Proprietà degli indicatori di qualità della ricerca scientifica

Michela Gnaldi e Maria Giovanna Ranalli[1]
Dipartimento di Economia Finanza e Statistica, Università di Perugia

Riassunto. I *ranking* delle università sono misure della performance relativa delle istituzioni universitarie basate su indicatori compositi (IC). La costruzione di IC comporta l'adozione di una serie di decisioni, dalla scelta degli indicatori semplici alla selezione dei criteri di normalizzazione, a quella dei pesi etc. Queste scelte sono fonti di incertezza che possono condizionare il valore e il conseguente ordinamento delle università. Inoltre, l'università è un'istituzione complessa e condensare in un unico valore numerico il giudizio di valore su una molteplicità di attività e funzioni può essere discutibile. In questa nota, s'illustrano alcune questioni metodologiche implicate nella costruzione di IC e si valuta l'opportunità dell'impiego di IC per il *benchmark* delle università italiane attraverso un'analisi di sensibilità.

Parole chiave: Valutazione della ricerca; Indicatori compositi; Ranking; Analisi di sensibilità.

1. Introduzione

Nel corso degli ultimi decenni, gli esercizi di valutazione della ricerca si sono diffusi a livello internazionale e la valutazione della qualità della produzione scientifica è diventata una priorità nell'agenda politica di molti paesi. I criteri impiegati per valutare la qualità della ricerca possono essere classificati in due categorie principali: criteri bibliometrici e valutazioni di pari (*peer-review*). Entrambi mostrano potenzialità e limiti, tanto che gli studiosi non accordano una preferenza esclusiva ad uno dei due.

Gli indicatori bibliometrici e i giudizi di qualità degli esperti sono impiegati, insieme ad altri indicatori di qualità della didattica, per produrre *ranking* di atenei,

[1] La presente nota è stata realizzata nell'ambito del PRIN 2007, cofinanziato dal MIUR, "Modelli, indicatori e metodi statistici per rappresentare l'efficacia formativa di corsi di laurea ai fini dell'accreditamento e del miglioramento", il cui coordinatore scientifico è Luigi Fabbris. La nota è stata redatta da: M. Gnaldi per i Paragrafi da 1 a 5 e 7 e da M.G. Ranalli per il Par. 6. Le autrici desiderano ringraziare gli anonimi referee per i preziosi suggerimenti migliorativi della prima forma della nota.

ovvero graduatorie basate su indicatori compositi. I *ranking* delle università sono una prassi diffusa che influenza la percezione della qualità delle istituzioni universitarie e le priorità e le scelte di governi, mercato del lavoro, studenti e loro famiglie. Nei paesi in cui il sistema di valutazione della qualità è consolidato, e i criteri di allocazione delle risorse sono basati su criteri trasparenti, le università utilizzano i *ranking* per individuare punti di forza e debolezza e vedono nel raggiungimento di alti posizionamenti un imperativo strategico (LERU, 2010).

Tuttavia, l'impiego di *ranking* per il *benchmark* delle università presenta alcune criticità. In primo luogo, la costruzione di indicatori compositi comporta l'adozione di una serie di decisioni – dalla selezione degli indicatori semplici, alla scelta dei criteri di normalizzazione, dei pesi e dei modelli di aggregazione – che rappresentano potenziali fonti di incertezza del *ranking* e dovrebbero essere prese in considerazione quando si costruisce un indicatore composito, poiché possono condizionarne il valore e, quindi, condizionare l'ordinamento delle unità.

L'analisi della sensibilità degli ordinamenti consente di valutare quanta parte della variabilità osservata nei *ranking* può essere attribuita alle fonti di incertezza nella loro costruzione.

Un'ulteriore criticità nell'impiego di *ranking* è legata al fatto che essi definiscono la qualità delle istituzioni universitarie attraverso una scala unica e monotòna, laddove invece le università adempiono a funzioni sia didattiche che di ricerca.

In questa nota, si vuole contribuire al dibattito sulla costruzione di *ranking* per il *benchmark* delle università italiane. Dopo una descrizione dei principali strumenti impiegati per la valutazione della qualità della ricerca (Par. 2), si offre un quadro teorico di alcune delle criticità legate all'impiego di *ranking* (Par. 3). Si impiega poi un insieme di indicatori di qualità della ricerca prodotti dal Comitato di Indirizzo per la Valutazione della Ricerca (Par. 4.1) per costruire due indicatori compositi e altrettanti *ranking* degli atenei italiani. La variazione in ciascuno dei due ranking è stata analizzata tramite un'analisi della sensibilità, per esplorarne la robustezza (Par. 4.2). Alcune considerazioni di sintesi concludono il lavoro (Par. 5).

2. Valutare la ricerca tramite *peer-review* o indicatori bibliometrici

Il metodo di valutazione *peer-review* consiste nell'analisi degli output della ricerca (articoli, monografie, brevetti, etc.) svolta nelle istituzioni universitarie o negli enti di ricerca, da parte di panel di esperti selezionati dall'autorità che presiede il processo di valutazione. La soggettività del metodo di composizione dei giudizi opera pertanto nelle fasi di (Abramo *et al.*, 2008):

i. selezione degli esperti che valuteranno i prodotti;
ii. valutazione della qualità dei prodotti, nel senso che il risultato non ha il carattere dell'universalità poiché i criteri di formazione dei giudizi sono definiti in modo autonomo dai panel e, quand'anche fossero esplicitati, sono suscettibili di soggettive interpretazioni;
iii. selezione dei prodotti da sottoporre al giudizio degli esperti.

I costi e i tempi del processo di *peer-review* sono altri critici: il sistema richiede impieghi di risorse umane ed economiche ingenti e tempi di realizzazione lunghi.

L'approccio bibliometrico alla valutazione della ricerca utilizza indicatori di tipo quantitativo legati a due direttrici principali: le pubblicazioni e le citazioni. Le riviste sono valutate in base all'*Impact Factor (IF)*, ottenuto ogni anno calcolando la media del numero di citazioni degli articoli pubblicati nella rivista nei due anni immediatamente precedenti da articoli presenti nell'archivio di *Thomson Scientific*.

Gli autori, invece, sono valutati in base alla quantità delle pubblicazioni e delle citazioni. I dati sono ottenuti facendo delle *query* ad una serie di banche dati, come quella progettata da E. Garfield e attualmente gestita da *Thomson Scientific*. Una delle misure recenti proposte per misurare la qualità della produzione scientifica è l'indice *h* di Hirsch (2005): un ricercatore ottiene un valore *h* dell'indice se *h* dei suoi *N* lavori hanno almeno *h* citazioni, mentre gli altri (*N*-*h*) non hanno più di *h* citazioni ciascuno (Lazaridis, 2010).

L'assunto di base degli approcci bibliometrici alla valutazione della ricerca è che i ricercatori, specialmente quelli affiliati ad una struttura pubblica, normalmente disseminano i risultati della loro ricerca attraverso pubblicazioni in riviste scientifiche, preferibilmente internazionali e con *IF,* e dunque incluse nella banca dati di *Thomson Scientific*.

