

Confessioni di un ricercatore non pentito

Anonymous

Perché servono i dati ?

⇒ **per studiare i comportamenti individuali e/o collettivi** (esempi: chi abbandona la scuola secondaria? chi si iscrive ad un corso universitario? chi non completa un corso universitario? chi fa domanda per abilitazione universitaria?)

⇒ **per monitorare i processi, possibilmente in tempo reale** (esempi: scelte di indirizzo scuola secondaria – transizioni scuola-mercato del lavoro – transizioni lauree triennali-magistrali – ricerca occupazione)

⇒ **per valutare efficacia delle politiche** (esempi: le assunzioni di insegnanti dovute alla legge 107/2015 ha migliorato gli apprendimenti ? la no-tax area ha fatto aumentare le iscrizioni ? la VQR ha aumentato la produttività scientifica ?)

I dati per rispondere a queste domande vengono

→ da fonti amministrative (anagrafi studenti, anagrafi docenti)

→ da indagini campionarie (PISA, PIAAC, ISTAT)

→ da combinazioni dei due (Invalsi, Almalaurea)

Vantaggi indagini campionarie:

- ✓ dati puliti e completi e possibilmente rappresentativi della popolazione

Limiti indagini campionarie:

- ✓ basati su autodichiarazione degli intervistati
- ✓ *non response rates* spesso opachi
- ✓ rappresentatività non sempre aggiustabile con i pesi campionari

Vantaggi fonti amministrative:

- ✓ copertura della popolazione amministrata

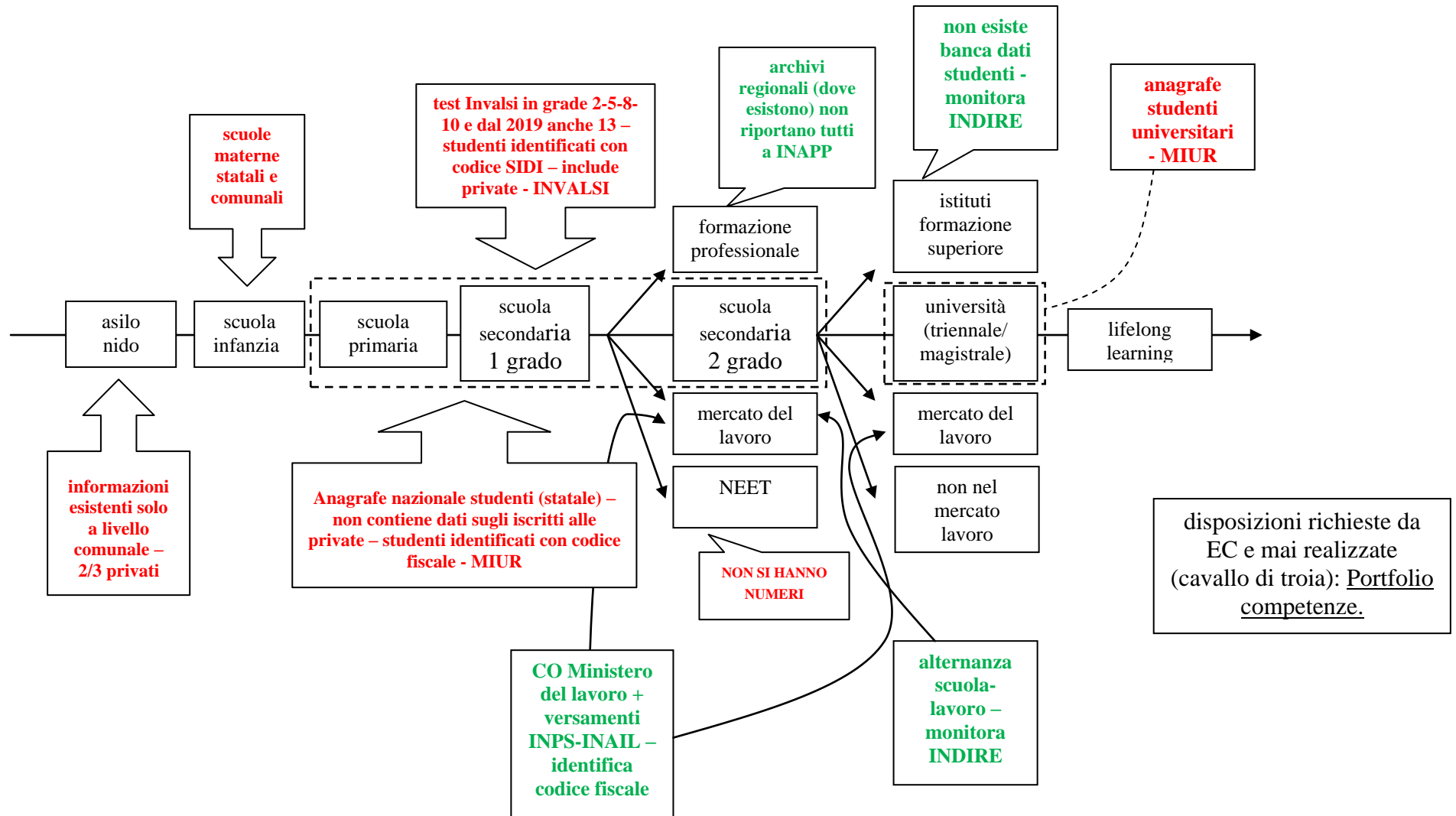
Limiti fonti amministrative:

- ✓ assenza informazioni non rilevanti per processo amministrativo
- ✓ dati incorporano imperfezioni/errori ente gestore
- ✓ tutela privacy impone accesso limitato solo alle finalità amministrative

Esempio ① - COME MONITORARE IL PROCESSO FORMATIVO DEI GIOVANI ITALIANI

In Italia non esiste la possibilità di conoscere i percorsi formativi delle nuove generazioni in una prospettiva longitudinale, dove cioè anno dopo anno viene registrato il livello di formazione frequentato e i risultati scolastici conseguiti.

