L'integrazione di differenti tecniche di rilevazione dei dati utilizzando la propensity score

F. Camillo[*], V. Conti[**], S. Ghiselli[**]
[*]Università di Bologna, [**]Consorzio Interuniversitario ALMALAUREA
furio.camillo@unibo.it; valentina.conti@almalaurea.it;
silvia.ghiselli@almalaurea.it

1. Introduzione

Negli ultimi decenni, prima negli Stati Uniti e più recentemente anche nel nostro Paese, si sono via via diffuse nuove e più potenti tecniche di rilevazione dei dati. Infatti, se fino ad una ventina di anni fa si poteva contare, principalmente, su tre metodi di indagine (face-to-face, via telefono o via posta), oggigiorno la diffusione di internet ha aperto nuovi orizzonti, inimmaginabili fino a qualche tempo fa. La tecnica CAWI (Computer-Assisted Web Interview), in particolare, pur costituendo ancora oggi una realtà di nicchia, assume maggiore rilevanza nel momento in cui la popolazione oggetto di studio è composta soprattutto da persone con elevato livello di istruzione, di età mediamente ridotta, che ha facile accesso ad internet e buona padronanza delle più moderne tecnologie. Una popolazione di certo avvantaggiata, sotto questi punti di vista, e che pertanto può essere facilmente raggiunta attraverso il web.

Resta però vero che le più recenti tecniche di rilevazione via internet, ed in particolare la metodologia CAWI, consentono solo raramente di raggiungere tassi di risposta soddisfacenti; per tale motivo, è spesso necessario ricorrere alla combinazione di tecniche di rilevazione differenti, così da garantire la necessaria copertura e rappresentatività del campione di intervistati. È ovvio che in tal caso il ricercatore deve risolvere un problema di non poco conto: l'omogeneizzazione di informazioni raccolte con strumenti di rilevazione differenti.

Il presente contributo si pone come obiettivo proprio quello di valutare se le risposte rese da intervistati sottoposti a diversa tecnica di rilevazione (nel caso in esame, CAWI o CATI - Computer-Assisted Telephone Interview) possono risultare differenti e se, in presenza di distorsioni, queste siano imputabili alla tecnica utilizzata.

È infatti verosimile che la metodologia di indagine possa influenzare, in modo non casuale, le risposte rese dai laureati. Ad esempio, la presenza/assenza del rilevatore è una determinante importante per la qualità delle informazioni raccolte, visto il suo ruolo nel fornire delucidazioni o integrazioni nel corso dell'intervista. D'altra parte, è noto che in taluni casi l'apporto dell'intervistatore può risultare contenuto, se non addirittura

controproducente, perché con il proprio atteggiamento rischia di influenzare le risposte degli intervistati.

2. Indagine CAWI+CATI per la rilevazione della condizione occupazionale dei laureati italiani

L'analisi cui si è appena fatto cenno è resa disponibile grazie all'indagine sulla condizione occupazionale dei laureati condotta, con cadenza annuale, dal Consorzio Interuniversitario ALMALAUREA. La rilevazione consente il monitoraggio delle più recenti tendenze del mercato del lavoro, verificando gli sbocchi occupazionali dei laureati delle università aderenti al Consorzio nel primo quinquennio successivo al conseguimento del titolo (tutti i laureati sono infatti intervistati dopo uno, tre e cinque anni dalla laurea).

Nello specifico, i dati provengono dall'ultima indagine condotta da ALMALAUREA nel corso del 2008: complessivamente, sono stati esaminati oltre 287mila laureati di 47 università aderenti al Consorzio. La rilevazione ha coinvolto anche tutti i laureati post-riforma (circa 140mila) di primo e di secondo livello dell'anno solare 2007[1], quelli di cui si rende conto in queste pagine. La scelta di coinvolgere un così ampio numero di laureati, senza ricorrere a tecniche di campionamento, deriva da numerose esigenze: prima tra tutte quella di rispondere alle richieste dei responsabili dei singoli corsi di laurea, che necessitano di una documentazione ricca ed articolata fino al dettaglio di corso, ma anche al Ministero dell'Istruzione, dell'Università e della Ricerca, il quale ha stabilito che tutti gli atenei italiani sono tenuti a valutare, fino a livello di singolo corso, gli esiti occupazionali dei propri laureati. Questo per varie ragioni: non solo monitorare lo stato di avanzamento della Riforma Universitaria ma anche programmare nuovi corsi di laurea.

L'elevato numero di laureati indagati ha imposto l'utilizzo di strumenti di indagine che garantissero costi e tempi di rilevazione il più possibile contenuti. Come già accennato in precedenza, ciò si è realizzato in particolar modo con l'introduzione di una doppia metodologia di rilevazione, CAWI e CATI. Il disegno di rilevazione ha previsto un'organizzazione dell'indagine su due periodi dell'anno; ciò al fine di garantire un analogo intervallo temporale tra laurea e intervista. I laureati del periodo gennaio-giugno 2007 sono infatti stati intervistati tra aprile e giugno 2008; i laureati del periodo luglio-dicembre sono invece stati intervistati tra settembre e novembre 2008.

[1] In seguito alla Dichiarazione di Bologna (1999), il sistema universitario italiano è stato profondamente modificato. Il nuovo sistema, avviato nei primi anni 2000, prevede tra l'altro l'organizzazione dei corsi di laurea su due livelli: un primo livello di durata triennale, ed un secondo livello di durata biennale.

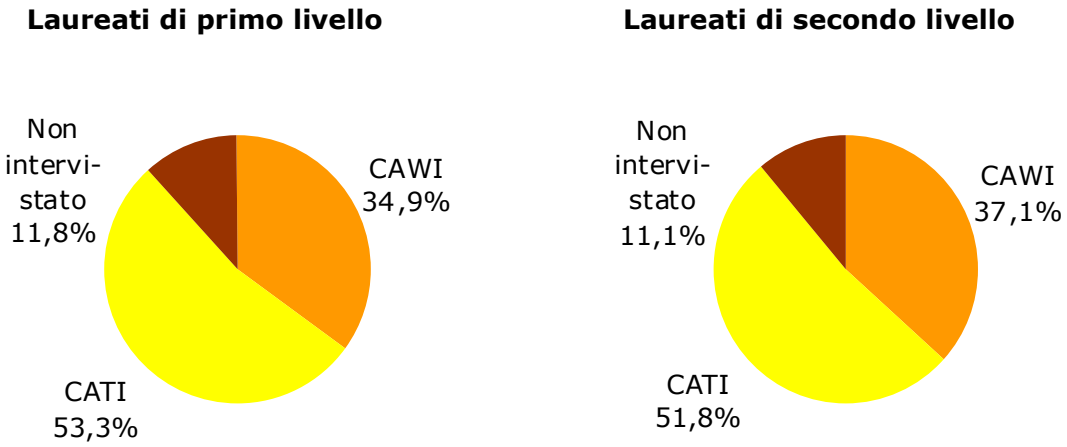

Fig. 1 – Tassi di risposta alle rilevazioni CAWI e CATI per tipologia di laureato
Fonte: ALMALAUREA, anno 2008.