Anche la valutazione della qualità dei prodotti della ricerca attraverso indicatori bibliometrici presenta limiti, alcuni di carattere generale e altri legati ai criteri di calcolo degli indicatori. Un limite generale è che le citazioni possono essere omaggi, spesso reciproci, oppure retorici (JCQAR, 2008). Le citazioni del primo tipo sono il riconoscimento di un "debito intellettuale" nei confronti di un autore, le seconde fanno riferimento agli articoli dai quali l'autore è partito e non servono per spiegare i risultati ma per far capire il punto di partenza dello sviluppo scientifico.

Per quanto riguarda le limitazioni specifiche (JCQAR, 2008):

i. il periodo di due anni utilizzato per il calcolo dell'*IF* è opportuno solo in alcuni settori (ad esempio quello delle scienze mediche), poiché la maggior parte degli articoli sono citati immediatamente dopo la pubblicazione; tuttavia, può non essere sufficiente per altri settori; ad esempio, è stato osservato come su tre milioni di citazioni in riviste di matematica (*Math Reviews Citation database*), circa il 90% cada oltre la finestra dei due anni;

ii. le pubblicazioni sono considerate una proxy dell'output di ricerca, sottovalutando altre forme di disseminazione (saggi, monografie, rapporti tecnici, brevetti, opere d'arte), prevalenti in alcuni settori scientifici, quali le scienze sociali e quelle umanistiche. Non pare, quindi, appropriato comparare riviste in settori diversi utilizzando l'*IF*;
iii. le riviste che pubblicano articoli in lingua diversa dall'inglese tendono a ricevere meno citazioni delle altre solo perché molti ricercatori non possono leggerli;
iv. le riviste che pubblicano prevalentemente articoli di revisione della letteratura tendono a ricevere più citazioni (e ad avere più alto *IF*) delle altre, quindi non è tanto la qualità della rivista a determinarne l'*IF,* quanto la sua tipologia;
v. l'indice *h* può non essere del tutto appropriato per confrontare la produzione scientifica di autori diversi[2].

Per superare i limiti dell'*IF* sono state proposte diverse varianti e integrazioni. Per esempio, l'*AR-index* (Jin *et al.*, 2007) e il *Dynamic h-index* (Egghe 2007) introducono correzioni tali da rendere l'*h-index* dipendente dal tempo e superare il limite dell'orizzonte temporale di due anni. Il *Disciplinary IF* è impiegato da Pudovkin e Garfield (2004) per tener conto delle diverse abitudini nelle citazioni dei diversi ambiti scientifici. L'*R-Factor*, o *Research Factor*, è un indice che tiene conto delle molteplici sfaccettature dell'attività scientifica di un ricercatore, vale a dire non solo pubblicazioni, ma anche organizzazione di conferenze, coordinamento di gruppi di ricerca, partecipazione a conferenze e gruppi di ricerca, attività di editor e referee, etc. (Tucci *et al*., 2010).

Sebbene l'uso dei tradizionali indici bibliometrici si sia consolidato negli anni, il riconoscimento dei limiti di entrambi gli strumenti di valutazione della ricerca non ha condotto, fino ad oggi, ad accordare una preferenza indiscussa all'uno o all'altro, né ad una convergenza verso un modello di valutazione predominante. Nel complesso, a livello internazionale, si riscontra una tensione continua tra due obiettivi divergenti: uno volto a incrementare l'accuratezza e il dettaglio nella rappresentazione delle attività di ricerca, l'altro attento agli aspetti pratici, di contenimento della spesa e di semplificazione del processo di valutazione.

Il perseguimento di questi obiettivi si traduce nell'adozione di sistemi di valutazione della ricerca caratterizzati da diversi livelli di complessità. Da una parte, i sistemi *formula-based*, come quello statunitense che adotta un unico indicatore composito ottenuto attraverso una combinazione pesata di indicatori bibliometrici.

[2] Ad esempio, supponiamo che l'autore A abbia pubblicato 30 articoli e i 20 più citati di essi abbiano ricevuto 20 citazioni ciascuno, quindi, *h*=20. Supponiamo che l'autore B abbia anch'esso pubblicato 30 articoli e che i suoi 20 più citati articoli abbiano ricevuto 50 citazioni. Anche per B, *h*=20, ma è chiaro che il suo *Impact Factor* è più alto di quello dell'autore A.

Dall'altra, contesti, come quello anglosassone, nei quali la valutazione della qualità della produzione scientifica è rimessa al giudizio di esperti[3]. Infine, contesti in cui la valutazione è basata sull'impiego congiunto di indicatori bibliometrici e del giudizio dei pari[4].

3. I *ranking* delle università

I *ranking* delle università, e gli indicatori compositi (IC) su cui sono costruiti, integrano una notevole quantità di informazioni in un formato accessibile da una platea anche non specializzata e rappresentano pertanto uno strumento che consente di posizionare agevolmente la performance di un'istituzione universitaria nel contesto nazionale (o internazionale).

Tuttavia, l'impiego di *ranking* per il *benchmark* e l'allocazione delle risorse presenta criticità di carattere metodologico. Saisana e D'Hombres (2008) rilevano che la validità dei *ranking* dipende dalla scelta degli indicatori elementari, dall'adeguatezza delle metodologie impiegate per il calcolo, dalla trasparenza del processo e dalla robustezza delle graduatorie.

La costruzione di un IC segue una sequenza obbligata di attività: la definizione del quadro teorico di riferimento, la scelta del dataset, la sostituzione dei dati mancanti, l'analisi statistica, la scelta dello schema di normalizzazione, pesi e modelli di aggregazione e, infine, l'analisi di sensibilità della graduatoria. La normalizzazione è indispensabile quando gli indicatori semplici sono su scale di misura diverse, per consentire aggiustamenti di scala. Il sistema di pesi, che va di pari passo con il criterio d'aggregazione, può essere scelto vuoi sulla base di tecniche statistiche, vuoi in base all'opinione di esperti.

Tutte queste scelte di metodo rappresentano potenziali fonti d'incertezza che dovrebbero essere tenute sotto controllo quando si costruisce un IC, poiché una piccola variazione in uno di essi può determinare variazioni importanti nel valore assunto dall'indicatore complesso (Saisana *et al*., 2005; Munda e Nardo, 2005; Saltelli, 2007).

---

[3] Va tuttavia notato che dal 2008, in Gran Bretagna, il sistema di *peer-review* è accompagnato da uno *shadow metrics exercise* (il cosiddetto RAE HERO), ovvero un sistema di valutazione basato su un indicatore composito costruito aggregando un insieme di variabili bibliometriche.

[4] Ad esempio, in Australia gli esperti sono chiamati ad esprimere il proprio giudizio su due scale distinte, in relazione sia agli output di ricerca che ad una serie di variabili bibliometriche rilevate per gruppo di ricerca.

