Appendice A. Il confronto tra valutazione peer e valutazione bibliometrica

A.1 Il campione casuale

Un campione casuale di 1.412 articoli su rivista passibili di valutazione bibliometrica è stato estratto dalla popolazione di 15.029 articoli indicizzati WoS e sottomessi alla valutazione nel GEV02. La popolazione è stratificata in base della distribuzione dei Prodotti all'interno dei Settori Scientifico-Disciplinari (SSD nel seguito) dell'Area, aggregando in un unico settore la Fisica applicata e la Didattica e storia della fisica. Gli SSD considerati sono: Fisica sperimentale; Fisica teorica, modelli e metodi matematici; Fisica della materia; Fisica nucleare e subnucleare; Astronomia e astrofisica; Fisica per il sistema terra e il mezzo circumterrestre; Fisica applicata, didattica e storia della fisica. La classificazione degli articoli all'interno dei SSD si basa sulla SC di WoS, ed è calcolata escludendo i casi di articoli duplicati presentati da diversi autori all'interno di uno stesso strato campionario. Il campione include il 9,1% dei Prodotti di Fisica sperimentale, il 9,3% di quelli di Fisica teorica, modelli e metodi matematici e di Fisica della materia, il 9,6% di Fisica nucleare e subnucleare, il 9,9% di Astronomia e astrofisica, l'8,5% di Fisica per il sistema terrestre e il mezzo circumterrestre e il 9,2% di Fisica applicata, didattica e storia della fisica (Tabella A.1). Il campione è stato estratto ai primi di Settembre 2012, prima dell'inizio del processo di revisione *peer*, mediante una procedura casuale con il vincolo di selezionare un campione significativo di Prodotti, percentualmente simile in ciascun SSD (estraendolo dai circa 14.000 Prodotti valutabili bibliometricamente che erano stati identificati in quel momento).

Tab. A.1: Distribuzione degli articoli su rivista nel campione e nella popolazione

SSD	*Popolazione*	*Campione*	*%*
Fisica sperimentale	1.531	139	9,1
Fisica teorica, modelli e metodi matematici	5.350	499	9,3
Fisica della materia	3.741	349	9,3
Fisica nucleare e subnucleare	467	45	9,6
Astronomia e astrofisica	2.719	270	9,9
Fisica per il sistema terra e il mezzo circumterrestre	329	28	8,5
Fisica applicata, didattica e storia della fisica	892	82	9,2
Totale	15.029	1.412	9,4

Tab. A.2: Distribuzione delle valutazioni bibliometriche nel campione e nella popolazione

Classe	*Popolazione*	*%*	*Campione*	*%*
Fisica sperimentale				
E	917	59,90	75	53,96
B	255	16,66	30	21,58
A	101	6,60	5	3,60
L	60	3,92	9	6,47
IR	198	12,93	20	14,39
Fisica teorica, modelli e metodi matematici				
E	2615	48,88	253	50,70
B	922	17,23	101	20,24
A	457	8,54	42	8,42
L	463	8,65	27	5,41
IR	893	16,69	76	15,23
Fisica della materia				
E	2.450	65,49	237	67,91
B	502	13,42	41	11,75
A	202	5,40	16	4,58
L	133	3,56	13	3,72
IR	454	12,14	42	12,03
Fisica nucleare e subnucleare				
E	373	79,87	40	88,89
B	38	8,14	0	0,00
A	9	1,93	1	2,22
L	13	2,78	0	0,00
IR	34	7,28	4	8,89
Astronomia e astrofisica				
E	1532	56,34	158	58,52
B	474	17,43	45	16,67
A	174	6,40	13	4,81
L	203	7,47	20	7,41
IR	336	12,36	34	12,59
Fisica per il sistema terra e il mezzo circumterrestre				
E	82	24,92	12	42,86
B	60	18,24	4	14,29
A	49	14,89	0	0,00
L	62	18,84	2	7,14
IR	76	23,10	10	35,71
Fisica applicata, didattica e storia della fisica				
E	401	44,96	33	40,24
B	159	17,83	18	21,95
A	96	10,76	9	10,98
L	109	12,22	8	9,76
IR	127	14,24	14	17,07
Totale				
E	8370	55,69	808	57,22
B	2410	16,04	239	16,93
A	1088	7,24	86	6,09
L	1043	6,94	79	5,59
IR	2118	14,09	200	14,16

La Tabella A.2 riporta la distribuzione nelle classi di valutazione VQR (Eccellente, Buono, Accettabile, Limitato, *Informed Review* o IR) ottenuta utilizzando la valutazione bibliometrica degli articoli su rivista nei sette SSD, per il campione e per la popolazione. Come si vede, la distribuzione delle valutazioni bibliometriche (E, B, A, L, IR) è sufficientemente vicina nella popolazione e nel campione, sia per il totale sia per i Settori Scientifico Disciplinari, così da concludere che il campione estratto è rappresentativo della popolazione di riferimento.

Per ciascun articolo su rivista incluso nel campione casuale sono disponibili le seguenti informazioni:

- Rapporto del primo revisore (P1)
- Rapporto del secondo revisore (P2)
- Rapporto di un eventuale terzo e quarto revisore (P3 e P4)
- Valutazione di sintesi dei giudizi del primo e secondo revisore (P)
- Valutazione bibliometrica (F).

Le variabili P e P1-P4 assumono come valore una delle 4 classi di valutazione E, B, A, L; la valutazione bibliometrica F ha come possibile risultato anche la classe di valutazione IR, ossia il suggerimento di procedere con la *informed peer review* nel caso di risultati molto diversi tra i due indicatori bibliometrici (*Impact Factor* e numero citazioni, rif. i criteri del GEV descritti nell'Appendice B). Le quattro classi, secondo il Bando VQR, sono definite con riferimento ai percentili della distribuzione della qualità degli articoli pubblicati nel mondo. In particolare, la qualifica di eccellente corrisponde a un articolo che si colloca nel 20% superiore della distribuzione della qualità degli articoli pubblicati nel mondo, quella di buono nel successivo 20%, di accettabile nel successivo 10% e, infine, quella di limitato nel 50% inferiore. Le variabili P1-P4 sono originariamente misurate su una scala numerica compresa tra 1 e 27, con un punteggio da 1 a 9 assegnato a 3 diversi criteri; tali punteggi sono successivamente utilizzati per determinare per ciascun Prodotto sottomesso a valutazione la classe di valutazione *peer* del Prodotto, sulla base dei criteri fissati dal GEV[1]; le variabili P ed F sono invece rispettivamente espresse in termini delle 4 o 5 classi di valutazione sopra elencate. Sulla base del Bando VQR, alle quattro classi E, B, A, L corrispondono rispettivamente i punteggi 1; 0,8; 0,5; 0.

La classificazione adottata nell'analisi bibliometrica si basa sui criteri descritti nell'Appendice B di questo rapporto. Nella revisione dei pari, ai revisori esterni è stato richiesto di valutare ciascun Prodotto sulla base della loro percezione soggettiva della qualità del Prodotto rispetto alla distribuzione mondiale dei Prodotti della ricerca nel settore scientifico a cui il Prodotto faceva riferimento. La valutazione dei revisori è stata quindi sintetizzata sulla base di un algoritmo specifico al GEV02, secondo il quale, rispettivamente: i Prodotti di classe E erano quelli con un punteggio complessivo pari a 23-27; i Prodotti di classe B avevano un punteggio complessivo pari a 18-22; i Prodotti di classe A un punteggio complessivo pari a 15-17 e i Prodotti di classe L un

[1] L'etichetta "P1", "P2", "P3" e "P4" assegnata ai revisori è puramente convenzionale e riflette esclusivamente l'ordine di accettazione della proposta di revisione avanzata al potenziale revisore.

punteggio compreso tra 3 e 14. Al fine di confrontare i risultati della valutazione bibliometrica e della revisione tra pari, si procede nel seguito a confrontare gli indicatori F e P. Anche altri confronti possono essere tuttavia di importanza significativa: in particolare il confronto tra le valutazioni tra pari P1 e P2 consente di valutare il grado di corrispondenza dei giudizi tra i due revisori[2].

A.2 Le distribuzioni F e P

Le distribuzioni F e P sopra descritte non sono immediatamente confrontabili, dato che la distribuzione F delle valutazioni bibliometriche comprende una classe IR che non è invece prevista nella valutazione dei pari. E' però possibile ipotizzare che una discordanza di almeno due classi tra la valutazione del primo e secondo revisore segnali un'incertezza della revisione dei pari del tutto analoga a quella che emerge dal confronto tra numero di citazioni e fattore di impatto della sede di pubblicazione nell'analisi bibliometrica: in analogia con la classificazione IR della valutazione bibliometrica. Si è definita dunque un'addizionale classificazione "Incerta *Peer*" (IP) per la valutazione dei revisori quando le valutazioni dei due revisori differivano di almeno per due classi di valutazione. Così facendo, è possibile effettuare il confronto tra le distribuzioni F e P. La Tabella A.3 mostra la distribuzione in numeri assoluti e percentuali degli indicatori F e P sopra descritti per il totale del campione del GEV02.

Tab. A.3: Confronto tra F e P – totale del campione

Valutazione bibliometrica (F)	***Valutazione peer (P)***					
	E	**B**	**A**	**L**	**IP**	**Totale**
E	**252**	**371**	**45**	**14**	**126**	**808**
% rispetto al totale delle valutazioni bibliometriche di classe E	*31,19*	*45,92*	*5,57*	*1,73*	*15,59*	*100*
B	**28**	**112**	**32**	**23**	**44**	**239**
% rispetto al totale delle valutazioni bibliometriche di classe B	*11,72*	*46,86*	*13,39*	*9,62*	*18,41*	*100*
A	**2**	**34**	**10**	**20**	**20**	**86**
% rispetto al totale delle valutazioni bibliometriche di classe A	*2,33*	*39,53*	*11,63*	*23,26*	*23,26*	*100*
L	**3**	**19**	**11**	**33**	**13**	**79**
% rispetto al totale delle valutazioni bibliometriche di classe L	*3,80*	*24,05*	*13,92*	*41,77*	*16,46*	*100*
IR	**15**	**79**	**23**	**36**	**47**	**200**
% rispetto al totale delle valutazioni bibliometriche IR	*7,50*	*39,50*	*11,50*	*18,00*	*23,50*	*100*
Totale	**300**	**615**	**121**	**126**	**250**	**1412**
% rispetto al totale delle valutazioni bibliometriche	*21,25*	*43,56*	*8,57*	*8,92*	*17,71*	*100*

[2] Nel GEV02, per 43 Prodotti è stato necessario procedere anche a una terza valutazione dei pari. La terza valutazione non è stata qui considerata.

Gli elementi sulla diagonale principale della Tabella A.3 corrispondono ai casi in cui la valutazione dei pari e quella bibliometrica coincidono. Gli elementi al di fuori della diagonale principale corrispondono invece ai casi di non coincidenza tra F e P, o perché la valutazione F è migliore della P (elementi al di sopra della diagonale principale) o viceversa (elementi al di sotto della diagonale). La Tabella A.3 mostra che nella maggior parte dei casi la discordanza tra la valutazione bibliometrica e quella dei pari è dovuta al fatto che la valutazione bibliometrica è più generosa. In particolare, gli articoli classificati come eccellenti sulla base degli indicatori bibliometrici sono 808, contro i soli 300 Prodotti eccellenti della valutazione tra pari: solo il 31,2% degli articoli classificati come E secondo la bibliometria ottiene E anche secondo la revisione tra pari, mentre rispettivamente nel 46%, 6% e 2% dei casi i Prodotti bibliometricamente eccellenti risultano buoni, accettabili o limitati nella valutazione tra pari. D'altro lato, il numero di articoli che sono classificati in B, A e L dalla valutazione tra pari (615, 121 e 126 articoli rispettivamente) è nettamente più elevato rispetto agli articoli che risultano in B, A ed L secondo la valutazione bibliometrica (239, 86 e 79 articoli rispettivamente). Infine, la numerosità di valutazioni incerte è maggiore nella revisione tra pari (250 articoli) rispetto a quella bibliometrica (200 articoli). Le valutazioni bibliometriche incerte hanno nel 47% dei casi una valutazione almeno pari a B nell'analisi *peer*, mentre in circa il 70% dei casi le valutazioni incerte *peer* ricevono un punteggio bibliometrico almeno pari a B.

Complessivamente, l'analisi bibliometrica e la revisione tra pari coincidono nel 32,1% dei casi. Se si sommano alle valutazioni coincidenti quelle che differiscono di una sola classe, si arriva al 67,3% del campione. Gli articoli con valutazioni che differiscono per due classi sono 89, il 6,3% del campione, quelli con massima discordanza (ossia, che differiscono per 3 classi) sono 17 (l'1,2% del campione). Il restante 25,2% del campione ha una assegnazione IR o IP con uno dei due metodi, e perviene ad una classe di assegnazione definita secondo l'altro metodo.

La Tabella A.4 mostra la distribuzione degli indicatori P1 e P2. Le valutazioni dei due revisori coincidono nel 41,5% dei casi, sono diverse per una classe di valutazione nel 40,8% dei casi e divergono invece rispettivamente per 2 o 3 classi di valutazione nel 14,1% e nel 3,6% dei casi. E' da notare anche che le valutazioni su un giudizio di assegnazione alla classe E sono convergenti in 221 casi, pari a poco meno della metà del totale delle valutazioni eccellenti fornite dal primo revisore e dal secondo revisore.