L'ampia disponibilità di indirizzi di posta elettronica (complessivamente pari, per i laureati post-riforma del 2007, all'85%) ha facilitato il ricorso alla rilevazione CAWI (compiuta in tutte le sue fasi all'interno della struttura di ALMALAUREA): i laureati sono così stati contattati via e-mail ed invitati a compilare un questionario ospitato sul sito web di ALMALAUREA. Il disegno di ricerca ha previsto anche due solleciti, inviati a circa una settimana di distanza uno dall'altro. La rilevazione CAWI, conclusasi dopo circa tre settimane dall'avvio, ha condotto a tassi di risposta inusualmente elevati per rilevazioni di questo tipo: se calcolato rispetto alle e-mail inviate, è risultato pari al 41% tra i laureati di primo livello e al 43% tra quelli di secondo livello (rispetto al complesso della popolazione indagata, invece, le percentuali sono rispettivamente 35 e 37% - Fig. 1).

Durante la seconda fase di rilevazione, tutti coloro che, per vari motivi, non avevano compilato il questionario on-line, sono stati contattati telefonicamente, al fine di raggiungere coperture e tassi di risposta soddisfacenti per i fini della ricerca (sul complesso dei laureati coinvolti nella rilevazione risulta superiore all'88%, senza significative differenze tra corsi di primo e secondo livello).

In aggiunta, per garantire stime rappresentative del complesso dei laureati italiani (e colmare quindi il gap dovuto all'assenza, come si è già detto, di una parte di atenei), i risultati presentati nei rapporti ALMALAUREA sono stati sottoposti ad una particolare procedura statistica di "riproporzionamento"[2].

[2] Si tratta di una procedura iterativa che attribuisce ad ogni laureato intervistato un "peso", in modo tale che le distribuzioni relative alle variabili oggetto del riproporzionamento siano -il più possibile- simili a quelle osservate nell'insieme dei laureati italiani. Le variabili considerate in tale procedura sono: genere, gruppo disciplinare, area geografica dell'ateneo, area di residenza alla laurea. Per ottenere stime ancora più precise è stata considerata l'interazione tra la variabile genere e tutte le altre sopraelencate (cfr. CISIA-CERESTA, 2001).

Lo strumento di rilevazione è stato il medesimo per entrambe le fasi di indagine, tranne qualche sporadico adattamento al fine di ottimizzarlo rispetto al tipo di contatto con il laureato (internet o telefono)[3]. Si tratta di un questionario relativamente "snello": nel complesso poco più di 30 domande, anche se ogni laureato poteva rispondere al massimo a 25-27 quesiti, il cui numero è variabile a seconda del tipo di laurea conseguita (primo o secondo livello) e alla condizione occupazionale dichiarata (coloro che lavorano devono rispondere a più domande). Il questionario di indagine ha consentito di rilevare le principali informazioni relative alle esperienze formative e lavorative compiute dopo il conseguimento del titolo: attività di qualificazione post-laurea seguite, condizione occupazionale al momento dell'intervista, tempi di ingresso nel mercato del lavoro, caratteristiche del lavoro svolto (contratto, ramo di attività economica, guadagno, utilizzo delle competenze acquisite all'università e richiesta formale e sostanziale della laurea), solo per citare le principali informazioni rilevate.

Le notizie relative agli esiti occupazionali sono state, successivamente, integrate con l'ampia documentazione già a disposizione di ALMALAUREA grazie al questionario completato alla vigilia della conclusione degli studi dai laureati. Essi lo compilano sia a fini statistici sia per rendere disponibile il proprio cv alle aziende: si tratta di informazioni relative alle caratteristiche socio-demografiche dei laureati (famiglia di origine, area di residenza, genere, età alla laurea), al percorso pre-universitario (tipo di diploma conseguito, voto di diploma) ed universitario compiuto (corso di laurea scelto, voto di laurea, regolarità negli studi), nonché alle ulteriori esperienze formative e di curriculum compiute durante gli anni di studio (conoscenza delle lingue e degli strumenti informatici, esperienze di stage, di studio all'estero e di lavoro).

3. Alcune analisi descrittive preliminari

Vista la complessità della tematica affrontata, si è ritenuto opportuno valutare, inizialmente grazie ad una semplice analisi descrittiva, se esistono differenze strutturali tra coloro che hanno compilato il questionario via web e quanti hanno invece fornito le proprie risposte nel corso della rilevazione telefonica.

Questa analisi preliminare è fondamentale perché permette di comprendere se esiste un'autoselezione del campione di rispondenti all'indagine CAWI rispetto a quella CATI, ovvero se è più probabile che alcuni gruppi di laureati siano più propensi a rispondere al primo stimolo rispetto al secondo. Se questo risulta confermato, è allora importante

[3] In particolare, si è trattato di variare alcune parole chiave, come ad esempio "di seguito troverà elencate alcune attività di formazione" nella versione CAWI è divenuto "di seguito le elencherò alcune attività di formazione" in quella CATI. In altri casi, invece, è stato lievemente modificato il meccanismo di controllo di coerenza tra le risposte date, in particolare alleggerendo queste fasi nella versione CAWI.

rimuovere questo effetto di distorsione prima di valutare se esistono differenze sistematiche legate allo strumento di indagine (ovvero se si rilevano discrepanze tra le risposte rese al questionario via web rispetto a quello telefonico semplicemente in virtù dello strumento utilizzato). In altri termini, se ad esempio rispondono all'indagine telefonica in particolare gli occupati è naturale aspettarsi un maggior numero di risposte "attualmente lavoro" nell'indagine CATI rispetto alla CAWI. In tal caso, però, non si tratterebbe di una distorsione legata al tipo di strumento di indagine, ma semplicemente alla diversa popolazione di rispondenti.