Ogni IC può essere considerato come un modello (OECD, 2008) in cui l'IC rappresenta la variabile risposta e le variabili esplicative sono le scelte metodologiche, esogene al fenomeno misurato e quindi fonti di incertezza spuria. L'analisi di sensibilità consente di misurare tale incertezza, ovvero di valutare quanta parte della variabilità osservata nei *ranking* può essere attribuita alle diverse fonti di incertezza. Poiché obiettivo primario dell'analisi di sensibilità è quello di quantificare l'incertezza totale nei *ranking* come risultato delle fonti di incertezza nel modello, essa può aiutare a giudicarne la robustezza e ad identificare quali unità di analisi sono favorite (o sfavorite) dagli assunti.

La seconda criticità nell'uso di *ranking* (e di IC) è legata al fatto che essi definiscono la qualità delle istituzioni universitarie attraverso una scala unica e monotòna, mentre le università adempiono tanto funzioni di didattica quanto di ricerca. Inoltre, anche all'interno delle tradizionali funzioni di didattica e ricerca, i ruoli e gli obiettivi perseguiti dalle istituzioni universitarie si sono fortemente diversificati e specializzati. Il confronto attraverso un unico indicatore composito di unità che differiscono sotto diverse dimensioni è discutibile perché condensa il giudizio su attività e missioni molto differenti in un unico valore numerico o in un'unica posizione all'interno di un *ranking*[5] (EUA, 2009; LERU, 2010).

4. Il caso di studio

4.1 Gli indicatori

L'applicazione descritta nel prosieguo della nota riguarda un insieme di indicatori di qualità della ricerca prodotti dal Comitato di Indirizzo per la Valutazione della Ricerca (CIVR) per il primo esercizio nazionale (VTR 2001-2003). Per l'esercizio, le strutture di ricerca hanno trasmesso al CIVR dati sulla produzione scientifica

[5] I limiti e le criticità legate all'adozione di indicatori compositi per il *benchmark* di istituzioni universitarie hanno sollecitato l'adozione di iniziative internazionali. Tra queste, le *Regole di Berlino*, proposte dall'International Ranking Expert Group, mirano a rendere conto della multidimensionalità delle istituzioni universitarie e la necessità di fornire chiare informazioni sulla scelta degli indicatori semplici e sui criteri impiegati per la costruzione dei ranking per garantire la trasparenza del processo. Va, inoltre, ricordata la sperimentazione di due progetti-pilota finanziati dall'Unione Europea: U-Map e U-Multirank (EUA, 2009; LERU, 2010). Il primo rappresenta un tentativo di classificazione delle università europee attraverso una mappatura delle loro attività e ha condotto all'identificazione di sei aree di attività (o dimensioni) delle università: didattica, profilo degli studenti, attività di ricerca, trasferimento di conoscenza, orientamento internazionale e rapporti col tessuto economico-sociale regionale/locale. Il secondo è un esperimento di valutazione della qualità delle istituzioni universitarie attraverso la costruzione di tanti ranking quante sono le dimensioni che le caratterizzano (le 6 aree di attività identificate dal progetto U-Map).

(articoli, brevetti, monografie, atti di convegni etc.) e dati sulla mobilità internazionale in entrata e in uscita, sui dottorandi, assegnisti e borsisti *post-doc* e sui finanziamenti di progetti di ricerca derivanti da bandi MIUR, da bandi UE, da altri soggetti e da risorse proprie. Questi ultimi, insieme con la valutazione dei prodotti, hanno composto il giudizio finale sulla struttura stessa. Il CIVR ha poi pubblicato l'insieme di queste variabili, alcune delle quali sono aggregate a livello di ateneo, altre a livello di area scientifico-disciplinare.

Ai fini dell'applicazione riportata in questa nota, gli indicatori CIVR espressi a livello di area scientifico-disciplinare sono riportati a livello di ateneo, calcolando la media aritmetica ponderata dei valori di area, con pesi dati dal numero di ricercatori ETP[6] delle 20 aree, ovvero:

$$x_{qu} = \sum_{h=1}^{H} x_{hqu} w_{hu},$$

dove x_{hqu} denota il valore del q-esimo indicatore elementare, per $q = 1, \ldots, Q$, associato alla h-esima area scientifico-disciplinare ($h = 3,\ldots, H = 20$) dell'università u-esima ($u = 1, \ldots, U = 61$) e w_{hu} è il peso in termini di ricercatori ETP dell'area scientifica h nell'università u.

Un'anomalia di molti indicatori proposti dal CIVR è la loro dipendenza dalla dimensione delle unità statistiche per le quali sono calcolate (Fabbris e Gnaldi, 2008). Gli indicatori assoluti sono quindi relativizzati, eliminando l'effetto "spurio" del numero di ricercatori che operano nelle diverse strutture, trovando indici riferiti al singolo ricercatore, ovvero:

$$x^{*}_{qu} = 100 \frac{x_{qu}}{x_{q+}},$$

dove x_{q+} è il valore medio nazionale dell'indice:

$$x_{q+} = \sum_{u=1}^{U} x_{qu} w_u$$

e w_u è il peso, espresso in ricercatori ETP, dell'ateneo u-esimo sul totale nazionale.

Gli indicatori sui quali si conduce l'analisi[7] sono 14:

---

[6] Secondo la definizione del CIVR, sono ricercatori ETP (Equivalenti a Tempo Pieno) i professori ordinari, i professori associati e i ricercatori a tempo indeterminato.

[7] Per una trattazione più ampia dei criteri di selezione dei 14 indicatori si vedano Fabbris e Gnaldi (2008).

- *Score (o punteggio) dei prodotti*. È la media aritmetica pesata dei punteggi di merito espressi dagli esperti sui prodotti selezionati: il CIVR ha dato peso 1 a quelli eccellenti, 0,8 a quelli buoni, 0,6 a quelli accettabili e 0,2 a quelli limitati e nullo a quelli non valutabili. Così costruito, l'indicatore ha la proprietà di "numero puro", variando tra 0, quando tutti i lavori presentati non sono valutabili, e 1, quando tutti i lavori presentati sono valutati come "eccellenti".
- *Numero di PRIN finanziati*. È il numero medio nel triennio 2001-2003 di PRIN finanziati; i dati sono stati ricavati dall'archivio MIUR-Cineca.
- *Percentuale di prodotti valutati eccellenti.* La *percentuale di prodotti valutati eccellenti* è ottenuta rapportando i prodotti valutati eccellenti al totale dei prodotti valutati e l'indicatore è fornito dal CIVR.
- *Percentuale di prodotti valutati almeno buoni*. Calcolato in modo analogo ai prodotti valutati eccellenti.
- *Percentuale di prodotti valutati accettabili*. Calcolato in modo analogo ai prodotti valutati eccellenti.
- *Percentuale di prodotti con Impact Factor.* La percentuale è ottenuta rapportando il numero di prodotti con IF, fornito dal CIVR, al numero di prodotti presentati.
- *Brevetti attivi all'estero.* È il numero di brevetti italiani depositati e attivati all'estero al 31/12/2003.
- *Spin-off attivati*. È il numero degli *spin-off* attivati nel triennio 2001-2003.
- *Partnership attivate*. È il numero di *partnership* che hanno originato entrate superiori a 500.000 euro per l'università.
- *Score dei brevetti*. È la media aritmetica pesata dei punteggi di merito espressi dagli esperti sui brevetti. Lo *score* è stato calcolato con un criterio analogo a quello suggerito dal CIVR per lo *score* dei prodotti scientifici, ponderando il punteggio di merito con i medesimi pesi.
- *Indice di valorizzazione economica della ricerca*. È la media aritmetica ponderata del numero di brevetti depositati all'estero nel triennio (con peso 1,5 quelli valutati eccellenti), del numero di brevetti attivi depositati all'estero al 31/12/2003 (con peso 1,5 quelli valutati eccellenti), dei ricavi ottenuti dalla vendita di brevetti o loro licenze, del numero degli *spin-off* attivati nel triennio e del numero di *partnership* che hanno originato entrate per la struttura superiori a 500.000 euro. Alle cinque voci, il CIVR ha applicato i pesi, rispettivamente, 1, 1, 2, 4 e 2.
- *Propensione al ringiovanimento dei ricercatori*. È la media nel triennio del numero di ricercatori in formazione, ovvero dottorandi, assegnisti e borsisti post-doc.