Tab. A.4: Confronto tra le valutazioni P1 e P2 – totale del campione

P1	***P2***				
	E	**B**	**A**	**L**	**Totale**
E	221	169	37	26	453
% rispetto al totale delle valutazioni di classe E del primo revisore	*48,8*	*37,3*	*8,2*	*5,7*	*100*
B	170	276	82	62	590
% rispetto al totale delle valutazioni di classe B del primo revisore	*28,8*	*46,8*	*13,9*	*10,5*	*100*
A	36	80	31	44	191
% rispetto al totale delle valutazioni di classe A del primo revisore	*18,8*	*41,9*	*16,2*	*23,0*	*100*
L	25	64	31	58	178
% rispetto al totale delle valutazioni di classe L del primo revisore	*14,0*	*36,0*	*17,4*	*32,6*	*100*
Totale	452	589	181	190	1 412
% rispetto al totale delle valutazioni del primo revisore	*32,0*	*41,7*	*12,8*	*13,5*	*100*

Le Tabelle A.5 e A.6 estendono i risultati delle Tabelle A.3 e A.4 ai singoli SSD. In particolare, dall'analisi dei dati della Tabella A.5 emerge che in tutti i SSD il numero di valutazioni eccellenti è nettamente maggiore secondo la valutazione bibliometrica rispetto alla *peer*. D'altra parte, il numero di valutazioni buone e accettabili è in genere maggiore secondo l'analisi *peer* rispetto a quella bibliometrica.

Tab. A.5: Confronto tra F e P per SSD

Fisica sperimentale						
Valutazione bibliometrica (F)	***Valutazione peer (P)***					
	E	**B**	**A**	**L**	**IP**	**Totale**
E	21	38	5	2	9	75
% rispetto al totale delle valutazioni bibliometriche di classe E	*28,0*	*50,7*	*6,7*	*2,7*	*12,0*	*100*
B	2	6	7	6	9	30
% rispetto al totale delle valutazioni bibliometriche di classe B	*6,7*	*20,0*	*23,3*	*20,0*	*30,0*	*100*
A	0	2	1	1	1	5
% rispetto al totale delle valutazioni bibliometriche di classe A	*0,0*	*40,0*	*20,0*	*20,0*	*20,0*	*100*
L	0	2	1	2	4	9
% rispetto al totale delle valutazioni bibliometriche di classe L	*0,0*	*22,2*	*11,1*	*22,2*	*44,4*	*100*
IR	2	9	2	5	2	20
% rispetto al totale delle valutazioni bibliometriche IR	*10,0*	*45,0*	*10,0*	*25,0*	*10,0*	*100*
Totale	25	57	16	16	25	139
% rispetto al totale delle valutazioni bibliometriche	*18,0*	*41,0*	*11,5*	*11,5*	*18,0*	*100*
Fisica teorica, modelli e metodi matematici						
Valutazione bibliometrica (F)	***Valutazione peer (P)***					
	E	**B**	**A**	**L**	**IP**	**Totale**
E	87	114	13	5	34	253

% rispetto al totale delle valutazioni bibliometriche di classe E	*34,4*	*45,1*	*5,1*	*2,0*	*13,4*	*100*
B	16	51	9	7	18	101
% rispetto al totale delle valutazioni bibliometriche di classe B	*15,8*	*50,5*	*8,9*	*6,9*	*17,8*	*100*
A	1	16	6	7	12	42
% rispetto al totale delle valutazioni bibliometriche di classe A	*2,4*	*38,1*	*14,3*	*16,7*	*28,6*	*100*
L	3	8	1	10	5	27
% rispetto al totale delle valutazioni bibliometriche di classe L	*11,1*	*29,6*	*3,7*	*37,0*	*18,5*	*100*
IR	9	28	6	13	20	76
% rispetto al totale delle valutazioni bibliometriche IR	*11,8*	*36,8*	*7,9*	*17,1*	*26,3*	*100*
Totale	116	217	35	42	89	499
% rispetto al totale delle valutazioni bibliometriche	*23,2*	*43,5*	*7,0*	*8,4*	*17,8*	*100*
Fisica della materia						
Valutazione bibliometrica (F)	***Valutazione peer (P)***					
	E	**B**	**A**	**L**	**IP**	**Totale**
E	59	113	10	4	51	237
% rispetto al totale delle valutazioni bibliometriche di classe E	*24,9*	*47,7*	*4,2*	*1,7*	*21,5*	*100*
B	2	16	9	7	7	41
% rispetto al totale delle valutazioni bibliometriche di classe B	*4,9*	*39,0*	*22,0*	*17,1*	*17,1*	*100*
A	0	4	2	8	2	16
% rispetto al totale delle valutazioni bibliometriche di classe A	*0,0*	*25,0*	*12,5*	*50,0*	*12,5*	*100*
L	0	2	2	7	2	13
% rispetto al totale delle valutazioni bibliometriche di classe L	*0,0*	*15,4*	*15,4*	*53,8*	*15,4*	*100*
IR	1	13	4	9	15	42
% rispetto al totale delle valutazioni bibliometriche IR	*2,4*	*31,0*	*9,5*	*21,4*	*35,7*	*100*
Totale	62	148	27	35	77	349
% rispetto al totale delle valutazioni bibliometriche	*17,8*	*42,4*	*7,7*	*10,0*	*22,1*	*100*
Fisica nucleare e subnucleare						
Valutazione bibliometrica (F)	***Valutazione peer (P)***					
	E	**B**	**A**	**L**	**IP**	**Totale**
E	22	9	4	1	4	40
% rispetto al totale delle valutazioni bibliometriche di classe E	*55,0*	*22,5*	*10,0*	*2,5*	*10,0*	*100*
B	0	0	0	0		0
% rispetto al totale delle valutazioni bibliometriche di classe B	*0,0*	*0,0*	*0,0*	*0,0*	*0,0*	*0*
A	0	0	0	1	0	1
% rispetto al totale delle valutazioni bibliometriche di classe A	*0,0*	*0,0*	*0,0*	*100,0*	*0,0*	*100*
L	0	0	0	0	0	0
% rispetto al totale delle valutazioni bibliometriche di classe L	*0,0*	*0,0*	*0,0*	*0,0*	*0,0*	*0*
IR	1	0	2	1	0	4

% rispetto al totale delle valutazioni bibliometriche IR	*25,0*	*0,0*	*50,0*	*25,0*	*0,0*	*100*
Totale	23	9	6	3	4	45
% rispetto al totale delle valutazioni bibliometriche	*51,1*	*20,0*	*13,3*	*6,7*	*8,9*	*100*
Astronomia e astrofisica						
Valutazione bibliometrica (F)	***Valutazione peer (P)***					
	E	**B**	**A**	**L**	**IP**	**Totale**
E	55	78	5	1	19	158
% rispetto al totale delle valutazioni bibliometriche di classe E	*34,8*	*49,4*	*3,2*	*0,6*	*12,0*	*100*
B	5	25	6	1	8	45
% rispetto al totale delle valutazioni bibliometriche di classe B	*0,0*	*0,0*	*0,0*	*0,0*	*0,0*	*0*
A	1	9	0	1	2	13
% rispetto al totale delle valutazioni bibliometriche di classe A	*7,7*	*69,2*	*0,0*	*7,7*	*15,4*	*100*
L	0	5	5	8	2	20
% rispetto al totale delle valutazioni bibliometriche di classe L	*0,0*	*0,0*	*0,0*	*0,0*	*0,0*	*0*
IR	2	18	4	3	7	34
% rispetto al totale delle valutazioni bibliometriche IR	*5,9*	*52,9*	*11,8*	*8,8*	*20,6*	*100*
Totale	63	135	20	14	38	270
% rispetto al totale delle valutazioni bibliometriche	*23,3*	*50,0*	*7,4*	*5,2*	*14,1*	*100*
Fisica per il sistema terra ed il mezzo circumterrestre						
Valutazione bibliometrica (F)	***Valutazione peer (P)***					
	E	**B**	**A**	**L**	**IP**	**Totale**
E	5	4	0	1	2	12
% rispetto al totale delle valutazioni bibliometriche di classe E	*41,7*	*33,3*	*0,0*	*8,3*	*16,7*	*100*
B	1	3	0	0	0	4
% rispetto al totale delle valutazioni bibliometriche di classe B	*25,0*	*75,0*	*0,0*	*0,0*	*0,0*	*100*
A	0	0	0	0	0	0
% rispetto al totale delle valutazioni bibliometriche di classe A	*0,0*	*0,0*	*0,0*	*0,0*	*0,0*	*0*
L	0	0	1	1	0	2
% rispetto al totale delle valutazioni bibliometriche di classe L	*0*	*0*	*50*	*50*	*0*	*100*
IR	0	5	3	1	1	10
% rispetto al totale delle valutazioni bibliometriche IR	*0,0*	*50,0*	*30,0*	*10,0*	*10,0*	*100*
Totale	6	12	4	3	3	28
% rispetto al totale delle valutazioni bibliometriche	*21,4*	*42,9*	*14,3*	*10,7*	*10,7*	*100*
Fisica applicata, didattica e storia della fisica						
Valutazione bibliometrica (F)	***Valutazione peer (P)***					
	E	**B**	**A**	**L**	**IP**	**Totale**
E	3	15	8	0	7	33
% rispetto al totale delle valutazioni bibliometriche di classe E	*9,1*	*45,5*	*24,2*	*0,0*	*21,2*	*100*
B	2	11	1	2	2	18

% rispetto al totale delle valutazioni bibliometriche di classe B	*11,1*	*61,1*	*5,6*	*11,1*	*11,1*	*100*
A	0	3	1	2	3	9
% rispetto al totale delle valutazioni bibliometriche di classe A	*0,0*	*33,3*	*11,1*	*22,2*	*33,3*	*100*
L	0	2	1	5	0	8
% rispetto al totale delle valutazioni bibliometriche di classe L	*0*	*25*	*13*	*63*	*0*	*100*
IR	0	6	2	4	2	14
% rispetto al totale delle valutazioni bibliometriche IR	*0,0*	*42,9*	*14,3*	*28,6*	*14,3*	*100*
Totale	5	37	13	13	14	82
% rispetto al totale delle valutazioni bibliometriche	*6,1*	*45,1*	*15,9*	*15,9*	*17,1*	*100*

Complessivamente, la tendenza della valutazione bibliometrica a essere più vantaggiosa rispetto a quella *peer* è comune a tutti i SSD (un test formale di tale ipotesi è presentato nella sezione A.3) La Tabella A.6 estende i risultati della Tabella A.4 ai singoli SSD.

Tab. A.6: Confronto tra le valutazioni P1 e P2 per SSD

Fisica sperimentale					
P1	***P2***				
	E	**B**	**A**	**L**	**Totale**
E	19	18	3	2	42
% rispetto al totale delle valutazioni di classe E del primo revisore	*45,2*	*42,9*	*7,1*	*4,8*	*100*
B	11	26	9	4	50
% rispetto al totale delle valutazioni di classe B del primo revisore	*22,0*	*52,0*	*18,0*	*8,0*	*100*
A	6	10	6	3	25
% rispetto al totale delle valutazioni di classe A del primo revisore	*24,0*	*40,0*	*24,0*	*12,0*	*100*
L	0	10	4	8	22
% rispetto al totale delle valutazioni di classe L del primo revisore	*0,0*	*45,5*	*18,2*	*36,4*	*100*
Totale	36	64	22	17	139
% rispetto al totale delle valutazioni del primo revisore	*25,9*	*46,0*	*15,8*	*12,2*	*100*
Fisica teorica, modelli e metodi matematici					
P1	***P2***				
	E	**B**	**A**	**L**	**Totale**
E	79	64	4	7	154
% rispetto al totale delle valutazioni di classe E del primo revisore	*51,3*	*41,6*	*2,6*	*4,5*	*100*
B	75	87	25	22	209

% rispetto al totale delle valutazioni di classe B del primo revisore	*35,9*	*41,6*	*12,0*	*10,5*	*100*
A	19	28	10	13	70
% rispetto al totale delle valutazioni di classe A del primo revisore	*27,1*	*40,0*	*14,3*	*18,6*	*100*
L	13	24	7	22	66
% rispetto al totale delle valutazioni di classe L del primo revisore	*19,7*	*36,4*	*10,6*	*33,3*	*100*
Totale	186	203	46	64	499
% rispetto al totale delle valutazioni del primo revisore	*37,3*	*40,7*	*9,2*	*12,8*	*100*
Fisica della materia					
P1	***P2***				
	E	**B**	**A**	**L**	**Totale**
E	49	35	17	9	110
% rispetto al totale delle valutazioni di classe E del primo revisore	*44,5*	*31,8*	*15,5*	*8,2*	*100*
B	42	65	19	24	150
% rispetto al totale delle valutazioni di classe B del primo revisore	*28,0*	*43,3*	*12,7*	*16,0*	*100*
A	6	17	4	16	43
% rispetto al totale delle valutazioni di classe A del primo revisore	*14,0*	*39,5*	*9,3*	*37,2*	*100*
L	5	16	10	15	46
% rispetto al totale delle valutazioni di classe L del primo revisore	*10,9*	*34,8*	*21,7*	*32,6*	*100*
Totale	102	133	50	64	349
% rispetto al totale delle valutazioni del primo revisore	*29,2*	*38,1*	*14,3*	*18,3*	*100*
Fisica nucleare e subnucleare					
P1	*P2*				
	E	**B**	**A**	**L**	**Totale**
E	20	6	3	0	29
% rispetto al totale delle valutazioni di classe E del primo revisore	*69,0*	*20,7*	*10,3*	*0,0*	*100*
B	3	3	1	0	7
% rispetto al totale delle valutazioni di classe B del primo revisore	*42,9*	*42,9*	*14,3*	*0,0*	*100*
A	0	2	2	2	6
% rispetto al totale delle valutazioni di classe A del primo revisore	*0,0*	*33,3*	*33,3*	*33,3*	*100*
L	0	1	1	1	3
% rispetto al totale delle valutazioni di classe L del primo revisore	*0,0*	*33,3*	*33,3*	*33,3*	*100*
Totale	23	12	7	3	45