Da una prima analisi descrittiva[4], compiuta distintamente per i laureati di primo e di secondo livello, è emerso chiaramente come siano generalmente più propensi a rispondere allo stimolo web i laureati che hanno una buona padronanza degli strumenti informatici, in particolare della navigazione su internet. Si tratta di studenti che hanno vissuto appieno la propria esperienza universitaria, frequentando regolarmente la maggior parte delle lezioni senza maturare esperienze di lavoro, se non occasionali, part-time; tra l'altro, hanno conseguito la laurea con votazioni mediamente superiori al complesso dei laureati. Il collettivo è composto inoltre da laureati di età mediamente inferiore rispetto alla media. La preparazione di questo gruppo di laureati è confermata, tra l'altro, dalla buona conoscenza della lingua inglese, sia scritta che parlata, e dal voto di laurea, generalmente superiore alla media. Sia i laureati di primo che di secondo livello che hanno optato per la compilazione del questionario CAWI hanno, più degli altri, conseguito un titolo in ingegneria o in un percorso geo-biologico, scientifico, agrario o linguistico. Tra le numerose altre informazioni analizzate è interessante rilevare che questi laureati provengono con maggiore probabilità da famiglie culturalmente più avvantaggiate (i genitori sono in possesso, in particolare, di un titolo di scuola secondaria superiore o di laurea), anche se in alcuni casi hanno un titolo di scuola media inferiore; infine, risiedono o hanno studiato in atenei del Nord.

I laureati che invece hanno accettato di rispondere all'indagine telefonica presentano caratteristiche altrettanto particolari: hanno generalmente un'età più elevata, votazioni inferiori alla media e sono, soprattutto, lavoratori-studenti, ovvero persone che durante gli studi universitari hanno prevalentemente lavorato. Anche per questi motivi, di solito non hanno compilato il questionario loro sottoposto al termine dell'università; ne deriva che molte informazioni di curriculum non sono disponibili. È interessante comunque sottolineare che si tratta soprattutto di laureati, sia di primo che di secondo livello, delle professioni sanitarie o dei gruppi giuridico, educazione fisica, insegnamento o

[4] Tale analisi è stata sviluppata attraverso una procedura statistica, denominata DEMOD, che permette di identificare le variabili, indipendentemente dalla loro natura, in grado di caratterizzare un determinato gruppo di persone. Attraverso opportuni test probabilistici (di fatto dei chi-quadrato), infatti, si individuano le modalità o le medie delle variabili (a seconda della tipologia di variabile in esame) che risultano significativamente diverse rispetto al complesso della popolazione. I risultati illustrati, pertanto, devono essere letti in termini probabilistici.

architettura. Con maggiore frequenza risiedono o hanno studiato in atenei del Sud (tra i laureati di secondo livello è significativa anche la quota di chi abita o ha studiato al Centro). Come ci si poteva attendere, infine, la padronanza della navigazione su internet è alquanto modesta.

4. Modello di regressione logistica per la valutazione delle variabili che incidono sulla partecipazione alla rilevazione CAWI rispetto alla CATI

La stima di una possibile distorsione dei dati dovuta allo strumento utilizzato ha comportato, successivamente all'analisi descrittiva, la valutazione simultanea di tutti i fattori che possono incidere sulla probabilità di reagire ad uno stimolo rispetto ad un altro.

Ciò è avvenuto attraverso l'adozione di un modello di regressione logistica che ha valutato la probabilità di partecipare all'indagine CAWI rispetto a quella CATI[5].

Le variabili considerate sono relative a informazioni *socio-demografiche* (genere, area geografica di residenza, classe sociale[6] e titolo di studio dei genitori), sulla *carriera universitaria* (gruppo disciplinare, area geografica dell'ateneo, informazioni su tipologia di laureato puro/ibrido[7], voto ed età alla laurea, regolarità degli studi[8], frequenza alle lezioni), sulle *conoscenze ed esperienze acquisite durante gli anni di studio* (competenze linguistiche e informatiche, esperienze di lavoro e di studio all'estero) nonché sulle *aspettative future* (intenzione di proseguire ulteriormente gli studi, disponibilità a trasferte per motivi lavorativi, ipotesi di reiscrizione all'università).

[5] Il modello valuta tale probabilità in funzione di una serie di variabili esplicative (definite *covariate*) la cui selezione è derivata dalle analisi descrittive preliminari. Per la selezione del modello è stata utilizzata la procedura *forward stepwise conditional process*, che consiste nell'introdurre una variabile alla volta nell'equazione di regressione. Ad ogni passo si inserisce la variabile che ha la maggiore capacità esplicativa; è inoltre possibile eliminare le variabili inserite precedentemente nel modello, le quali divengono non significative dopo l'introduzione di ulteriori fattori esplicativi.

[6] Per la classe sociale dei laureati si è adottato lo schema proposto da A. Schizzerotto (2002). La classe sociale, definita sulla base del confronto fra la *posizione socioeconomica* del padre e quella della madre del laureato, corrisponde alla posizione di livello più elevato fra le due.

[7] ALMALAUREA definisce "puro" il laureato che si è immatricolato ad un corso riformato (ovvero organizzato su due livelli); "ibrido" è invece colui che ha concluso un corso riformato con il contributo di crediti formativi maturati all'interno di percorsi pre-riforma. Come ci si può facilmente attendere, quest'ultima tipologia di laureati presenta generalmente performance di studio più modeste.

[8] Si tratta di una variabile che considera il tempo realmente impiegato dal laureato per terminare l'università. È un aspetto importante, soprattutto tenendo conto dell'elevata quota di laureati italiani (circa il 60% nella generazione del 2008) che termina il proprio percorso formativo oltre i tempi previsti.