- *Capacità dei ricercatori di finanziare le ricerche*. È l'ammontare dei finanziamenti per progetti di ricerca provenienti dal MIUR, dall'Unione Europea e da altri soggetti.
- *Propensione dei ricercatori alla internazionalizzazione*. È il rapporto tra la somma del numero di anni-persona in cui i ricercatori affiliati alla struttura sono stati in mobilità all'estero (per periodi superiori a tre mesi) e quello dei ricercatori residenti all'estero, trasformati in anni-persona, che hanno operato nella struttura con contratti almeno trimestrali.

Fabbris e Gnaldi (2008) dimostrano che esistono due dimensioni latenti negli indicatori di valutazione della qualità ricerca, delle quali una esprime la qualità scientifica dei prodotti, l'altra il valore sociale ed economico della ricerca. La qualità scientifico-accademica dei prodotti della ricerca è saturata principalmente dal rating dei prodotti valutati, cioè dal giudizio di qualità dei prodotti espresso dai panel di esperti CIVR, composto dallo *scoring* dei prodotti, dalla percentuale di prodotti eccellenti, dalla percentuale di prodotti almeno buoni, da quella dei prodotti almeno accettabili, dallo *scoring* dei brevetti e dalla propensione alla mobilità dei ricercatori. Il secondo fattore è correlato, invece, ai brevetti attivi all'estero, alle imprese generate come *spin-off*, alle *partnership* attivate e ad ogni valorizzazione applicativa della ricerca *latu senso*, nonché al numero di Prin finanziati, alla percentuale di prodotti con *IF*, al ringiovanimento dei ricercatori e all'ammontare di finanziamenti per progetti di ricerca.

Nell'applicazione che segue, i due gruppi di variabili che identificano i fattori sono stati impiegati per costruire due IC, uno di qualità della ricerca scientifica e uno di valorizzazione applicativa della ricerca, e, di conseguenza, due ranking degli atenei. Le università private e le scuole speciali (Sissa di Trieste, Sant'Anna di Pisa e Normale di Pisa) sono state escluse dalle analisi poiché costituiscono gruppi a se stanti.

4.2 Indicatori compositi e analisi dei ranking

Ciascun IC dipende dalla scelta di: (a) metodo di normalizzazione degli indicatori semplici, (b) criterio d'aggregazione, (c) pesi da applicare ai singoli indicatori, (d) indicatori da usare nella costruzione dell'IC (OECD, 2008). Secondo i valori degli indicatori elementari e il metodo di combinazione degli stessi, ad ogni ateneo sarà assegnato un valore dell'IC e, di conseguenza, una posizione in graduatoria.

Per valutare la robustezza delle graduatorie ottenute sui fattori "Qualità" e "Valorizzazione", sono state esaminate varie scelte metodologiche in combinazione. Sono state combinate le seguenti scelte (Gnaldi e Ranalli (2009):

(a) cinque schemi di normalizzazione (ranking, standardizzazione, minimo-massimo, distanza dall'ateneo di riferimento, scale categoriche),
(b) due schemi di aggregazione (lineare e geometrico),
(c) il sistema di pesi proveniente dall'analisi fattoriale.

Fra i cinque schemi di normalizzazione, la standardizzazione e il metodo del minimo-massimo si configurano come trasformazioni lineari che mantengono invariata la distribuzione dell'indicatore fra i diversi atenei. La standardizzazione produce anche valori negativi, mentre il metodo del minimo-massimo produce tutti valori compresi fra zero e uno. Anche la distanza da un ateneo di riferimento si configura come una trasformazione lineare ed è calcolata come rapporto tra il valore che l'indicatore assume per un ateneo e quello che assume per un ateneo scelto come riferimento. Se, come ateneo di riferimento, si sceglie il migliore (peggiore), si avranno solo valori minori (maggiori) di uno, mentre se si sceglie l'ateneo medio o mediano si avranno valori che oscillano attorno ad uno. Il ranking, che sostituisce al valore dell'indicatore per un ateneo la sua posizione in graduatoria, elimina le differenze in termini assoluti fra gli atenei livellando eccellenze e deficienze. Questa tendenza è accentuata dall'uso di scale categoriche che raggruppano gli atenei in gruppi (cinque nel nostro caso) secondo i quantili.

In entrambi i sistemi di aggregazione scelti, i pesi sono l'espressione di rapporto di scambio fra gli indicatori: un deficit in un indicatore è compensato da un surplus in un altro indicatore. Per quanto riguarda il sistema di pesi, la letteratura offre un'ampia gamma di strumenti (OECD, 2008). In questo lavoro, ci si limita all'uso dell'analisi fattoriale. I pesi sono ricavati tramite stime di massima verosimiglianza che permettono di testare la significatività dei fattori e impiegarne, caso per caso, un numero adeguato. Questi pesi correggono, compensandolo, il ruolo di due o più indicatori correlati. Questa metodologia per la determinazione dei pesi non è adeguata quando s'intende attribuire ad ogni singolo indicatore una misura di importanza teorica o soggettiva, indipendentemente dal fatto che la dimensione latente che si sta misurando possa essere rilevata con più di un indicatore.

I cinque schemi di normalizzazione, il sistema di pesi e i due sistemi di aggregazione sono combinati producendo nove combinazioni[8] e altrettanti IC. Per ognuna di queste nove combinazioni, è calcolato l'IC con gli indicatori associati ai due fattori (6 per il fattore qualità e 8 per quello di valorizzazione della ricerca), escludendo un indicatore semplice alla volta. Dopo l'eliminazione, per la determinazione dell'IC, si aggiornano i pesi impiegando l'analisi fattoriale su, rispettivamente, 5 e 7 indicatori elementari.