% rispetto al totale delle valutazioni del primo revisore	*51,1*	*26,7*	*15,6*	*6,7*	*100*
Astronomia e astrofisica					
P1	***P2***				
	E	**B**	**A**	**L**	**Totale**
E	48	35	9	8	100
% rispetto al totale delle valutazioni di classe E del primo revisore	*48,0*	*35,0*	*9,0*	*8,0*	*100*
B	26	72	18	7	123
% rispetto al totale delle valutazioni di classe B del primo revisore	*21,1*	*58,5*	*14,6*	*5,7*	*100*
A	3	15	3	5	26
% rispetto al totale delle valutazioni di classe A del primo revisore	*11,5*	*57,7*	*11,5*	*19,2*	*100*
L	6	5	5	5	21
% rispetto al totale delle valutazioni di classe L del primo revisore	*28,6*	*23,8*	*23,8*	*23,8*	*100*
Totale	83	127	35	25	270
% rispetto al totale delle valutazioni del primo revisore	*30,7*	*47,0*	*13,0*	*9,3*	*100*
Fisica per il sistema terra e il mezzo circumterrestre					
P1	***P2***				
	E	**B**	**A**	**L**	**Totale**
E	3	3	1	0	7
% rispetto al totale delle valutazioni di classe E del primo revisore	*42,9*	*42,9*	*14,3*	*0,0*	*100*
B	6	6	0	1	13
% rispetto al totale delle valutazioni di classe B del primo revisore	*46,2*	*46,2*	*0,0*	*7,7*	*100*
A	0	2	2	1	5
% rispetto al totale delle valutazioni di classe A del primo revisore	*0,0*	*40,0*	*40,0*	*20,0*	*100*
L	0	1	1	1	3
% rispetto al totale delle valutazioni di classe L del primo revisore	*0,0*	*33,3*	*33,3*	*33,3*	*100*
Totale	9	12	4	3	28
% rispetto al totale delle valutazioni del primo revisore	*32,1*	*42,9*	*14,3*	*10,7*	*100*
Fisica applicata, didattica e storia della fisica					
P1	***P2***				
	E	**B**	**A**	**L**	**Totale**
E	3	8	0	0	11
% rispetto al totale delle valutazioni di classe E del primo revisore	*27,3*	*72,7*	*0,0*	*0,0*	*100*
B	7	17	10	4	38

% rispetto al totale delle valutazioni di classe B del primo revisore	*18,4*	*44,7*	*26,3*	*10,5*	*100*
A	2	6	4	4	16
% rispetto al totale delle valutazioni di classe A del primo revisore	*12,5*	*37,5*	*25*	*25*	*100*
L	1	7	3	6	17
% rispetto al totale delle valutazioni di classe L del primo revisore	*5,9*	*41,2*	*17,6*	*35,3*	*100*
Totale	13	38	17	14	82
% rispetto al totale delle valutazioni del primo revisore	*15,9*	*46,3*	*20,7*	*17,1*	*100*

A.3 Il confronto tra le distribuzioni di F e P

Il confronto tra la valutazione dei pari e quella bibliometrica si può basare su due criteri fondamentali:

1. Grado di concordanza tra la distribuzione F e la distribuzione P, ossia se F e P tendono ad assegnare lo stesso punteggio ad ogni articolo
2. Grado di differenza sistematica esistente tra F e P misurata mediante la differenza media del punteggio assegnato da F e P sulla base dei pesi attribuiti alle classi della VQR.

Ovviamente, una perfetta concordanza implica anche la non esistenza di differenze sistematiche tra F e P, ma il contrario non è necessariamente vero, e in generale i due criteri misurano due diversi aspetti della differenza esistente tra le due distribuzioni. Si consideri ad esempio una distribuzione con un basso grado di concordanza tra F e P (molti articoli ricevono differenti valutazioni F e P). Anche in tale caso può accadere che, in media, F e P forniscano un punteggio complessivo simile. Questa distribuzione sarebbe caratterizzata da un basso livello di concordanza e da un basso grado di differenza sistematica: adottare uno dei due metodi di valutazione (per esempio quella bibliometrica, F) comporterebbe una frequente differenza di valutazione degli articoli sulla base della bibliometria e della valutazione *peer* (ossia, si avrebbero molti articoli con una buona valutazione in base a F, ma una peggiore valutazione in base a P, o viceversa).

Alternativamente, si consideri un caso di elevata (ma non perfetta) concordanza tra F e P. In questo caso, potrebbe ancora succedere che, per esempio, il numero di articoli con classificazione elevata sia sistematicamente maggiore in F che in P. In questo caso si avrebbe un elevato grado di concordanza, ma anche un alto grado di differenza sistematica tra le due distribuzioni, dato che il punteggio medio attribuito da F differirebbe dal punteggio medio di P. Adottare uno dei due metodi di valutazione può risultare in una sopravalutazione (o sottovalutazione) in relazione all'altro criterio: ossia, gli articoli riceverebbero un punteggio notevolmente diverso se valutati con F o con P.

Da un punto di vista statistico, il grado di concordanza tra F e P può essere misurato utilizzando la statistica K di Cohen; differenze sistematiche tra F e P possono invece essere misurate guardando alle differenze tra le medie delle distribuzioni e valutandone la significatività con un test *t* di *Student*.

A.3.1 Il grado di concordanza tra le distribuzioni F e P

La Tabella A.7 riporta i valori della statistica K di Cohen, calcolati per l'intero campione e separatamente per ciascun SSD. I risultati sono riferiti a campioni omogenei (*paired sample*), ossia ai Prodotti del campione per i quali sono disponibili sia i risultati della valutazione *peer* sia quelli relativi alla valutazione bibliometrica, eliminando cioè dal campione i Prodotti per i quali la valutazione bibliometrica fornisce come risultato una classificazione IR; nel caso del GEV di Fisica, le osservazioni a disposizione si riducono a 1212. La statistica K è costruita in modo tale da essere pari a zero quando la concordanza tra due (o più) valutazioni è del tutto casuale, vale a dire nel caso in cui le valutazioni siano indipendenti l'una dall'altra; la statistica assume invece valore pari ad 1 nel caso in cui ci sia perfetta concordanza. E' possibile calcolare il test utilizzando una matrice standard di pesi lineari (1; 0,67; 0,33; 0) attribuiti ai casi di concordanza, discordanza di una classe e così via, rispettivamente. In questo caso, nel totale del campione, K è uguale a 0,23, un valore statisticamente diverso da zero agli usuali livelli di significatività. Il grado di concordanza negli settori scientifici disciplinari presenta valori generalmente vicini a quelli registrati per l'intero campione: a parziale eccezione, in Fisica nucleare e subnucleare la statistica K è pari a 0,9 mentre è pari a 0,37 in Fisica per il sistema terrestre e circumterrestre. In ogni caso, il test z conferma che i valori della statistica K di Cohen sono sempre statisticamente diversi da zero a un livello di confidenza dell'1% (ovvero la probabilità che siano uguali è inferiore all'1%).

Come accennato sopra, il calcolo di K riportato nella prima riga della tabella usa pesi lineari. È possibile argomentare che nel nostro caso i pesi appropriati da utilizzare debbano però essere quelli suggeriti dalle regole della VQR. In particolare, è possibile calcolare la distanza tra le valutazioni utilizzando i punteggi numerici della VQR (1; 0,8; 0,5; 0), associati con le valutazioni qualitative (E; B; A; L). La seconda riga della Tabella A.7 riporta i valori della statistica K calcolati utilizzando i pesi della VQR. I risultati mostrano che in questo caso la concordanza è sempre maggiore rispetto alle valutazioni basate su pesi lineari, a rafforzamento dell'ipotesi dell'esistenza di un buon grado di concordanza sia nel totale del campione che in ciascun SSD.

Tabella A.7: Statistica K di Cohen sul grado di concordanza: La tabella riporta la statistica K e in parentesi il valore z[3] ad essa associato. La presenza di uno o due asterischi indica la significatività del test al livello rispettivamente del 5% e dell'1%.

[3] Il test z verifica se K è statisticamente pari a zero assumendone una distribuzione Gaussiana, o normale. Si calcola dividendo il valore di K per il suo errore standard. Se il valore di z è superiore al valore-soglia della distribuzione normale corrispondente a una certa probabilità, si conclude che la statistica K è statisticamente diversa da zero, ossia che le due valutazioni non sono indipendenti l'una dall'altra e mostrano quindi un grado statisticamente significativo di concordanza.

Test	*Totale campione*	*Fisica sperimentale*	*Fisica teorica, modelli e metodi matematici*	*Fisica della materia*	*Fisica nucleare e subnucleare*	*Astronomia e astrofisica*	*Fisica per il sistema terra e il mezzo circumterrestre*	*Fisica applica, Didattica e Storia della fisica*
F e P, pesi lineari	0,2302 (14,26)**	0,1957 (4,00)**	0,2428 (8,18)**	0,1862 (7,26)**	0,951 (1,83)*	0,2708 (7,19)**	0,3671 (2,37)**	0,2153 (3,33)**
F e P, pesi VQR	0,2515 (15,10)**	0,2049 (3,95)**	0,2559 (8,59)**	0,2099 (7,70)**	0,1001 (1,96)*	0,3048 (8,00)**	0,3975 (2,52)**	0,2715 (3,77)**
P1 e P2, pesi lineari	0,2337 (11,65)**	0,2761 (4,41)**	0,1968 (5,81)**	0,2038 (5,17)**	0,4225 (3,70)**	0,2461 (5,46)**	0,3963 (2,45)**	0,1785 (2,14)**
P1 e P2, pesi VQR	0,2271 (11,33)**	0,2719 (4,30)**	0,1979 (5,86)**	0,1932 (4,85)**	0,4216 (5,36)**	0,2342 (5,36)**	0,4387 (2,72)**	0,1831 (2,10)**

La Tabella A.7 riporta anche la statistica K per il grado di concordanza tra i due revisori (P1 e P2), sia per il totale del campione che per i singoli SSD. Nel complesso del campione, il grado di concordanza tra la valutazione bibliometrica (F) e la revisione *peer* (P) è simile a quello esistente tra i giudizi formulati tra i due revisori esterni: in quest'ultimo caso, la statistica K calcolata con pesi lineari è pari a 0,23 (0,27 con i pesi della VQR). Analoghi risultati si hanno a livello dei singoli SSD. Il test z associato conduce sempre a rifiutare l'ipotesi nulla di non concordanza.

A.3.2 Il grado di differenza sistematica tra le distribuzioni F e P

La Tabella A.8 riporta il punteggio medio risultante dalle valutazioni F e P. I valori numerici sono ottenuti sommando i pesi assegnati dalla VQR alle quattro classi di merito e dividendo per il numero degli articoli valutati. Si noti ancora una volta come, date le regole della VQR, gli scarti tra F e P non abbiano lo stesso peso: ad esempio, la differenza tra L e A ha un peso di 0,5, mentre la differenza tra E e B ha un peso pari solo a 0,2. Come nel caso delle analisi contenute nella sezione precedente, i risultati riportati sono riferiti a campioni omogenei (*paired sample*), ossia ai Prodotti del campione per i quali sono disponibili sia i dati della valutazione *peer* sia quelli relativi alla valutazione bibliometrica, eliminando cioè dal campione i Prodotti per i quali la valutazione bibliometrica fornisce come risultato una classificazione IR. Come ricordato sopra, gli articoli a disposizione in questo caso sono 1212.

La terza colonna mostra che il punteggio medio finale della revisione *peer* (punteggio P) è pari a 0,72, la quarta colonna contiene il punteggio medio ottenuto nella valutazione bibliometrica. Il risultato più interessante dell'analisi è mostrato nella sesta colonna, che presenta la differenza tra valutazione *peer* e valutazione bibliometrica, con le colonne 8-9 che riportano il risultato del test *t* (e del corrispondente *p-value*), per verificare se la differenza tra P e F è significativa. Nel totale del campione, emerge una differenza sistematica tra la valutazione bibliometrica e la valutazione *peer*: più precisamente, la valutazione media ottenuta con l'analisi bibliometrica è migliore di quella ottenuta con la valutazione *peer*. Il risultato è confermato anche per i dati riferiti ai sette SSD.

Tabella A.8: Test t sulla differenza tra i punteggi bibliometrici e peer review

SSD	*Punteggio P1*	*Punteggio P2*	*Punteggio P*	*Punteggio F*	*Diff F-P*	*# Osservazioni*	*Test t*	*p-value*
Fisica sperimentale	0,69	0,72	0,68	0,85	0,17	119	5,9488	0,00
Fisica teorica, modelli e metodi matematici	0,74	0,77	0,73	0,84	0,11	423	7,125	0,00
Fisica della materia	0,74	0,69	0,69	0,91	0,21	307	13,788	0,00
Fisica nucleare e subnucleare	0,87	0,83	0,83	0,99	0,16	41	4,491	0,00
Astronomia e astrofisica	0,79	0,76	0,78	0,85	0,07	236	4,237	0,00
Fisica per il sistema terra e il mezzo circumterrestre	0,76	0,74	0,74	0,84	0,10	18	1,401	0,18
Fisica applicata, didattica e storia della fisica	0,61	0,66	0,60	0,76	0,16	68	4,103	0,00
Totale	0,74	0,74	0,72	0,86	0,14	1212	16,407	0,00

A.3.3 Prime conclusioni

Nel totale del campione dei Prodotti del GEV02 conferiti per la valutazione, si riscontra una più che adeguata concordanza tra valutazioni effettuate con il metodo della revisione tra pari e con quello bibliometrico. Inoltre, il grado di concordanza tra valutazione finale bibliometrica e *peer* è molto simile al grado di concordanza tra le due valutazioni *peer*. D'altro canto, però, emerge evidenza di differenze sistematiche tra i punteggi corrispondenti alle valutazioni *peer* e bibliometriche. In effetti, è possibile osservare che il numero di Prodotti della ricerca classificati

come eccellenti (E) con l'algoritmo di valutazione bibliometrica sia superiore a quello dei Prodotti eccellenti secondo la valutazione tra pari.

