Quale profilo per gli statistici italiani?

Silvia Salini, Francesca De Battisti
Silvia.salini@unimi.it, Francesca.debattisti@unimi.it

Il crescente interesse nei confronti del tema della valutazione in tutti gli ambiti della vita pubblica ha coinvolto anche la produzione scientifica di docenti universitari e ricercatori, spesso misurata tramite gli indicatori bibliometrici.

Viene dunque naturale interrogarsi riguardo alla statistica: quali sono i database da cui è possibile trarre informazioni sulla produzione scientifica degli statistici italiani e quali sono le caratteristiche delle diverse fonti di dati, in termini di coerenza, di correlazione tra gli indicatori ottenuti o di possibilità di sintetizzare mediante un indice opportuno le informazioni. Cercheremo di dare una risposta a queste domande.

Abbiamo considerato le seguenti fonti di dati:

1) Current Index to Statistics, American Statistical Association and Institute of Mathematical Statistics (http://www.statindex.org/) contiene un indice bibliografico relativo a pubblicazioni in statistica, probabilità e argomenti affini (CIS).
2) Web of Science, Institute for Scientific Information e Thomson Reuters (distribuzione) (http://WoSwebofknowledge.com/)contiene un repertorio bibliografico commerciale con una copertura non estensiva ma selettiva delle riviste (e di altre fonti bibliografiche) più rilevanti (WOS).
3) Scopus, Elsevier (www.info.Scopus.com) contiene un repertorio bibliografico commerciale con un carattere più estensivo rispetto all'iniziativa WOS in quanto include più riviste (SCO).
4) Google Scholar, versione per la ricerca scientifica di Google, contiene vari tipi di prodotti di ricerca, è più ricco delle fonti dati commerciali citate in precedenza; tuttavia presenta una qualità dei dati decisamente peggiore; un'interfaccia consigliabile per l'interrogazione di Google Scholar, che permette un'appropriata pulizia dei dati di partenza, è Publish or Perish, sviluppato da Anne-Wil Harzing (http://www.harzing.com/PoP.htm) (POP).

La lunga e laboriosa attività di interrogazione ha messo in luce i punti di forza e di debolezza di ciascuna fonte; una breve sintesi è riportata in Tabella 1.

Tabella 1

	Punti di forza	Punti di debolezza
CIS	- circoscritta - copertura temporale dal 1974	- non gratuita - non aggiornata - criteri di selezione delle riviste non rigorosi - interrogazione solo mediante cognome, con conseguenti errori di omonimia
WOS	- aggiornata - criteri di inclusione severi - interrogazione mediante filtri per campi, per affiliazione, per tipologia di lavoro	- non gratuita - periodo di copertura: dipende dalla licenza acquistata (per l'Università di Milano si parte dal 1990) - interrogazione per cognome e iniziale - problemi di omonimia - problemi legati all'affiliazione
SCO	- parzialmente gratuita - copertura temporale dal 1970 - aggiornata - vantaggi operativi:	- copertura temporale delle citazioni dal 1996 - problemi di omonimia e di errato matching fra autore e affiliazione

	automaticamente fornisce la storia dell'affiliazione di ogni autore	
POP	- gratuita - criteri di inclusione: tutto ciò che è presente sul Web - copertura temporale illimitata - più esteso delle altre banche dati	- non è un database - produce dati di pessima qualità