[8] Si noti, infatti, che dalle dieci combinazioni originarie è stata esclusa quella tra aggregazione geometrica e standardizzazione poiché quest'ultima produce valori negativi.

L'eliminazione di un indicatore alla volta permette di analizzare il ruolo di ciascun indicatore nella performance degli atenei[9]. In questo modo si ottengono, per ciascun ateneo, 63 IC per il fattore inerente alla qualità della produzione scientifica e 81 per quello di valorizzazione applicativa della ricerca e sono create altrettante posizioni in graduatoria.

La variazione in ciascuno dei due ranking ottenuti con la costruzione degli IC è analizzata graficamente per esplorarne la robustezza. Un primo output dell'analisi ordina gli atenei in base alla loro posizione in graduatoria mediana fra tutte quelle ottenute con i diversi IC (Figure 1 e 2). La posizione mediana di un ateneo è individuata ordinando le posizioni dello stesso nelle graduatorie ottenute dalle diverse combinazioni dei fattori, prendendo la posizione che divide in due parti di uguale numerosità tale graduatoria. L'intervallo dal 5° al 95° percentile è individuato in modo analogo.

In generale, il fattore valorizzazione si dimostra complessivamente più variabile del fattore qualità e le ultime posizioni per ambedue le graduatorie sono occupate in modo stabile dal medesimo gruppo di atenei, mentre per le altre posizioni c'è una certa variabilità. Inoltre, le due graduatorie differiscono in maniera sostanziale, con atenei che, pur essendo nelle prime posizioni nella graduatoria valorizzazione, non hanno buone performance in termini di qualità scientifica. È questo il caso di alcuni politecnici. Se, infatti, si considera per ciascun ateneo la posizione mediana nelle due graduatorie, si ottiene un coefficiente di correlazione ρ di Spearman pari a 0,37, valore che indica una debole concordanza fra le due graduatorie.

Al fine di confrontare la posizione di ogni ateneo nelle due diverse tipologie di graduatorie, si riportano nella Fig. 3 gli atenei ordinati per graduatoria media ottenuta usando i 14 indicatori semplici (Gnaldi e Ranalli, 2009). Il valore riportato, quindi, è simile a quelli delle Figure 1 e 2, solo che, invece della mediana, si riporta la media e la graduatoria si basa su tutti gli indicatori invece che sui due sottoinsiemi. Per ciascun ateneo, quindi, si stimano le mediane dei due sottoinsiemi e la media di tutti gli indicatori. I tre punti di ciascun ateneo sono uniti da una linea per facilitare la lettura del grafico. Atenei con una performance analoga nei due fattori avranno una linea più corta di quella degli atenei con una performance piuttosto diversa.

Si nota che i primi cinque e gli ultimi otto atenei hanno una performance abbastanza stabile secondo tutti i fattori. Altri atenei (Trieste, Torino, Chieti-Pescara, Venezia Ca' Foscari, Siena e Perugia Stranieri, Napoli Orientale e Roma IUSM) presentano, invece, una buona prestazione in termini di qualità cui non corrisponde un'altrettanto buona posizione in termini di valorizzazione economico-produttiva della ricerca. Presentano un comportamento opposto i politecnici di Torino, Milano, Marche e Bari e le università di Cagliari, Firenze, Perugia, Brescia e Catania.

[9] È possibile anche eliminare due o più indicatori per volta, quando si ritenga importante valutare il ruolo di gruppi di indicatori.

Figura 1. *Ranking dei punteggi del fattore "qualità della ricerca" dei 61 atenei pubblici, escluse le scuole speciali, e intervalli del 90% di variazione intorno al valore mediano*

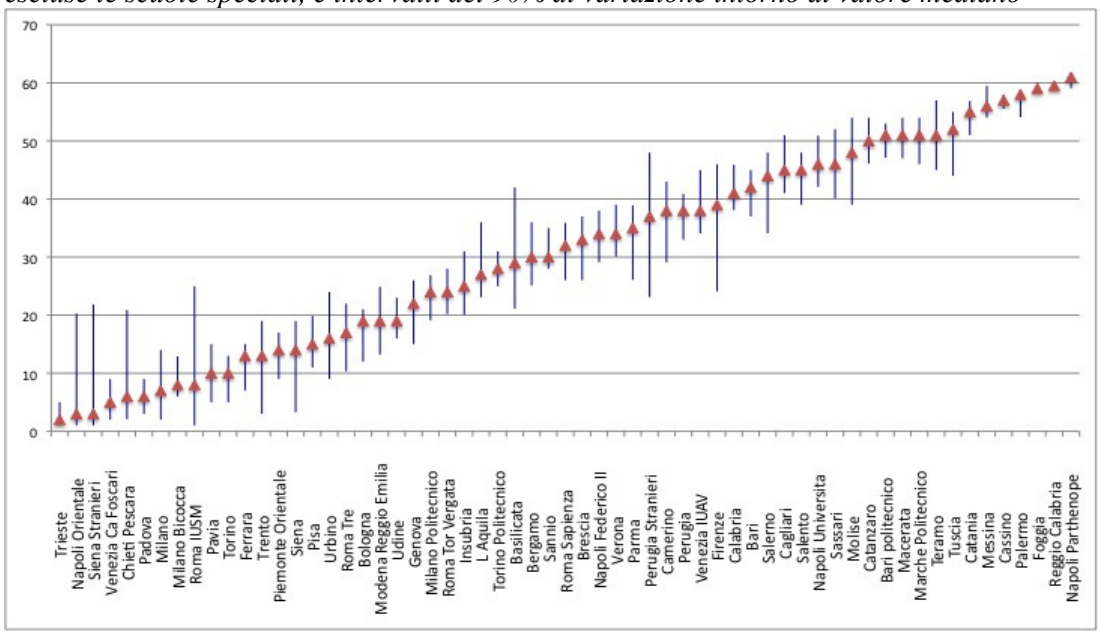


Figura 2. *Ranking dei punteggi del fattore "valorizzazione della ricerca" dei 61 atenei pubblici, escluse le scuole speciali, e intervalli del 90% di variazione intorno al valore mediano*