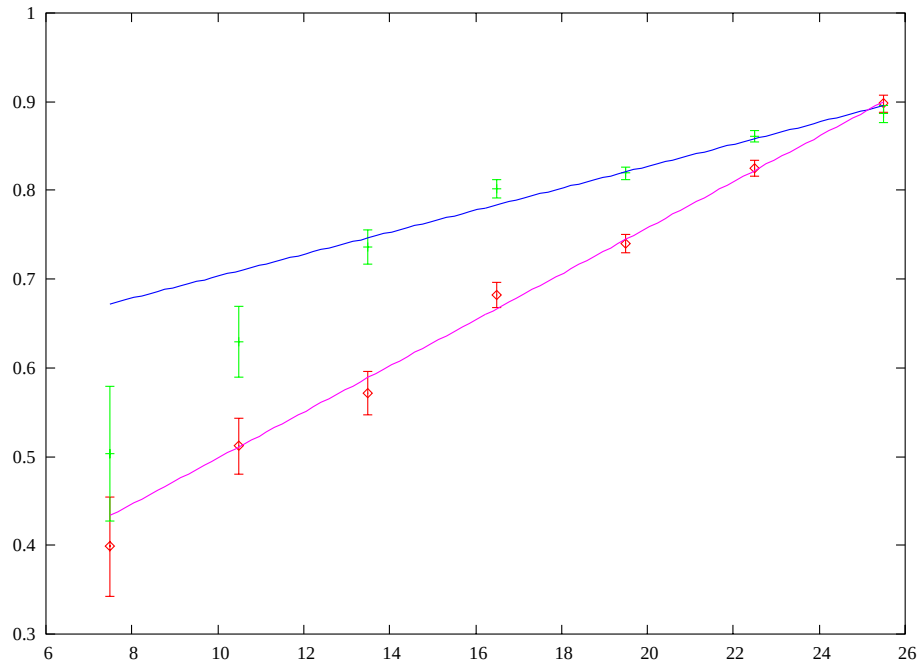
Il grado di concordanza tra valutazioni *peer* e valutazioni bibliometriche è elevato in tutti gli SSD. Le differenze sistematiche tra i punteggi medi sono statisticamente significative e sempre di segno positivo (ossia, la valutazione bibliometrica è significativamente più favorevole in media rispetto a quella *peer*).

A.4 Un'analisi sintetica sui dati grezzi

Come si è visto precedentemente, la classificazione di un Prodotto nelle classi di merito (E, B, A, L, IR, IP) influenza il confronto tra bibliometria e PR. Conviene quindi effettuare il confronto direttamente sui dati grezzi, anche allo scopo di estrarre stime possibili dei vari contributi agli errori. Utilizzeremo le valutazioni dei revisori nell'intervallo 3-27 e il percentile dell'*impact factor* e delle citazioni, normalizzato per anno e SC WoS. Avendo riscontrato nella sezione precedente l'assenza di variazioni significative tra vari SSD, ci limiteremo a un'analisi globale.

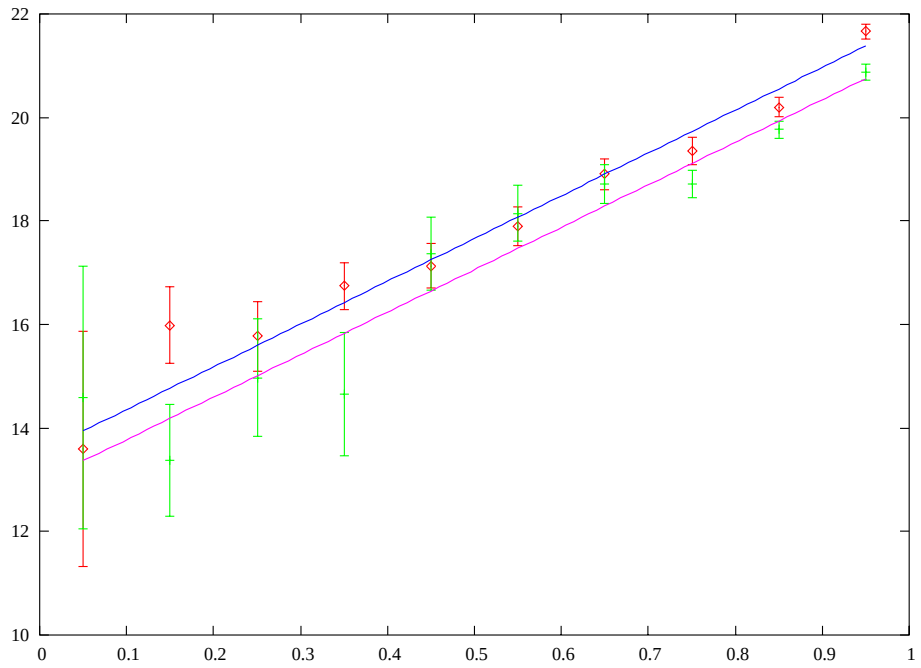
Presentiamo per primo il grafico del percentile dell'*impact factor* e del percentile delle citazioni come funzione della valutazione media (Fig. A.1). È evidente la forte correlazione fra le due quantità. Le barre d'errore indicano l'errore statistico standard. Un *fit* lineare funziona bene nel caso delle citazioni ($\chi^2 = 0{,}5$ per grado di libertà) mentre si osservano delle evidenti variazioni dalla linearità nel caso dell'*impact factor* (un *fit* lineare dà un χ^2 di 2,5 per grado di libertà, dominato dai Prodotti con bassa valutazione *peer*).

Fig 1: Percentile delle citazioni (rosso) e dell'impact factor (verde) come funzione della valutazione media peer raggruppata in sette classi.



Lo stesso grafico può essere fatto all'inverso, ovvero scambiando gli assi e raggruppando i dati a seconda del percentile (o dell'*impact factor* o della PR). In Fig. A.2 presentiamo il grafico della media della valutazione *peer* versus il percentile dell'*impact factor* e delle citazioni. In questi grafici *inversi* l'ipotesi lineare funziona lievemente meglio per l'*impact factor* (il *fit* lineare ha un $\chi^2 = 1$ per il percentile delle citazioni e un $\chi^2 = 2$ per il percentile dell'*impact factor*). Tuttavia in questa seconda versione la non linearità per il percentile dell'*impact factor* si manifesta principalmente nella regione ad alto *impact factor*.

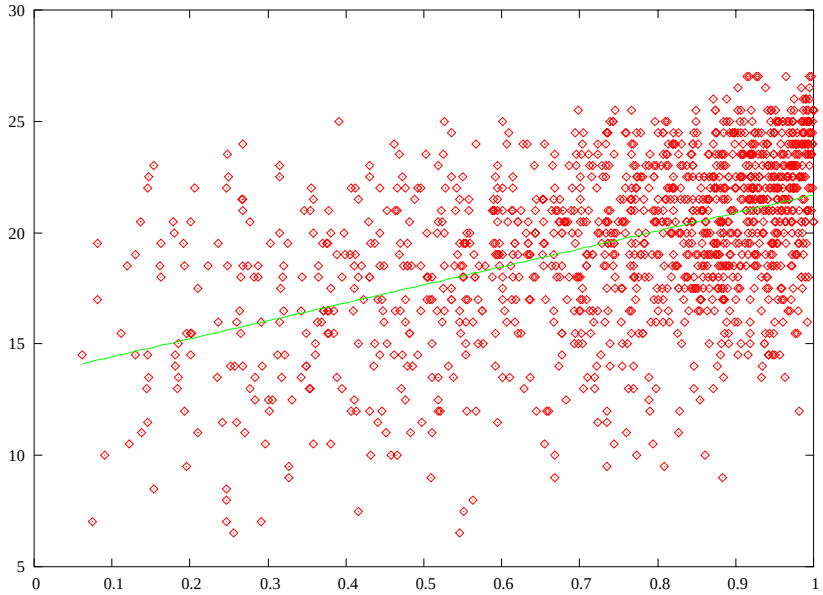
Fig. A.2: Valutazione media PR come funzione del percentile delle citazioni (rosso) e del percentile dell'impact factor (verde) raggruppati in 10 intervalli del corrispondente percentile.



L'ipotesi di linearità della valutazione come funzione del numero delle citazioni è ben soddisfatta, mentre ci sono problemi con l'*impact factor* (che pur rimanendo ben correlato, non ha probabilmente una dipendenza lineare). La parte restante della sezione è dedicata all'analisi del percentile delle citazioni.

È possibile ridisegnare la figura A.2, per quanto riguarda il percentile delle citazioni, senza raggruppare i dati in intervalli di percentile. Il *fit* lineare risulta molto simile a quello della figura precedente.

Fig. A.3: Valutazione media come funzione del percentile delle citazioni (ogni Prodotto è un punto).



A questo punto sussistono tutti gli elementi per fare un'analisi degli errori. Infatti, se si studiano gli scarti quadratici medi σ^2 tra la valutazione media dei due Revisori e il *fit* lineare, si trova $\sigma^2 = 11$. In altri termini, l'informazione sulle citazioni ci permette di ricostruire la valutazione media con un errore $\sigma = 3{,}3$ (nella scala da 3 a 27).

Ma quest'errore è dovuto a un'inaccuratezza della previsione basata sulle citazioni o a un errore dovuto alla variabilità di valutazione tra i revisori? È possibile dare una risposta (sorprendente) a questa domanda facendo un semplice un esperimento concettuale: si supponga di mandare lo stesso articolo a M revisori. Si supponga, inoltre, che nel limite di M che tende a infinito la media della valutazione dei revisori vada al valore intrinseco dell'articolo (questa è più una definizione del valore intrinseco che un'ipotesi).

Si ottiene banalmente che

$$\sigma^2(M) = A + \frac{B}{M} \ , \qquad (A.1)$$

dove A è il contributo allo scarto quadratico medio dovuto alla valutazione basata sulle citazioni e B è il contributo dei Revisori. Fortunatamente si dispone anche dei dati per $M = 1$ (è sufficiente ridisegnare la figura precedente utilizzando il dato relativo a un solo revisore). In questo caso si trova che $\sigma^2 = 18$. Ne deriva che $A = 4$ e $B = 14$.

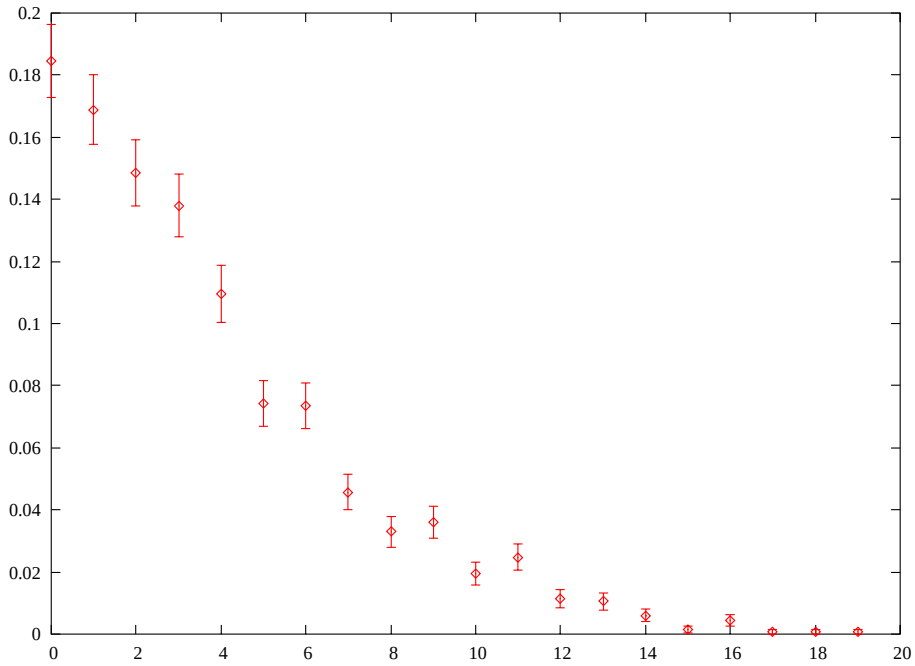
Possiamo quindi attribuire un errore dovuto alla valutazione citazionale pari a 2 mentre l'errore della valutazione media dei revisori è $\sqrt{B/2} = 2{,}7$. Una conferma della correttezza della stima di B può ottenersi guardando il valor medio del quadrato della differenza tra le valutazioni dello stesso Prodotto di due revisori. Questa quantità (Q) ha un valore di 27 (con un errore di 1),

mentre l'analisi precedente predice *Q = 2*B*. Le due analisi, come ci si poteva aspettare, sono dunque consistenti.

Si noti che l'analisi è stata fatta senza includere né l'informazione sull'*impact factor* né le correzioni dipendenti dai *PACS Numbers*, che dovrebbero rendere l'analisi bibliometrica più accurata, diminuendo l'errore. Già senza tali fattori migliorativi da questa analisi risulta che sarebbero necessari quattro revisori per Prodotto per avere una stima PR affetta dallo stesso errore della (cioè, precisa quanto la) bibliometria.

In conclusione, viene mostrato in Fig. A.4 l'andamento della grandezza "frazione di Prodotti per cui la differenza (sempre normalizzata tra 3 e 27) tra le valutazioni dei due revisori differisce di *x"*, in funzione di |*x*|. La distribuzione non è compatibile con una gaussiana e un *fit* ragionevole (χ^2 = 1,5) è dato da $f_{a,b}(x) = a\,(1 + b * x)e^{-b*x}$ con $b = 0{,}38$ e con errore pari a 0,01.

Fig. A.4: Frazione di Prodotti con una data differenza nelle valutazioni dei due revisori come funzione del valore assoluto della differenza. Le barre di errore indicano gli errori statistici standard.



National Agency for the Evaluation of Universities and Research Institutes
anvur
Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca

Evaluation of Research Quality
VQR
Valutazione Qualità della Ricerca

Appendice B. Il Documento Criteri dell'Area 2

B1. Introduzione

Questo documento ha come oggetto l'organizzazione, il funzionamento e le modalità valutative del Gruppo di Esperti della Valutazione (GEV) istituito dall'ANVUR per l'Area 02, Scienze Fisiche.

