LA DEONTOLOGIA DI CHI PRODUCE E DETIENE DATI STATISTICI: DALLA POSSIBILITÀ ALLA CERTEZZA DELL'ACCESSO

T. Boeri, M. Pellizzari

1. INTRODUZIONE

Con l'approvazione del "*Codice di deontologia e buona condotta per il trattamento di dati personali per scopi statistici e scientifici*" si conclude il lungo iter che ha portato alla modifica della legge 675/1996 in direzione di un più facile accesso ai dati statistici per scopi scientifici.

Nonostante molte delle perplessità espresse da Andrea Ichino nel suo intervento su questa rivista siano condivisibili, riconosciamo il grande passo avanti che il Codice rappresenta per la definizione di un quadro normativo che consenta, anche nel nostro paese, procedure di accesso ai dati statistici compatibili con le esigenze della ricerca scientifica. Non intendiamo quindi soffermarci in questo breve saggio sulle questioni ampiamente dibattute su questa rivista, ma anche in altre sedi[1], da Andrea Ichino e Ugo Trivellato riguardo ai meriti e ai difetti della nuova normativa. Vorremmo, invece, segnalare che, accanto ad un quadro normativo che consenta ai ricercatori di accedere ai dati, è necessario che presso le istituzioni che raccolgono e/o detengono questi dati si sviluppi una prassi, ben diversa da quella corrente, di apertura verso il mondo della ricerca e di facilitazione dell'accesso ai dati.

Per questo proponiamo l'introduzione dell'obbligo – almeno per gli enti pubblici – di condividere con i ricercatori accreditati – almeno quelli di ruolo presso le università – i dati statistici che questi richiedono per svolgere lavori di ricerca senza scopi di lucro.

In Italia, anche prima dell'adozione del Codice, molti dati sarebbero potuti essere utilizzati, molte e interessati ricerche avrebbero potuto essere condotte e tuttavia, spesso a causa delle resistenze degli enti che detenevano queste statistiche, ciò non è stato possibile. Oggi, la nuova normativa ci permette di fare molte più cose, di accedere a molte più basi di dati, più complete e di migliore qualità. Tuttavia, qualora la prassi di gestione di questi dati da parte delle istituzioni che li detengono rimanesse quella attuale, i vantaggi per la comunità scientifica sarebbero limitati.

---

[1] Si veda l'interessante e costruttivo dibattito su www.lavoce.info

In altre parole, crediamo che il fatto che in Italia non esista un sito come quello del *UK Data Archive* inglese (www.data-archive.ac.uk) o della *Current Population Survey* americana (www.bls.census.gov/cps) sia dovuto solo in parte alla normativa, ma anche e soprattutto all'atteggiamento di molte istituzioni, spesso riluttanti, se non apertamente ostili, a investire nella diffusione dei dati elementari per la ricerca.

L'atteggiamento dei possessori dei dati nei confronti dei ricercatori è tanto più importante in quanto la nuova normativa semplicemente consente l'accesso ai dati, ma non garantisce che questi vengano effettivamente messi a disposizione dei ricercatori. In altre parole, oggi il nuovo codice offre ai ricercatori un'autorizzazione legale a utilizzare molte basi di dati, ma nulla vieta a chi detiene questi dataset di negarne l'accesso. Si tratta solo di un'autorizzazione virtuale.

Inoltre, come ha ricordato Ichino nel suo intervento su questa rivista, il fatto che la definizione di dato personale sia affidata ad un criterio largamente soggettivo (*identificabili con mezzi ragionevoli*), lascia ampio spazio a chi, e crediamo siano molti, non è intenzionato a impiegare, anche modeste risorse, per rendere la basi di dati esistenti più facilmente accessibili alla comunità scientifica. Quest'ultima spesso non ha proprie rappresentanze nelle amministrazioni che producono i dati, le quali rischiano perciò di percepire solo i costi (modesti) associati alla loro diffusione e non i (grandi) benefici derivanti dalle analisi e valutazioni di politiche che ne potrebbero scaturire.

Con questo breve contributo intendiamo stimolare un mutamento delle prassi che troppo spesso impediscono ai ricercatori di accedere a dati statistici fondamentali per la valutazione delle politiche pubbliche. Siamo certi che, come noi, anche molti dei lettori di questa rivista abbiano avuto in passato gli stessi problemi che denunciamo in questo saggio. Quanti progetti di ricerca sono rimasti nel cassetto perché, pur esistendo i dati necessari per portarli a termine, non è stato possibile accedervi? E quanti mesi di lavoro abbiamo speso nel convincere questo o quell'istituto, questo o quell'ente a condividere con noi informazioni indispensabili per approfondimenti e studi di grande rilevanza, spesso anche per le attività dello stesso istituto?

Inoltre, molte amministrazioni pubbliche (es. ISTAT, INPS, ministeri, comuni, centri per l'impiego, etc.) già provvedono per proprie esigenze operative a raccogliere e organizzare dati individuali. In questi casi, i costi che si sosterrebbero per distribuire queste informazioni sarebbero davvero minimi. Tuttavia, anche l'accesso ai dati amministrativi è limitatissimo in Italia.