Tab. 1 – Modelli di regressione logistica, stimati per tipologia di laureato, per valutare la probabilità di partecipare alla rilevazione CAWI rispetto alla CATI

	Laureati di primo livello			Laureati di secondo livello		
	B	**Exp(B)**	**Sig.**	**B**	**Exp(B)**	**Sig.**
Genere				*ne*		
Donne	*mr*					
Uomini	-0,102	0,903	0,000			
Area geografica di residenza			0,001	*ne*		
Nord	*mr*					
Centro	-0,137	0,872	0,000			
Sud e Isole	-0,121	0,886	0,000			
Estero	-0,011	0,989	0,912			
Classe sociale dei genitori			0,000			0,000
Borghesia	-0,117	0,889	0,000	-0,154	0,858	0,000
Classe media impiegatizia	0,002	1,002	0,938	-0,058	0,944	0,126
Piccola borghesia	-0,093	0,911	0,000	-0,112	0,894	0,008
Classe operaia	*mr*			*mr*		
Non risponde	-0,182	0,833	0,001	-0,359	0,698	0,000
Titolo di studi dei genitori			0,008	*ne*		
Nessun titolo	-0,255	0,775	0,043			
Licenza elementare	-0,037	0,964	0,282			
Licenza media inferiore	-0,029	0,971	0,135			
Scuola secondaria superiore	*mr*					
Un solo genitore con laurea	0,022	1,022	0,326			
Entrambi i genitori con laurea	0,042	1,043	0,145			
Non risponde	-0,221	0,802	0,004			
Gruppo disciplinare			0,000			0,000
Agrario	-0,223	0,800	0,000	0,042	1,043	0,715
Architettura	-0,593	0,553	0,000	-0,495	0,610	0,000
Chimico-farmaceutico	-0,315	0,730	0,000	0,117	1,124	0,430
Economico-statistico	-0,344	0,709	0,000	-0,340	0,711	0,000
Educazione fisica	-0,614	0,541	0,000	-0,475	0,622	0,001
Geo-biologico	-0,190	0,827	0,000	0,034	1,035	0,665
Giuridico	-0,546	0,579	0,000	-0,694	0,500	0,000
Ingegneria	-0,115	0,892	0,009	-0,067	0,935	0,337
Insegnamento	-0,560	0,571	0,000	-0,390	0,677	0,001
Letterario	-0,443	0,642	0,000	-0,216	0,806	0,007
Linguistico	-0,298	0,742	0,000	-0,095	0,909	0,314
Medico	-0,931	0,394	0,000	-0,322	0,725	0,008
Politico-sociale	-0,358	0,699	0,000	-0,192	0,825	0,008
Psicologico	-0,346	0,707	0,000	-0,143	0,867	0,066
Scientifico	*mr*			*mr*		
Area geografica dell'Ateneo			0,000			0,004
Nord	*mr*			*mr*		
Centro	0,285	1,330	0,000	0,088	1,092	0,009
Sud e Isole	0,159	1,173	0,000	-0,044	0,957	0,241
Puro/Ibrido			0,000			0,000
Puro	*mr*			*mr*		
Ibrido	0,048	1,049	0,046	0,011	1,011	0,764
Non risponde	-1,918	0,147	0,000	-1,844	0,158	0,000
Voto di laurea	0,009	1,009	0,000	0,009	1,009	0,001
Età alla laurea	-0,005	0,995	0,032	-0,011	0,989	0,004
Regolarità negli studi			0,000	*ne*		
In corso, in età canonica	*mr*					
In corso, oltre l'età canonica	-0,140	0,869	0,000			
1 anno fuori corso	-0,123	0,884	0,000			
2 anni fuori corso e oltre	-0,138	0,871	0,000			

segue

segue Tab. 1 – Modelli di regressione logistica, stimati per tipologia di laureato, per valutare la probabilità di partecipare alla rilevazione CAWI rispetto alla CATI

	Laureati di primo livello			Laureati di secondo livello		
	B	**Exp(B)**	**Sig.**	**B**	**Exp(B)**	**Sig.**
Frequenza regolare alle lezioni			0,000			0,019
Meno del 25% degli insegnamenti previsti	-0,103	0,902	0,005	-0,119	0,888	0,094
Tra il 25 e il 50%	-0,081	0,922	0,008	-0,132	0,876	0,050
Tra il 50 e il 75%	-0,110	0,896	0,000	-0,113	0,893	0,006
Oltre il 75% degli insegnamenti	*mr*			*mr*		
Non risponde	-0,161	0,851	0,058	-0,128	0,880	0,466
Conoscenza della lingua inglese parlata			0,000			
Nessuna	*mr*			0,090	1,094	0,715
Limitata	-0,029	0,972	0,792	0,441	1,554	0,000
Discreta	-0,132	0,876	0,230	0,157	1,170	0,000
Buona	-0,238	0,789	0,033	*mr*		
Ottima	-0,333	0,717	0,004	-0,093	0,911	0,067
Madrelingua	-0,480	0,619	0,105	0,193	1,213	0,792
Non risponde	-0,203	0,816	0,135	0,152	1,164	0,384
Conoscenza della lingua inglese scritta			0,001			
Nessuna	*mr*			0,007	1,007	0,981
Limitata	0,248	1,282	0,051	-0,294	0,745	0,001
Discreta	0,292	1,339	0,024	-0,151	0,860	0,000
Buona	0,338	1,403	0,009	*mr*		
Ottima	0,414	1,512	0,002	0,050	1,051	0,283
Madrelingua	0,285	1,330	0,348	0,277	1,319	0,715
Non risponde	0,133	1,142	0,395	-0,241	0,786	0,232
Conoscenza nella navigazione internet			0,000			
Nessuna	-0,556	0,574	0,000	-0,499	0,607	0,002
Limitata	-0,593	0,553	0,000	-0,635	0,530	0,000
Discreta	-0,450	0,637	0,000	-0,440	0,644	0,000
Buona	-0,204	0,815	0,000	-0,166	0,847	0,000
Ottima	*mr*			*mr*		
Non risponde	-0,449	0,638	0,000	-0,267	0,766	0,012
Esperienze lavorative durante gli studi			0,000			
Lavoratore-studente	*mr*			*mr*		
Studente-lavoratore	0,008	1,008	0,808	0,132	1,141	0,027
Nessuna esperienza di lavoro	-0,070	0,933	0,064	0,173	1,189	0,007
Non risponde	0,310	1,363	0,128	0,125	1,133	0,520
Lavorava alla laurea?			0,001	*ne*		
Sì	-0,068	0,934	0,000			
No	*mr*					
Non risponde	-0,152	0,859	0,419			
Ha studiato all'estero?			0,006			0,000
Sì	0,081	1,084	0,002	0,227	1,254	0,000
No	*mr*			*mr*		
Non risponde	-0,045	0,956	0,700	-0,358	0,699	0,114
Intende proseguire gli studi?			0,000			0,000
Sì	*mr*			0,115	1,122	0,000
No	-0,142	0,867	0,000	*mr*		
Non risponde	-0,437	0,646	0,000	-0,541	0,582	0,010
Disponibilità a trasferte			0,000			0,000
Sì, anche con trasferimenti di residenza	*mr*			*mr*		
Sì, anche frequenti	0,086	1,090	0,000	0,095	1,100	0,002
Sì, ma solo limitate	0,031	1,031	0,120	0,042	1,043	0,261
Non disponibile a trasferte	-0,130	0,878	0,001	-0,204	0,816	0,029
Non risponde	-0,367	0,693	0,000	-0,166	0,847	0,130