Il documento è suddiviso in tre parti: la prima parte (sezioni 2 e 3) precisa le modalità organizzative del GEV 02, la seconda (sezioni 4-7) stabilisce i criteri e le modalità a cui il GEV 02 si atterrà ai fini dell'esercizio di valutazione, la terza parte (sezione 8) è dedicata alle regole per i conflitti d'interesse.

In linea con gli obiettivi generali dell'esercizio di Valutazione della Qualità della Ricerca (VQR) 2004-2010, i criteri adottati del GEV 02 mirano a fornire un panorama qualitativo della ricerca nel campo delle Scienze Fisiche in Italia. Nel complesso, i prodotti che il GEV 02 esaminerà per questo esercizio saranno principalmente articoli su riviste scientifiche. Questi articoli saranno classificati prevalentemente utilizzando indicatori bibliometrici, integrati allo scopo di valutare sia la qualità della rivista in cui sono pubblicati sia il loro impatto specifico, quantificato in base al numero citazioni ricevute dall'articolo fino al 31 dicembre 2011. Chiaramente, articoli pubblicati all'inizio del periodo di valutazione hanno avuto più tempo per accumulare citazioni e raggiungere un valore statisticamente significativo dell'impatto, rispetto ai lavori pubblicati negli ultimi anni del periodo di valutazione. Questi ultimi potranno richiedere valutazioni aggiuntive secondo la metodologia dell'*informed peer review*.

B2. Delimitazione dell'area GEV 02

Area 02, Scienze Fisiche:

- **FIS/01** (Fisica Sperimentale),
- **FIS/02** (Fisica Teorica, Modelli e Metodi Matematici),
- **FIS/03** (Fisica della Materia),
- **FIS/04** (Fisica Nucleare e Subnucleare),
- **FIS/05** (Astronomia e Astrofisica),
- **FIS/06** (Fisica per il Sistema Terra e il Mezzo Circumterrestre),
- **FIS/07** (Fisica Applicata a Beni Culturali, Ambientali, Biologia e Medicina),
- **FIS/08** (Didattica e Storia della Fisica).

B3. Organizzazione del GEV 02

B3.1 sub-GEV

I vari Sub-GEV hanno competenze prevalenti ma non esclusive nei seguenti Settori Scientifico-Disciplinari (SSD). **NB: per la composizione aggiornata del GEV e dei subGEV si faccia riferimento alle Tabelle 1.4 e 1.5, sezione 1 del Rapporto finale di Area.**

Sub-GEV I	FIS/01 (Fisica Sperimentale) FIS/04 (Fisica Nucleare e Subnucleare)	Battiston, **Bracco**, Calabrese, De Palma, Diemoz, Gianotti
Sub-GEV II	FIS/02 (Fisica Teorica, Modelli e Metodi Matematici) FIS/03 (Fisica della Materia) FIS/04 (Fisica Nucleare e Subnucleare)	Andreoni, **Guerra**, Parisi, Parrinello, Zwirner
Sub-GEV III	FIS/05 (Astronomia e Astrofisica) FIS/06 (Fisica per il Sistema Terra e il Mezzo Circumterrestre)	De Bernardis, Matteucci, **Matarrese**
Sub-GEV IV	FIS/01 (Fisica Sperimentale) FIS/03 (Fisica della Materia) FIS/07 (Fisica Applicata a Beni Culturali, Ambientali, Biologia e Medicina) FIS/08 (Didattica e Storia della Fisica)	Arimondo, Di Fabrizio, **Ferdeghini**, Siciliano

Il coordinatore del Sub-GEV è indicato in grassetto.

B3.2 Allocazione dei prodotti all'interno del GEV

L'allocazione dei prodotti ai Sub-GEV avverrà sulla base del SSD e dei *PACS Numbers* 2010 (http://www.aip.org/pacs/) indicati dal soggetto valutato e trasmessi dalle Strutture. Il significato del SSD associato al prodotto, che può anche essere diverso dal SSD di appartenenza del soggetto valutato, si riferisce al GEV che con maggior competenza, secondo il soggetto valutato, può valutare il prodotto in questione.

Nel caso in cui un prodotto sia assegnato a più GEV per il suo carattere interdisciplinare saranno adottati identici criteri di valutazione concordati tra i vari GEV. A tale scopo, i Presidenti dei GEV interessati istituiscono specifici Gruppi di Consenso Inter-Area.

B3.3 Regole di funzionamento del GEV

- La convocazione del GEV avviene almeno quindici giorni prima della riunione. La riunione è convocata dal Presidente, che ne fissa l'ordine del giorno.
- Le decisioni all'interno del GEV vengono prese a maggioranza semplice dei presenti. Per partecipare alla votazione non è necessario essere fisicamente presenti alle riunioni, purché presenti in modalità telematica.
- Alle riunioni del GEV partecipa, con funzioni di segretario senza diritto di voto, anche l'assistente attribuito dall'ANVUR al GEV, Dott. Ric. Alberto Francesco Anfossi. Al termine di ciascuna riunione viene redatto un resoconto della seduta in italiano e un verbale succinto che riporta le conclusione principali in lingua italiana e in lingua inglese. I verbali vengono fatti circolare tra i membri del GEV e approvati tramite email o utilizzando l'ambiente *software* predisposto dal CINECA.

B4. Mix valutativo

Salvo restando che la responsabilità finale della valutazione è affidata al GEV, il GEV 02 utilizzerà per la valutazione la tecnica della *informed peer review*, vale a dire una combinazione di criteri bibliometrici e di revisione *peer*.

I prodotti di cui alla tipologia "a" della sezione 2.3 del Bando ANVUR VQR 2004-2010 identificati nelle banche dati *Web of Science* di *Thomson Reuters* (WoS) e/o Scopus di *Elsevier B. V.* (SCOPUS) verranno valutati utilizzando i criteri bibliometrici descritti nella sezione 6.5.

I prodotti delle tipologie "b", "c", "d", "e" elencate nella sezione 2.3 del Bando ANVUR VQR 2004-2010 verranno valutati utilizzando la revisione *peer*.

I prodotti valutati mediante *peer review* (che prevede l'invio a revisori esterni oppure, in alcuni casi limitati per i quali vi siano le competenze richieste all'interno, la valutazione diretta *peer* all'interno del GEV) appartengono dunque a quattro categorie:

1. prodotti di ricerca di cui alle tipologie "b", "c", "d", "e" elencate nella sezione 2.3 del Bando ANVUR VQR 2004-2010;
2. articoli che sono indicizzati nelle banche dati WoS e/o SCOPUS che richiedono la *peer review* in base all'algoritmo bibliometrico descritto nella sezione 6.5;
3. articoli che sono indicizzati nelle banche dati WoS e/o SCOPUS e che saranno valutati utilizzando l'algoritmo bibliometrico e la *peer review* al fine di studiare la correlazione tra i due metodi di valutazione: tali articoli saranno individuati tramite un algoritmo di campionamento casuale stratificato studiato da un Gruppo di Lavoro dell'ANVUR;
4. articoli pubblicati su riviste non indicizzate.

B5. *Peer review*

Nel caso di utilizzo della *peer review* i prodotti saranno inviati a due revisori esterni oppure, in un numero limitato di casi, valutati, sussistendo le condizioni di competenza e di assenza di conflitti, all'interno del GEV utilizzando le stesse procedure e la stessa scheda di revisione. La scelta dei revisori esterni verrà effettuata evitando conflitti di interesse tra i revisori stessi e gli autori e/o la struttura di affiliazione. Inoltre, verrà garantita l'indipendenza dei revisori ponendo attenzione alla sede di affiliazione, alla collaborazione scientifica e, ove possibile, alla nazionalità. Per minimizzare i conflitti di interesse si privilegeranno i revisori operanti al di fuori dei confini nazionali.

B5.1 L'individuazione dei revisori *peer*

Il GEV intende coinvolgere revisori esterni con un profilo di ricerca internazionale, un curriculum di altro profilo, testimoniato, in particolare negli ultimi anni, da un elevato numero di pubblicazioni nelle sedi di riferimento della comunità scientifica internazionale del settore, un significativo numero di citazioni e la necessaria competenza nella specifica area di valutazione. Il GEV preparerà un elenco di revisori esterni, stabilendo *standard* minimi di qualità scientifica, d'impatto sulla comunità scientifica internazionale e di esperienza nella valutazione.

Grande attenzione verrà posta al mantenimento dell'anonimato dei revisori, sia nella fase di predisposizione dell'elenco dei revisori sia nella fase operativa di valutazione.

Per quanto attiene alla prima, il Presidente GEV consulterà la lista di revisori della propria area resa disponibile dal CINECA e chiederà ai componenti GEV, tramite i coordinatori dei Sub-GEV, di suggerire un numero significativo di esperti che soddisfano ai criteri indicati nel paragrafo precedente.

Il Presidente GEV raccoglierà le indicazioni corredate d'informazioni fornite sulla base di una scheda condivisa e, anche con l'ausilio dei coordinatori dei Sub-GEV, provvederà a modificare la lista CINECA con integrazioni e/o cancellazioni.

Il processo di integrazione della lista continuerà per tutta la durata dell'esercizio di valutazione, sulla base delle necessità che dovessero emergere a valle della trasmissione dei prodotti da parte delle Strutture.

B5.2 Assegnazione dei livelli di merito VQR sulla base delle valutazioni *peer*

La valutazione dei revisori *peer* si baserà su un'apposita scheda revisore predisposta dal GEV, costituita da una serie di domande a risposta multipla e da un campo libero con numero limitato di parole. Il GEV trasformerà le indicazioni contenute nella scheda revisore in una delle quattro classi finali di merito. Nel caso di valutazioni non convergenti dei revisori *peer* o, nel caso di disponibilità di entrambe, tra *peer review* e analisi bibliometrica, il Sub-GEV creerà al suo interno un Gruppo di Consenso con il compito di proporre al GEV il punteggio finale del prodotto oggetto del giudizio

difforme dei revisori esterni mediante la metodologia del *consensus report*. Il Gruppo di Consenso potrà avvalersi anche del giudizio di un terzo esperto. In ogni caso la responsabilità della valutazione conclusiva è dell'intero GEV.

B6. Analisi bibliometrica

B6.1 Basi di dati

Il GEV utilizzerà come base di dati di riferimento principale WoS, opportunamente integrata dalla banca di dati SCOPUS per confronti specifici.

B6.2 Finestra temporale delle citazioni

Nel calcolo dell'indicatore bibliometrico il GEV utilizzerà le citazioni fino al 31 Dicembre 2011.

B6.3 Autocitazioni

L'opportunità di includere o escludere le autocitazioni nella valutazione bibliometrica è tuttora oggetto di dibattito nella comunità scientifica. Nella VQR 2004-2010 non verranno escluse le autocitazioni per motivi di ordine tecnico legati soprattutto all'uso di WoS, che non lo consente direttamente, e ai problemi legati alla disambiguazione dei nomi degli autori, se lo si volesse realizzare a partire dai dati grezzi.

B6.4 Gli indicatori bibliometrici

La valutazione si baserà su una classificazione della rivista su cui il prodotto è stato pubblicato e su un indicatore bibliometrico che misura l'impatto del prodotto nel periodo che va dalla data di pubblicazione al 31 Dicembre 2011.

In particolare verranno considerati:

- l'*Impact Factor di Journal Citation Reports* di *Thomson Reuters* (IF) della rivista nell'anno di pubblicazione;
- il numero di citazioni ricevute dal prodotto fino al 31 dicembre 2011;

B6.5 Determinazione della classe finale di merito

La determinazione della classe finale di merito, tra le quattro previste dalla VQR 2004-2010 (eccellente, buono, accettabile, limitato) sarà funzione delle seguenti quantità:

- l'IF della rivista e/o l'analogo indicatore SCOPUS *SCImago Journal Rank* (SJR);
- le citazioni ricevute e la data di pubblicazione del prodotto;
- i *PACS Numbers* del prodotto indicati da chi sottopone il prodotto alla valutazione;

- la *Science Category* (SC) WoS e/o l'analogo indicatore SCOPUS *All Science Journals Classification* (ASJC).

I parametri che controllano la dipendenza della funzione dai *PACS Numbers* saranno decisi dal GEV dopo un'ampia sperimentazione di calibrazione dell'algoritmo, al termine della quale il GEV si impegna a rendere pubblici i parametri utilizzati per ciascun *PACS Number*.

La valutazione utilizzerà, per tutti gli articoli pubblicati su riviste indicizzate nelle basi di dati WoS e SCOPUS, un algoritmo che terrà conto sia del numero di citazioni sia dell'indicatore bibliometrico della rivista ospitante (IF, SJR o combinazioni di diversi indici). Tale scelta è dettata dalle seguenti considerazioni:

a. il puro indicatore citazionale assume valori anche molto piccoli a seconda della disciplina e dell'età dell'articolo, rendendo difficile una discriminazione effettiva tra classi di merito; queste considerazioni sono il risultato di una significativa sperimentazione effettuata utilizzando le basi di dati acquisite per l'esercizio VQR 2004-2010;

b. l'uso del solo indicatore citazionale costituisce un obiettivo facile per possibili future manipolazioni, inducendo comportamenti fuorvianti quali l'estensiva autocitazione e la citazione mutua all'interno di un gruppo ristretto, non giustificate da considerazioni di natura scientifica.

