sia per l'introduzione di criteri bibliometrici sia in relazione all'uso dei risultati, che formeranno la base per la distribuzione agli Atenei di una importante frazione di risorse (Art.5 c.3 (e) L. 240/2010).
La VQR 2004-2010 è stata oggetto di numerose critiche da parte della comunità scientifica di riferimento per una serie di motivi che vanno dalle scelte politiche di fondo, al metodo di raccolta dei dati, all'uso combinato di indicatori di impatto dei singoli e delle riviste, ai criteri utilizzati nella costruzione degli indicatori finali utilizzati per l'erogazione dei fondi.
A fronte di queste critiche, ci si attendeva una correzione di rotta in occasione del successivo esercizio di valutazione VQR 2011-2014 che risultò invece presentare la stessa struttura di fondo del precedente.
La presente sintesi, ampiamente ispirata dal lavoro di Franceschini e Maisano del 2017, e in larga misura fondata sul lavoro di Abramo e collaboratori, presenta una panoramica delle critiche rivolte sia all'approccio metodologico sia all'implementazione pratica del più importante momento di valutazione dello stato della ricerca in Italia.

Breve descrizione della VQR 2011-2014

Come accennato, la VQR 2011-2014, il cui bando approvato il 30 luglio 2015 è stato modificato definitivamente l'11 novembre 2015 (Anvur, 2015), rappresenta la terza incarnazione della valutazione della ricerca in Italia e si fonda, nelle aree di interesse per la nostra Scuola, su una valutazione in larga misura bibliometrica dei prodotti della ricerca. La valutazione di una struttura segue come composizione di un certo numero di indicatori basati al 75% sul punteggio ottenuto dai prodotti esposti e per il restante 25% su altri parametri quali internazionalizzazione, capacità di attrarre risorse ed altro.
In questa sede ci si limiterà a considerare la valutazione di prodotti quali articoli su rivista e *proceedings* di congressi, richiamati nel seguito come "pubblicazioni" o "lavori" o "prodotti".
Per la valutazione dei prodotti esposti dagli "addetti" alla ricerca, ANVUR si è servita di 16 panel di esperti, i cosiddetti GEV (Gruppi di Esperti della Valutazione), che hanno gestito 16 aree di ricerca alcune considerate *non-bibliometriche* (scienze umanistiche, giuridiche e sociali) e le restanti *bibliometriche* (scienze, ingegneria, medicina). Ci concentreremo su queste ultime per evidenti ragioni.

La procedura prevede che ogni addetto alla ricerca selezioni un certo numero di pubblicazioni (in dipendenza da vari fattori), appartenenti al quadriennio 2011-2014. Nel nostro caso il numero di riferimento è due, almeno nella grande maggioranza dei casi. Un lavoro non potrà essere presentato che da un coautore all'interno di una assegnata istituzione. Le pubblicazioni sono scelte e proposte all'istituzione dall'addetto alla ricerca che li seleziona fra quelli presenti nel sito docente - login MIUR.

A norma del DM 27 giugno 2015 (Miur, 2015), la procedura di valutazione deve assegnare, al termine di una analisi di tipo bibliometrico o di una *peer-review*, una precisa classe di merito ad ogni pubblicazione presentata secondo la seguente scala di qualità: A-eccellente (peso 1), B-elevata (0.7), C- discreta (0.4), D- accettabile (0.1), E-limitata (0), F-non valutabile (0).

Semplificando, ogni istituzione sottomette i lavori da valutare specificando per ognuno la *subject category* , SC, più opportuna (con riferimento a Wos o a Scopus) e il GEV di riferimento.

Il processo di valutazione bibliometrica procede associando ad ogni lavoro (i) due indicatori

- C_i –il numero di citazioni raccolte ad una certa data (Wos o Scopus)
- J_i -un indice della qualità delle sede di pubblicazione, distinto per SC e anno. Il tipo di metrica, suggerito dall'addetto, è determinato dal GEV anche in dipendenza dal database di riferimento: Wos – IF e 5YIF AI, Scopus - IPP, SNIP, SJR.
  AI e SJR sono indici di prestigio, gli altri di impatto. (in casi particolari sono stati scelti altri indicatori)

Quindi si procede nel modo seguente:

1) Una volta scelto il database e la metrica, ad ogni lavoro, appartenente ad un SC e pubblicato in un dato anno, è assegnata una coppia di valori C_i e J_i.
2) Si normalizza C_i considerando il rank percentile $Fc(C_i)$ compreso fra 0 e 100% relativo alla distribuzione dei valori di C_i di tutta la produzione internazionale in quella SC e in quell'anno.
3) Si normalizza J_i considerando il rank percentile $Fj(J_i)$ compreso fra 0 e 100% relativo alla distribuzione dei valori di J_i di tutta la produzione internazionale in quella SC e in quell'anno.
4) Si posiziona la coppia normalizzata $Fc(C_i)$, $Fj(J_i)$ su apposita mappa ove sono state preventivamente individuate delle regioni idonee alla valutazione del prodotto. Tali regioni (che formano la cosiddetta "cravatta") dipendono dalla SC dall'anno e sono individuate dal GEV.

5) Per individuare tali regioni si definisce un indicatore aggregato formato mediante combinazione lineare di $Fc(C_i)$, $Fj(J_i)$

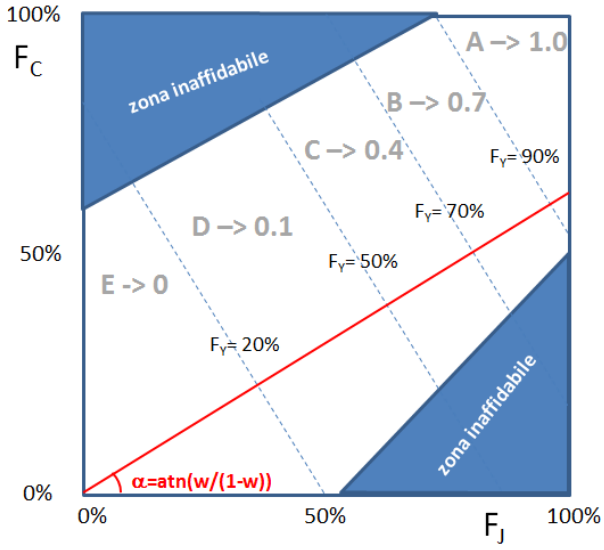
$$Yi= w* Fc(C_i) + (1-w)* Fj(J_i)$$

Con w compreso fra 0 e 1 peso variabile per dare più o meno importanza alle citazioni o alla sede di pubblicazione. E' scelto dal GEV. Dipende anche dall'anno di pubblicazione con raccomandazione ANVUR di aumentare w per i lavori vecchi e diminuirlo per quelli recenti

6) Si normalizza Y_i considerando il rank percentile $Fy(Y_i)$ compreso fra 0 e 100% relativo alla distribuzione dei valori di Y_i di *tutta* la produzione internazionale in quella SC e in quell'anno.