La diversa struttura delle varie fonti fa supporre che si possano individuare situazioni diverse per lo stesso studioso. La nostra analisi si propone di valutare la coerenza delle informazioni ottenute. L'analisi è stata effettuata su tutti i 444 ricercatori e docenti di statistica del settore scientifico disciplinare SECS/S01 a livello nazionale[1], per ogni autore sono stati rilevati: il numero di pubblicazioni, il corrispondente periodo temporale e, dove disponibili, il numero totale di citazioni e il valore dell'indice *h* (Hirsch Index[2]). Per 29 autori non è stato possibile costruire il record corrispondente (a causa di POP) mentre per altri 13 (SECS/S-01) si è riscontrato il valore 0 per tutte le variabili considerate nelle banche dati. È stata poi applicata una procedura di analisi per l'individuazione di outlier multivariati (Filtzmoser et all. 2005, 2008[3]), ovvero record per i quali i valori delle diverse variabili presentavano anomalie, cioè incoerenze fra i diversi database. In particolare sono stati isolati 23 ulteriori soggetti, sui quali è stata fatta un'attenta revisione. Alcuni di questi sono effettivamente ricercatori con un profilo di ricerca particolare, altri invece sono autori per i quali le banche dati producono output errati. In particolare 9 soggetti presentano valori elevati sulle variabili ottenute secondo alcune fonti rispetto alle altre; ad esempio, autori di libri (valori maggiori in POP), autori di articoli su vecchie riviste di statistica a diffusione nazionale (valori maggiori in CIS), autori in discipline con elevato impatto (valori maggiori in WOS e Scopus). Sono invece stati corretti 14 record che per motivi diversi (un carattere speciale nel nome, un'omonimia, un cambio di affiliazione o un errato record nel database) erano stati costruiti in modo errato.

Tramite una analisi dei gruppi[4] sono stati individuati 5 insiemi con diverse numerosità e diverse caratteristiche. La Tabella 2 riporta le medie e le mediane per cluster di alcune delle variabili considerate, facilmente descritte tramite il nome assegnato.

Tabella 2

		Cluster					
		1	**2**	**3**	**4**	**5**	**Total**
N		211	141	17	41	5	415
NpubCIS	*Media*	3.51	9.6	13.06	22.34	39	8.26
	Mediana	2	8	12	20	48	5
NpubSCO	*Media*	1.28	6.87	5.94	18.54	32.8	5.46
	Mediana	0	7	5	18	29	3
hindexSCO	*Media*	0.3	2.12	1.88	5.17	8.6	1.57

[1] Fonte MIUR docenti.

[2] Un ricercatore ottiene un valore ***h*** dell'indicatore se ha h paper con almeno h citazioni ciascuno e i rimanenti (N-h) paper non hanno più di h citazioni ciascuno.

[3] P. Filzmoser, R.G. Garrett, C. Reimann. *Multivariate outlier detection in exploration geochemistry*. Computers & Geosciences, 31:579-587, 2005.
P. Filzmoser, R. Maronna, M. Werner. *Outlier identification in high dimensions*, Computational Statistics and Data Analysis, 52, 1694-1711, 2008.

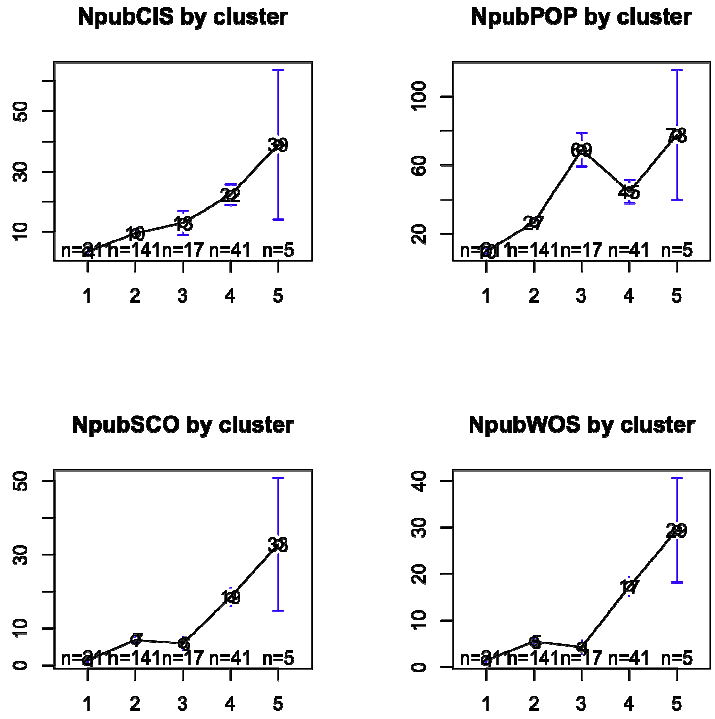
[4] È stata applicata una cluster gerarchica, utilizzando la distanza euclidea quadratica e il metodo di Ward (Zani S. (2000), *Analisi dei Dati Statistici II*, Giuffrè, Capitolo V)