Il vantaggio di un monitoraggio longitudinale delle carriere formative è quello della individuazione degli snodi problematici (e dei conseguenti abbandoni), che fanno sì che l'Italia sia il paese in Europa a più basso tasso di scolarità (oltre che di basso livello di competenza nella popolazione adulta).



La ragione di tale assenza è duplice: la molteplicità degli attori istituzionali (i comuni presidiano la formazione prescolare, l'amministrazione statale presidia scuola primaria e secondaria, le regioni autonome presidiano i propri sistemi formativi, le regioni a statuto ordinario presidiano la formazione professionale, le università presidiano la formazione terziaria,) e una interpretazione eccessivamente protettiva della privacy degli studenti.

Il primo fattore ha dato origine ad una molteplicità di software gestionali non intercomunicanti adottati dai vari attori e gestiti normalmente con contratti di appalto esterni, che non hanno alcun interesse ad interfacciarsi tra loro.

Il secondo fattore produce l'impossibilità formale di usare il codice fiscale (unico identificativo univocamente determinato) come chiave di collegamento tra le informazioni delle diverse banche dati, con l'esito di perdere frazioni significative di informazione a seguito della ricodifica degli studenti (basti pensare alla mobilità territoriale degli studenti di origine straniera).

Il risultato principale di questa carenza è l'impossibilità di ottenere misure corrette di frequenza scolastica per coorte di nascita, unico strumento adeguato per monitorare l'assolvimento dell'obbligo scolastico e per localizzare in modo puntuale interventi contro la dispersione scolastica. Poiché almeno tre obiettivi della strategia europea di Lisbona 2020 sono definiti in termini di partecipazione scolastica (*childcare, early school leaving e tertiary educational attainment*), appare evidente l'importanza di poter adeguatamente monitorare questo processo, per la tempestività e l'efficacia degli interventi adottabili.

Il nucleo centrale delle informazioni sulle carriere scolastiche è rappresentato dall'**Anagrafe nazionale degli studenti** (<http://www.miur.gov.it/anagrafe-nazionale-studenti>) alimentata dalle iscrizioni online delle famiglie e convalidato dalle segreterie scolastiche, su software gestito centralmente da parte di HP, che gestisce tutta l'infrastruttura informatica del MIUR-scuola.

Questo gestionale delle carriere scolastiche nel sistema statale (che esclude quindi Valle d'Aosta, Trentino, tutto il settore privato nella scuola e tutta la formazione professionale – complessivamente quasi un quarto della popolazione scolastica) non è interfacciato con la **rilevazione condotta da Invalsi sul livello di apprendimento degli studenti** (in seconda e quinta primaria, terza secondaria di primo grado, seconda e quinta secondaria di secondo grado), in quanto MIUR gestisce i propri record con codice fiscale e Invalsi gestisce con codici SIDI → diventa così impossibile poter tracciare il percorso individuale dello studente in termini di efficacia dell'apprendimento.

Quando uno studente si diploma e si iscriva all'università, entra nei gestionali dell'università presso cui si iscrive (tutti rigorosamente diversi l'uno dall'altro, in ossequio all'autonomia universitaria). Tuttavia le università statali (di nuovo perdiamo l'informazione di chi si iscrive in un ateneo privato) sono obbligate a riversare i dati sulle carriere dei propri studenti in un archivio gestito da Cineca, su commessa del dipartimento università del MIUR-università (anch'esso chiamato **Anagrafe Nazionale degli Studenti** - http://www.miur.gov.it/web/guest/anagrafe-studenti_rilevazioni-universita).

Ovviamente le due anagrafi degli studenti non comunicano tra loro, per cui non è possibile conoscere quante bocciature abbia subito uno studente iscritto all'università (l'insuccesso scolastico è materia coperta da privacy, e quindi la privacy pone ostacoli al collegamento delle due anagrafi – senza considerare che due providers esterni non hanno interesse a dialogare, ma a spartirsi il mercato).

Un sogno possibile: se ogni nuovo cittadino che entra nel sistema formativo nazionale (per nascita o per immigrazione) venisse dotato di una identità digitale (analogamente a quanto accade con gli esami medici sulla tessera sanitaria), su cui fosse obbligatorio registrare i passaggi nel sistema formativo a tutti i livelli, inclusi i risultati conseguiti in termini di votazione e di test sulle competenze, ottenendo il **portfolio** che certifica la storia formativa e permette nel contempo di studiarne la dinamica nella popolazione.