Vogliamo, in un certo senso, soffiare sul fuoco dell'indignazione che, come comunità scientifica, ci ha portato a chiedere a gran voce la modifica della legge 675/96 e, attraverso la denuncia collettiva dei numerosi "muri di gomma" che ognuno di noi ha incontrato nel corso del proprio lavoro, provare a modificare la prassi di gestione dei dati statistici. Ma vorremmo spingerci oltre l'indignazione. La cultura della ricerca scientifica 3 nel nostro paese, crediamo, riuscirà ad affermarsi solo quando vi sarà una massa critica di studi che ne dimostrerà la grande rilevanza pratica e l'utilità nell'evitare sprechi di risorse. Affinché si raggiunga questa massa critica di studi, è necessario che molti più dati siano resi accessibili e richiamino l'attenzione degli studiosi.

Di qui la necessità di introdurre *l'obbligo per le amministrazioni pubbliche che detengono dati utilizzabili per fini di ricerca scientifica di fornirli ai ricercatori che li richiedano per svolgere analisi senza scopo di lucro*. E perché non si tratti di uno dei tanti obblighi virtuali, occorrerebbe anche prevedere l'imposizione di "sanzioni" agli enti che non mettano a disposizione dei ricercatori dei "file standard", con dati anonimi ma individuali, in tempi ragionevoli. Occorrerebbe che questi file standard fossero accessibili a costi di produzione, pienamente documentati e distribuiti sull'intera platea dei ricercatori che ne fanno richiesta. Per i molti dati che le amministrazioni pubbliche già raccolgono nell'esercizio delle loro funzioni il costo di formattare queste informazioni già esistenti è molto contenuto, e il costo marginale di metterle a disposizione di un altro ricercatore è pressoché nullo.

Sarà necessario definire molti altri dettagli ma in questa sede vorremmo solo creare il consenso sul principio secondo cui un ente pubblico che detiene dati statistici è obbligato a condividerli, a costi di produzione, con il ricercatore che ne faccia richiesta per lavori di ricerca scientifica senza scopo di lucro.

Non vorremmo con questo fare un processo sommario alle molte istituzioni che nel nostro paese detengono dati statistici e che spesso non fanno nulla nel facilitarne l'accesso. È un problema di cultura, di diffidenza verso la ricerca scientifica, che si estende ben al di là delle amministrazioni direttamente coinvolte e dei casi che documentiamo in questo saggio. È la stessa cultura che ha animato coloro che si sono battuti contro la sperimentazione e valutazione del cosiddetto "metodo Di Bella", o di quei politici e amministratori che hanno introdotto misure sperimentali come il "Reddito Minimo d'Inserimento" senza preoccuparsi di raccogliere i dati elementari necessari a valutare la sperimentazione. Vorremmo allora che si cogliesse l'occasione offerta dall'approvazione del codice deontologico anche per proporre una *deontologia della produzione e disseminazione dei dati*, in grado di permettere agli enti coinvolti di soddisfare quegli obblighi di informazione che, peraltro sono spesso già contenuti nelle leggi istitutive di molte amministrazioni.

L'obbligo di produzione di file standard potrebbe rivelarsi, tra l'altro, un utile strumento per queste amministrazioni per iscrivere a bilancio capitoli di spesa finalizzati alla distribuzione dei dati (come, per esempio, la costruzione di siti internet simili a quelli citati in precedenza[2]). Inoltre, tale obbligo, qualora fosse esteso anche a tutti i produttori di dati, anche quelli non ufficiali, permetterebbe alla statistica ufficiale di confrontarsi apertamente sul piano della qualità e della affidabilità dei numeri. Si pensi, per esempio, al recente conflitto tra Istat e Eurispes sulla misurazione dell'inflazione[3]. Se i ricercatori italiani avessero tutti avuto accesso ai micro dati di prezzi e consumi utilizzati per produrre le stime dell'inflazione, ISTAT avrebbe probabilmente trovato più facilmente un maggiore sostegno da parte della comunità scientifica in quel dibattito.

La disponibilità di microdati da parte di tutte le fonti non può che stimolare e premiare chi produce i dati più affidabili e di migliore qualità.

---

[2] Da non confondere con siti finalizzati alla distribuzione di dati aggregati, poco utili alla ricerca scientifica. Si veda il dibattito tra Pirrone *et al.* (2004) e Pellizzari (2004) sul sito www.lavoce.info.

[3] Si veda il comunicato ISTAT del 30 dicembre 2002 e il dibattito sul sito www.lavoce.info "Statistiche, errori e speculazioni".

Nelle pagine che seguono descriviamo alcuni episodi in cui abbiamo dovuto rinunciare a progetti di ricerca a nostro avviso interessanti perché non ci è stato consentito l'accesso ai dati necessari. Nella maggior parte dei casi, si vedrà, si trattava di dati che avremmo potuto consultare anche in base alla normativa precedente al codice deontologico. Dunque non si tratta di un problema di norme, ma di applicazione delle stesse. In tutti i casi l'esistenza di un obbligo come quello che proponiamo avrebbe fornito, a noi e ai nostri interlocutori all'interno degli enti produttori dei dati, un potente strumento per permettere la produzione di file standard e l'accesso ai dati.

2. IL CASO DELLA SPERIMENTAZIONE DEL REDDITO MINIMO D'INSERIMENTO

Cominciamo da una vicenda emblematica, una clamorosa occasione mancata per la ricerca sociale e economica nel nostro paese, quella della cosiddetta sperimentazione del Reddito Minimo di Inserimento (RMI).