7) Per ogni SC e anno, la distribuzione degli Y_i rappresenta la "*distribuzione della produzione scientifica internazionale*" ed è quindi idonea ad assegnare una delle 5 classi di merito al prodotto i-esimo in funzione del suo valore di $Fy(Y_i)$

A-eccellente (peso 1) se	$0.9<Fy(Y_i)<1.0$
B- elevato (0.7) se	$0.7<Fy(Y_i)<0.9$
C- discreto (0.4) se	$0.5<Fy(Y_i)<0.7$
D- accettabile (0.1) se	$0.2<Fy(Y_i)<0.5$
E- limitato (0) se	$0<Fy(Y_i)<0.2$



8) Secondo ANVUR la valutazione ottenuta non è affidabile se l'indicatore Y_i è ottenuto in corrispondenza di valori troppo alti di $Fc(C_i)$ e contemporaneamente bassi di $Fj(J_i)$ e viceversa. In tal caso il GEV può decidere di applicare una ulteriore fase di valutazione con *peer-review* informata.

   Nota: Con riferimento specifico alle aree 09 e 08b, tutti i lavori sono stati controllati dai GEV con particolare riguardo alla correttezza della *Subject Category* e alla coerenza fra indicatore relativo all'impatto e indicatore relativo alla sede di pubblicazione. In caso di inconsistenza anche lieve si è proceduto ad una lettura attenta del lavoro. Nel documento ufficiale del GEV09 si legge infatti

   <<La valutazione dei prodotti da parte del GEV09, in armonia con quanto effettuato dai GEV1-GEV8b (e in parte dal GEV8a) oltre che dal GEV11b e dal GEV13, ha seguito la procedura della informed peer review (IPR)….

   E' infine molto importante sottolineare come i prodotti di ricerca per cui era disponibile una analisi bibliometrica non sono mai stati attribuiti automaticamente….>>

Ottenuto il peso e il voto di ciascun prodotto, sia questo derivato dalla procedura vista o attraverso *informed peer-review*, tutti i voti dei prodotti sottoposti da una istituzione vengono sommati e combinati con altri indicatori (comunque meno pesanti) per determinare il voto complessivo dell'istituzione (Miur, 2015, art.2).

In realtà, con specifico riferimento alla valutazione dei dipartimenti, è stata messa a punto una ulteriore procedura che passa attraverso il cosiddetto "voto standardizzato" ed il confronto con il "dipartimento virtuale". Se ne parlerà in seguito.

Analisi delle criticità

Nella procedura di valutazione anzi descritta sono presenti diverse fasi caratterizzate da peculiari criticità. Le esamineremo una ad una riportando i commenti più diffusi in letteratura. Di seguito gli aspetti che verranno evidenziati:

In neretto si riportano suggerimenti per possibili procedure alternative. Questi suggerimenti non sono necessariamente in accordo col pensiero di chi scrive ma sono consistenti con la critica di volta in volta sollevata.

Metodo
- Approccio misto bibliometrico/*peer-review*
- Valutazione di un numero limitato di prodotti
- Valutazione di aggregati partendo dalla prestazioni dei singoli

Raccolta dati
- Raccolta dei prodotti non automatizzata ma affidata a strutture e addetti

Indicatori (scelta)
- Uso di indicatori "per rivista" quali IF e SJR per valutare singoli articoli

Indicatori (combinazione/somma)
- Uso di una combinazione di indicatori di impatto di singolo articolo con indicatore di impatto di rivista
- Indicatore complessivo come media pesata di percentili
- Arbitrarietà delle funzioni, dei pesi e dei limiti (cravatte)
- Discrezionalità decisionale dei GEV

Indicatori per strutture (ranking, distribuzione dei fondi premiali)
- Imbuti e de-imbutizzazione
- Voto standardizzato

Metodo

1) Anvur non prende posizione netta a favore della valutazione mediante analisi bibliometrica o a favore di una valutazione *peer-review*. Cerca un compromesso che viene risolto in parte a favore di quest'ultima per le aree non bibliometriche (anche se con introduzione di scale di qualità nelle riviste) mentre per le aree bibliometriche permane un mix di metodologie il cui peso relativo dipende dal particolare GEV e altri fattori quali l'età del prodotto.

Le conseguenze negative dovute all'uso della *peer-review* nelle valutazioni a larga scala è stato evidenziato da Abramo, et al. (2013). Da questa decisione deriveranno una serie di scelte obbligate, prima fra tutte quella del ridotto numero di lavori assoggettati al vaglio valutativo.

Sebbene sia assodato che esiste una correlazione fra risultati della *peer-review* e indicatori citazionali per i singoli prodotti e pure una correlazione fra ranking ottenuti tramite *peer-review* e metodi bibliometrici a livello individuale, nel caso di valutazioni comparative su larga scala di aree bibliometriche, la superiorità dell'analisi bibliometrica in termini di accuratezza e costi è ampiamente acclarata (Abramo e al., 2011).

Con riferimento alla VQR, la asserita confrontabilità dei risultati ottenuti da analisi bibliometrica e *peer-review* dichiarata in (Bertocchi et al., 2015) è stata contestata da altri studiosi (vedi ad esempio Baccini e De Nicolao, 2017) che rilevano come gli autori appartenessero al GEV oggetto dell'indagine (area 13), e come nel caso in esame (VQR) la *peer review* sia di tipo "*informed*" e non "doppio cieco".