Esempio ② - Valorizzazione del merito degli insegnanti

Unione di quattro base dati:

- ⇒ banca dati associata alla compilazione dei RAV per l'anno 2015-16 (informazioni relative agli insegnanti a tempo indeterminato – età e titolo di studio – esperienza dei dirigenti);
- ⇒ banca dati dei risultati delle prove SNV-Invalsi
- ⇒ totale delle schede compilate dalle scuole per il Monitoraggio della valorizzazione professionale (schede “merito” - MIUR)
- ⇒ banca dati delle retribuzioni degli insegnanti (MEF)

Il comportamento delle scuole potrà comunque essere studiato nel dettaglio solo quando si disponga sia delle informazioni del RAV sia delle schede sul merito, condizione che si verifica solo per **8311** scuole corrispondenti a **664905** insegnanti. Ma quante sono le scuole che hanno aderito ?

da archivi MIUR

regione	banca dati RAV			banca dati Merito			incrocio RAV pubblicato e scheda merito
	informazioni presenti in banca dati RAV ma RAV non pubblicato (p.e. CPIA)	RAV pubblicato (insieme ristretto)	Totale scuole di cui si ha informazione statistica (insieme esteso)	scuole di cui si hanno informazioni RAV ma non scheda merito	scuole di cui si ha scheda merito ma assenti in banca dati RAV	scheda di valorizzazione del merito e presenza in banca dati RAV	
Piemonte	56	572	628	62	1	565	554
Liguria	11	189	200	8	2	190	186
Lombardia	53	1132	1185	54	2	1129	1109
Veneto	23	604	627	30	3	594	593
Friuli V.G.	12	168	180	12	0	168	164
Emilia Romagna	29	532	561	21	3	537	527
Toscana	31	480	511	28	1	482	471
Marche	10	243	253	11	0	242	240
Umbria	17	141	158	19	0	139	138
Lazio	52	728	780	61	6	713	705
Abruzzo	28	200	228	26	1	201	199
Molise	18	53	71	16	1	54	53
Campania	67	1009	1076	76	2	998	988
Basilicata	34	122	156	33	2	121	120
Puglia	47	669	716	41	4	671	665
Calabria	35	382	417	34	1	382	376
Sicilia	45	864	909	55	3	851	844
Sardegna	65	276	341	66	1	274	270
Totale	633	8364	8997	653	33	8311	8202

da Archivi MEF

	numero scuole	insegnanti ruolo	insegnanti tempo determinato	Totale insegnanti	insegnanti medi per scuola	quota insegnanti tempo determinato
assenza erogazioni	2738	152928	45211	198139	72.37	22.8%
erogazioni incomplete	3024	268385	63886	332271	109.88	19.2%
erogazioni coerenti	3436	303089	79809	382898	111.44	20.8%
Totale	9198	724402	188906	913308	99.29	20.7%

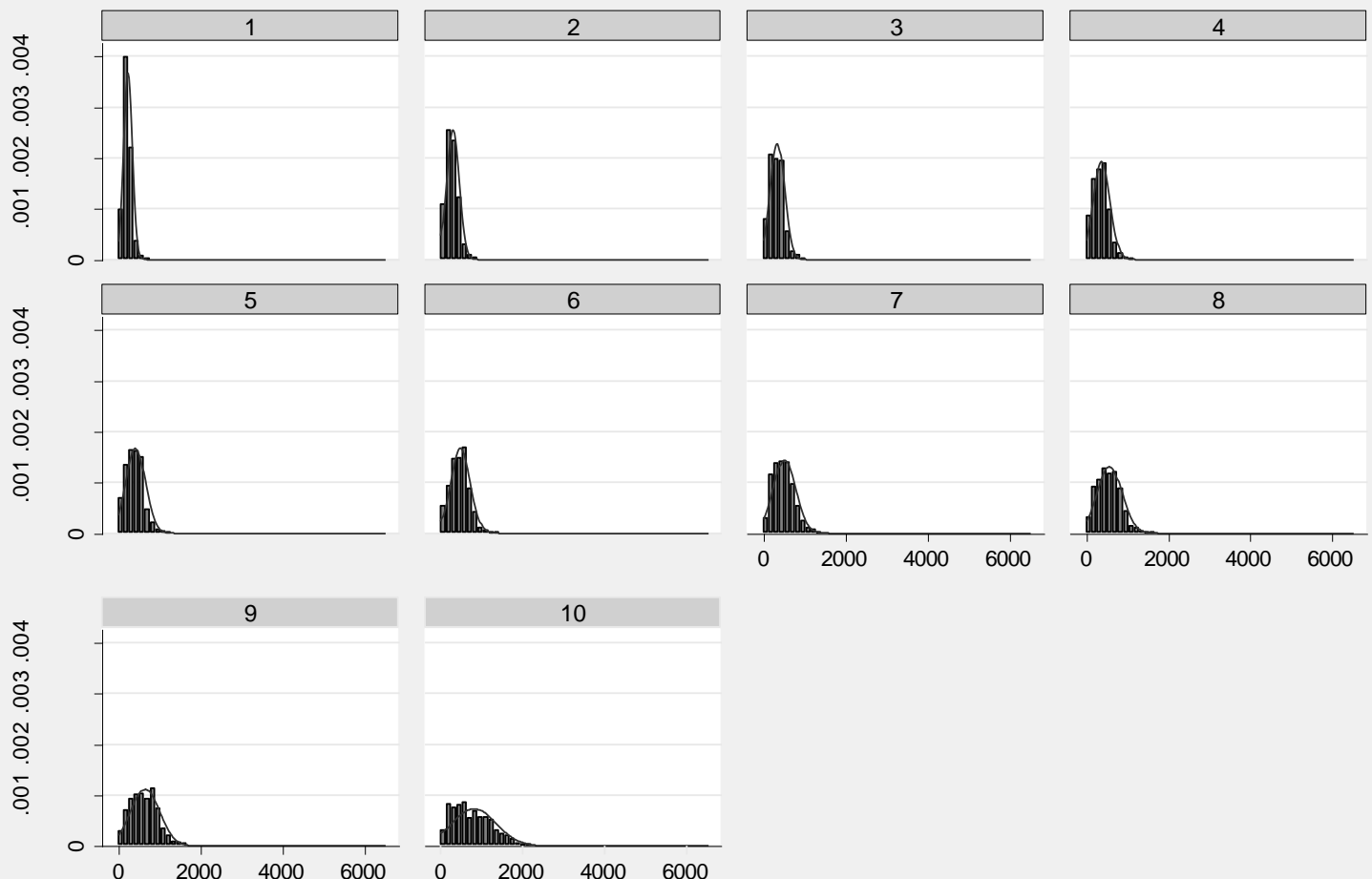
In questo universo di scuole si osservano tre tipi di comportamenti:

- ⇒ scuole che non hanno avviato il processo di valorizzazione del merito, non costituendo neppure il comitato di valutazione dei docenti (pari al **4.9%** delle scuole dell'insieme ristretto, ma **10%** dell'insieme esteso)
- ⇒ scuole che hanno costituito il comitato, ma che per varie ragioni non è poi arrivato a deliberare la distribuzione del premio tra gli insegnanti (pari al **8.0%** delle scuole e all'**8.1%** degli insegnanti);
- ⇒ scuole che hanno ripartito i premi al merito (pari al **87.1%** delle scuole e all'**87.0%** degli insegnanti).

- ⇒ 578203 insegnanti hanno partecipato al processo e 237032 hanno ricevuto un premio

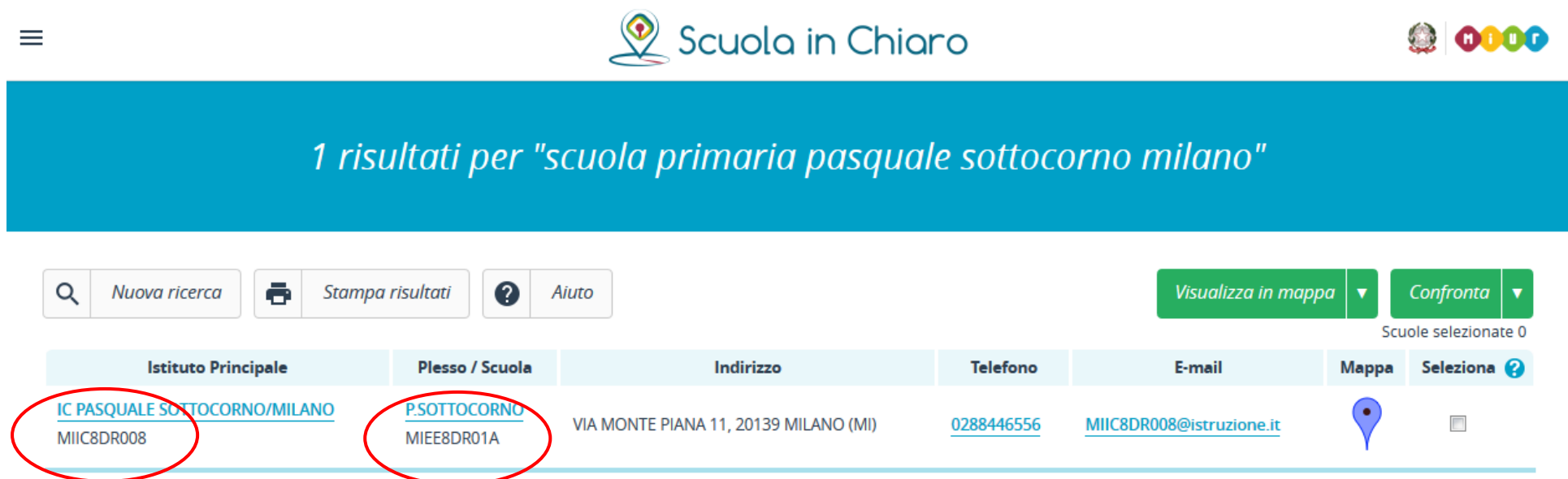
Ma quante sono le scuole ? ovvero la misura come è stata recepita ?

Distribuzione di frequenza tra gli insegnanti per gruppi di scuole ordinate secondo la disuguaglianza nella distribuzione interna del premio – 3436 scuole con distribuzione completata




distribuzione della valorizzazione del merito per decili della disuguaglianza nelle distribuzioni

L'origine del problema, ovvero la madre di tutti i mali ?
Il codice meccanografico che identifica le scuole.



The screenshot shows the 'Scuola in Chiaro' website interface. At the top, there is a navigation menu with a hamburger icon, the logo 'Scuola in Chiaro', and the MIUR logo. A blue banner displays the search results: '1 risultati per "scuola primaria pasquale sottocorno milano"'. Below the banner, there are buttons for 'Nuova ricerca', 'Stampa risultati', 'Aiuto', 'Visualizza in mappa', and 'Confronta'. A table lists the search results with columns for 'Istituto Principale', 'Plesso / Scuola', 'Indirizzo', 'Telefono', 'E-mail', 'Mappa', and 'Seleziona'. The first result is circled in red, showing the code 'MIIC8DR008' under 'Istituto Principale' and 'P.SOTTOCORNO MIEE8DR01A' under 'Plesso / Scuola'.