Come noto ai lettori di questa rivista, una delle tecniche più accreditate per la valutazione delle politiche sociali consiste nel disegnare esperimenti naturali in cui ad un gruppo casualmente selezionato si applica la politica da sperimentare, il cui effetto viene quindi identificato confrontando gli esiti di questo gruppo sperimentale con quelli di un "gruppo di controllo", composto da individui in tutto simili a quelli che hanno preso parte alla sperimentazione, ma ai quali non è stata applicata la nuova politica.

Esperimenti di questo tipo sono molto diffusi come strumento per lo sviluppo delle politiche attive del lavoro nei paesi anglosassoni e del nord Europa. Un famoso esempio è quello dei premi al reimpiego nel design dei sussidi di disoccupazione negli Stati Uniti. L'esperimento, in questo caso, consisteva nel selezionare casualmente un gruppo di beneficiari del sussidio e proporre loro un consistente premio monetario se fossero riusciti a trovare un impiego nell'arco di qualche settimana. Gli esiti occupazionali dei beneficiari del premio, sia di breve che di lungo termine, sono stati quindi valutati confrontandoli con quelli di un gruppo di controllo.[4].

In Europa continentale, e in Italia in particolare, esperimenti sociali di questo tipo sono sempre stati difficili da attuare per le resistenze dei molti che li considerano discriminatori: perché due persone simili dovrebbero ricevere, anche se per un periodo limitato, sussidi diversi solo perché sono stati selezionati in gruppi diversi? Queste perplessità sono, in una certa misura, comprensibili. Ma ciò che ha impedito l'effettiva sperimentazione del RMI non è tanto il rifiuto morale di attuare asimmetrie di trattamento quanto il fatto che, una volta presa la decisione di accettare l'"ingiustizia" di attivare l'RMI solo in un numero ristretto di Comuni, non si siano poi usati i dati sugli individui coinvolti per valutare l'esperimento a beneficio di tutti. Esperimenti di questo tipo servono proprio per meglio capire gli effetti di una riforma, per migliorarla e per poi introdurla per tutti nella miglior forma possibile.

[4] Per una rassegna sull'argomento si veda Meyer (1995).

Occorre, inoltre, notare che anche le politiche e le misure che, a norma di legge, non dovrebbero creare disparità di trattamento fra i beneficiari, finiscono inevitabilmente per creare delle asimmetrie. A seconda della dislocazione dell'amministrazione che eroga il sussidio, dei costi (in termini di tempo) che i diversi potenziali beneficiari devono sostenere per accedere al programma, dell'informazione disponibile circa il programma stesso, si creano delle asimmetrie fra la platea dei beneficiari, asimmetrie che riducono spesso in modo molto forte il grado di copertura di queste misure. La letteratura sul cosiddetto "take-up" delle politiche sociali documenta proprio queste asimmetrie "involontarie", che spesso svantaggiano i gruppi più deboli e bisognosi di aiuto, ad esempio perché poco informati sulla nuova normativa[5]. Nel caso degli esperimenti naturali, le asimmetrie sono limitate nel corso del tempo (quello necessario per la sperimentazione), sono del tutto casuali e sono finalizzate al miglioramento di un intervento pubblico a beneficio di tutti.

La sperimentazione del Reddito Minimo d'Inserimento cominciò nel 1998 in 39 comuni (peraltro selezionati in modo assolutamente non casuale). Questo nuovo strumento di protezione sociale prevedeva il pagamento di un sussidio a tutte le famiglie il cui reddito fosse inferiore ad una soglia di povertà definita con criteri oggettivi. Il sussidio era calcolato in modo da riportare la famiglia alla soglia di povertà ed era condizionale alla partecipazione a programmi di reinserimento, lavorativo o sociale a seconda dei casi. Il RMI rappresentava un programma importante nel panorama dello stato sociale italiano perché si configurava come l'unico sussidio universale, ovvero pagato a tutte le famiglie in condizione di povertà, indipendentemente dalla condizione lavorativa, dalla composizione famigliare e dall'età. Si trattava, inoltre, del primo esperimento sociale mai condotto in Italia. La sperimentazione doveva durare inizialmente fino al 2000 e fu poi estesa (allargando, ancora una volta in modo non casuale, il numero dei Comuni coinvolti) fino al 2002.

I comuni interessati erano tenuti ad inviare alla Commissione d'Indagine sull'Esclusione Sociale, l'organo promotore e coordinatore della sperimentazione del RMI, i dati (in parte in forma aggregata) relativi ai beneficiari e ai loro esiti reddituali e occupazionali. Su questi dati fu preparata una relazione di valutazione. La conclusione dell'esperimento e la pubblicazione della relazione, purtroppo, coincisero con il cambio di governo. Il nuovo governo si mostrò subito poco propenso a proseguire l'esperimento e proibì alla commissione di rendere pubblica la relazione sulla valutazione del RMI (che non fu nemmeno presentata al Parlamento!).

Provammo allora a chiedere alla Commissione di avere accesso ai dati individuali sui beneficiari per provare noi a condurre una ricerca che valutasse l'esperimento. Ritenevamo utile investire risorse intellettuali in questo sforzo perché gli esiti occupazionali e reddituali dei beneficiari ci sarebbero stati di grande utilità nel capire i disincentivi all'offerta di lavoro associati all'introduzione di un reddito minimo garantito in Italia. Molte altre questioni interessanti si sarebbero potute

[5] Si veda Hernanz *et al.* (2004).

investigare con questi dati: il grado di targeting delle politiche, il peso del settore informale nel condizionare l'allocazione del sussidio, etc.