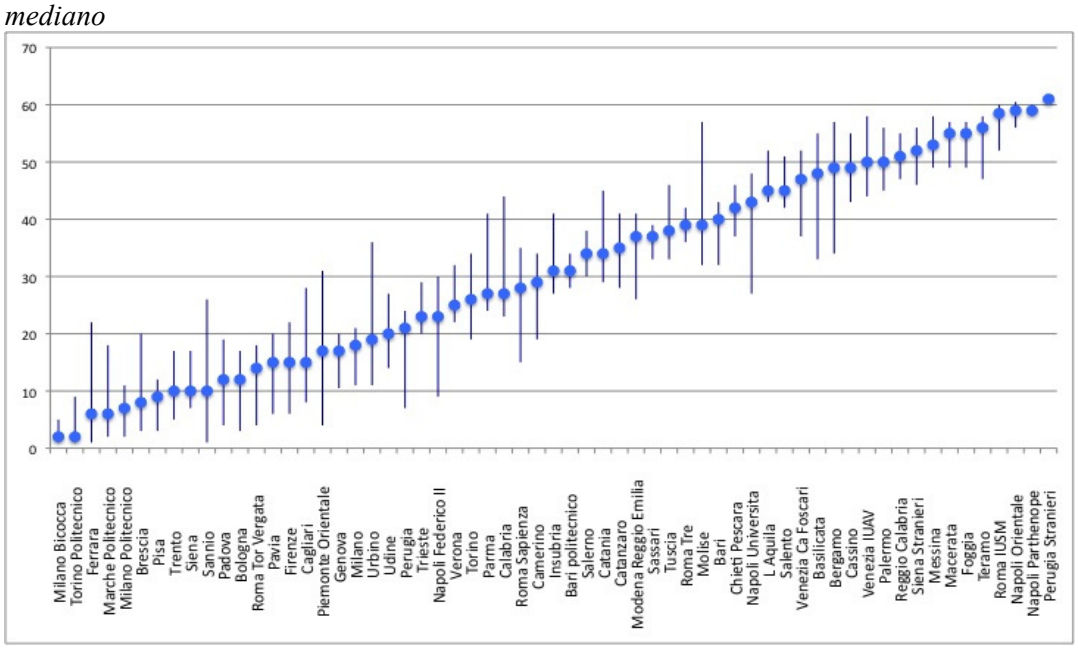
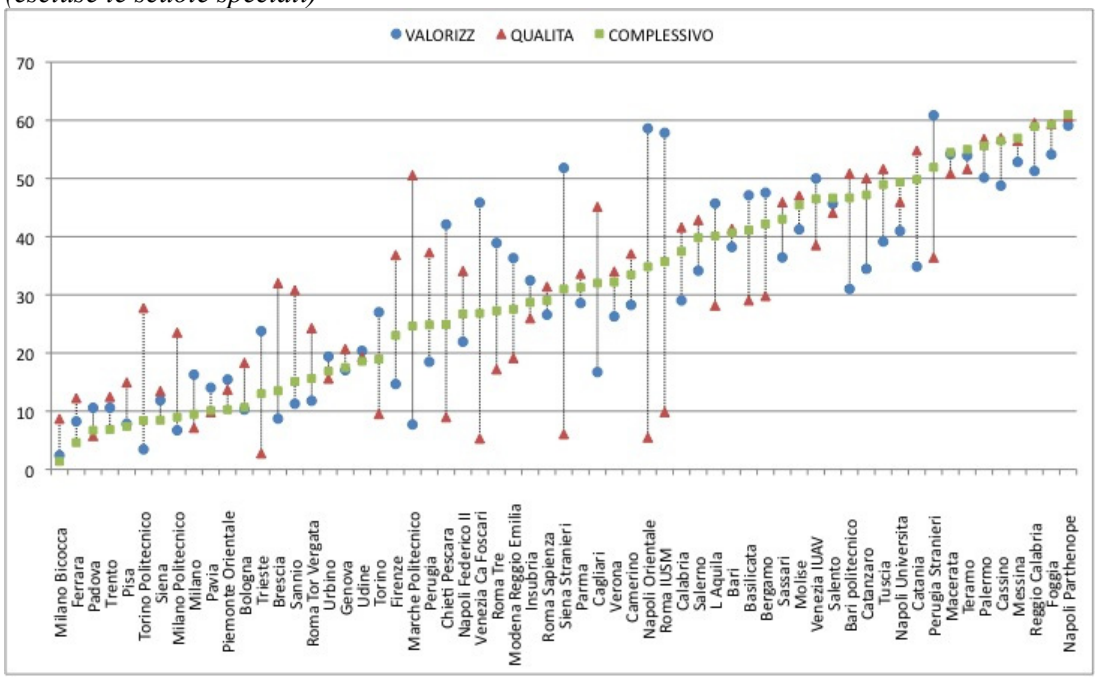


Figura 3. *Ranking medio dei 61 atenei pubblici secondo tutti gli indicatori, secondo gli indicatori del fattore "valorizzazione della ricerca" e del fattore "qualità della ricerca" (escluse le scuole speciali)*



La Fig. 3 fornisce uno strumento di analisi della dipendenza limitata alla media fra le graduatorie ottenute nei *setting* rappresentati dai tre sottoinsiemi di indicatori. Nella Tab. 1 quest'analisi è estesa alla dipendenza in distribuzione. Per ciascun ateneo (in ordine alfabetico) e fattore (qualità – valorizzazione) si riportano le distribuzioni percentuali per classi di posizione in graduatoria. La frequenza percentuale più alta in ciascuna riga (distribuzione) è evidenziata in grassetto. Per ciascun ateneo, distribuzioni simili (diverse) indicano una performance simile (diversa) secondo i due fattori considerati.

Tabella 1. *Distribuzione percentuale per ateneo delle posizioni in graduatoria secondo le classi di rank e i due fattori analizzati (sono omesse le frequenze pari a zero; in grassetto i valori maggiori per ciascuna riga)*

Ateneo	**Fattore**	**Rank 1—5**	**6—10**	**11—15**	**16—20**	**21—25**	**26—30**	**31—35**	**36—40**	**41—45**	**46—50**	**51—55**	**56—61**
Bari	Qualità							3	27	**67**	3		
	Valorizz.						4	23	30	**43**			
Bari	Qualità									2	43	**56**	

Ateneo	**Fattore**	**Rank 1—5**	**6—10**	**11—15**	**16—20**	**21—25**	**26—30**	**31—35**	**36—40**	**41—45**	**46—50**	**51—55**	**56—61**
Politecnico	Valorizz.					2	37	**60**					
Basilicata	Qualità				2	**38**	24	14	14	8			
	Valorizz.					4		4	10	15	28	**35**	5
Bergamo	Qualità					6	**60**	25	8				
	Valorizz.					4		5	5	16	**32**	22	16
Bologna	Qualità	2	2	14	**62**	21							
	Valorizz.	**36**	10	31	23								
Brescia	Qualità				2	2	30	**43**	24				
	Valorizz.	**37**	32	23	4	4							
Cagliari	Qualità								3	**57**	32	8	
	Valorizz.		14	**42**	15	21	5	1	2				
Calabria	Qualità							2	21	**71**	6		
	Valorizz.					**42**	32	9	5	9	4		
Camerino	Qualità						8	22	**49**	21			
	Valorizz.			4	2	10	**53**	31					
Cassino	Qualità											3	**97**
	Valorizz.									12	**60**	25	2
Catania	Qualità										5	**52**	43
	Valorizz.						10	**63**	12	11	4		
Catanzaro	Qualità									2	**65**	33	
	Valorizz.					1	26	23	**40**	10			
Chieti - Pescara	Qualità	**35**	30	16	13	5	2						
	Valorizz.							2	9	**80**	9		
Ferrara	Qualità	2	17	**81**									
	Valorizz.	**48**	20	9	17	2	4						
Firenze	Qualità				2	11	8	11	**41**	17	10		
	Valorizz.	5	25	21	16	**33**							
Foggia	Qualità												**100**
	Valorizz.										10	**52**	38
Genova	Qualità		3	6	24	**59**	8						
	Valorizz.		5	21	**74**								
Insubria	Qualità				6	**44**	37	11	2				
	Valorizz.						**47**	27	17	7	1		
L'Aquila	Qualità					29	**48**	16	8				
	Valorizz.									**57**	37	6	
Macerata	Qualità										38	**60**	2
	Valorizz.									1	12	**48**	38
Marche Politecnico	Qualità									5	33	**60**	2
	Valorizz.	**43**	28	22	6								