L'algoritmo utilizzato per la classificazione degli articoli nelle quattro classi di merito della VQR 2004-2010 sarà di norma il seguente:

1. dato l'articolo e la rivista che l'ha pubblicato si identifica la corrispondente SC in WoS o la ASJC in SCOPUS; nel seguito l'algoritmo verrà descritto con riferimento alle sole SCs e all'IF, essendo implicito che identica procedura potrà essere seguita per le ASJC e per altri indicatori bibliometrici;

2. l'articolo è anche caratterizzato dai *PACS Numbers* indicati da chi sottomette l'articolo;

3. se la rivista appartiene a più di una SC si utilizza, ai fini dell'individuazione univoca della SC, l'indicazione del soggetto valutato che ha proposto l'articolo o, se necessario, l'eventuale modifica da parte del GEV;

4. sia in WoS sia in SCOPUS esiste la categoria *multidisciplinary science*, che include riviste caratterizzate da una pluralità di argomenti scientifici, quali per esempio *Nature* e *Science;* gli articoli pubblicati su una rivista di tale categoria saranno riassegnati a un'altra SC sulla base delle citazioni contenute nell'articolo. In particolare, per ognuno degli articoli pubblicati sulle riviste citate si individuerà una (o più) SC di appartenenza, e si sceglierà la SC finale con una regola di decisione maggioritaria. Nell'assegnazione alla nuova SC, l'articolo porterà con sé l'IF della rivista e il numero di citazioni ricevute;

5. nel caso in cui le SC siano eccessivamente ampie verranno introdotte delle *Subject Sub-Categories* (SSC) sulla base dei *PACS Numbers*, che costituiscono una classificazione più fine di quella fornita dalle SCs;

6. si calcola la funzione di distribuzione cumulativa empirica dell'IF delle riviste appartenenti alla SC individuata per l'anno di pubblicazione dell'articolo da valutare. Nel caso in cui la SC sia stata suddivisa in SSC il GEV determinerà un'opportuna classificazione delle riviste per la SSC, sempre rispettando l'ordinamento per IF;

7. si divide la funzione di distribuzione cumulativa in quattro classi di IF decrescente, caratterizzate dai valori di probabilità 0.2, 0.2, 0.1, 0.5;

8. si calcola la funzione di distribuzione cumulativa empirica del numero di citazioni di tutti gli articoli (dalla data di pubblicazione al 31 dicembre 2011) pubblicati dalle riviste appartenenti alla SC individuata per l'anno di pubblicazione dell'articolo da valutare. Nel caso in cui la SC sia stata suddivisa in SSC il GEV individuerà degli opportuni fattori correttivi che tengano conto delle differenti pratiche citazionali nelle SSC;

9. si divide la funzione di distribuzione cumulativa del numero di citazioni in quattro classi di numero di citazioni decrescente, caratterizzate dai valori di probabilità 0.2, 0.2, 0.1, 0.5;

10. dati l'IF e il numero di citazioni dell'articolo da valutare, esso viene attribuito ad una delle sedici coppie di classi rappresentate in una matrice 4x4, che contiene in colonna le classi della distribuzione dell'IF e in riga le classi della distribuzione del numero di citazioni;

11. l'attribuzione della classe finale di merito avviene secondo l'algoritmo seguente, nel quale la lettera A si riferisce alla classe finale "eccellente", la B a "buono", la C ad accettabile, e la D a "limitato". Gli elementi etichettati "IR" si riferiscono ai casi nei quali il GEV valuterà direttamente l'articolo, sulla base dell'insieme dei dati bibliometrici, della data di pubblicazione oppure lo affiderà alla *peer review* esterna.

a. Quando le coordinate dell'articolo lo posizionano in una dei quattro elementi della diagonale principale, e quindi le due indicazioni basate su IF e su citazioni coincidono, la classe finale è la stessa (Figura 1);

b. quando le coordinate dell'articolo lo posizionano nella prima riga, seconda colonna, ovvero classe A per le Citazioni, classe B per IF, l'articolo è assegnato alla classe di merito A;

c. quando le coordinate dell'articolo lo posizionano nella seconda riga, terza colonna, ovvero classe B per le Citazioni, classe C per IF, l'articolo è assegnato alla classe di merito B;

d. in tutti gli altri casi l'articolo è assegnato alla categoria IR e il GEV valuterà direttamente l'articolo, sulla base dell'insieme dei dati bibliometrici, data di pubblicazione oppure lo affiderà alla *peer review* esterna.

Fig. B.1: Matrice di corrispondenza tra classi iniziali di IF, citazioni e classe finale VQR 2004-2010.

Indicatore bibliometrico

N. di citazioni	1	2	3	4
1	A	A	IR	IR
2	IR	B	B	IR
3	IR	IR	C	IR
4	IR	IR	IR	D

B6.6 Gestione dei conflitti

Nel caso di disponibilità di valutazione *peer* e bibliometrica per lo stesso prodotto, eventuali conflitti di attribuzione verranno risolti dal GEV su proposta del Sub-GEV tramite un gruppo di consenso.

B6.7 Assenza di indicatori di citazione

Tutti i prodotti non contenuti nelle basi di dati citazionali WoS e SCOPUS saranno sottoposti a valutazione diretta da parte del GEV o a *peer review* da parte di revisori esterni selezionati dal GEV.

B7. Altri prodotti

Gli *abstract* relativi ad atti di congresso pubblicati su riviste con codice ISSN saranno sottoposti a *peer review*, ma non potranno ricevere una classificazione VQR 2004-2010migliore di C.

I brevetti saranno sottoposti a *peer review* da parte di esperti esterni, anche stranieri. I livelli di merito VQR 2004-2010 A e B potranno essere assegnati esclusivamente a brevetti internazionali e/o che siano già stati ceduti o dati in licenza.

B8. Conflitti di interesse

I membri dei GEV si asterranno dal valutare o dall'assegnare ad altri membri GEV o a esperti esterni:

a. prodotti di cui siano autori o co-autori;

b. prodotti di cui siano autori o co-autori parenti o affini fino al quarto grado;

c. prodotti presentati da Università presso cui i membri stessi abbiano o abbiano avuto un rapporto di lavoro o con le quali abbiano svolto incarichi o collaborazioni ufficiali negli anni a partire dal 1/1/2007;

d. prodotti presentati da Enti di Ricerca vigilati dal MIUR e da altri soggetti pubblici o privati sottoposti volontariamente alla VQR 2004-2010 presso cui i membri stessi abbiano o abbiano avuto un rapporto di lavoro o con cui abbiano svolto incarichi o collaborazioni ufficiali, inclusa l'affiliazione a Enti di Ricerca, negli anni a partire dal 1/1/2007.

Nei casi di cui al punto d) precedente esiste conflitto d'interesse:

i. nel caso in cui la Struttura abbia una permanente strutturazione interna di tipo territoriale o disciplinare (es. Sezione locale di Ente di Ricerca, Istituto, Dipartimento), limitatamente ai prodotti presentati dalla stessa articolazione;

ii. nel caso in cui la Struttura non abbia una permanente strutturazione interna di tipo territoriale o disciplinare (es. Sezione locale di Ente di Ricerca, Istituto, Dipartimento), in riferimento a tutti i prodotti presentati nei limiti in cui ciò sia possibile senza precludere la possibilità di valutare il prodotto;

iii. nel caso in cui la strutturazione interna abbia luogo a più livelli gerarchici (es. più Istituti riuniti sotto un Dipartimento) il conflitto d'interesse sorge al livello più basso (i.e., membri GEV affiliati a Istituti diversi di uno stesso Dipartimento sono in conflitto di interesse soltanto rispetto a prodotti presentati da autori appartenenti allo stesso Istituto).

Nei casi di conflitto d'interesse il Presidente del GEV incaricherà delle procedure di valutazione un altro membro del GEV per cui non sussistano conflitti di interesse.

Nel caso di conflitti d'interesse che coinvolgano il Presidente del GEV, l'assegnazione dei prodotti relativi sarà fatta dal Coordinatore VQR 2004-2010 o da persona da lui incaricata.

National Agency for the Evaluation of Universities and Research Institutes
anvur
Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca

Evaluation of Research Quality
VQR
Valutazione Qualità della Ricerca

Appendice C. Considerazioni e analisi specifiche dell'Area 2

Premessa

Sia il DM sia il conseguente Bando VQR erano caratterizzati da una forte rigidità: successivamente i GEV hanno fatto la scelta di stabilire criteri molto dettagliati per la valutazione prima delle presentazioni dei Prodotti, senza lasciare spazio a successivi aggiustamenti in corso d'opera. In questa situazione è stato molto difficile (e in alcuni casi impossibile) porre rimedio ad alcune pecche nella progettazione dell'esercizio di valutazione, pecche che pur non inficiando il processo complessivo, contribuiscono all'aumento dell'inevitabile errore statistico sulla valutazione, diminuendone l'accuratezza.

Il GEV ritiene necessario sottolineare queste mancanze sperando che possano essere corrette nel successivo esercizio di valutazione. Dove possibile, presenteremo alcune proposte operative per il futuro. È cruciale avviare fin da adesso una riflessione sul prossimo esercizio valutativo. Sarebbe deplorevole se le Strutture di ricerca modificassero negativamente il loro comportamento nell'ipotesi che la prossima valutazione abbia esattamente gli stessi criteri dell'attuale.

Il GEV ha molto chiaro che alcune delle seguenti considerazioni possono essere specifiche dell'Area 2 (ma questi sono i soli dati che abbiamo analizzato) e che le problematiche possono essere molto diverse in altre Aree – specialmente in quelle non bibliometriche – ritiene tuttavia di essere un potenziale buon osservatore per vari motivi:

- l'Area 2 è molto omogenea e permette più facilmente un'analisi globale;
- l'Area 2 riceve un forte contributo dagli Enti di ricerca permettendo di esaminare bene i problemi valutativi degli Enti;
- è tradizione della Fisica analizzare grandi masse di dati sperimentali, variando durante l'esperimento i criteri di analisi dei dati cercando di minimizzare tutte le fonti di errore sia sistematico sia statistico[4]. L'impossibilità di fare questi successivi aggiustamenti nei criteri di analisi risulta particolarmente fastidiosa per un fisico abituato a procedere in maniera opposta nella sua vita professionale.

[4] La parola errore è usata con il significato tradizionale dell'analisi dei dati e non con il significato del linguaggio comune (ovvero sbaglio). Si parlerà di errore statistico, dovuto a un non completo campionamento e di errore sistematico in caso di distorsioni. Gli errori statistici sono inevitabili: non è per esempio possibile avere il parere di *tutti* gli esperti del campo su ogni Prodotto, usare un numero limitato di revisori per ciascun Prodotto è assolutamente necessario, ma induce un errore statistico non trascurabile. Inoltre, ci sono discrepanze sistematiche inevitabili tra la valutazione PR e quella bibliometrica, ovvero errori sistematici che possono essere corretti solo dopo essere stati osservati.

Nel presentare queste osservazioni cercheremo di seguire un ordine logico dividendole per argomenti. Lo svantaggio di procedere in questo modo è di mescolare osservazioni più importanti con altre meno importanti.

C.1 Prodotti valutabili e loro valutazione

Il meccanismo di valutazione adottato vuole arrivare alla valutazione delle Strutture facendo la media di valutazioni di un numero piccolo di pubblicazioni per singolo ricercatore (trascuriamo qui la problematica connessa a pubblicazioni in comune tra vari ricercatori che sarà discussa in seguito). Questo GEV, dopo aver analizzato i risultati della valutazione, ritiene che tale scelta metodologica, nella sua forma attuale, mescoli eccessivamente la produttività del singolo ricercatore con quella della Struttura valutata (Ente, Dipartimento, ecc.) e necessiti di alcuni piccoli ma significativi correttivi nella prossima VQR.

Per ogni Ricercatore universitario i Prodotti richiesti erano di norma 3 e non più di 3 (6 nel caso di un Ricercatore universitario con incarico di ricerca presso un Ente o di un dipendente di un Ente). Con questa richiesta la Struttura valutata non si avvantaggia di singoli molto produttivi (per esempio giovani che non abbiano altro carico di lavoro se non quello di fare ricerca e pubblicare, in grado di produrre più di 1 articolo/anno) e viene svantaggiata dalla presenza di singoli meno produttivi per motivi non necessariamente legati alla mediocrità (per esempio ricercatori maturi che abbiano responsabilità gestionali).

Questa scelta è diversa da quella fatta dal CIVR per la VTR 2001-2003. Infatti, nel precedente esercizio valutativo le Strutture erano state lasciate libere di selezionare indipendentemente dai nomi degli autori singoli i Prodotti da presentare. In questo modo possibili sacche improduttive sfuggivano alla valutazione precedente. La scelta della VQR attuale è dovuta alla giusta volontà di evidenziare e penalizzare tali sacche.

Tuttavia il risultato finale è che il paragone tra la valutazione finale di Strutture di grande qualità è spesso dominato dalla presenza casuale di un numero piccolo di Prodotti limitati o assenti. La presenza di un piccolo numero di Ricercatori inattivi scientificamente o con una produzione scientifica di basso valore è assolutamente fisiologica tenuto conto dei carichi gestionali e dell'età, a volte avanzata, di alcuni di essi. È necessario introdurre piccoli correttivi per evitare che fluttuazioni fisiologiche sul numero di ricercatori scientificamente poco attivi aumentino l'errore sul risultato finale.

Una soluzione molto semplice, che raccomandiamo fortemente, consiste nell'introdurre una piccola franchigia sia sui Prodotti mancanti sia sui Prodotti limitati (per esempio pari al 5% in ciascuna categoria) in maniera tale che percentuali fisiologiche inferiori alla franchigia non siano penalizzanti.

Nel caso poi degli Enti di ricerca, nello stabilire il numero di Prodotti dovuti, i Tecnologi dovevano contribuire con 3 Prodotti ed erano esclusi coloro che facevano attività di servizio. Dato

che non è sempre semplice distinguere tra chi fa servizio e chi fa ricerca tecnologica e, comunque, la produttività scientifica in termini di pubblicazioni di un Tecnologo non è comparabile con quella di un Ricercatore, sarebbe forse opportuno ripensare il ruolo dei Tecnologi nell'ambito della prossima valutazione.