La risposta fu che presso la poi disciolta Commissione d'Indagine sull'Esclusione Sociale erano stati raccolti solo i dati semi-aggregati ricevuti dai comuni e che, anche per dati in quella forma aggregata, avevano ricevuto dal Governo il divieto di distribuzione. Tuttavia, ci fu fatto presente che, nonostante questo divieto, la proprietà dei dati restava dei singoli Comuni e che la Commissione non avrebbe imposto alcun vincolo a questi ultimi circa la distribuzione di dati anonimi sui beneficiari del RMI.

Da allora ci siamo messi in contatto con molti dei 39 comuni che hanno partecipato all'esperimento, in alcuni casi ricevendo grande attenzione e supporto, in altri incontrando il "muro di gomma". Ad oggi, con molta fatica e grazie al lavoro di molte persone, sia all'interno dei Comuni che dell'università, siamo riusciti ad ottenere i dati individuali sui beneficiari del RMI nei comuni di Foggia, Napoli, Genova e Rovigo[6]. Al momento stiamo ancora lavorando su queste informazioni, ma è già evidente che si tratta di dati che avrebbero permesso una valutazione più oggettiva e circostanziata di quella sommaria che portò all'abbandono del progetto RMI (oggi peraltro sostituito sulla carta da un programma gemello, il Reddito di Ultima Istanza).

Durante il difficile processo di raccolta di questi dati, uno dei problemi più spinosi che anche i Comuni più disponibili hanno dovuto affrontare riguarda l'impegno del personale dell'amministrazione comunale nella sistemazione, raccolta e distribuzione dei dati. Com'è facile immaginare, i dati, tipicamente tenuti presso gli assessorati ai servizi sociali, non erano stati raccolti in modo facilmente trasferibile. Le informazioni anagrafiche erano spesso su supporto informatico, mentre quelle relative alle successive assegnazioni ai programmi di inserimento, alle sospensioni, alle esclusioni, etc. erano solitamente contenute in forma cartacea nelle personali cartelle di lavoro dei singoli assistenti sociali. Si rendeva, quindi, necessario un certo lavoro di riorganizzazione delle informazioni, che poteva essere svolto quasi esclusivamente presso l'amministrazione stessa e con la collaborazione di qualcuno dei dipendenti del Comune. Le resistenze cominciavano a questo punto: anche quando l'assessore si mostrava disponibile alla distribuzione dei dati, la riorganizzazione delle informazioni toglieva tempo prezioso che gli impiegati avrebbero dovuto dedicare ad altre mansioni. Infatti, la gestione delle informazioni statistiche per fini di ricerca non rientra nelle mansioni degli impiegati comunali.

Se anche i Comuni fossero soggetti all'obbligo di fornire i dati statistici in loro possesso per fini di ricerca, sarebbe forse più semplice ottenere la collaborazione fattiva delle amministrazioni e, soprattutto, prevedere già nella fase di raccolta del dato individuale il suo caricamento su supporti informatici, che ne facilitino poi l'elaborazione a fini statistici.

Tornando alla nostra esperienza, nei comuni di Foggia, Napoli, Genova e Ro-

[6] I dati sono disponibili, previa autorizzazione del Comune interessato, a chiunque li volesse utilizzare sul sito www.frdb.org

vigo è stata la disponibilità degli amministratori, degli assistenti sociali e degli assessori, insieme all'impegno di alcuni bravissimi e volenterosi studenti a permetterci di ottenere i dati. Ci piacerebbe che per raggiungere questo risultato non dovessimo dipendere dalla cortesia – peraltro molto apprezzata - di queste persone. In molti altri casi, infatti, non siamo stati così fortunati. La selezione dei Comuni che rientreranno nel nostro lavoro sul RMI sulla base della disponibilità e cortesia degli amministratori non sarà certo casuale!

Ci rendiamo perfettamente conto che l'organizzazione delle informazioni su file standard a disposizione dei ricercatori può essere dispendiosa, sia in termini di tempo che monetari, e che, quindi tale obbligo dovrà prevedere norme che regolino la distribuzione di questi oneri tra ente e ricercatore. È peraltro possibile che siano i ricercatori stessi – come avvenuto nel nostro caso – a farsi carico dei costi di organizzazione dei dati. Si tratta di costi relativamente contenuti, soprattutto se i dati vengono già raccolti dalle amministrazioni in forma trasferibile (cosa utile anche per elaborazioni al loro interno).

Se tutto venisse fatto in modo trasparente, su siti internet, sarebbe peraltro possibile distribuire questi costi su di una platea relativamente estesa di ricercatori, riducendo i costi per ciascuno. I costi sono, in ogni caso, troppo modesti per costituire il problema vero. Il fatto è che non viene rispettato il principio secondo cui un ente pubblico che detiene dati utili per la ricerca scientifica è tenuto a trasferirli in forma anonima al ricercatore che ne faccia richiesta per svolgere analisi senza scopo di lucro.

Si noti, inoltre, che grazie alla nuova normativa, i dati sui beneficiari del RMI potrebbero facilmente essere classificati come anonimi. Una volta oscurati nomi e cognomi, nonostante l'ampiezza campionaria sia limitata, identificare le persone presenti nell'archivio richiederebbe, a nostro avviso, costi molto elevati, mezzi quindi "non ragionevoli" nella terminologia del codice deontologico.