Istituto Principale	Plesso / Scuola	Indirizzo	Telefono	E-mail	Mappa	Seleziona
<u>IC PASQUALE SOTTOCORNO/MILANO</u> MIIC8DR008	<u>P.SOTTOCORNO</u> MIEE8DR01A	VIA MONTE PIANA 11, 20139 MILANO (MI)	<u>0288446556</u>	<u>MIIC8DR008@istruzione.it</u>		<input type="checkbox"/>

Nato come codice parlante (provincia/tipologia istituto/tipologia gestione) non si è previsto l'accorpamento delle scuole, per cui ad ogni ondata di accorpamento viene cambiato il codice.

Ma quale scuola ha frequentato tuo figlio e quali insegnanti ha avuto ? Non lo saprai mai ... invece in US

Econometrica, Vol. 73, No. 2 (March, 2005), 417–458

TEACHERS, SCHOOLS, AND ACADEMIC ACHIEVEMENT

BY STEVEN G. RIVKIN, ERIC A. HANUSHEK, AND JOHN F. KAIN¹

This paper disentangles the impact of schools and teachers in influencing achievement with special attention given to the potential problems of omitted or mismeasured variables and of student and school selection. Unique matched panel data from the UTD Texas Schools Project permit the identification of teacher quality based on student performance along with the impact of specific, measured components of teachers and schools. Semiparametric lower bound estimates of the variance in teacher quality based entirely on within-school heterogeneity indicate that teachers have powerful effects on reading and mathematics achievement, though little of the variation in teacher quality is explained by observable characteristics such as education or experience. The results suggest that the effects of a costly ten student reduction in class size are smaller than the benefit of moving one standard deviation up the teacher quality distribution, highlighting the importance of teacher effectiveness in the determination of school quality.

KEYWORDS: Student achievement, teacher quality, school selection, class size, teacher experience.

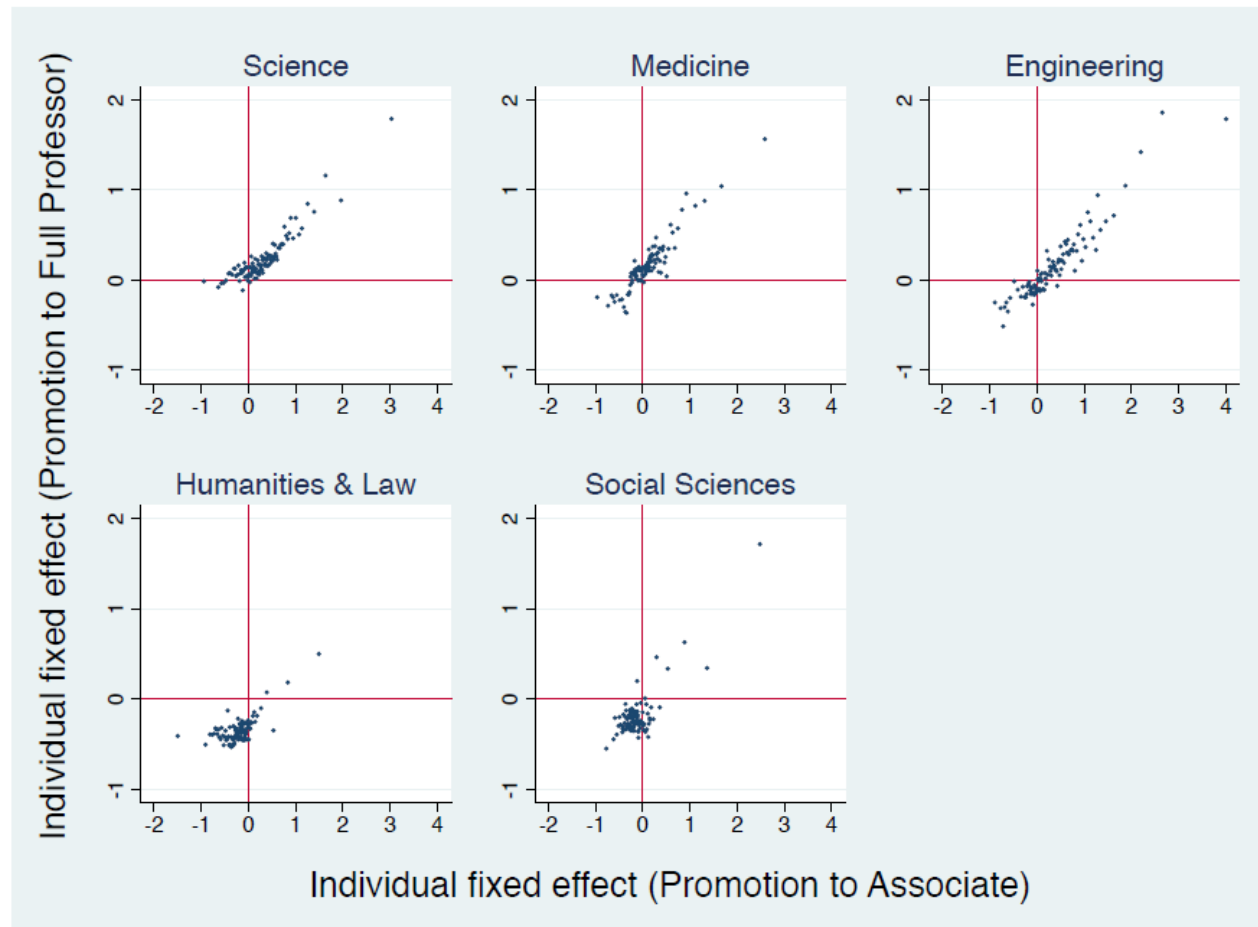
Esempio ③ - Studio degli incentivi alla produzione scientifica degli accademici italiani