In altre parole il *referee* revisiona un lavoro già pubblicato ed è quindi a conoscenza dell'impatto in termini di citazioni. Il giudizio conseguente potrà verosimilmente essere influenzato da, e correlato a, tali informazioni.

Ma il problema vero è la promiscuità, la miscelazione dei due metodi, con pesi e modalità discrezionali. **Meglio sarebbe stato dividere a monte i finanziamenti in due lotti anche in funzione della numerosità delle aree rispettivamente bibliometriche (cosiddette *hard sciences*) e non bibliometriche, e consistentemente applicare due modalità differenti di valutazione** (100% biblio e 100% *peer-review*) in modo da, fra l'altro

- non mettere in competizione aree valutate in modo sostanzialmente diverso.
- non permettere l'insorgenza fra GEV e GEV (bibliometrici) di differenze dovute al diverso peso attribuito alla *peer-review*. Quest'ultimo punto come vedremo, si rivelerà particolarmente critico nella valutazione degli aggregati, in quanto porterà all'introduzione del voto standardizzato.

2) Anvur decide di valutare un autore attraverso un numero limitato di lavori**,** mediamente due per le strutture universitarie, invece che attraverso la sua produzione complessiva (ancorché in un assegnato intervallo di tempo). Questa scelta è obbligata dalla presenza dell'approccio *peer-review* del punto precedente.

Le conseguenze (che potrebbero essere anche il frutto di una precisa scelta politica) sono importanti (Franceschini e Maisano, 2017). E' infatti evidente che il numero di prodotti valutati, oltre a tutto uguale per tutte le aree e SSD,

- è troppo piccolo per una emersione completa dell'eccellenza.
- è inidoneo per una misura della performance media in termini di produttività e impatto in quanto non considera una porzione significativa di lavori e questi non sono raccolti con un metodo a campione significativo. Per analogo motivo non è in generale possibile ricavare una distribuzione attendibile della qualità dei prodotti (nel loro complesso).
- pone l'accento esclusivamente sulle criticità. Tale critica trova tutti concordi e, sebbene questo aspetto sia a volte interpretato come una precisa scelta di fondo, appare conseguenza probabile del limitato numero di lavori valutato, a sua volta

conseguenza del voler lasciare uno spazio alla valutazione *peer-review*, cosa di cui si è già detto. Una scelta politica indirizzata alla riduzione delle criticità appare sicuramente condivisibile, ma di efficacia solo parziale. Essendo l'obiettivo ultimo quello del rafforzamento complessivo della ricerca italiana nel panorama competitivo internazionale, non si vede come non possa essere altrettanto efficace uno stimolo diretto dell'eccellenza, essendo il cumulo dei prodotti o se si vuole la prestazione media, il vero output del Paese.

Al limite si potrebbe ben considerare che fra due situazioni equivalenti in senso medio sia da preferire, in un'ottica competitiva, quella che vede emergere idee eccellenti , pur a fronte di criticità serie, rispetto ad una prestazione senza lode e senza infamia. **Un'alternativa sicuramente migliore sarebbe stata quella di stabilire un numero di prodotti attesi più alto (ad esempio pari al quadruplo degli addetti) direttamente per il dipartimento, senza o con meno vincoli sui singoli autori.** Una tal soluzione avrebbe permesso di risolvere anche altre criticità come vedremo in seguito.

3) Scopo dichiarato della VQR è *la valutazione dei risultati della ricerca scientifica di istituzioni e dipartimenti***.** Di aggregati quindi. Sembrerebbe naturale posizionarsi a livello dipartimentale e procedere ad una raccolta dati già aggregata per la struttura e valutare l'aggregato piuttosto che i singoli lavori. In altre parole assegnare il numerino fatidico direttamente ai dipartimenti. Si eviterebbe in questo modo di ricorrere a espedienti per correggere e tener conto degli effetti della grande variabilità della qualità dei singoli lavori (e.g. regole ad hoc sulla forma e la pendenza delle "bande di merito" (Franceschini, 2017), da cui il nome di cravatta data al piano Fj-Fc, e alla definizione di zone inaffidabili nel medesimo piano)

   Non solo. Si nota che l'operatore f() che porta dall'indicatore della singola pubblicazione (citazioni, IF, ecc.) al voto finale (eccellente, buono…) e infine all'indicatore aggregato è ***fortemente discretizzato, non lineare e non invertibile***. E' noto che in questo caso si osserva una non trascurabile differenza se un operatore quale ad esempio E(.) (expectation) è applicato a monte o a valle del processo, ovvero E[f()] <> f(E[.]). Se applicato a valle (i.e. E[f()]) esso conserverà traccia esplicita (in forma di *bias*) della varianza dei dati iniziali**. Sarebbe quindi auspicabile (un'ipotesi da vagliare) procedere in ordine inverso: si raccolgano prima i dati per dipartimento, ad esempio, e si ottengano valori medi o integrali aggregati, e solo a valle si applichi il processo (cravatte, ecc.) che porta al voto finale.**

Naturalmente il legame fra Fy e classe di merito (pt.7) andrebbe definito in modo continuo.

Si potrebbe fissare il numero complessivo di prodotti per dipartimento (200-300 o proporzionale al numero degli addetti come già accennato) e ricavare la prestazione, quindi il voto e infine il rank, partendo dall'impatto aggregato, che non sarebbe soggetto a quella aleatorietà caratteristica dei singoli lavori e al fenomeno sopra accennato.

I vincoli sui lavori da presentare potrebbero essere più elastici e comunque a livello di aggregato, permettendo ad esempio la presenza di un autore su un ampio numero di lavori.

D'altro canto, in una procedura siffatta occorrerebbe superare le criticità legate alla confrontabilità della prestazione fra gli ssd interni al dipartimento. Cosa non semplice da implementare senza ricadere nell'errore di fondo che si cerca di superare

Raccolta dati

1) Anvur decide in modo avverso ad una raccolta dati automatizzata (direttamente dalle basi dati internazionali) e opta invece per un meccanismo che richiede l'intervento delle strutture e dei singoli docenti. Questo modo di procedere, che verosimilmente evita ad ANVUR contenziosi di varia natura, comporta un elevato tasso di errori (Abramo et al., 2014) riguardanti sia la presunta qualità del lavoro presentato (scelta di presentare un lavoro quando in realtà ve n'era un altro migliore a disposizione, con conseguente scatto di voto) che la scelta del *dataset* (Scopus o Wos) e della SC più conveniente in fase di caricamento.