3. I DATI RISERVATI DELLE RILEVAZIONI TRIMESTRALI SULLE FORZE DI LAVORO (RTFL) DELL'ISTAT

È noto che l'indagine sulle forze di lavoro ISTAT è costruita utilizzando un campione longitudinale a rotazione: una volta entrata nel campione ogni famiglia vi rimane per due trimestri consecutivi, ne esce per altri due e vi rientra a distanza di un anno dalla prima intervista per altri due trimestri consecutivi. Questa struttura 8 longitudinale del campione non era resa pubblica fino a pochissimi mesi fa. Oggi è possibile ottenere da ISTAT i file standard longitudinali, ma solo relativi ad intervalli temporali di 12 mesi (ovvero informazioni relative agli individui intervistati ad aprile di ogni anno).

Questa è un'ottima notizia. Per quanto sappiamo, si tratta al momento del primo effetto concreto della nuova normativa sulla privacy. Non dovrebbero quindi sussistere problemi alla distribuzione anche delle informazioni longitudinali tra trimestri. Speriamo che ISTAT provveda al più presto a fornire file standard anche per questi dati.

Non crediamo sia necessario descrivere quali e quante ricerche si potrebbero condurre, quali e quanti indicatori si potrebbero costruire conoscendo la struttura longitudinale completa delle RTFL. Solo a titolo di esempio, la misura dei flussi sul mercato del lavoro – quante persone trovano un lavoro, quante lo lasciano, quante entrano sul mercato del lavoro, quante ne escono – è possibile solo con questi dati e ad oggi solo chi ne ha accesso può produrre queste statistiche.

La rilevazione dei flussi del mercato del lavoro è essenziale per analizzare gli effetti di molte politiche, soprattutto quelle che cambiano le regole "al margine", solo per i nuovi assunti, come il Pacchetto Treu o la Legge Biagi.

È importante notare però che se anche ci fosse, o ci fosse stato in passato, un problema di privacy le soluzioni adottate, per esempio, in Inghilterra ci sembrano molto più sensate del modo con cui l'accesso a questi dati è stato centellinato in Italia. Il problema di riservatezza nasce perché unendo due indagini successive, il vettore di caratteristiche associate ad un singolo individuo raddoppia, rendendone, in linea teorica, l'identificazione più semplice[7]. L'Office of National Statistics inglese risolve il problema distribuendo, accanto ai file contenenti tutti gli intervistati in ogni trimestre e privi di identificativo personale, anche i file longitudinali per ogni coppia di trimestri che possono essere associati longitudinalmente, di nuovo senza identificativi personali. Ogni record del file longitudinale contiene una serie di variabili relative al primo trimestre, rinominate con suffisso 1, e le stesse variabili relative al trimestre successivo, con suffisso 2. Il problema della privacy è risolto eliminando o ri-aggregando nel file longitudinale alcune variabili disponibili invece nei file originali. Paradossalemente un ricercatore italiano accede con più facilità ai dati sulle forze di lavoro inglesi che a quelle del suo paese (si provi a verificare sul sito www.data-archive.ac.uk).

In Italia si è seguito fino ad oggi un metodo molto più macchinoso, che speriamo non venga mantenuto per accedere ai file longitudinali trimestrali. L'ISTAT era disponibile a distribuire i dati previa presentazione e successiva approvazione di un progetto di ricerca. La ragione di questa procedura altamente discrezionale (chi valuta il progetto? Sulla base di quali parametri?) ci è oscura. Anche se non siamo al corrente di casi in cui una richiesta di questo tipo sia stata respinta, siamo consapevoli di innumerevoli casi in cui questa richiesta non è nemmeno stata fatta. Il processo che conduce alla nascita di un progetto di ricerca passa, infatti, anche attraverso l'esplorazione dei dati disponibili, a caccia di fenomeni nuovi, anomalie, particolarità da spiegare o che spieghino aspetti ancora oscuri o che, ancora, supportino teorie non ancora corroborate. Questo è ancora più vero per quanto riguarda informazioni elementari, quali le forze di lavoro e in particolare i flussi sul mercato del lavoro. Quanti progetti di ricerca sono rimasti nel cassetto perché non valeva la pena, per un'idea vaga, che magari si sarebbe rivelata sbagliata, imbarcarsi nel processo di richiesta dei file longitudinali? Tanti, a giudicare dalla mole di elaborazioni finite nel cestino quando l'idea poteva essere esplorata con dati liberamente e facilmente disponibili.

Un secondo aspetto importante dell'Indagine sulle Forze di Lavoro riguarda le

[7] Oggi ci chiediamo se questo sia possibile con "mezzi ragionevoli"!

variabili oscurate. Il questionario su cui si basa la rilevazione contiene, ad esempio, informazioni riguardo alla provincia di residenza, alla provincia di lavoro, alla cittadinanza, agli anni di residenza nella provincia. Queste variabili non sono distribuite nei file standard, per motivi o di privacy[8] o di bontà del dato. Seguendo lo stesso strano principio adottato per i file longitudinali, sottoponendo all'ISTAT un dettagliato progetto di ricerca si può chiedere di ottenere queste variabili. In questo caso siamo al corrente, anche per esperienza diretta, di richieste che sono state respinte dall'ISTAT.