Segue

segue Tab. 1 – Modelli di regressione logistica, stimati per tipologia di laureato, per valutare la probabilità di partecipare alla rilevazione CAWI rispetto alla CATI

	Laureati di primo livello			Laureati di secondo livello		
	B	**Exp(B)**	**Sig.**	**B**	**Exp(B)**	**Sig.**
Si riscriverebbe all'università?			0,000	*ne*		
Stesso corso e stesso ateneo	*mr*					
Altro corso dello stesso ateneo	0,059	1,061	0,010			
Stesso corso ma presso altro ateneo	0,177	1,194	0,000			
Altro corso e altro ateneo	0,096	1,100	0,001			
Non si iscriverebbe più all'università	-0,146	0,864	0,022			
non risponde	0,097	1,101	0,278			

Ne= variabile non entrata nel modello
Mr= modalità di riferimento nel calcolo dei coefficienti b
Fonte: ALMALAUREA, anno 2008.

Come si può notare, si tratta in tutti i casi di variabili strutturali e di percorso formativo e lavorativo precedenti all'acquisizione del titolo. La scelta di includere solo questo tipo di variabili è ovvia: occorre selezionare i soli fattori "pre-trattamento", ovvero precedenti alla rilevazione CAWI o CATI.

L'analisi è stata effettuata distintamente per tipologia di laureato, ovvero è stato adottato un modello di regressione logistica per i laureati di primo livello e uno per quelli di secondo livello. Anche se nel modello relativo ai laureati di secondo livello sono entrate meno variabili, le stime ottenute sono molto simili, a conferma che l'autoselezione del campione di rispondenti non è tanto legata al tipo di laurea conseguita, quanto ad un altro ordine di fattori: dimestichezza con l'uso di strumenti informatici e maggiore disponibilità di tempo libero, ad esempio, sono elementi che favoriscono la partecipazione alla rilevazione via web. I risultati tra l'altro confermano, salvo qualche eccezione, quanto emerso nel corso dell'analisi descrittiva preliminare. Le differenze rilevate rispetto all'analisi descrittiva sono tra l'altro da attribuire alla maggiore accuratezza del modello di regressione che, si ricorda, considera simultaneamente tutti i fattori in gioco.

In particolare, nella tabella 1 sono riportati i coefficienti stimati per ciascuna modalità (colonne B e Exp(B)) e la relativa significatività (colonna Sig.). Tra i risultati più interessanti si rileva che, sia tra i laureati di primo che tra quelli di secondo livello, rispondono con maggior probabilità alla rilevazione CAWI i laureati del gruppo scientifico (meno probabile invece la risposta di quelli dei gruppi medico, giuridico, architettura, educazione fisica ed insegnamento). Si nota infatti che tutti i coefficienti stimati sono negativi, indicando pertanto una probabilità minore di rispondere alla rilevazione via web rispetto alla modalità di riferimento, rappresentata dai laureati del gruppo scientifico.

Inoltre, al diminuire delle capacità di navigare con dimestichezza su internet diminuisce corrispondentemente la probabilità di partecipare alla rilevazione CAWI:

rispetto a chi possiede un'ottima conoscenza, infatti, coloro che dichiarano di non avere nessuna competenza hanno il 40% in meno di probabilità di compilare il questionario CAWI.

La percentuale di corretta classificazione è complessivamente pari al 62,6% per i laureati di primo livello e al 61,9% per quelli di secondo; percentuali che si riducono al 50% nel caso di classificazione di coloro che partecipano alla rilevazione via web. Tali valori non consentono di escludere l'idea che l'autoselezione degli intervistati dipenda dalle loro caratteristiche pre-trattamento, ma neppure di esplicitare appieno il meccanismo di autoselezione. In casi come questi è doveroso quindi confrontare il collettivo di rispondenti all'indagine CAWI con quello CATI, ponendo però entrambi "a parità di condizioni" rispetto alle informazioni pre-trattamento. Tutto ciò, si ricorda, al fine di ottenere una misura netta della distorsione dei risultati legata all'utilizzo di diverse tecniche di rilevazione.

5. Propensity score matching per la valutazione delle differenze tra risposte CAWI e CATI

Come già anticipato, è ragionevole pensare che trattamenti diversi possano dar luogo a risposte diverse, a parità di tutte le condizioni rilevanti, e questo proprio a causa della diversa metodologia di indagine.

Domande come quella relativa al tipo di contratto, ad esempio, possono dare origine a interpretazioni diverse semplicemente in virtù del tipo di somministrazione cui ciascun laureato è sottoposto. Nell'ambito della rilevazione CAWI il laureato può scorrere visivamente sullo schermo tutte le modalità di risposta previste e riflettere accuratamente su quale sia la forma contrattuale corrispondente alla propria attività lavorativa. Viceversa, la rilevazione telefonica, seppure affetta dal problema di non riuscire a presentare visivamente e simultaneamente tutte le modalità di risposta, presenta un importante punto di forza: il rilevatore, che può in ogni momento fornire al laureato delucidazioni finalizzate al miglior riconoscimento delle forma contrattuale più corretta.

Per eliminare l'autoselezione nella partecipazione ad un'indagine rispetto all'altra, i rispondenti sono stati classificati in gruppi omogenei di "propensione" alla rilevazione CAWI rispetto alla CATI. In sostanza, il modello di regressione logistica illustrato in precedenza ha permesso di assegnare a ciascun laureato una probabilità (variabile tra 0 e 1) di partecipazione all'indagine CAWI, date le variabili pre-trattamento a disposizione: più è alto il valore associato ad un laureato più è probabile, date le caratteristiche considerate nel modello, che egli partecipi alla rilevazione via web rispetto a quella

telefonica. È ovvio pertanto che laureati cui si associano valori di probabilità simili saranno altrettanto simili anche dal punto di vista della propensione a rispondere all'una rispetto che all'altra rilevazione. In altri termini, se all'interno di uno di questi gruppi di laureati due persone sono state sottoposte a diverso trattamento (ovvero diversa tecnica di rilevazione) ciò può essere imputato al caso e non alle caratteristiche personali e di curriculum; questo perché di tali aspetti si è già tenuto conto nel modello di regressione logistica sviluppato. Di conseguenza, eventuali differenze nelle risposte rese dai laureati possono essere imputate esclusivamente al tipo di strumento di indagine utilizzato e non alle caratteristiche personali.