Ateneo	**Fattore**	**Rank 1—5**	**6—10**	**11—15**	**16—20**	**21—25**	**26—30**	**31—35**	**36—40**	**41—45**	**46—50**	**51—55**	**56—61**
Messina	Qualità											37	**63**
	Valorizz.										20	**67**	14
Milano	Qualità	37	**44**	17	2								
	Valorizz.		2	37	**49**	11							
Milano Bicocca	Qualità	3	**78**	17	2								
	Valorizz.	**99**	1										
Milano Politecnico	Qualità			2	10	**73**	16						
	Valorizz.	22	**69**	9									
Modena - Reggio Emili	Qualità			16	**43**	38	3						
	Valorizz.					4	4	26	**60**	6			
Molise	Qualità							2	14	21	**32**	**32**	
	Valorizz.							16	**37**	23	9	7	7
Napoli Federico II	Qualità						11	**57**	32				
	Valorizz.		7	4	16	**44**	28						
Napoli Orientale	Qualità	**78**	6	3	6	5	2						
	Valorizz.											2	**98**
Napoli Parthenope	Qualità												**100**
	Valorizz.												**100**
Napoli Università	Qualità								2	**46**	**46**	6	
	Valorizz.				4		6	9	20	**36**	26		
Padova	Qualità	37	**63**										
	Valorizz.	**38**	9	25	28								
Palermo	Qualità											25	**75**
	Valorizz.									12	38	**40**	10
Parma	Qualità					2	22	**38**	37	2			
	Valorizz.				1	20	**52**	17	4	6			
Pavia	Qualità	8	**56**	32	5								
	Valorizz.	2	20	30	**43**	2	2						
Perugia	Qualità						3	21	**70**	6			
	Valorizz.		16	6	26	**48**	4						
Perugia Stranieri	Qualità				2	6	16	19	**27**	16	11	3	
	Valorizz.												**100**
Piemonte Orientale	Qualità		17	**48**	35								
	Valorizz.	6	**35**	7	27	11	7	5	1				
Pisa	Qualità		3	**62**	33	2							
	Valorizz.	**40**	22	38									
Reggio - Calabria	Qualità												**100**
	Valorizz.										38	**60**	1
Roma	Qualità	**43**	17	16	11	8	5						

Ateneo	**Fattore**	**Rank 1—5**	**6—10**	**11—15**	**16—20**	**21—25**	**26—30**	**31—35**	**36—40**	**41—45**	**46—50**	**51—55**	**56—61**
IUSM	Valorizz.									4		7	**89**
Roma Sapienza	Qualità					3	30	**60**	6				
	Valorizz.			6	11	14	**49**	17	2				
Roma Tor Vergata	Qualità				6	**65**	29						
	Valorizz.	7	35	**42**	14	2							
Roma Tre	Qualità	2	5	16	**60**	17							
	Valorizz.							2	**80**	17			
Salento	Qualità							3	14	33	**49**		
	Valorizz.								2	**56**	33	9	
Salerno	Qualità						2	8	14	**49**	27		
	Valorizz.					1	7	**64**	27				
Sannio	Qualità						**52**	44	3				
	Valorizz.	**31**	22	14	20	7	4		2				
Sassari	Qualità						2		5	38	**43**	13	
	Valorizz.						2	31	**67**				
Siena	Qualità	6	17	**40**	37								
	Valorizz.		**52**	26	22								
Siena Stranieri	Qualità	**68**	14	5	5	6	2						
	Valorizz.									5	31	**53**	11
Teramo	Qualità									8	**35**	33	24
	Valorizz.										23	25	**52**
Torino	Qualità	10	**46**	44									
	Valorizz.				6	30	**35**	30					
Torino Politecnico	Qualità				2	10	**76**	13					
	Valorizz.	**79**	21										
Trento	Qualità	13	22	29	**35**	2							
	Valorizz.	6	**51**	32	11								
Trieste	Qualità	**95**	5										
	Valorizz.				7	**75**	15	2					
Tuscia	Qualità									8	19	**73**	
	Valorizz.							23	**47**	17	9	4	
Udine	Qualità		2	3	**59**	37							
	Valorizz.			23	30	**31**	16						
Urbino	Qualità		16	30	**38**	16							
	Valorizz.		4	**30**	25	25	7	4	6				
Venezia. Ca' Foscari	Qualità	**62**	37	2									
	Valorizz.						2	1	10	25	**51**	11	
Venezia IUAV	Qualità							19	**59**	17	5		
	Valorizz.									**30**	25	27	19

Ateneo	**Fattore**	**Rank 1—5**	**6—10**	**11—15**	**16—20**	**21—25**	**26—30**	**31—35**	**36—40**	**41—45**	**46—50**	**51—55**	**56—61**
Verona	Qualità						6	**68**	21	5			
	Valorizz.				1	**51**	41	2	4	1			

L'analisi dei ranking evidenzia gli atenei che sono in situazioni di particolare eccellenza o deficienza. Ha una variabilità elevata l'Ateneo di Cagliari, che mostra una buona performance in termini di valorizzazione (si posiziona fra l'11° e il 15° posto il 42% delle volte) ma anche, nell'8% dei casi, una posizione peggiore della 26°. Analizzando per quali combinazioni dei fattori questo accade[10], si nota che la performance peggiore si ha quando si esclude dagli indicatori semplici l'indice di valorizzazione della ricerca, e ciò indica che la valorizzazione della ricerca è un punto di forza dell'Ateneo. Se si esclude l'attivazione degli *spin-off*, la performance dell'Ateneo di Cagliari migliora, arrivando tra le prime dieci. Ciò indica che gli *spin off* sono un suo punto debole.

5. Considerazioni di sintesi

In questa nota, attraverso una selezione d'indicatori di qualità della ricerca prodotti dal CIVR abbiamo costruito due indicatori compositi e due *ranking* degli atenei italiani, uno di qualità della ricerca scientifica e l'altro di valorizzazione applicativa della ricerca. La variazione in ciascuno dei due *ranking* è stata analizzata con un'analisi di sensibilità per esplorarne la robustezza. Entrambe le graduatorie si dimostrano sensibili a variazioni nella scelta degli indicatori semplici, dei criteri di normalizzazione e di quelli di aggregazione. Tuttavia, la graduatoria costruita sul fattore *valorizzazione* è stata complessivamente più variabile di quella costruita sul fattore *qualità*. Le due graduatorie differiscono, anche notevolmente, con i politecnici e con altri atenei che, pur presentandosi ai primi posti nella graduatoria della valorizzazione, hanno performance minori in termini di qualità. Ciò costituisce una conferma empirica della non adeguatezza di valutazioni basate su un unico indicatore composito che finisce col fotografare una performance media, nascondendo deficit e surplus in altre dimensioni.