Gli universitari incaricati di ricerca presso gli Enti dovevano fornire 6 Prodotti, 3 tramite l'Università e 3 tramite l'Ente[5]. Mentre in alcuni Enti esisteva la figura dell'Incaricato di Ricerca, tale figura non esisteva in altri Enti. In questi casi si è dovuto procedere con soluzioni improvvisate. Questo GEV raccomanda che ben in anticipo sulla prossima VQR tutti gli Enti si dotino di una regolamentazione omogenea per quanto riguarda gli incarichi di ricerca.

Un discorso a parte è connesso alle grandi collaborazioni, che sono caratteristiche dell'Area 2 nella loro forma più estrema. È certamente ragionevole pensare che la produzione scientifica di una collaborazione di due persone possa essere circa il doppio di quella di una persona singola. La regola del Bando che prevedeva che ogni Struttura non potesse presentare lo stesso Prodotto più di una volta è sensata, ma si basa implicitamente su un'ipotesi di linearità della produzione scientifica come funzione del numero degli autori. È difficile pensare che la produttività di una collaborazione di dieci persone possa essere dieci volte quella di una persona singola, mentre è certo che una collaborazione di 2.000 persone non pubblica (per fortuna!) 2.000 volte di più di una persona singola.

Le grandi collaborazioni di migliaia di scienziati non si formano con lo scopo di permettere al singolo ricercatore di firmare svariate migliaia di articoli l'anno, ma con quello di risolvere problemi estremamente difficili. In questi casi l'ipotesi della dipendenza lineare della produzione scientifica dal numero di ricercatori che collaborano, che è alla base delle scelte del DM[6] mostra tutti i suoi limiti. L'INFN e l'INAF (ma in minore misura) sono le due istituzioni maggiormente danneggiate dagli effetti di questa ipotesi di linearità in quanto sono frequenti collaborazioni di un centinaio di ricercatori appartenenti allo stesso Ente di ricerca, afferenti a sottostrutture sparse per tutta l'Italia.

Una soluzione che ridurrebbe notevolmente i danni consiste nel permettere a sottostrutture dello stesso Ente (per esempio la Sezione di Catania e la Sezione di Padova dell'INFN) di presentare lo stesso Prodotto per la valutazione, esattamente come attualmente possono fare le Università di Catania e di Padova.

In realtà la situazione è ancora più complicata: gli incaricati di ricerca di un Ente appartenenti a due diverse Strutture universitarie possono presentare lo stesso Prodotto solo se a carico delle Strutture stesse. Al contrario, il Prodotto non può essere presentato due volte se è a carico dell'Ente, anche se ha un impatto sulla valutazione delle Strutture universitarie. Questo divieto produce un

[5] Il raddoppio del numero dei Prodotti per gli incaricati di ricerca è una richiesta che dovrebbe essere ripensata in quanto un incarico di ricerca in un Ente dovrebbe contribuire a migliorare la qualità dei Prodotti, ma non la quantità.
[6] In futuro sarebbe auspicabile un DM che non entri nei dettagli del processo di valutazione, ma che ne dia solo gli indirizzi e le linee politiche generali.

accoppiamento[7] tra la valutazione relativa delle Università e le scelte dell'Ente, che, ripetiamo, si eliminerebbe permettendo a diverse Sottostrutture dello stesso Ente di presentare lo stesso Prodotto.

Il DM, sulla scia della VTR 2001-2003, ha deciso di dividere i Prodotti valutabili in quattro categorie (decisione ineccepibile per la presentazione dei risultati) e di calcolare il voto della Struttura assegnando un valore a ciascuna classe (1 alla classe E, 0,8 alla classe B, 0,5 alla classe A, 0 alla classe L). Questa seconda decisione non è ottimale. Infatti, il risultato della valutazione bibliometrica è essenzialmente continuo (le due coordinate nel quadrato) e quella della valutazione PR (nel caso di due Revisori, prima dell'intervento del gruppo di consenso) è un numero intero nell'intervallo 6-54. Passare da una valutazione continua a una discreta per successivamente fare la media delle valutazioni discrete introduce un aumento dell'errore statistico siccome un Prodotto può facilmente passare da una classe a un'altra per effetto di irrilevanti variazioni nella valutazione. Sarebbe stato preferibile calcolare direttamente il contributo al voto finale, un numero compreso tra zero e uno, senza passare tramite una discretizzazione intermedia. La scelta fatta introduce un piccolo, non necessario, aumento dell'errore statistico.

Questa Sezione si conclude con un'ultima osservazione marginale. Sarebbe stata anche necessaria una maggior precisione nella definizione dei Prodotti non valutabili. Specialmente nelle monografie, ma anche in parte negli articoli su rivista. Ci sono Prodotti in cui il contenuto di ricerca originale si mescola con finalità didattiche, divulgative o di altro tipo. Il confine tra i Prodotti che contengono una quantità di ricerca originale sufficiente e quelli che non lo contengono è a volte labile. Le regole attuali che richiedono una scelta secca tra una valutazione 0 (limitato) e una valutazione -1 (non valutabile) in certi casi non sono facilmente applicabili e una gradualità nella regione di valutazione negativa sarebbe stata auspicabile.

C.2 Il processo di valutazione

Abbiamo già descritto le distorsioni che la procedura di valutazione bibliometrica definita nel Documento Criteri produce tra le varie SC e la necessità di fare una regolazione accurata dei parametri in maniera da avere una funzione che trasferisca la valutazione bibliometrica bidimensionale in un singolo numero che sia *equa* rispetto al variare della SC. In una futura valutazione la scelta di tale funzione potrà essere fatta in maniera controllata anche effettuando un confronto dell'analisi bibliometrica con quella della PR. Per esempio, questo GEV ha ritenuto che la valutazione bibliometrica di Prodotti *recenti* fosse meno affidabile di quella di Prodotti *più vecchi* e ha cercato di correggere questo effetto aumentando la percentuale di Prodotti inviati in PR pubblicati nel 2009 e nel 2010. Un confronto tra la valutazione bibliometrica e la PR (nell'ipotesi ragionevole che l'accuratezza della valutazione PR dipenda meno dall'età del Prodotto, se paragonata con l'accuratezza della valutazione bibliometrica) sarebbe estremamente utile per

[7] Tale accoppiamento è difficile da stimare quantitativamente con le informazioni attualmente in possesso di questo GEV. Allo scopo di arrivare a una stima quantitativa di questo accoppiamento il GEV intende studiare gli effetti di regole diverse sulla valutazione degli Enti e delle Università qualora si renda disponibile in maniera informatizzata una lista completa e verificata di tutti i Soggetti Valutati autori dei Prodotti (cioè non solo di coloro a cui il Prodotto è associato per la VQR) e delle loro affiliazioni.

verificare quantitativamente di quanto diminuisce la validità dell'analisi bibliometrica con l'età della pubblicazione e se, effettivamente (in accordo con il senso comune e con quanto fatto dalla maggior parte dei GEV "bibliometrici"), sia meglio privilegiare l'IF della rivista per Prodotti recenti, mentre per Prodotti *più vecchi* sia meglio privilegiare il numero di citazioni ricevute.

Ovviamente è anche necessario fare un aggiustamento fine dei parametri, di modo che non ci siano differenze sistematiche tra la valutazione bibliometrica e la PR e, quindi, non sia a priori più conveniente per un Prodotto bibliometrico essere valutato bibliometricamente piuttosto che in PR.

Un problema di risoluzione più difficile è legato all'errore statistico dovuto a discordanze tra i due o più Revisori dello stesso Prodotto. Al momento attuale, come mostrato nell'Appendice A, usando 2 Revisori per Prodotto l'errore sulla valutazione PR è maggiore di quello sulla valutazione bibliometrica. Le cose cambierebbero aumentando sensibilmente il numero di Revisori per Prodotto, ma questa soluzione non è praticabile.

Dato che la quasi totalità dei Prodotti conferiti sono articoli su riviste indicizzate WoS (con l'eccezione di FIS/08, che è molto simile a un'Area non bibliometrica) si può concludere che la valutazione attuale sarebbe stata più accurata se si fosse utilizzata la sola bibliometria, senza nemmeno fare le correzioni dipendenti dalla SC (che sono state fatte) e la suddivisione delle SC in SSC utilizzando i *PACS Numbers* (che non è stata fatta).

Questa conclusione fortemente negativa sulla valutazione PR è inevitabile? Un'analisi a campione sembra indicare che una grande sorgente degli errori statistici nella PR sia ascrivibile al fatto che alcuni Revisori sono più generosi e tendono a dare voti elevati mentre altri tendono a tenersi bassi con la valutazione; la valutazione del Prodotto dipende quindi in maniera non trascurabile, ma casuale, dal tipo di Revisore a cui il Prodotto è stato assegnato.

La decisione se un Prodotto si collochi al di sopra o al di sotto della media mondiale dei Prodotti del settore dipende moltissimo da quale il Revisore ritenga sia la media mondiale. Questa valutazione cambia in maniera incontrollata da Revisore a Revisore; al contrario, le differenze di valutazione di Prodotti diversi da parte dello stesso Revisore sono probabilmente molto più significative. L'effetto della diversa normalizzazione dei giudizi del Revisore e del conseguente errore statistico deve essere studiato quantitativamente e bisogna capire se si possono prendere misure correttive: una soluzione potrebbe consistere nel normalizzare i voti di ciascun Revisore in modo da minimizzare le discrepanze di giudizio con gli altri Revisori.

Questo GEV nota inoltre che sembrerebbe ragionevole per la prossima VQR accettare come Prodotti per la valutazione solo gli articoli su rivista WoS. Infatti, gli articoli su rivista non WoS sono una piccola frazione: l'1,6% del totale. Considerando solo i prodotti indicizzati WoS la percentuale di Articoli su rivista cala da 93,4% a 91,8% e non vale certamente la pena complicare il processo di valutazione per tener conto di questi prodotti che mediamente hanno avuto una valutazione scarsa (voto medio pari a 0,33). Altri Prodotti, come Monografie, Contributi in volume, Manufatti e Brevetti dovrebbero essere conferiti dalle Strutture in una quantità molto limitata ed

entrare nella valutazione con un contributo positivo solo se di grande qualità. Per esempio per i brevetti potrebbe essere previsto un indicatore addizionale con peso molto piccolo. Questa proposta potrebbe essere applicata a tutta l'Area, con l'eccezione di FIS/08, che, come tipologia di Prodotti, è molto simile a un'Area non bibliometrica.

Una valutazione puramente bibliometrica, fatta tuttavia con un attento bilanciamento di Prodotti che vengono da differenti comunità scientifiche con differenti prassi citazionali, ha il suo fascino. Da un lato le Strutture si potrebbero limitare a riempire un *database* con tutti i Prodotti bibliometrici, dal quale l'ANVUR selezionerebbe i Prodotti in maniera ottimale per le singole Strutture[8], dall'altro il compito dell'ANVUR consisterebbe solo nella scelta dei criteri di valutazione. Il lavoro necessario per portare a termine la valutazione sarebbe quindi sensibilmente ridotto rispetto all'attuale. Questa proposta è formulata nell'ambito dell'Area 2 e non è detto che sia applicabile a tutte le Aree bibliometriche. In futuro si raccomanda che le procedure di raccolta dei prodotti e della loro valutazione possano essere maggiormente adattate alle Aree e non debbano essere per forza uniformi, pur dovendo restare fermo un riferimento a *database* mondiali (per esempio WoS) in maniera da garantire l'uniformità della calibrazione assoluta inter Area.

L'autorevolezza di questa VQR sarebbe stata certamente maggiore se si fosse fatta una stima (almeno approssimata) di *tutti* i contributi all'errore statistico insito nella valutazione di ogni singola Struttura. Questo avrebbe permesso di conoscere il grado di affidabilità della valutazione e avrebbe contribuito a smussare possibili critiche. Tuttavia non c'è stato il tempo per fare quest'analisi degli errori statistici e questo GEV auspica che questa possa essere fatta[9] in un prossimo futuro. Il gran numero di Prodotti conferiti alla VQR, circa venti volte più grande di quello della VTR, rende realistico l'ambizioso obiettivo di fare una stima accurata e completa dell'errore statistico.

In conclusione, questo GEV ritiene che un'attenta analisi statistica dei risultati della valutazione sia necessaria prima di stabilire in forma definitiva i criteri per il prossimo esercizio valutativo. Fortunatamente, con questa valutazione l'ANVUR è in possesso di un *database* impressionante, unico al mondo, di Prodotti scientifici valutati sia bibliometricamente sia mediante PR e quindi ha in mano tutte le informazioni necessarie per poter affrontare questo compito con successo.

Anche a questo scopo il GEV02 auspica fortemente che l'ANVUR renda pubblico il suo *database*, dopo averlo ovviamente depurato dei dati sensibili (informazione sugli autori, sulla Struttura, nome dei Revisori, ecc.). Da un lato sarebbe un'operazione di grande trasparenza, fatta ovviamente nel rispetto della *privacy*, dall'altro sarebbe un grande regalo alla comunità scientifica mondiale in quanto permetterebbe di approfondire a partire da dati reali tutto il dibattito in corso tra

[8] La conoscenza di tutti i Prodotti della ricerca di una Struttura permetterebbe una stima statistica degli errori più accurata.

[9] Purtroppo non *tutti* i contributi all'errore statistico sono stimabili con le informazioni in possesso dell'ANVUR.

i vantaggi e gli svantaggi della valutazione bibliometrica e dalla PR. Gli studi risultanti potrebbero essere molto utili per i futuri esercizi di valutazione in Italia e nel mondo.

Sarebbe inoltre auspicabile che l'ambiente software e l'accesso ai vari database fossero resi disponibili ai Nuclei di Valutazione delle Strutture per scopi valutativi interni onde evitare duplicazioni di lavoro e dispersione di risorse. Infatti, sarebbe da perseguire l'obiettivo di una stretta collaborazione tra i Nuclei di Valutazione delle Strutture e l'ANVUR.