Formalmente, il metodo di valutazione delle differenze di risposta legate al trattamento (CAWI o CATI) è sviluppato seguendo un approccio legato a *impact evaluation in observational studies*. L'obiettivo consiste nel valutare se su una variabile target *Y* (le risposte alle varie domande del questionario somministrato) esercita un'influenza significativa il tipo di trattamento *T* subìto (CAWI e CATI).

Nello specifico, il trattamento *T* ha un effetto causale sulla variabile target *Y* per l'individuo *i*-esimo se il risultato in caso di trattamento (T_1) è diverso dal caso di assenza di trattamento (o di trattamento differente, T_0), ossia se viene verificata la seguente relazione:

$$\Delta_i \equiv Y_i(T_1) - Y_i(T_0) \neq 0$$

Nella realtà questa relazione non è osservabile, poiché ciascun individuo *i* subisce il trattamento T_1 o T_0. Ma, formulando adeguate ipotesi, è possibile identificare e misurare un effetto causale per l'individuo medio della popolazione in esame.

$$E\{\Delta\} \equiv E\{Y(T_1)\} - E\{Y(T_0)\} \neq 0$$

In tal modo è però necessario selezionare due campioni indipendenti e casuali, quello sottoposto al trattamento T_1 e quello di controllo T_0, nei quali si osserva la variabile target *Y*. È però vero che nel caso in esame ciò risulta più difficile, in quanto, come si è già detto, il trattamento stesso è legato a fenomeni di autoselezione della popolazione, che per definizione allontanano dall'ipotesi di casualità.

Tale problema, allora, può essere affrontato facendo riferimento all'approccio proposto da Rosembaum e Rubin (1983), noto come *propensity score*, che è definito come la probabilità condizionata, calcolata per ogni individuo, di ricevere un trattamento dato un vettore di covariate. La formula con cui viene spesso stimato è la seguente:

$$ln\left(\frac{e(x)}{1-e(x)}\right) = a + \beta^T f(x)$$

dove e(x) = Pr (T_1|x) è la probabilità di ricevere un trattamento (T_1) data una serie di covariate **x** e f(**x**) è una funzione di covariate. Dato un fissato propensity score, le covariate e i risultati dell'indagine diventano indipendenti dal tipo di trattamento subito.

Riassumendo quanto detto sopra ed adattandolo al caso concreto qui in esame, il modello logico che si intende utilizzare è così strutturato:

- *X* = matrice delle caratteristiche individuali e di curriculum disponibili sugli individui, tutte relative alla fase pre-trattamento e stimate attraverso il modello di regressione logistica;
- *T* = variabile di trattamento, che definisce l'appartenenza a una delle categorie di trattamento (CAWI o CATI);
- *Y* = variabile target, ovvero un insieme di variabili relative all'indagine sulla condizione occupazionale (ad esempio: condizione occupazionale e formativa, tempi di ingresso nel mondo del lavoro, guadagno, ecc.).

A partire dal modello di regressione logistica, pertanto, sono stati costruiti gruppi omogenei di laureati, all'interno dei quali risulta verificata la probabilità di bilanciamento, ovvero l'indipendenza rispetto alle variabili pre-trattamento X[9]. La verifica di tale probabilità è fondamentale perché solo in tal modo vi è la certezza di aver eliminato il problema dell'autoselezione.

All'interno di questi sottoinsiemi di popolazione è possibile valutare le discrepanze nelle risposte rese dai laureati e legate al diverso trattamento, semplicemente calcolando la differenza, per ciascuna variabile target, tra la distribuzione osservata e quella attesa in caso di indipendenza in distribuzione (ossia di indipendenza tra il tipo di intervista e la variabile target).

Gli effetti sulle stime delle percentuali di risposta alle specifiche modalità non sono mai superiori ai 2 punti percentuali in termini di scostamento, risultato questo molto incoraggiante perché conferma che le due tecniche di rilevazione non generano differenze

[9] Tale proprietà è stata verificata, per ciascuna variabile considerata nel modello di regressione, attraverso i test χ^2. Per i laureati di primo livello si è scomposta la popolazione in esame in quintili e successivamente, visto che all'interno del primo gruppo la proprietà di bilanciamento non era verificata, si è ulteriormente suddiviso il primo quintile in terzili. All'interno dei primi due terzili la proprietà di bilanciamento continuava a non risultare verificata, per tali motivi questa parte di popolazione (pari al 13,3%) non è stata considerata nelle successive analisi. Per i laureati di secondo livello la suddivisione è avvenuta dapprima in quartili e poi, all'interno del primo gruppo, in terzili. Al termine delle opportune verifiche sono stati esclusi i primi due terzili, pari al 16,7% del complesso della popolazione. I laureati esclusi dalle successive analisi mostrano caratteristiche molto particolari; tanto particolari da rendere praticamente impossibile l'eliminazione dell'autoselezione legata al tipo di rilevazione e da giustificare la loro esclusione dagli approfondimenti di seguito sviluppati. Inoltre, questi laureati generalmente non hanno compilato il questionario necessario per il recupero di molte informazioni relative alle esperienze pre-trattamento.

significative in termini di risposte fornite. Fanno eccezione due soli casi, che hanno permesso di individuare alcune anomalie nella formulazione dei relativi quesiti nella versione CAWI del questionario.

Il primo caso fa riferimento alla domanda relativa alla tipologia dell'attività lavorativa: come si può notare dalla tabella 2, nel caso della modalità "altro contratto a tempo determinato" si rileva uno scostamento pari o superiore a 2 punti percentuali.

Si ritiene che ciò sia dovuto all'uso della parola "altro" che può confondere il laureato, in particolare nella compilazione del questionario CAWI: il termine "altro", infatti, potrebbe spingere il laureato a ritenere tale modalità residuale, mentre in realtà essa identifica il vero e proprio contratto a termine.