Combina 3 banche dati:

- a) archivio amministrativo [docenti in servizio per ogni anno, identificati da nome/cognome/fascia/ssd – manca genere, età e cittadinanza (informazioni contenute nel codice fiscale protetto per privacy) – problema omonimia]
- b) archivio amministrativo [passaggi di carriera interpretati come concorsi – nonostante obbligo legislativo, verbali concorsi non sempre online – impossibile ricostruire i commissari – manca diritto accesso ai dati amministrativi retroattivi]
- c) archivio delle pubblicazioni scaricate da web of Science [match basato su nome e subject categoryim - possibile accedere a LoginMiur (anche se informazioni non certificate) perché nessuno ha mai firmato liberatoria ad utilizzo dei dati contenuti che si riferiscono a pubblicazioni!].

Tutto questo per arrivare ad una misura affidabile della capacità individuale

Individual Fixed Effects for two Promotions.



Binscatter plots showing individual fixed effect for promotion to associate professor (the first column of Table 5) and individual fixed effect for promotion to full professor (the second column) for academics who have been promoted twice in the 21 year period, in 100 equally sized bins of the x -axis variable.

Quali feedbacks per il policy makers ?

① necessità di una banca dati sulla produttività scientifica di accademici e ricercatori che abbia almeno tre caratteristiche:

- ✓ aggiornata periodicamente senza adempimenti ad hoc (VQR impone ogni cinque anni due prodotti migliori)
- ✓ certificata da autorità esterna (per evitare inserimento di lavori non scientifici, quali articoli giornale)
- ✓ visibile a tutti (per evitare attribuzioni inappropriate o comportamenti elusivi tipo caricamento capitoli monografie come articoli)

② per la stragrande maggioranza degli accademici (ma non per i ricercatori) esistono i *repositories* di ateneo [costruiti con architetture non comparabili per rispetto della autonomia universitaria] e l'obbligo di messa online dei cv [non comparabili in assenza di common template: per esempio non si può sapere chi abbia PhD e chi no]

③ per la stragrande maggioranza degli accademici (ma non per i ricercatori) esiste un *repository* nazionale [per altro usato per definire le soglie per ASN] senza che nessuno possa ispezionare cosa contenga [tranne che nel caso di contenziosi giudiziari] perché non autorizzato esplicitamente [nel caso del FABR i singoli che facevano domanda hanno dovuto autorizzare per poter processare la domanda].

④ Esiste una legge non attuata che crea l'Anpreps: legge 1/2009, art. 3bis *“Anagrafe nazionale dei professori ordinari e associati e dei ricercatori*
***A decorrere dall'anno 2009, con decreto del Ministro dell'istruzione, dell'università e della ricerca sono individuati modalità e criteri per la costituzione, senza nuovi o maggiori oneri per la finanza pubblica, presso il Ministero, di una Anagrafe nazionale nominativa dei professori ordinari e associati e dei ricercatori, contenente per ciascun soggetto l'elenco delle pubblicazioni scientifiche prodotte. L'Anagrafe è aggiornata con periodicità annuale.*”**

Esempio ④ - Sbocchi lavorativi dei laureati

Anvur deve fornire agli atenei indicatori di efficacia della didattica, tra cui il tasso di occupazione dei laureati per corso di laurea. Tale dato non esiste in natura.

Sostituti parziali:

→ tasso occupazione a tre anni da indagine campionaria Istat su sbocchi lavorativi dei laureati (disponibile solo per macroaree disciplinari e non per corso di laurea – si ignora distorsione a causa di non risposta – misurato a tre anni dalla laurea)

→ indagine Almalaurea (elevati tassi di risposta per via della obbligatorietà introdotta dagli atenei – copertura incompleta perché diversi atenei non aderiscono)

Sperimentazione MIUR-INPS:

① durata del processo: 2 anni – collaborazione uffici legali MIUR e INPS – sperimentazione come anticipazione del PSN – informativa a tutti gli interessati tramite post nei websites.

① Incrocio dei dati sui laureati 2010-11-12 (forniti da Cineca) con 8 banche dati INPS riferite al 2010-11-12-13-14 (dipendenti – autonomi – parasubordinati – parasubordinati professionisti – operai agricoli – autonomi agricoli – spettacolo – collaboratori domestici).

② si è preso come riferimento l'attività registrata nei due anni successivi all'anno di laurea (quindi 2011-12 per laureati 2010, 2012-13 per laureati 2011 e 2013-14 per laureati 2012).

③ due indicatori generali che forniscono informazioni sulla condizione lavorativa dei laureati:

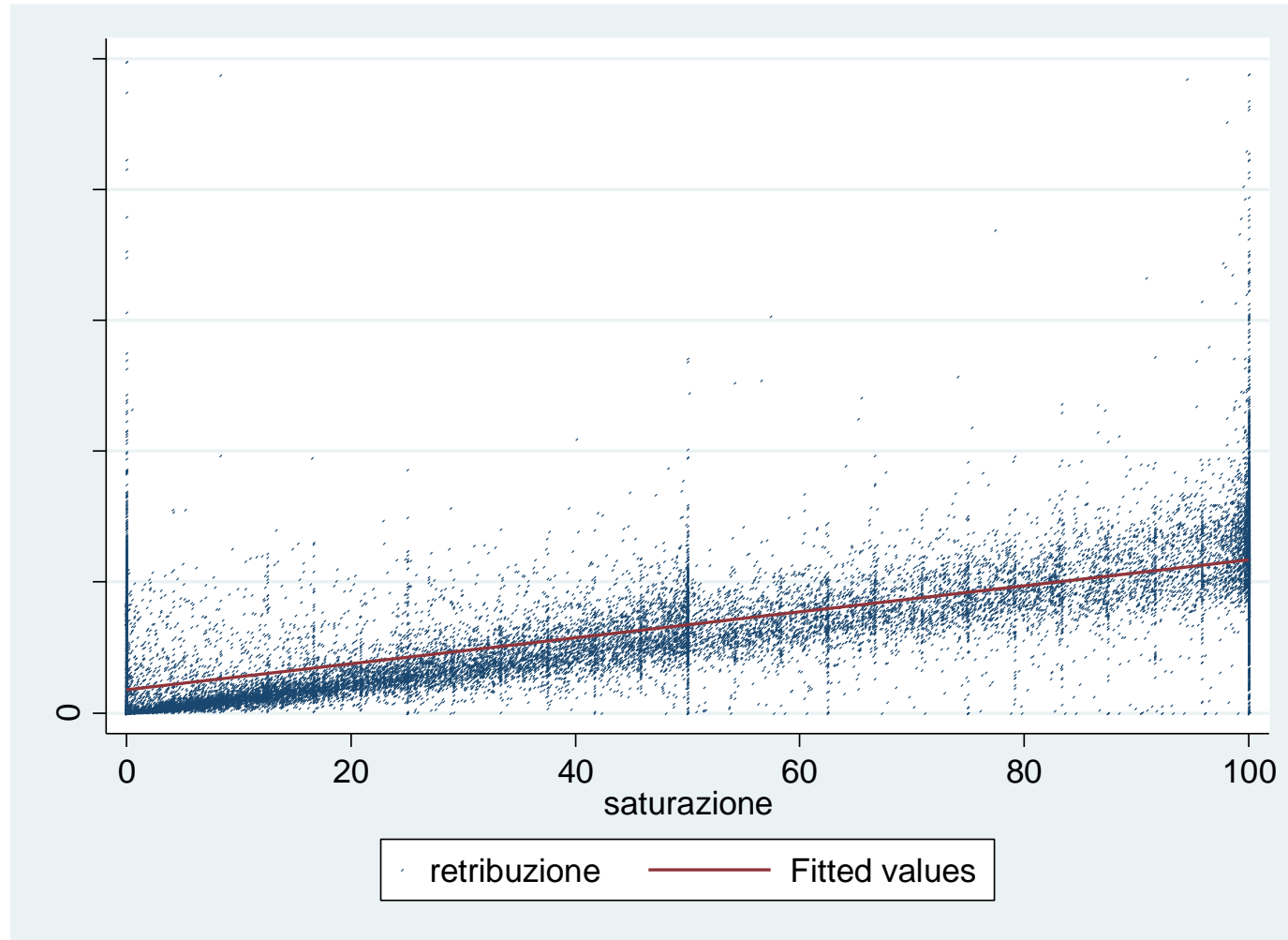
→ retribuzione lorda annua (come sommatoria di quanto percepito a qualunque titolo e in qualunque posizione lavorativa dal laureato nella media dei due anni seguenti alla laurea)

→ saturazione del tempo, ovvero la quota di tempo lavorato complessivamente rispetto al tempo potenzialmente occupabile (definito come 312 giornate lavorative annue a tempo pieno ovvero 12 mesi lavorativi a tempo pieno – le posizioni a tempo parziale sono conteggiate a metà).