Nel nostro caso particolare, ci sarebbe piaciuto lavorare su almeno due progetti che avrebbero richiesto alcune variabili oscurate. Uno riguardava la correlazione tra mobilità regionale e pendolarismo. Esiste in molti paesi e in molti studiosi il sospetto che il calo della mobilità inter-regionale registrato nei paesi europei negli ultimi decenni sia stato accompagnato da uno speculare aumento del pendolarismo. Questo sarebbe dovuto all'aumento dei prezzi delle case e degli affitti nelle zone a bassa disoccupazione e al contemporaneo miglioramento dei mezzi e delle infrastrutture di trasporto (automobili e strade). Ci sembrava un argomento importante perché riguarda sia la definizione dei confini di un mercato del lavoro locale che l'interazione tra politiche sociali di sostegno abitativo e politiche dei trasporti, due questioni che spesso vengono analizzate disgiuntamente. Naturalmente, per uno studio di questo tipo l'informazione relativa alla provincia di lavoro e di residenza, nonché agli anni di residenza, era cruciale.

Il secondo progetto riguardava, invece, il rapporto tra welfare locale e migrazione. Ci sarebbe piaciuto rispondere in modo convincente a molte domande sul cosiddetto *welfare shopping*. Si trattava di stabilire se le persone che si spostano da un'area all'altra del paese, o che arrivano nel nostro paese dall'estero, scelgano la loro destinazione in base, oltre che alle condizioni del mercato del lavoro, anche alla generosità e alla composizione del welfare locale. Per questo progetto, un volenteroso studente ha raccolto un ottimo dataset[9] contenente la spesa sociale comunale e regionale, divisa per funzioni, per un periodo di tempo sufficientemente lungo per permettere un'analisi interessante. Anche per questa ricerca ci sarebbero servite le variabili relative alla provincia[10] di residenza e di lavoro, e la cittadinanza. Nonostante la presentazione di un dettagliato progetto di ricerca che sembrava aver riscosso l'interesse dell'Istituto, la nostra richiesta ha avuto esito negativo. Ci è stato detto che quelle variabili non sono ritenute affidabili a causa delle procedure di riporto all'universo e che, quindi, l'istituto non poteva rischiare che con i suoi dati venissero prodotte stime distorte.

La qualità dei dati statistici è certamente un valore per noi ricercatori come e quanto per l'ISTAT, e tuttavia la bontà e affidabilità di un dato dipende anche, e in modo cruciale, dall'uso che se ne fa. In particolare, in questo caso non era necessariamente nostra intenzione produrre stime del numero di immigrati arrivati in ogni provincia in ogni anno. Il punto cruciale invece è che non crediamo che

---

[8] Questi dovrebbero tutti cadere alla luce della nuova normativa. Quindi ci aspettiamo l'inclusione di almeno alcune di queste variabili nei files standard in tempi brevissimi.

[9] Questo dataset sarà presto disponibile liberamente sul sito www.frdb.org

[10] Idealmente avremmo preferito il comune ma non ci abbiamo nemmeno provato.

l'ISTAT possa garantire la bontà di tutte le stime prodotte con i suoi dati. È certamente possibile produrre stime sbagliate anche con i dati contenuti nei file standard regolarmente distribuiti dall'ISTAT. Se incrociamo tutti i lavoratori residenti in una regione, che lavorano in un certo settore, che hanno un certo grado di istruzione, etc. sicuramente otterremo una cella di dimensioni troppo ridotte per essere utilizzata significativamente. Spetta però al ricercatore decidere se una stima si può fare o meno, se un sottocampione è troppo piccolo o troppo distorto. Spetta alla comunità scientifica, attraverso il complesso sistema di pubblicazioni, referaggi, seminari, reputazione, incentivare la produzione di stime corrette. Se invece il dato è irrimediabilmente sbagliato per un difetto della rilevazione, allora ha ragione l'ISTAT a non distribuirlo. Ma ci auguriamo che non sia questo il caso, almeno per l'informazione relativa alla provincia di residenza.

4. MICRODATI SULLE IMPRESE

In Italia non esistono microdati di impresa facilmente accessibili. Ciò non significa che non esistano fonti di dati individuali sulle imprese italiane, anzi, il censimento delle imprese ISTAT, i dati della Centrale dei Bilanci, l'Archivio Statistico delle Imprese Attive (ASIA), le inchieste mensili sulle imprese dell'ISAE, gli archivi amministrativi DM10 dell'INPS, sono solo alcune. Questi dati, tuttavia, non sono utilizzabili al di fuori delle istituzioni che li producono.

Solo all'interno di queste istituzioni è possibile analizzare, per esempio, quante imprese cessino le loro attività e quante ne vengano create ogni anno in Italia, in quali settori, con quale dimensione, con quali tecnologie, con che apertura internazionale, etc.

Crediamo che questa sia un'anomalia gravissima, che verrebbe superata dall'introduzione dell'obbligo di produzione, da parte delle amministrazioni interessate, di file standard accessibili dai ricercatori. L'ISAE, per esempio, che da anni produce indagini mensili sulle imprese, tra l'altro confrontabili con indagini simili svolte in altri paesi europei, dovrebbe essere vincolato a condividere questi dati con i ricercatori.

Nel caso dei dati d'impresa non sussiste nemmeno il problema della privacy. Non sono dati personali. Esiste, invece, secondo alcuni un problema di reputazione nei confronti delle imprese che rispondono alle indagini. Si argomenta che, se i dati raccolti in queste indagini fossero liberamente distribuiti, le imprese sarebbero scettiche sull'opportunità di partecipare ai sondaggi.