Tab. 2 – Stima delle differenze rilevate tra le risposte rese alla domanda sulla tipologia dell'attività lavorativa nell'indagine CAWI rispetto alla CATI, calcolate attraverso il propensity score matching, distintamente per tipologia di laureato

	Laureati primo livello		Laureati secondo livello	
	CAWI	**CATI**	**CAWI**	**CATI**
Tempo indeterminato	-0,543	0,462	1,793	-1,093
Contratto di inserimento	1,598	-1,358	1,000	-0,610
Apprendistato	0,329	-0,279	0,806	-0,491
Lavoro interinale	0,235	-0,199	0,911	-0,555
Lavoro a progetto	1,083	-0,921	0,274	-0,167
Collaborazione occasionale	-0,140	0,119	-0,753	0,459
Prestazione d'opera	0,842	-0,716	0,747	-0,455
Lavoro socialmente utile/di pubblica utilità	0,064	-0,055	0,092	-0,056
Piano di inserimento professionale	0,071	-0,060	0,117	-0,071
Lavoro intermittente o a chiamata	0,063	-0,053	0,205	-0,125
Job sarin	0,027	-0,023	0,074	-0,045
Lavoro occasionale accessorio	0,181	-0,154	0,580	-0,354
Associazione in partecipazione	0,052	-0,044	0,147	-0,090
Altro contratto a tempo determinato	-3,102	2,638	-3,114	1,898
Lavoro autonomo effettivo	-1,056	0,898	-1,330	0,811
Lavoro senza contratto	0,272	-0,231	-1,559	0,951
Non risponde	0,026	-0,022	0,011	-0,006

Fonte: ALMALAUREA, anno 2008.

Il secondo caso individuato riguarda invece la domanda relativa alla ricerca di un lavoro. Il quesito è in sé molto semplice e inequivocabile, ma l'assenza, nella versione CAWI, di alcuni controlli di coerenza ha determinato una certa discrepanza tra le risposte rese dai laureati che hanno optato per la compilazione via web rispetto a quella telefonica. Di seguito si riporta, per semplicità di esposizione, la sezione di questionario ora in esame.

D1. Lei sta cercando attivamente lavoro? Ai nostri fini la ricerca deve essere attiva, ovvero deve aver compiuto almeno un'azione concreta, ad es. inviando un curriculum.
[01] sì
[02] no

[da porre solo se D1=01]
D2. Quando ha compiuto l'ultima iniziativa per cercare un lavoro? Le ricordo che la ricerca del lavoro deve essere attiva ovvero deve aver compiuto almeno un'azione concreta di ricerca.
[01] negli ultimi 15 giorni
…
[04] oltre 6 mesi fa
[05] non ha ancora compiuto azioni concrete. Questa non è una riposta valida, perché la ricerca deve essere attiva. Tornare indietro e correggere le risposte *(modalità prevista solo per la CATI)*

Come si può notare, solo la versione CATI prevede la modalità 5 nella domanda D2: questa opzione di risposta non è cliccabile dall'intervistatore, serve solo per evidenziare un'incoerenza tra le risposte rese dal laureato. Come può infatti l'intervistato cercare attivamente un lavoro e non aver compiuto alcuna iniziativa per trovarlo? Nella versione CAWI questo controllo di coerenza non è stato di fatto previsto; seguendo quanto previsto sia dalla letteratura scientifica che dalla pratica operativa, infatti, ci si è mossi con l'intento di rendere il più semplice possibile la compilazione del questionario e dare massima fluidità al questionario.

Ma l'omissione di questo controllo di coerenza, ha generato, come già anticipato, una differenza sostanziale tra le risposte rese alla domanda D1 nelle due tecniche di rilevazione (tabella 3).

Tab. 3 – Stima delle differenze rilevate tra le risposte rese alla domanda sulla ricerca di lavoro nell'indagine CAWI rispetto alla CATI, calcolate attraverso il modello di propensity score matching, distintamente per tipologia di laureato

	Laureati primo livello		**Laureati secondo livello**	
	CAWI	**CATI**	**CAWI**	**CATI**
Cerca lavoro	10,705	-9,883	9,890	-8,020
Non cerca lavoro	-10,705	9,883	-9,891	8,021

Fonte: ALMALAUREA, anno 2008.

Tuttavia, ciò non ha generato problemi di particolare entità in fase di elaborazione dati, in quanto l'informazione contenuta nella domanda D1 è utilizzata solo per coloro che non si dichiarano occupati (tra coloro che hanno partecipato alla CAWI è pari al 10,9% tra i laureati di primo livello e al 41,3% tra quelli di secondo livello).

Tuttavia, per un'analisi più accurata, si è ritenuto opportuno utilizzare un sistema di pesi che ha permesso di correggere e valutare la distorsione generata. Seguendo lo

schema proposto da Lee (2006), il peso corretto per l'unità j della classe c[10] del campione CAWI diventa:

$$d_j^{W.PSA} = f_c d_j^w = \frac{\hat{N}_c^R / \hat{N}^R}{\hat{N}_c^W / \hat{N}^W} d_j^w$$

where

d_j^w = base peso, necessario per avere stime rappresentative della popolazione italiana

$\hat{N}_c^R$ = numero delle unità della classe c nell'indagine di riferimento (in questo caso l'indagine CATI)

$\hat{N}^R$ = numero totale delle unità nell'indagine di riferimento (in questo caso l'indagine CATI)

$\hat{N}_c^W$ = numero delle unità nella classe c dell'indagine CAWI

$\hat{N}^W$ = numero totale delle unità nell'indagine CAWI

Da notare che ciascun laureato di ogni gruppo c avrà il medesimo peso f_c.

Nella tabella 4 sono riportati i risultati rilevati per la condizione occupazionale dei laureati utilizzando il semplice peso d_j^w (prima colonna), necessario per ottenere stime rappresentative della popolazione italiana, e utilizzando il peso corretto $d_j^{W.PSA}$ (seconda colonna). Come si nota le differenze sono decisamente contenute, nell'ordine di un paio di punti percentuali. Ciò conferma che, nonostante il problema di rilevazione dell'informazione relativa alla ricerca di un lavoro, la distorsione generata è di fatto irrilevante.

Tab. 4 – Condizione occupazionale, distintamente per tipologia di laureato: confronto tra valori pesati e valori pesati e corretti con il sistema di pesi proposto da Lee

	Condizione occupazionale	
	Valori pesati	Valori pesati e corretti
Laureati di primo livello		
Lavora	29,2	31,5
Lavora ed è iscritto alla laurea di secondo livello	16,1	16,0
E' iscritto alla laurea di secondo livello (e non lavora)	44,6	42,0
Non cerca lavoro	2,8	2,9
Cerca lavoro	7,3	7,6
Laureati di secondo livello		
Lavora	61,7	62,1
Non cerca lavoro	17,7	17,4
Cerca lavoro	20,6	20,5

Fonte: ALMALAUREA, anno 2008.