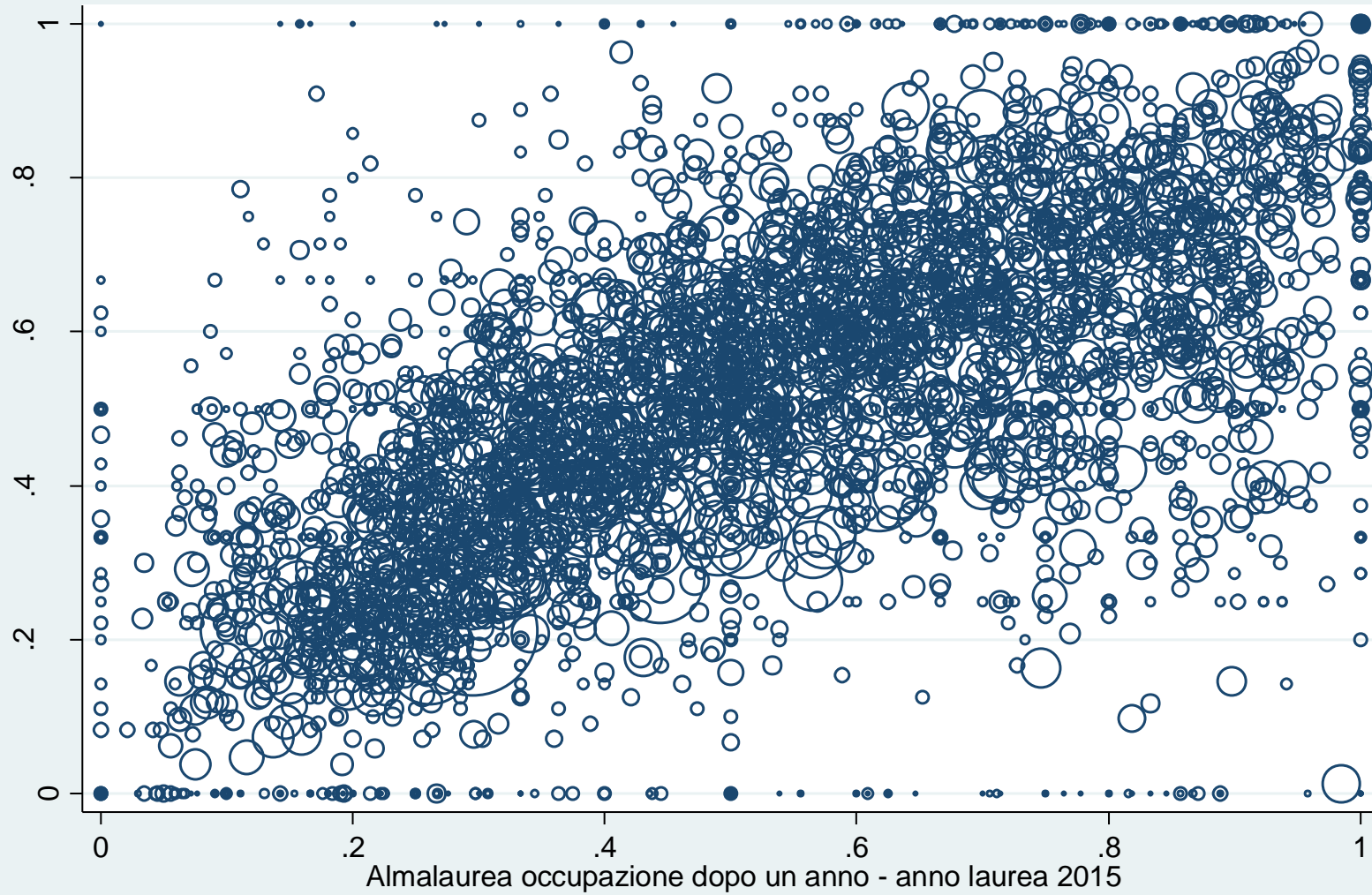
A titolo esemplificativo, ecco i dati relativi ai laureati dell'università di Bologna, per raggruppamento disciplinare delle lauree e anno di laurea

area didattica	2010	2011	2012	Total	area didattica	2010	2011	2012	Total
AGRARIA	8646	9281	8668	8850	INSEGNAMENTO	11916	10864	10284	10991
	34.88	38.72	39.22	37.51		47.40	44.79	42.53	44.82
	260	225	240	725		702	804	774	2280
ARCHITETTURA	8776	7949	8586	8394	LETTERARIA	7372	7200	6673	7081
	28.28	24.50	27.87	26.69		25.34	26.33	25.27	25.66
	132	186	169	487		1224	1314	1257	3795
CHIMICO-FARMACEUTICA	13927	13498	13715	13708	LINGUISTICA	8179	8142	8158	8160
	41.04	42.35	40.60	41.30		31.24	32.15	31.41	31.60
	297	320	358	975		526	534	556	1616
ECONOMICO-STATISTICA	13666	13043	11653	12803	MEDICA	15724	14539	12363	14219
	46.67	46.45	42.84	45.35		37.65	35.36	30.91	34.66
	1378	1320	1310	4008		808	705	777	2290
EDUCAZIONE FISICA	6620	6177	6713	6481	POLITICO-SOCIALE	9023	8589	8139	8591
	31.98	24.75	28.99	28.48		31.46	30.24	29.58	30.44
	151	158	118	427		1128	1164	1075	3367
GEO-BIOLOGICA	8585	8612	8272	8484	PSICOLOGICA	7328	6897	5499	6589
	17.76	21.85	19.32	19.67		26.84	26.01	20.62	24.55
	308	323	342	973		273	287	266	826
GIURIDICA	8289	8195	8483	8316	SCIENTIFICA	13088	12587	11227	12334
	28.47	28.62	28.52	28.54		38.35	37.03	30.42	35.41
	410	449	391	1250		334	317	300	951
INGEGNERIA	16423	17276	16570	16751	retribuzione media	11316	10855	10300	10823
	51.70	53.37	52.10	52.38	saturatione tempo	37.01	36.58	35.00	36.20
	1055	1041	1116	3212	numero laureati	8986	9147	9049	27182

La relazione tra saturazione del tempo lavorato e retribuzione è crescente e sensata.



Tassi occupazione per corso di laurea, pesati per numero laureati
casi 3365 - correlazione 0.687



Alcuni spunti conclusivi

- ① un dato amministrativo è tanto più utilizzabile a fini di monitoraggio/ricerca quanto più la progettazione della procedura amministrativa è improntata alla valutazione dell'efficienza e dell'efficacia del processo (devo domandarmi in anticipo di quali informazioni avrò bisogno a valle del processo).
- ② l'accessibilità pubblica del dato elementare è una garanzia di controllo democratico (accountability)
- ③ la protezione della privacy del dato elementare deve essere contemperata con l'interesse pubblico: se io MIUR/ateneo sussidio per 6000 euro annui la formazione universitaria di un soggetto ho diritto a conoscere il rendimento sul mercato del lavoro dell'investimento, indipendentemente dalla protezione della privacy dei guadagni individuali
- ④ per garantire l'interconnessione delle banche-dati, la codificazione degli identificativi deve essere progettata/monitorata dal centro (attuazione Agenda digitale?)