Troviamo questa argomentazione davvero poco convincente per due ordini di motivi. In primo luogo, con la sola esclusione delle imprese di grandissime dimensioni (in effetti in tutti queste banche dati la Fiat non può che essere facilmente identificabile), una volta oscurata la ragione sociale, i record non sarebbero identificabili con mezzi ragionevoli. In secondo luogo, nella grande maggioranza dei casi, le informazioni utili alla ricerca non sarebbero tali da suscitare la resistenza delle imprese. Normalmente i ricercatori sono interessati a conoscere il settore industriale, la composizione della forza lavoro (uomini, donne, laureati, etc.), la

localizzazione. Nulla di particolarmente sensibile o segreto. Esiste poi tutta una serie di informazioni di bilancio che sono pubbliche per legge.

5. LO STRANO CASO DEGLI ARCHIVI AMMINISTRATIVI DELL'INPS

Esiste in Italia una base di dati dalle enormi potenzialità, costruita con perizia e pazienza da un gruppo di ricercatori esperti, che permetterebbe di svolgere innumerevoli ricerche se fosse utilizzata intensivamente da tutta la comunità scientifica, in Italia e all'estero. Si tratta degli archivi amministrativi dell'INPS.

Il valore innovativo di questi dati, oggi ancora scarsamente utilizzati dalla comunità scientifica, consiste nella possibilità di associare alle informazioni individuali sui lavoratori quelle relative alle imprese presso le quali svolgono la propria attività. I dati si riferiscono, infatti, ad un campione 1:90 dell'universo dei lavoratori italiani iscritti all'INPS e seguiti nel tempo, dal 1985 al 2001 (sono previsti aggiornamenti periodici dei dati per gli anni più recenti). Ad ogni lavoratore vengono associate le informazioni presenti negli archivi dei datori di lavoro ed è quindi possibile, per esempio, sapere se una persona lascia a poi rientra presso la stessa impresa (un aspetto ritenuto molto rilevante nel valutare la durata della disoccupazione, vedi Katz *et al.*; 1990).

Dati con queste caratteristiche, detti anche *matched employer-employee data*, sono relativamente rari (esistono per la Francia e pochi altri paesi) e permetterebbero di condurre analisi molto importanti sull'andamento della spesa pensionistica, sugli effetti delle politiche di sostegno all'offerta di lavoro, etc. Se incrociati, con dati di impresa, permetterebbero di capire molte cose sulle determinanti dei differenziali salariali, la copertura della contrattazione collettiva, etc.

Solo pochissimi centri di ricerca[11] hanno ottenuto, tramite una convenzione ad hoc con l'INPS, la possibilità di utilizzare questa banca dati. Queste convenzioni impongono al beneficiario il divieto di distribuire i dati a terzi. L'interpretazione di questo vincolo è però ambigua. Alcuni dei centri che hanno accesso ai dati quindi ne consentono l'utilizzo presso le loro sedi (ISFOL e LaboRR).

Inoltre, il LaboRR ha in programma la distribuzione a titolo gratuito di un file standard. Tuttavia, comprensibili esigenze di finanziamento e le resistenze dell'INPS limitano notevolmente le informazioni che presumibilmente potranno essere incluse nei file standard.

A chi fa ricerca sarà evidente che nessuna di queste soluzioni garantisce la diffusione e l'uso intensivo che una banca dati come quella degli archivi INPS meriterebbe.

Per questo motivo, da anni stiamo tentando di avere accesso ai dati originari e solo dopo un lungo percorso di ripetute richieste, scambi epistolari, riunioni, etc., abbiamo finalmente ricevuto anche noi (come Università Bocconi) una proposta

---

[11] Al momento siamo a conoscenza di tre soli centri che hanno accesso ai dati INPS: il LaboRR di Torino, l'Isfol di Roma e il centro studi della Banca d'Italia. A breve anche l'Università Bocconi dovrebbe entrare in questo ristretto gruppo.

di convenzione. In accordo con una politica di più intenso utilizzo di questi dati da parte della comunità scientifica, l'INPS ha inoltre abbassato notevolmente i costi inizialmente richiesti per la fornitura di questi dati.

Ci auguriamo che i tanti sforzi fatti per ottenere questo importante risultato siano di beneficio per chi in futuro voglia utilizzare questi dati.

6. CONCLUSIONI

In questo articolo abbiamo documentato le molte resistenze, gli ostacoli, spesso insormontabili, che si trova di fronte chi vuole condurre analisi con microdati in Italia. Si tratta di ostacoli che il codice deontologico riesce solo a scalfire. Abbiamo proposto di introdurre l'obbligo per tutte le amministrazioni pubbliche che detengono dati utili per la ricerca scientifica di renderli accessibili nell'ambito di file standard anonimi. Questo servirebbe a stimolare una massa critica di studi che svilupperebbero anche nel nostro paese una cultura della ricerca scientifica.

Un obbligo di questo tipo servirebbe anche a proteggere le amministrazioni contro le forti pressioni politiche cui sono soggette, contribuendo indirettamente a rafforzare la credibilità delle loro rilevazioni. Già il caso della sperimentazione del RMI testimonia di come in alcuni casi le interferenze della politica impediscano di cedere i dati anche da parte di amministrazioni che sarebbero disponibili a farlo.