   Sebbene questo punto possa apparire, e forse sia, di secondaria importanza, le simulazioni effettuate da Abramo,(2014), sui prodotti esposti da tre sedi universitarie (VQR2004-2010) mettono in luce, ad esempio, che circa un terzo di prodotti con punteggio nullo avrebbe potuto essere sostituito con altri a punteggio non nullo. La situazione è particolarmente gravosa per quella università che si è limitata a chiedere agli addetti solo i tre lavori necessari e non l'intera lista. Meno lavoro da parte dei responsabili VQR ma di certo un risultato assai deludente. **Si consiglia quindi la completa automatizzazione della raccolta e una selezione (ottimizzata) dei prodotti direttamente fra quelli esposti nel sito loginmiur.cineca. Questo comporterebbe inoltre un crollo dei costi e la possibilità di avere una VQR annuale, a finestra mobile.**

Tale procedura non parrebbe di difficile implementazione se si pensa alle modalità con cui abbiamo scelto su IRIS i lavori da presentare. Il processo potrebbe essere semplicemente automatizzato previa una semplificazione delle regole. Forse gli aspetti giuridico-legali andrebbero approfonditi per evitare il proliferare di contenziosi.

Indicatori (scelta)

1) Si è già accennato al fatto che la scelta degli indicatori da utilizzare nella valutazione di un prodotto è in parte lasciata alle strutture e in parte ai GEV con prevedibili disomogeneità di giudizio.
   Ma la critica più severa viene mossa contro la scelta di utilizzare nella valutazione della qualità di un singolo prodotto anche indicatori di impatto (una pletora) nati per la valutazione delle riviste (IF, 5YIF, AI,IPP, SNIP, SJR).
   Contro questa pratica esiste una bibliografia immensa (Seglen, 1997; IEEE, 2013). Anthony van Raan ha detto:
   "***if there is one thing every bibliometrician agrees, it is that you should never use the journal impact factor to evaluate research performance for an article or for an individual — that is a mortal sin***" (Marx and Bornmann, 2013).
   E' opinione universalmente accettata che la variabilità del numero di citazioni ricevute dagli articoli pubblicati da una determinata rivista è molto alta; ne segue che l'uso di indicatori statistici quali la media e la mediana riferiti alla rivista abbia poca rilevanza nel predire l'impatto (il numero di citazioni attese ad esempio) di un singolo lavoro ivi pubblicato. Un po' come voler prevedere l'altezza futura di un neonato guardando l'altezza media nazionale. Magari il valore atteso è quello ma la varianza è così ampia da raccomandare altri metodi (es. l'altezza dei genitori). E poco importa che il risultato della valutazione non sia quello di valutare i singoli lavori o i singoli individui ma istituzioni e aggregati (Ancaiani et al., 2015): l'uso di metriche nate per le riviste non può essere usato per valutare singoli prodotti.
   Si sconsiglia l'uso degli indicatori IF, 5YIF, AI,IPP, SNIP, SJR e similari in quanto idonei solo come indicatori statistici riguardanti le riviste e non come proxy dell'impatto atteso di un singolo prodotto.

Indicatori (combinazione)

Si è visto, punto 5 del paragrafo di sintesi VQR, come l'indicatore globale aggregato sia formato mediante combinazione lineare di $Fc(C_i)$, $Fj(J_i)$

$$Yi= w* Fc(C_i) + (1-w)* Fj(J_i)$$

Contro questo modo di procedere vengono sollevate diverse critiche

1) La prima è quella di combinare indici diversi (questo va oltre il fatto che uno di questi indici è pure usato in senso improprio come da punto precedente) nella costruzione dell'indicatore. Il parametro peso che servirà a bilanciarli, e che rappresenta la pendenza delle fasce di merito, dovrebbe derivare da considerazioni scientifiche. Di fatto, invece, è assegnato senza la copertura di un criterio uniforme fra i GEV. Infatti il metodo, proposto in (Anfossi et al, 2016), non si giustifica sulla base della letteratura corrente né è basato su un rigoroso metodo scientifico (Abramo and D'Angelo, 2016). L'effetto dell'aleatorietà di tale fattore peso, w, rappresenta un serio fattore di inattendibilità della procedura.
   Verosimilmente, risultati meno incerti e più consistenti, si otterrebbero rinunciando alla valutazione dei prodotti più recenti (e.g. ultimi due anni) usando una finestra con *lag* e, contemporaneamente utilizzando esclusivamente un indicatore bibliometrico di impatto. In questo modo troverebbe soluzione anche il problema sollevato nel punto precedente.

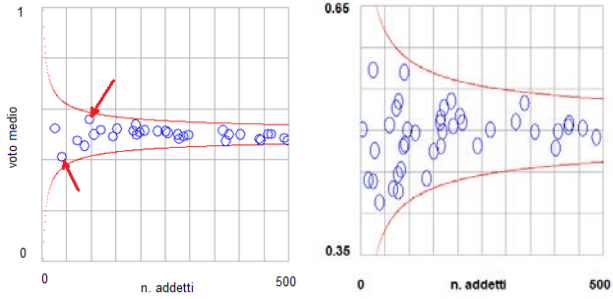
2) Non vengono miscelati direttamente gli indicatori bensì le loro controparti normalizzate in *rank* percentili; ***un'autentica sciocchezza.*** Nonostante la strenua difesa dell'algoritmo ad opera di Benedetto et al., (2016) resta il fatto che basta andare su *google* per trovare numerosi esempi in cui gli studenti delle *High Schools* vengono messi in guardia contro l'uso di tali pratiche. Si legga al proposito il lavoro di Thomson, 1993.
   Si suggerisce di abbandonare l'uso dei percentili durante il processo di costruzione del voto e, comunque, evitare combinazioni, lineari o meno, di percentili.