	Mediana	0	2	2	5	9	1
NpubPOP	*Media*	9.93	26.78	69	44.68	77.8	22.33
	Mediana	9	25	61	41	61	18
hindexPOP	*Media*	1.62	4.72	10.18	7.88	14.6	3.8
	Mediana	2	5	10	8	13	3
NpubWOS	*Media*	1.45	5.46	4.24	17.34	29.4	4.83
	Mediana	1	6	3	17	26	3
hindexWOS	*Media*	0.51	2.06	1.71	5.41	8.8	1.67
	Mediana	0	2	2	5	9	1

Analizzando la tabella è possibile isolare un gruppo molto numeroso di ricercatori (211) che presentano valori molto bassi per tutti gli indici. Il 50 per cento di questi ricercatori ha al massimo un lavoro presente nella banca dati WOS, anche se ha più di 2 articoli di statistica o probabilità nella banca dati CIS e partecipa a convegni e produce *working paper*, come mostra la media degli indici di POP. È presente un secondo gruppo abbastanza numeroso (141) composto da ricercatori che per il 50 per cento hanno più di 6 lavori sia su WOS sia su Scopus e che svolgono una rilevante attività produttiva, evidenziata anche da valori elevati su CIS e POP. Il terzo gruppo, composto da soli 17 studiosi, ha i valori in assoluto più elevati per POP, sia per la produttività sia per la diffusione. Un'analisi dettagliata ha mostrato che i componenti di questo gruppo sono studiosi italiani, autori di importanti libri, che partecipano a comitati scientifici di convegni e di riviste, sono stati editor di *special issue* e tendono ad avere numerosi incarichi accademici. Questi studiosi hanno valori degli indici per WOS e SCO più bassi degli studiosi del cluster 2. Il cluster 4 è speculare al cluster 3: elevati valori per SCO e WOS e valori più bassi per POP. I 41 studiosi che lo compongono sembrano orientare il loro lavoro su articoli per riviste, come mostrano gli elevati valori per produttività e diffusione per WOS e SCO, puntando anche su riviste di alto prestigio. I 5 studiosi appartenenti al cluster 5 sono studiosi eccellenti che hanno valori davvero inusuali per tutte le banche dati.

La Figura 1 mostra graficamente quanto rilevato in merito alla produttività. In tale figura è interessante notare come i pattern di WOS e SCO siano molto simili.

Figura 1



Mettiamo ora in evidenza la ripartizione per ruolo nei gruppi. La Tabella 3 mostra che nel gruppo 1 sono presenti prevalentemente ricercatori. I gruppi 3, 4, 5 sono composti, per la maggior parte, da ricercatori senior, anche se nel cluster 5 degli eccellenti è presente un ricercatore che ha, senza dubbio, un curriculum positivamente anomalo rispetto a tutti gli statistici italiani e in particolare alla sua categoria.