Le ultime posizioni, in ambedue le graduatorie, sono occupate dai medesimi atenei. Il confronto grafico tra le posizioni medie in graduatoria degli atenei secondo i fattori *qualità* e *valorizzazione* ha confermato che gli atenei che occupano le prime

[10] I dati dettagliati per ateneo non possono essere riportati per motivi di spazio, ma sono resi disponibili su richiesta.

e le ultime posizioni mostrano una performance coerente (secondo tutti i fattori), mentre gli atenei che occupano le posizioni centrali nella graduatoria sono più suscettibili al metodo di misura della performance.

L'analisi delle distribuzioni percentuali per classi di posizione occupate in graduatoria dagli atenei ha inoltre permesso di evidenziare situazioni di eccellenza (e grave deficienza) ed atenei che presentano performance molto diversificate in funzione dei criteri di calcolo dell'IC. Nell'insieme, se si considera il fattore *qualità*, la posizione in graduatoria di solo 38 atenei su 61 rimane stabilmente (in almeno il 50% dei casi) nella medesima classe di *ranking*; invece, se si considera il fattore *valorizzazione*, gli atenei che occupano una posizione stabile scendono a 30. Ne consegue che (i) tanto per il fattore *qualità*, quanto (e in misura maggiore) per il fattore *valorizzazione*, le posizioni in graduatoria di molti atenei sono molto dipendenti dalle scelte metodologiche adottate per la costruzione dell'indicatore composito; (ii) la stabilità nelle posizioni occupate in graduatoria da circa la metà delle università considerate non è sufficiente per poter dare giudizi sulla performance di questi atenei.

Nel complesso, data la forte sensibilità riscontrata in ambedue i *ranking*, si ritiene che essi siano uno strumento grezzo di informazione per il *policy maker* sulla posizione relativa occupata dalle istituzioni universitarie, poiché la posizione dipende in modo rilevante dalle scelte di metodo assunte nella fase di costruzione del ranking.

In secondo luogo, poiché le scelte metodologiche influiscono sulla stabilità delle graduatorie, è opportuno associare ad ogni *ranking* una misura di sensibilità che ne evidenzi la robustezza, se si vogliono trarre indicazioni consapevoli sulle differenze osservate tra le performance delle università. A tal fine, l'impiego di rappresentazioni grafiche o di distribuzioni percentuali per classi di posizione occupate in graduatoria dagli atenei – simili a quelle riportate nella presente nota – può risultare utile per dare un'idea della robustezza dell'indicatore composito costruito.

Riferimenti bibliografici

ABRAMO G., D'ANGELO C.A., PUGINI F. (2008) The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology, *Scientometrics,* **76(2)**: 225–244

EGGHE L. (2007) Dynamic h-index: The Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, **58(3)**: 452–454

EUROPEAN UNIVERSITY ASSOCIATION (EUA) (2009) *Institutional Diversity in European Higher Education. Tensions and Challenges for Policy Makers and Institutional Leaders*: http://www.eua.be/Libraries/Publications/Institutional_Diversity_in_European_Higher_Education.sflb.ashx

FABBRIS L., GNALDI M. (2008) Indicatori di valutazione della qualità della ricerca negli atenei: sensibilità, sostituibilità e capacità discriminatoria. In: FABBRIS L., BOCCUZZO G., MARTINI M.C. (a cura di) *Professionalità nei servizi innovativi per studenti universitari*, CLEUP, Padova: 139-171

GNALDI M., RANALLI M.G. (2009) Composite indicators of scientific research, *Quaderni di Statistica*, Liguori Editore, **11**: 165-181

HIRSCH J.E. (2005) An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences of the United States of America*, **102(46)**: 16569-16572

JIN B.H., LIANG L.M., ROUSSEAU R., EGGHE L. (2007) The R- and AR-indices: Complementing the h-index, *Chinese Science Bulletin*, **52(6)**: 855–863

JOINT COMMITTEE ON QUANTITATIVE ASSESSMENT OF RESEARCH (JCQAR) (2008) *Citation Statistics*, A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS): www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf

LAZARIDIS T. (2010) Ranking university departments using the mean h-index, *Scientometrics*, **82:** 211–216

LEAGUE OF EUROPEAN RESEARCH UNIVERSITIES (LERU) (2010) *University Rankings: Diversity, Excellence and the European Initiative*, Advice Paper NR.3: http://www.leru.org/files/publications/LERU_AP3_2010_Ranking.pdf

MUNDA G., NARDO M. (2005) *Constructing Consistent Composite Indicators: the Issue of Weights*, Joint Research Centre, Ispra

OECD (2008) *Handbook on Constructing CIs – Methodology and User Guide,* (http://composite-indicators.jrc.ec.europa.eu/Handook.htm)

PUDOVKIN A.I., GARFIELD E. (2004) Rank normalized impact factor: A way to compare journal performance across subject categories. In: *Proceedings of the 67th Annual Meeting of the American Society for Information Science & Technology*, **41**: 507-515

SAISANA M., D'HOMBRES B. (2008) *Higher Education Rankings: Robustness Issues and Critical Assessment*, Joint Research Centre, Ispra

SAISANA M., TARANTOLA S., SALTELLI A. (2005) Uncertainty and sensitivity techniques as tools for the analysis and validation of CIs, *Journal of the Royal Statistical Society, Series* A, **168**: 307-323

SALTELLI A. (2007) Composite indicators between analysis and advocacy, *Social Indicators Research*, **81**: 65-77

TUCCI M.P., FONTANI S., FERRINI S. (2010) L'*R-Factor*: un nuovo modo di valutare l'attività di ricerca, *Studi e Note di Economia,* **XV(1)**: 103-140

Properties of Scientific Research Quality Indicators

Summary. *University rankings and Composite Indicators (CIs) are useful tools in policy analysis and public communication. However, rankings are subject to a plethora of criticism. CI development involves stages where a number of judgements have to be made, from the selection of individual indicators, to the choice of normalisation methods, weighting schemes, aggregation models etc. All these choices are potential sources of uncertainty which should be addressed when constructing a CI because they can have a significant impact on the composite index and rankings of the individual units considered. Besides, university is a multidimensional phenomenon, which makes it difficult to condense the diversified work going on within universities into a single score or ranking. In this work - after a revision of the main university research evaluation instruments - we address some of the methodological challenges implied in the construction of CIs and investigate the degree to which composite measures are an appropriate metric for evaluating and ranking the research performances of Italian universities.*

Keywords*. Research quality assessment; Composite Indicators; Ranking; Sensitivity analysis.*