Indicatori per strutture e aggregati

Alla conclusione di ogni esercizio VQR vi è sempre una attesa conferenza ove i risultati più appetibili da un punto di vista mediatico vengono presentati alla stampa. E' il fatidico momento dei *Ranking*; tutti potranno leggere sulla stampa nazionale chi ha vinto o come si è piazzato l'Ateneo sotto casa rispetto alla concorrenza. E' il fascino delle classifiche. Per arrivare a questo importante momento, Anvur a messo a punto indicatori specifici, anche complessi.

1) La prima critica mossa a queste classifiche è quella di voler confrontare entità di dimensione differente. E' evidente che la classifica non potrà basarsi su un punteggio cumulato in quanto questo sarà legato alla numerosità dell'aggregato. Ma neppure utilizzare valori medi risolve il problema. Quando uscirono i primi risultati aggregati per Ateneo della VQR 2004-2010 ci si accorse che i grandi Atenei non raggiungevano mai posizioni di eccellenza, riservate invece a quelli piccoli. Il fenomeno è facilmente spiegabile da un punto di vista statistico ma anche dal buon senso: assumendo che la qualità degli "addetti" sia casuale, nei grandi atenei una buona prestazione del singolo ha maggiore probabilità di essere compensata dalla prestazione scadente di un altro addetto. La prestazione media è caratterizzata infatti da una varianza pari a quella, σ_0^2 , della distribuzione originale (qualità del singolo addetto) divisa per il numero degli addetti considerati nella media, cioè la grandezza dell'ateneo considerato.
Nella figura seguente è riportata una semplice simulazione ove di assume che la prestazione media sia 0.5 e che la deviazione standard originale σ_0 (distribuzione gaussiana) sia 0.33. Vengono riportati in rosso i limiti di confidenza per un ricoprimento del 99% la cui equazione, nel caso di distribuzione normale, risulta

$$cb(99\%) = m_0 \pm \frac{2.57\sigma_0}{\sqrt{N}}$$



Uno zoom (destra) rende più evidente il fenomeno. Si è utilizzato un generatore gaussiano.
Il fenomeno è molto noto in letteratura, grafici come quelli riportati si chiamano **"*funnel plot*"** e sono usati spesso, ad esempio in campo medico.
La contromisura adottata da ANVUR è stata quella di suddividere gli atenei in tre classi funzioni della numerosità. Questo accorgimento (che come vedremo non

afferra comunque il nocciolo del problema) oltre a rappresentare ovviamente una accorgimento dell'ultima ora, introduce ulteriori difficoltà come è facile immaginare. Per un Ateneo di numerosità prossima a quella di una soglia, lo stare da una parte o dall'altra può fare una grande differenza. Si è infatti assistito a spostamenti sospetti delle soglie fra le due VQR che hanno permesso ad alcuni Atenei di cambiare il proprio *ranking* (Baccini et al, 2017).

E' stata anche proposta la *de-imbutizzazione* dei dati, moltiplicando i voti per la radice del numero degli addetti. Secondo i proponenti questo raddrizzamento dell'imbuto permetterebbe finalmente un confronto equo fra aggregati di numerosità differente. De Nicolao mette in guardia rispetto sia alla divisione in zone che alla de-imbutizzazione ed il perché è presto detto. Nella sostanza, e questo aspetto non è stato ben evidenziato in (Abramo et al., 2015a), i *funnel plot* vengono utilizzati abitualmente per rendere conto dell'incertezza legata alla misura o alla rappresentazione di una serie di dati. In questo modo misure effettuate su campioni di numerosità diversa risultano più facilmente confrontabili. La variabilità di un dato però può anche rappresentare, per una certa quota, un reale sparpagliamento dei dati non dovuto ad incertezza o errori di misura. Il piccolo aggregato che si vede assegnato un punteggio alto può essere davvero composto da studiosi di alta qualità. La variabilità nei risultati dei piccoli aggregati non è solo dovuta ad una maggiore incertezza, ad un errore di misura non compensato nell'operazione di media, *ma al naturale e non compensato sparpagliamento della prestazione stessa*. Non è affatto detto quindi che Il risultato sia un artefatto statistico (sulla misura) da compensare matematicamente: se quel piccolo aggregato è formato da addetti davvero bravi (o scadenti) abbassare (o alzare) artificiosamente il loro punteggio è un errore. Ed è altrettanto vero che nell'ateneo di grandi dimensioni le eccellenze, sicuramente presenti, vengono compensate e nascoste dalle criticità, rendendo il punteggio prossimo al valore atteso. Ma sarebbe ingiusto castigare tale Ateneo dimenticandosi delle eccellenze.

Ciò che emerge è che la media non è idonea per confrontare la prestazione di aggregati di dimensioni diverse, specialmente se i risultati servono ad assegnare fondi**. Tali fondi potrebbero ad esempio (solo un'ipotesi da vagliare) essere assegnati attribuendoli direttamente ai prodotti eccellenti**, cosa peraltro simile a ciò che viene fatto, su scala ridotta, col recente **"**Fondo per il finanziamento delle attività base di ricerca" (Anvur, 2017a)**.** Oppure se, come alcuni dicono, la VQR è pensata per evidenziare e punire le criticità, l'assegnazione potrebbe

complementare le frazioni di prodotti accettabili, limitati e mancanti (voto 0.1 e 0.0).

Suggerimento: abolire comunque le classifiche. Non servono confronti fra gli aggregati per distribuire le risorse.

2) Gli originari indicatori pensati da ANVUR per la valutazione degli aggregati sono stati in un secondo tempo sostituiti da altri, più complessi, elaborati da G. Poggi, 2014, basati sul concetto di standardizzazione dei voti e di aggregato virtuale o specchio e già sperimentati nella VQR2004-2011. Nell'ultima VQR, il risultato di questa elaborazione sfocia nella definizione dell'indicatore ISPD per la valutazione dei dipartimenti.