[10] Si intende ciascun gruppo individuato all'interno del quale è verificata la proprietà di bilanciamento.

6. Conclusioni

Nel presente contributo, dopo una breve analisi descrittiva, si è fatto ricorso al *propensity score matching method* per valutare se, a parità di quesito, esistono differenze nelle risposte rese ad indagini CAWI rispetto a quelle fornite attraverso metodologia CATI. Nel caso specifico, l'approfondimento ha riguardato circa 140mila laureati italiani del 2007 che sono stati coinvolti nel 2008 nell'indagine che il Consorzio ALMALAUREA conduce annualmente al fine di valutarne gli esiti occupazionali.

Le analisi sono state compiute distintamente per tipologia di laureato (di primo o di secondo livello), senza peraltro evidenziare particolari differenze tra i due collettivi. Ciò significa che l'autoselezione del campione di rispondenti è legata non tanto al tipo di laurea conseguita, quanto ad una serie di fattori relativi soprattutto alla dimestichezza con l'uso di strumenti informatici e alla disponibilità di tempo libero. Fattori, questi, che innalzano considerevolmente la probabilità di partecipare alla rilevazione on-line.

Attraverso alcuni modelli di regressione logistica si sono stimate le variabili che incidono con maggiore forza sulla propensione a rispondere alla rilevazione via web rispetto a quella telefonica. È così emerso che è più probabile che rispondano i laureati del gruppo scientifico, al contrario di quelli dei gruppi medico, giuridico, architettura, educazione fisica ed insegnamento. Inoltre, solo per citare i risultati più interessanti, la probabilità di rilasciare l'intervista via web cresce al crescere delle competenze informatiche, in particolare alla capacità di navigare con dimestichezza su internet.

Grazie ai modelli di regressione logistica, inoltre, è stato possibile attribuire a ciascun laureato la relativa probabilità di risposta alla rilevazione CAWI condizionatamente alle variabili pre-trattamento disponibili. Successivamente si sono aggregati i laureati in gruppi omogenei di probabilità. Ciò significa che, se entro un determinato gruppo di laureati, una persona ha partecipato all'indagine via web e una a quella telefonica, ciò può essere attribuito esclusivamente al caso, dato che in termini probabilistici il modello di regressione li rende confrontabili. È per tale motivo che entro ciascun gruppo di laureati è stato possibile valutare le differenze tra le risposte rese all'indagine CAWI rispetto alla CATI.

Le discrepanze tra le risposte rese sono decisamente contenute, nell'ordine di qualche punto percentuale, salvo due eccezioni che hanno permesso di individuare altrettante anomalie nella formulazione del questionario CAWI. In un caso si tratta della domanda relativa alla tipologia dell'attività lavorativa svolta dove, utilizzando la parola "altro" in corrispondenza di una modalità di risposta, è stata generata una sottostima, nella versione CAWI, di tale scelta da parte dei laureati. Nel secondo caso, invece, l'assenza di un controllo di coerenza tra alcune risposte rese dall'intervistato ha di fatto aumentato la quota di laureati che dichiara di cercare un lavoro. È verosimile che in questi casi ciò sia

imputabile in particolare all'assenza del rilevatore, figura molto importante in quanto capace di fornire ragguagli e delucidazione nel corso dell'intervista. Sarà pertanto premura del Consorzio ALMALAUREA, in vista delle future rilevazioni, correggere tali limiti nella versione CAWI del proprio questionario. Fortunatamente, questi problemi non hanno generato errori di particolare entità in fase di analisi dei dati: le variazioni sono pari a circa 2 punti percentuali. Ciò è stato valutato adottando un sistema di pesi che di fatto ha permesso di correggere la distorsione imputabile allo strumento di rilevazione.

Il sistema messo a punto e qui presentato consente di disporre di uno strumento decisamente importante per valutare la portata dei risultati presentati negli annuali rapporti ALMALAUREA. Ciò assume ancora maggiore rilievo considerando che in futuro la combinazione dei due strumenti di indagine, CAWI e CATI, sarà potenziato all'interno del Consorzio, al fine di poter raccogliere un'ampia mole di dati a costi relativamente contenuti.

In futuro si intende arricchire ulteriormente tale valutazione attraverso alcune tecniche che realizzano le condizioni di comparabilità fra unità sottoposte a diversi metodi di indagine utilizzando, in termini multivariati, tutta l'informazione disponibile nel set di variabili pre-trattamento. Tali tecniche (Camillo, D'Attoma 2009) consentono inoltre di generare un vero e proprio sistema semi-automatico di **impact evaluation**, visto che sfruttano, mediante un approccio di tipo *data mining* e interamente *data driven*, le proprietà geometriche degli spazi multivariati generati dall'interazione eventualmente esistente fra le variabili pre-trattamento.

Alcune analisi preliminari[11], comunque, risultano confortanti e confermano quanto qui esposto.

[11] Si tratta di un approfondimento presentato dagli autori al convegno ITACOSM09, First ITAlian COnference on Survey methodology, Siena, 10-12 giugno 2009.

Bibliografia

Camillo F., Girotti C. (2007) L'impatto dell'integrazione di tecniche multiple di rilevazione nell'indagine sulla condizione occupazionale AlmaLaurea: una misura di propensity score in spazi condizionati multivariati, in: *IX Profilo dei laureati italiani. La riforma allo specchio*, Consorzio Interuniversitario AlmaLaurea (a cura del), Il Mulino, 289-309.

Camillo F., D'Attoma I. (2009) A New Data Mining Approach to Estimate Causal Effects of Policy Interventions, Expert Systems with Applications doi:10.1016/j.eswa.2009.05.072

Cammelli A. (2009) *XI rapporto sulla condizione occupazionale dei laureati. Occupazione e occupabilità dei laureati a dieci anni dalla Dichiarazione di Bologna*, disponibile su www.almalaurea.it/universita/occupazione/occupazione07.

CISIA-CERESTA (2001) Manuale di SPAD. Versione 4.5, Parigi.
Lee S. (2006) Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *Journal of Official Statistics*, Vol. 22, No 2, 329-349.

Rosenbaum, P.R. and Rubin, D.B. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects*. Biometrika*, 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984) Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, *Journal of the American Statistical Association*, 79, 516-524.

Schizzerotto A., a cura di, (2002) *Vite ineguali. Disuguaglianze e corsi di vita nell'Italia contemporanea*, Bologna, Il Mulino.