Un altro eclatante esempio viene proprio dall'INPS, ed è stato ampiamente denunciato sul sito www.lavoce.info. Circa un anno fa, durante l'estate del 2003, proprio nel mezzo del dibattito sulla riforma delle pensioni che sembrava allora in procinto di essere approvata, è stata casualmente resa pubblica una circolare interna del Commissario Straordinario dell'INPS datata 6 agosto 2003 che, su diretta richiesta del ministro del Welfare, proibiva a tutte le strutture dell'Istituto "…di fornire dati, stime o analisi sulle questioni o sui conti dell'Istituto" a chiunque. Fortunatamente la cosa fece il giusto clamore e l'INPS si affrettò a precisare che il divieto era finalizzato a non influenzare il dibattito sulle pensioni. Come se un dibattito non informato potesse essere più obiettivo, più costruttivo, più sereno!

Un obbligo come quello che proponiamo servirebbe anche a evitare queste dannosissime interferenze della politica nella distribuzione delle statistiche, a tutto vantaggio dell'autonomia e dall'autorevolezza delle amministrazioni pubbliche che le producono.

Università Bocconi TITO BOERI

Fondazione Rodolfo Debenedetti MICHELE PELLIZZARI

BIBLIOGRAFIA

V. HERNANZ, F. MALHERBET, M. PELLIZZARI (2004), *Take-up of welfare benefits in OECD countries: a review of the evidence*, OECD Social, Employment and Migration Working Papers n. 17.

A. ICHINO (2003), *Statistica e privacy*, 20 novembre 2003, www.lavoce.info

A. ICHINO (2003), *Le perplessità di un utilizzatore di dati di fronte al «Codice di deontologia e buona condotta per il trattamento di dati personali per scopi statistici e scientifici»*, "Statistica", 4, pp. 673-684.

A. ICHINO (2004), *Un codice non fa primavera*, 25 maggio 2004, www.lavoce.info

L. F. KATZ, B. D. MEYER, (1990) *Unemployment insurance, recall expectations and unemployment outcomes*, "Quarterly Journal of Economics", vol. 105(4), 973-1002

LAVOCE.INFO, AA.VV., (2003), *Statistiche, errori e speculazioni*, 27 febbraio 2003, www.lavoce.info

LAVOCE.INFO, REDAZIONE, (2003), *Trasparenza, previdenza e…influenza*, 4 settembre 2003, www.lavoce.info

B.D. MEYER (1991), *Lessons from the U.S. unemployment insurance experiments*, "Journal of Economic Literature", vol. 31(1), 91-131.

M. PELLIZZARI (2004), *Dati aggregati o dati individuali*, 21 ottobre 2004, www.lavoce.info

S. PIRRONE, P. SESTITO (2004), *Valutare in trasparenza*, 21 ottobre 2004, www.lavoce.info

U. TRIVELLATO (2003), *Protezione dei dati personali e ricerca scientifica*, "Statistica", 4, pp. 627-648.

U. TRIVELLATO (2004), *Un codice per tutelare privacy e ricerca*, 20 maggio 2004, www.lavoce.info

U. TRIVELLATO (2004), *Fare di ogni erba un fascio non aiuta*, 25 maggio 2004, www.lavoce.info

RIASSUNTO

La deontologia di chi produce e detiene dati statistici: dalla possibilità alla certezza dell'accesso

L'approvazione del "*Codice di deontologia e buona condotta per il trattamento di dati personali per scopi statistici e scientifici*" rappresenta un notevole miglioramento delle possibilità offerte ai ricercatori italiani di accedere a microdati per lavori scientifici. Tuttavia, la nuova normativa semplicemente autorizza il ricercatore ad accedere alle banche dati, ma non impone alle istituzioni che raccolgono o detengono i dati di garantirne l'accesso ai ricercatori. Per questo motivo, in questo articolo proponiamo l'introduzione dell'obbligo – almeno per gli enti pubblici – di condividere con i ricercatori accreditati – almeno quelli di ruolo presso le università – i dati statistici che questi richiedono per svolgere lavori di ricerca senza scopi di lucro. Nell'articolo documentiamo la nostra personale l'esperienza di rapporti spesso difficili con le amministrazioni che detengono dati statistici, testimonianza di un atteggiamento spesso riluttante, se non apertamente ostile, a investire nella diffusione dei dati elementari per la ricerca. L'obbligo che proponiamo consentirebbe inoltre agli enti produttori o detentori di dati di iscrivere a bilancio fondi (minimi) finalizzati alla distribuzione delle statistiche e di proteggersi dalle interferenze della politica.

SUMMARY

A code of conduct for data production and dissemination

The recently approved "*Codice di deontologia e buona condotta per il trattamento di dati personali per scopi statistici e scientifici*" greatly improves the Italian legislation on the privacy of personal data towards easier access to databases by researchers who intend to use them for non-commercial scientific purposes. However, the new legislation simply allows the researcher to access more and better data, it does not guarantee that the institutions that collect and posses these databases will share them with the scientists. To this end, we propose the introduction of a legal obligation – at least for public institutions – to share with researchers – at least those employed by universities – all statistical data that might be requested for non-commercial scientific studies. In this article, we document our own personal experience of requesting data from public administrations, showing how they are often reluctant, if not openly refuse, to invest resources in the dissemination of basic statistical information for research purposes. Not only would the legal obligation we propose allow these institutions to devote specific resources for the dissemination of statistical data, but it would also protect them from political pressure.