La prima operazione sicuramente esposta a critiche è la standardizzazione. Ogni prestazione viene riferita a quella del settore scientifico corrispondente la cui distribuzione dei voti è descritta in termini di media e varianza. Il voto del singolo addetto quindi è standardizzato sottraendo la media di SSD e dividendo per la deviazione standard. Questa operazione si basa **sull'aprioristico assunto della equivalenza qualitativa di tutti i settori scientifici**. Tale assunto non trova riscontro nella realtà; il nostro paese, come altri, eccelle in alcuni campi mentre è solo discreto o addirittura mediocre in altri. Un'informazione vitale questa per l'identificazione di punti di forza e criticità nelle varie discipline a sostegno di una seria politica di finanziamento alla ricerca.

Ma procediamo per gradi. Mediante la elaborata procedura descritta a pag. 3 e 4, i prodotti di ogni autore sono stati messi a confronto con la produzione internazionale in quella *subject category* risultando in un indicatore (i suoi difetti li abbiamo visti) che gode almeno del pregio della non autoreferenzialità. Se il voto medio di un settore scientifico bibliometrico è basso significa che nelle SC tipiche di quel settore, quell'SSD sfigura a livello internazionale. ***Non considerare questa informazione rappresenta un'occasione perduta e una lacuna nel metodo***.

Certo, questa argomentazione è in parte indebolita dalla frazione di valutazione dovuta a peer- review, ma è irragionevole che le differenze fra i settori siano in via esclusiva imputate a differenze di *metodo (*e.g. differente peso della peer-review*) e giudizio* (soggettività) dei GEV. Certamente è una decisione politicamente comoda perché evita di distinguere esplicitamente fra ssd meritevoli e ssd non meritevoli ma, allo stesso tempo, viene meno uno degli scopi primari di ogni politica di valutazione della ricerca, ovverosia quello di ***identificare le aree disciplinari critiche e provvedere in merito con politiche adeguate.***

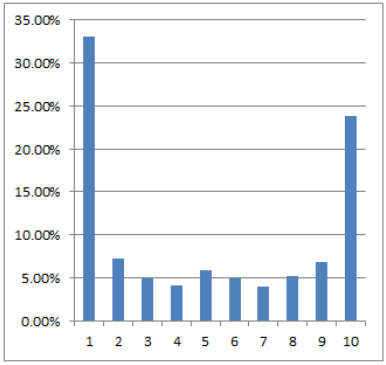
Questo a monte di ogni successiva politica di finanziamento delle discipline che dovrà sicuramente anche tenere conto della differente strategicità delle stesse rispetto allo sviluppo del Paese (Abramo et al., 2015b).

Tutti uguali invece.

Si passa poi all'applicazione tecnica del metodo con assegnazione di un indice ai dipartimenti. Il metodo del dipartimento virtuale è sicuramente idoneo a comporre le diverse realtà dei dipartimenti. Si ripresenta comunque il problema della differente numerosità degli aggregati , qui risolta esplicitamente mediante **de-imbutizzazione** (vedi nota metodologica in Anvur, 2017b). Inoltre, sfugge in genere che i risultati numerici derivanti dalla procedura non rappresentano un indicatore proporzionale alla performance e sono quindi inidonei a distribuire finanziamenti in modo equo. Per rendersene conto in modo semplice basta simulare dipartimenti composti da un unico SSD caratterizzato da prestazione distribuita uniformemente in un intervallo molto ridotto. In tali condizioni i dipartimenti mostrano una prestazione sostanzialmente equivalente e un finanziamento uniforme rappresenterebbe un'equa distribuzione di fondi.

Con il metodo Poggi invece, quello che si ottiene è un ordinamento dei dipartimenti con valori dell'indicatore linearmente variabile da zero a uno. Quindi sostanzialmente un *ranking* che spazza via la prestazione sottostante.

Anvur sembra essere al corrente di questo fatto. Nell'utilizzare l'ispd per i finanziamenti premiali ai dipartimenti stabilisce infatti una quota uguale per tutti i vincitori, un premio appunto. L'ispd funziona in questo caso come una espansore di dinamica utile a discriminare fra ***"buoni"*** e ***"cattivi"*** o se vogliamo ***"eccellenti" e "fannulloni".*** Ma i termini, a parte il richiamo mediatico, sono completamente fuorvianti; basta un nulla per precipitare di decine se non ci centinaia di posizioni, come si evince dal grafico a fianco che mostra la distribuzione dei valori di P_inf per gli 845 dipartimenti valutati nella VQR 1.0. Come si evince oltre il 33% è nel primo decile e circa il 24% nell'ultimo. Ci si aspetterebbe ragionevolmente una distribuzione della performance fatta a campana, con una consistente parte centrale dovuta ad una maggioranza senza lode e senza infamia e due code minori di eccellenti



e fannulloni. Invece viene una U, e viene così perché l'istogramma **non è *un indicatore di performance***, ed è quindi inidoneo ad una distribuzioni equa di risorse in modo proporzionale ad essa.

Suggerimento 1: abolire la peer-review nei settori bibliometrici e di conseguenza la necessità di un voto "standardizzato". I voti sono già standardizzati rispetto alla prestazione internazionale.
Suggerimento 2: abolire le classifiche. Non servono confronti fra gli aggregati per distribuire le risorse. In ogni caso, dare indicatori finali proporzionali alla performance. Cardinali, non ordinali.

Conclusioni e qualche proposta

Ne è passata di acqua sotto i ponti da quando il programma del centro-sinistra per le elezioni politiche del 2006 (pp. 236-237) affermava che

"…Occorre orientare le strategie di riforma verso:
- il miglioramento del nostro modello universitario non dualista, in cui l'integrazione tra ricerca e didattica è la caratteristica fondante di ogni ateneo e di ogni carriera docente."

Qualcosa di certo deve essere cambiato nel frattempo se in una intervista al quotidiano La Repubblica pubblicata il 4 febbraio 2012, il Prof. Sergio Benedetto, autorevole membro dell'ANVUR, dichiarava, peraltro mai smentito dal ministero o dai membri del governo:

«Tutte le università dovranno ripartire da zero. E quando la valutazione sarà conclusa, avremo la distinzione tra researching university e teaching university. Ad alcune si potrà dire: tu fai solo il corso di laurea triennale. E qualche sede dovrà essere chiusa. Ora rivedremo anche i corsi di dottorato, con criteri che porteranno a una diminuzione molto netta».