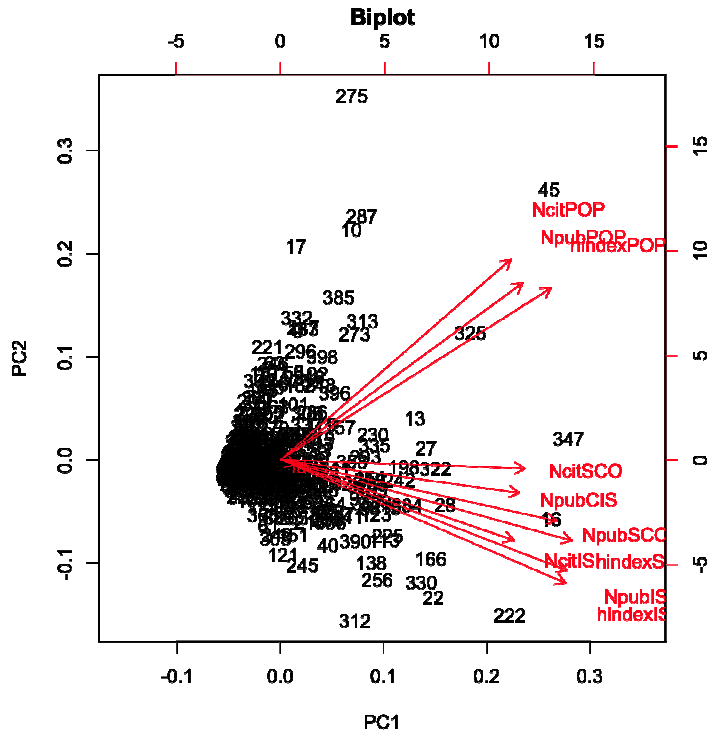
Tabella 3

Cluster	Ordinario	Associato	Ricercatore
1	48	60	103
2	50	43	48
3	11	3	3
4	30	9	2
5	4	0	1

Ci si può ora chiedere se sia possibile comprimere l'informazione ottenuta dalle 4 banche dati in alcuni indicatori sintetici. Un analogo studio nell'ambito delle scienze sociali[5], applicando l'analisi delle componenti principali[6], ha individuato tre dimensioni: produzione, internazionalizzazione, diffusione. Nel nostro caso, come mostra la Figura 2, l'applicazione dell'analisi delle componenti principali non ha evidenziato le stesse dimensioni; ne è invece emersa una , la prima, correlata positivamente con tutte le variabili considerate, che da sola spiega il 69 per cento della varianza totale e una seconda , legata principalmente ai tre indicatori di POP (Npub, Ncit, Hindex), confermando di fatto quanto emerso dalla *cluster analysis*: POP evidenzia una seconda componente presente nel profilo di ricerca degli statistici italiani, legata alla pubblicazione di libri, alla partecipazione/organizzazione di convegni, alla direzione di collane e numeri speciali.

[5] http://www.sociol.unimi.it/papers/2010-02-09_Ferruccio%20biolcati-Rinaldi.pdf
[6] Zani S. (2000), *Analisi dei Dati Statistici II*, Giuffrè, Capitolo III

Figura 2



A conclusione, in merito al confronto tra le banche dati, si può dire che SCO e WOS forniscono informazioni simili per quanto riguarda gli statistici, CIS non usa criteri selettivi di inclusione ed è molto datata (potremmo pensare in futuro a MathSciNet come alternativa) e POP sembra misurare un diverso aspetto.

Le importanti lezioni che abbiamo appreso da questo esercizio sono due:

- non c'è un unico profilo di 'buon ricercatore'. Il confronto è difficile, ognuno ha politiche di ricerca diverse e ottiene perciò un profilo diverso;
- l'uso di una singola fonte non è consigliabile. Questa affermazione può sembrare banale, ma è sempre più frequente che le aree scientifiche disciplinari, non soltanto quella di statistica, scelgano di utilizzare ai fini della valutazione una sola banca dati, prendendo per buoni gli output che produce; come abbiamo messo in evidenza il confronto tra banche dati diverse è essenziale per controllare i risultati ottenuti.

Da ultimo sembra utile suggerire ai lettori interessati di controllare il proprio record e di segnalare eventuali errori e correzioni ai gestori delle banche dati stesse. Per richieste di correzione degli errori nei record della banca dati WOS si utilizza il sito http://scientific.thomsonreuters.com/techsupport/datachange/; in SCO si utilizza il pulsante "Feedback" nella schermata dei risultati della ricerca per la correzione delle intestazioni autore e legame con gli articoli; in CIS si possono effettuare segnalazioni utilizzando il sito https://secure.imstat.org/secure/orders/ciscontact.asp oppure scrivendo a question@statindex.org. Publish or Perish invece è un'interfaccia per l'interrogazione di Google Scholar, non è una banca dati, quindi non esiste un record relativo all'autore che possa essere modificato.

Per saperne di più:

Adler R., Ewing J. e Taylor P. (2008), "Citation Statistics", *Statistical Science*, 24(1), 1-14.

Bollen J., Van de Sompel H., Hagberg A. e Chute R. (2009b) "A Principal Component Analysis of 39 Scientific Impact Measures" *PLoS ONE* 4(6): e6022. doi:10.1371/journal.pone.0006022.

Franceschet M. (2009), A cluster analysis of scholar and journal bibliometric indicators, *Journal of the American Society for Information Science and Technology*,60(10), 1950-1964.