Appendice D. Il bilanciamento tra le *subject categories*

I Prodotti classificati IR dall'algoritmo descritto nella sezione 2.3.1 sono stati oggetto di un'analisi preliminare volta a evitare che l'uso di criteri bibliometrici per la loro attribuzione a una delle quattro classi finali (o la loro effettiva attribuzione alla classe IR e, quindi, alla valutazione *peer*) favorissero una *Subject Category* (SC) rispetto alle altre. Nella sezione 2.3.1 sono trattate in dettaglio sia le cause sia i vincoli connessi a questa procedura; in questa appendice ci si limita a fornirne i dettagli. Si è proceduto a:

- individuare le N_{SC} Subject Categories principali o significative per il GEV (ovvero le SC in cui comparivano il maggior numero di articoli indicizzati conferiti dalle strutture per la valutazione, rif. Tab. 1.3, integrate dalle SC di didattica della Fisica);

- applicare i criteri di valutazione bibliometrica (cioè la matrice in fig. 2.1) alla totalità degli Articoli presenti nel *database* WoS mondiale per tali SC negli anni 2004-2010; si tratta complessivamente di più di 1.400.000 Articoli;

- individuare criteri di valutazione bibliometrica dipendenti dalla SC e verificare che le percentuali di Prodotti del *database* mondiale attributi a ciascuna classe VQR per ciascuna SC, *i.e.*, E_{SC}, B_{SC}, A_{SC}, L_{SC} fossero il più possibile vicine le une alle altre. Per semplicità si è proceduto confrontando il *punteggio* VQR (v_{SC}) di ogni SC sulla base del *database* mondiale:

$$v_{SC} = 1.0 * E_{SC} + 0.8 * B_{SC} + 0.5 * A_{SC} + 0.0 * L_{SC}\ , \qquad (D.1)$$

con l'obiettivo che la scelta dei criteri bibliometrici dipendenti dalla SC portassero alla seguente situazione:

$$v_{SC1} \cong v_{SC2} \cong \cdots \cong v_{SCN}; \qquad (D.2)$$

Per "criteri di valutazione bibliometrica dipendenti dalla SC" si intende quanto riportato nelle figure D.1 e seguenti, cioè l'individuazione delle regioni – e delle relative soglie – nel piano cartesiano [*Impact Factor*, Citazioni] valutate come E, B, A, L o IR.

Al fine di ottenere valori di v_{SC} paragonabili tra le varie *SC* è stato necessario attribuire, unicamente ai fini dell'analisi qui descritta, i Prodotti valutati IR in ogni SC (zone grigie nelle figure D.1 e seguenti) alle quattro classi E, B, A, L. A tal fine, il GEV ha deciso di fare ricorso ai Prodotti estratti in modo casuale per l'analisi di correlazione tra valutazione bibliometrica e valutazione *peer,* derivando cioè per le regioni del quadrato classificate IR le percentuali di assegnazione alle classi risultanti dalla valutazione *peer* dei Prodotti del campione. Questa

procedura è molto simile a quella seguita dall'ANVUR per la calibrazione degli algoritmi bibliometrici tra i vari GEV (Rif. Appendice A del Rapporto Finale ANVUR).

Nelle figure D.1, D.2, D.3 vale il seguente codice colore:

- Giallo: valutazione bibliometrica "Eccellente";
- Blu: valutazione bibliometrica "Buono";
- Rosso: valutazione bibliometrica "Accettabile";
- Verde: valutazione bibliometrica "Limitato";
- Grigio: IR, valutazione *peer*.

Tab. D.1: Le Subject Categories (SC) oggetto dell'analisi preliminare. Nome abbreviato ed esteso delle SC; numero di articoli su rivista indicizzati presentati dalle Strutture al GEV02 per la valutazione suddiviso per SC; punteggio VQR come definito nella (D.1) risultante dalla valutazione dell'intero database WoS degli anni 2004-2010. NB: la media è pesata sul numero totale di Prodotti presenti in ciascuna SC.

SC	***Nome esteso SC***	***# Prodotti***	***Punteggio VQR DB WoS Mondo 2004-10***
BU	*Astronomy & Astrophysics*	3.718	0,562
UP	*Physics, Particles & Fields*	2.904	0,563
UK	*Physics, Condensed Matter*	1.612	0,567
UN	*Physics, Nuclear*	1.261	0,556
UB	*Physics, Applied*	1.152	0,559
RY	*Nuclear Science & Technology*	1.128	0,575
SY	*Optics*	963	0,565
OA	*Instruments & Instrumentation*	899	0,568
UR	*Physics, Mathematical*	488	0,572
UH	*Physics, Atomic, Molecular & Chemical*	421	0,575
PM	*Materials Science, Multidisciplinary*	392	0,568
EI	*Chemistry, Physical*	361	0,547
UF	*Physics, Fluids & Plasmas*	321	0,584
DA	*Biophysics*	186	0,570
IQ	*Engineering, Electrical & Electronic*	165	0,577
GC	*Geochemistry & Geophysics*	135	0,555
VY	*Radiology, Nuclear Medicine & Medical Imaging*	135	0,558
LE	*Geosciences, Multidisciplinary*	113	0,550
QG	*Materials Science, Coatings & Films*	110	0,655
QQ	*Meteorology & Atmospheric Sciences*	109	0,555
XQ	*Spectroscopy*	100	0,535
NS	*Nanoscience & Nanotechnology*	82	0,559
PU	*Mechanics*	49	0,559
HB	*Education, Scientific Disciplines*	37	0,572
HA	*Education & Educational Research*	8	0,572
		Media pesata	**0,565**

Fig D.1: Nel piano cartesiano [Impact Factor, Citazioni] e riferendosi alla rappresentazione in percentili della Fig. 2.2, per ogni Subject Category elencata in Tab. D.1 vengono indicate graficamente le regioni corrispondenti alle classi E (giallo), B (blu), A (rosso), L (verde) o IR (grigio).

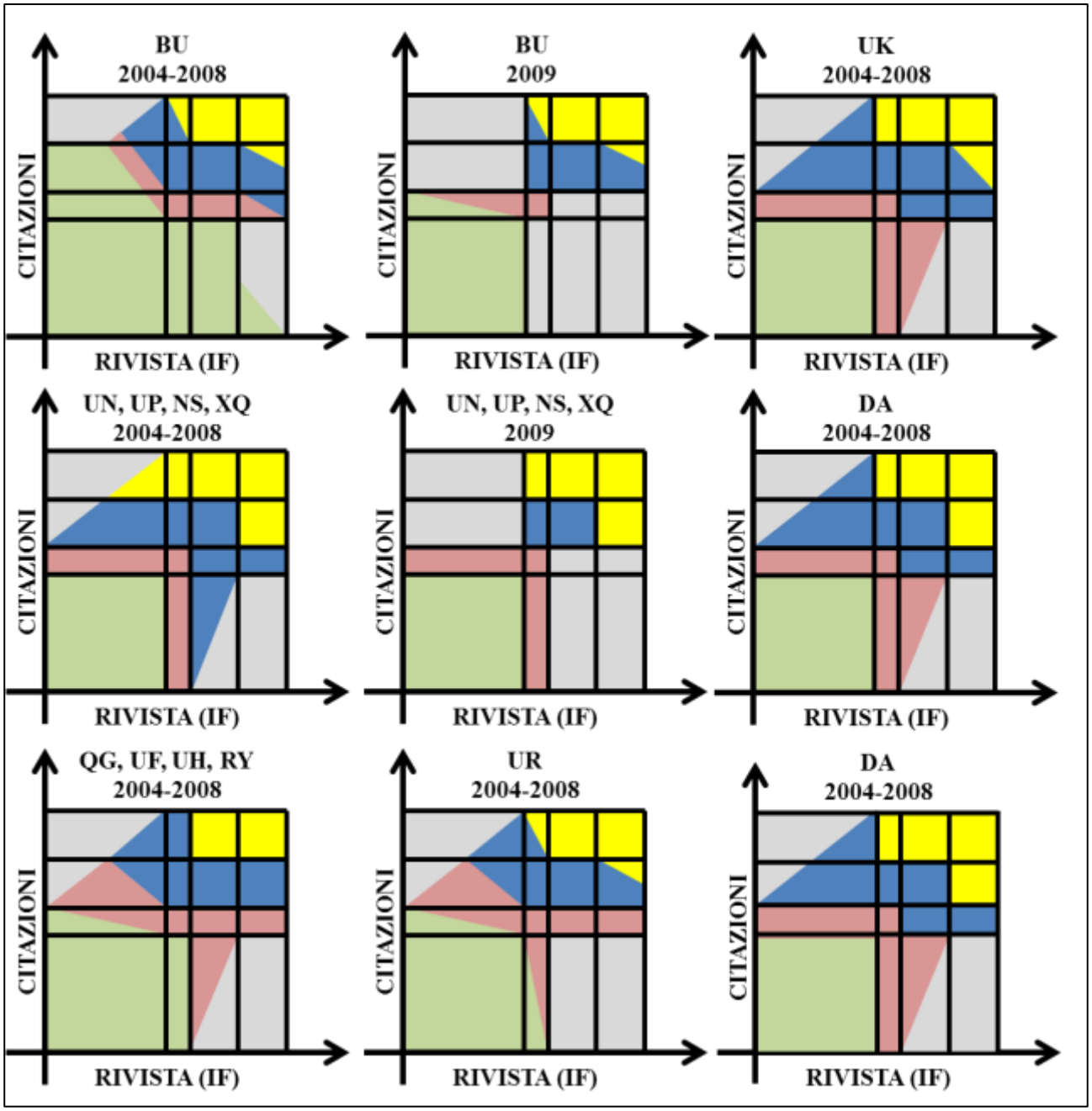


Fig D.2: Idem.

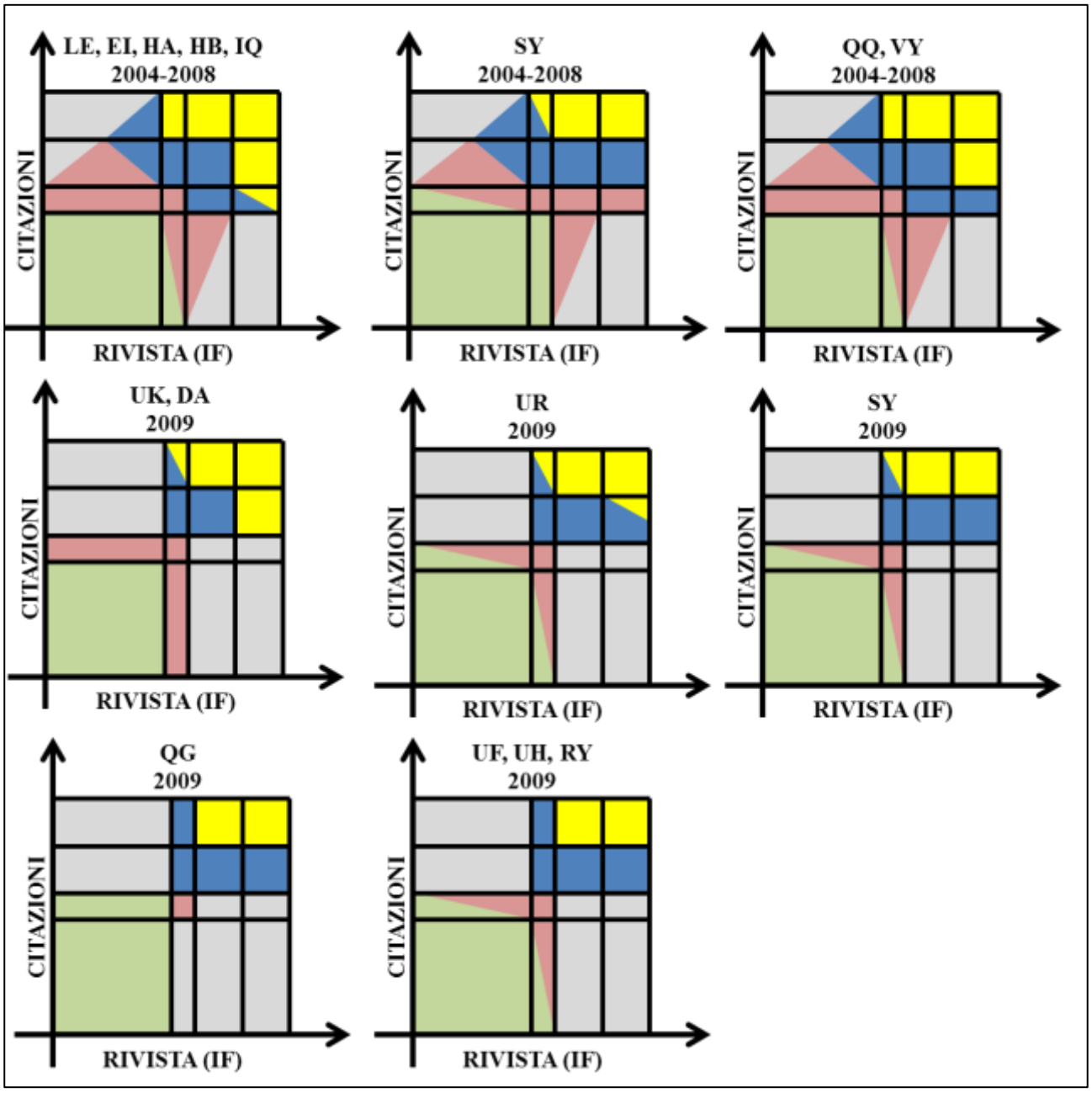

LE, EI, HA, HB, IQ
2004-2008
SY
2004-2008
QQ, VY
2004-2008
UK, DA
2009
UR
2009
SY
2009
QG
2009
UF, UH, RY
2009
CITAZIONI
RIVISTA (IF)

Fig D.3: Idem, per tutte le SC non presenti nelle Figg. D.1 e D.2, incluse quelle non oggetto dell'analisi preliminare.