E' a partire dal 2010, dalla L.240 (Gelmini), che varie personalità e opinionisti dell'accademia, o di estrazione economica, confindustriale e della politica intervengono a favore di un superamento del modello di università generalista e integrato a favore di uno più polarizzato.
Queste tesi sono state perfettamente assimilate dal governo Renzi. L'ex premier infatti affermava esplicitamente a Torino in occasione dell'inaugurazione dell'anno accademico a.a. 15-16:

«*Ci sono già università di serie A e di serie B in Italia e rifiutare la logica del merito dentro le università e pensare che tutte siano brave è quanto di più antidemocratico vi possa essere […] Bisogna saper riconoscere il merito, non possiamo pensare di portare tutte le 90 università nella competizione globale, allora ci spazzeranno via tutti quanti*».

Il significato è chiaro: **finanziare solo un numero ridotto di poli di eccellenza**.

Questo preambolo, forse eccessivo, per ricordare che la ristrutturazione del finanziamento alle università e del sistema stesso della ricerca viene da lontano, è cominciata, non si fermerà ed è e sarà attuata di fatto **attraverso i dipartimenti**.
Una quota sempre maggiore del finanziamento alla ricerca seguirà canali diretti, i dipartimenti appunto o anche i singoli ricercatori, come il già citato "Fondo per il finanziamento delle attività base di ricerca" dimostra.
A prescindere dal fatto che tale ristrutturazione sia o meno condivisibile, è certamente della massima importanza fare sì che non ci travolga e sfruttare al massimo i meccanismi tecnici con cui viene attuata per attrezzare le nostre strutture e metterle in grado di esprimersi al meglio.
Se sono i Dipartimenti ad essere messi sotto ai riflettori è lì che occorrerà intervenire. I Dipartimenti, in tutto o in parte, andranno riprogettati forse "anche" in funzione della VQR. E' semplicemente quanto ci viene richiesto.
Come possono essere ripensati i nostri Dipartimenti? In accordo con quanto emerso dalla pur superficiale analisi dei meccanismi VQR, **una riorganizzazione dei Dipartimenti di tipo funzionale piuttosto che culturale** potrebbe risultare maggiormente adatta a conseguire risultati importanti in termini di VQR e quindi di finanziamenti. Fermo restando il fatto che gli **aspetti culturali non possono essere accantonati** semplicemente e che i risultati numerici della VQR vanno analizzati a fondo proprio per capire se e dove esistono spazi di manovra nella nostra Scuola.

Nei paragrafi precedenti si sono indicati alcuni elementi che sarebbero utili al miglioramento complessivo del processo di valutazione del sistema della ricerca in termini di equità e consistenza. Difficilmente, a ben vedere quanto è accaduto negli ultimi anni, ANVUR implementerà correzioni significative al meccanismo ormai in atto. **Non è questo il punto.** Come sottolineato nel "*nota bene*" in testa al presente documento, la VQR ha sempre comunque molto da offrire se sapremo coglierlo.

Cose da fare

- organizzare e analizzare i dati grezzi mediante messa a punto di opportuni algoritmi da validare simulando e riottenendo i risultati VQR2011-2014
- saggiare la sensibilità dei risultati dei dipartimenti (della Politecnica) rispetto ad interventi puntuali, sui singoli prodotti o addetti. Ad esempio, valutare gli effetti di differenti strategie di scelta e attribuzione dei prodotti a disposizione fra i coautori. Valutare l'effetto del miglioramento nel voto (di un gradino) nei prodotti risultati più scarsi.
- saggiare la sensibilità dei risultati rispetto ad interventi più importanti di riaggregazione dipartimentale. Ad esempio con la inaugurazione di un Dipartimento che raccolga le competenze organizzative e gestionali così preziose per la Scuola ma che "non fanno punti alla VQR".
- Valutare, sempre mediante simulazione, come accorpamenti diversi dei settori scientifici possano influire sulla performance collettiva sempre con l'intento di portare "alla Scuola" maggiori finanziamenti (premi).

Di esempi se ne possono fare altri e come sempre il lavoro del Gruppo di Monitoraggio e Valutazione vuole aprire temi di discussione sia nell'ambito dipartimentale sia nei singoli macrosettori nonché nei singoli SSD.

Ma, a prescindere dagli interventi tecnici, occorre innanzitutto una azione di sensibilizzazione culturale. Occorre dire con chiarezza

- **che non si possono più tollerare "addetti" privi di prodotti,**
- **che per soddisfare il proprio compito istituzionale di ricerca dovremmo tutti avere almeno 2 prodotti su rivista in quattro anni,**
- **che la "nostra comunità" scientifica viene danneggiata da quei "pochi" che non producono,**
- **che questo deve responsabilizzare i settori al loro interno e verso i dipartimenti e i dipartimenti nei confronti della Scuola.**

L'Ateneo si è mosso con qualche ritardo rispetto agli esercizi di valutazione VQR e, quando ormai volge al termine il terzo atto di questo processo, ancora si limita all'analisi dei risultati del secondo. **Azione meritoria e necessaria ma da noi considerata ancora insufficiente**.

Pur ritenendo che gli interventi "sul campo" debbano essere appannaggio delle strutture periferiche, è fuori di dubbio che dal centro si attendano una guida e un coordinamento maggiori.

In attesa che l'Ateneo metta in campo una serie di auspicate iniziative sul tema, non ci resta che muovere un invito a Dipartimenti e Settori ad avanzare proposte

concrete e credibili in un'ottica di responsabilità e di miglioramento globale delle prestazioni per il "bene" di tutti.

Bibliografia

Abramo, G., D'Angelo, C.A., Refrain from adopting the combination of citation and journal metrics to grade publications, as used in the Italian national research assessment exercise (VQR 2011–2014), (2016) Scientometrics, 109 (3), pp. 2053-2065.

Abramo, G., D'Angelo, C.A., Grilli, L., Funnel plots for visualizing uncertainty in the research performance of institutions, (2015a) Journal of Informetrics, 9 (4), pp. 954-961.

Abramo, G., D'Angelo, C.A., The VQR, Italy's second national research assessment: Methodological failures and ranking distortions, (2015b) Journal of the Association for Information Science and Technology, 66 (11), pp. 2202-2214.

Abramo, G., D'Angelo, C.A., Di Costa, F., Inefficiency in selecting products for submission to national research assessment exercises, (2014) Scientometrics, 98 (3), pp. 2069-2086.

Abramo, G., Cicero, T., D'Angelo, C.A., National peer-review research assessment exercises for the hard sciences can be a complete waste of money: The Italian case, (2013) Scientometrics, 95 (1), pp. 311-324.

Abramo, G., D'Angelo, C.A., Evaluating research: From informed peer review to bibliometrics, (2011) Scientometrics, 87 (3), pp. 499-514.

Ancaiani, A., Anfossi, A.F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciolfi, A., Costa, F., Colizza, G., Costantini, M., Di Cristina, F., Ferrara, A., Lacatena, R.M., Malgarini, M., Mazzotta, I., Nappi, C.A., Romagnosi, S., Sileoni, S., Evaluating scientific research in Italy: The 2004-10 research evaluation exercise, (2015) Research Evaluation, 24 (3), pp. 242-255.

Anfossi, A., Ciolfi, A., Costa, F., Parisi, G., & Benedetto, S. Large-scale assessment of research outputs through a weighted combination of bibliometric indicators. (2016) Scientometrics, 107(2), 671–683.

Anvur, 2015. Bando di partecipazione, Versione riveduta e approvata per la pubblicazione dal Consiglio Direttivo ANVUR 11 Novembre 2015,
http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf
(14.07.2017)

Anvur, 2017a. AVVISO PUBBLICO PER IL FINANZIAMENTO DELLE ATTIVITÀ BASE DI RICERCA,DI CUI ALL'ART. 1, COMMI 295 E SEGUENTI, DELLA LEGGE 11 DICEMBRE 2016 N. 232 (GU n.297 del 21-12-2016 - Suppl. Ordinario n. 57), http://www.anvur.org/attachments/article/1204/Avviso_pubblico_Procedura~.pdf (16.07.2017)

Anvur, 2017b. Nota metodologica sul calcolo dell'indicatore ISPD. http://hubmiur.pubblica.istruzione.it/alfresco/d/d/workspace/SpacesStore/a8a56378-d9f4-44cd-b0dd-7bce0d2f1b7d/Nota_metodologica_ISPD_ANVUR.pdf (15.07.2017)

Baccini A., De Nicolao G. Peer review e bibliometria non concordano. Neanche in Italia, 2017, https://www.roars.it/online/peer-review-e-bibliometria-non-concordano-neanche-in-italia/ (15.07.2017)

Baccini A., De Nicolao G. VQR: le classifiche prêt-à-porter confezionate dai GEV. And the winners are …, 2017, https://www.roars.it/online/vqr-le-classifiche-pret-a-porter-confezionate-dai-gev-and-the-winners-are/ (06.07.2017)

Benedetto S., Abate M., Armanini A., Cubelli R., Guerra G, Scanziani E, Setti G, Tramontano A, Volpe M, Zecchina R, Valutazione della ricerca, quell'algoritmo è affidabile, 2016, http://www.lavoce.info/archives/41481/valutazione-della-ricerca-quellalgoritmo-e-affidabile/ (06.07.2017)

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C.A., Peracchi, F., Bibliometric evaluation vs. informed peer review: Evidence from Italy, (2015) Research Policy, 44 (2), pp. 451-466.

Franceschini F., Maisano D., Critical remarks on the Italian research assessment exercise VQR 2011-2014, 2017, Journal of Informetrics, 11, pp. 337-357.

IEEE, Appropriate Use of Bibliometric Indicators for the Assessment of Journals, Research Proposals, and Individuals, (2013), https://www.ieee.org/publications_standards/publications/rights/ieee_bibliometric_statement_sept_2013.pdf (19.06.2017)

Marx Werner, Bornmann Luts, Journal Impact Factor: "the poor man's citation analysis" and alternative approaches, (2013), European Science Editing, 39 (3), pp. 62-63. http://www.lutz-bornmann.de/icons/declaration.pdf (19.06.2017).

Miur,2003. Decreto Ministeriale di organizzazione del processo di valutazione indicato nelle Linee guida del CIVR consultabile anche al sito http://www.civr.it, Decreto Ministeriale 16 dicembre 2003 n. 2206/RIC, http://attiministeriali.miur.it/anno-2003/dicembre/dm-16122003-n-2206ric.aspx (14.07.2017).

Miur, 2011. Decreto Ministeriale di disciplina del processo di valutazione dei risultati della Ricerca, prot. 17 del 15.07.2011, reperibile al seguente link http://www.anvur.org/attachments/article/122/vqr_d.m._n._17_del_15_07_2011_firmato.pdf (14.07.2017).

Miur, 2015. Linee guida valutazione qualita' della ricerca (VQR) 2011 - 2014. Decreto Ministeriale 27 giugno 2015 n. 458. http://attiministeriali.miur.it/anno-2015/giugno/dm-27062015.aspx (04.07.2017)

Poggi G., Il confronto basato sul Dipartimento Virtuale Associato e sul "Voto standardizzato", 2014, http://www.anvur.org/attachments/article/609/Dipartimento%20virtuale%20associato%20e%20voto%20standardizzato%20FINALE.pdf (04.07.2017).

Scopus, 2017. List of Scopus Subject Areas and All Science Journal Classification Codes (ASJC). https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/ (15.07.2017)

Seglen, P.O., Why the impact factor of journals should not be used for evaluating research, (1997) British Medical Journal, 314 (7079), pp. 498-502.

Thomson Reuiter – InCites, Anvur Cathegory Scheme. http://ipscience-help.thomsonreuters.com/incitesLive/globalComparisonsGroup/globalComparisons/subjAreaSchemesGroup/anvurCategoryScheme.html (15.07.2017)

Thomson, B., GRE 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. (1993), Paper presented at the Annual Meeting of the Mid-South Educational Research Association (New Orleans, LA, November 12, 1993), disponibile al link http://files.eric.ed.gov/fulltext/ED363637.pdf

Villa Giustiniani-Cambiaso, Genova