

RENATA CLERICI

ARCHIVI AMMINISTRATIVI E STUDIO DELLA POPOLAZIONE

Esperienze e prospettive di sfruttamento di archivi amministrativi e di indagini campionarie, di natura non strettamente demografica, per lo studio della popolazione^{*}

1. A partire dalla istituzione del SISTAN¹ si vanno diffondendo una nuova consapevolezza della ricchezza informativa inutilizzata, distribuita un po' in tutti i settori della Pubblica Amministrazione, e l'urgenza di un suo sfruttamento, innanzi tutto per il miglioramento dell'attività pubblica, ma anche al fine di raggiungere obiettivi più generali di conoscenza.

Nei più recenti programmi statistici nazionali compaiono vari progetti, formulati da diverse amministrazioni, volti a migliorare ed ampliare l'utilizzo a fini statistici degli archivi esistenti. Esaminando gli elenchi dei lavori per settore e per fonte di riferimento, se ciascun settore vede la presenza di più fonti di dati, non sembra invece essere perseguito pienamente il proposito, per altro verso continuamente rimarcato, della loro integrazione. Si prospetta, infatti, tutt'al più, la creazione di banche dati, in cui non verrebbero integrate le informazioni originali (o una loro selezione), ma sarebbero raggruppati gli indicatori ricavati dalle differenti rilevazioni dei diversi organismi.

L'auspicio secondo cui *"laddove esistano diversi registri relativi alle stesse unità di analisi, questi siano messi in grado di colloquiare, ad esempio inserendo, in ogni registro, le variabili che identificano le unità negli altri archivi, qualora siano diverse"* (SISTAN-ISTAT, 1994a)² pare

^{*} Testo rivisitato del lavoro presentato alle Giornate di Studio sulla Popolazione, Roma, 7-9 gennaio 1997, organizzato dal Gruppo di Coordinamento per la Demografia della S.I.S. Sessione "Fonti amministrative per lo studio della popolazione", organizzatore prof. Bernardo Colombo.

¹ Il d.l. 322/89 ha prodotto il riordino della statistica ufficiale in Italia, affiancando all'ISTAT, nell'ambito del SISTAN, una molteplicità di enti ed organismi pubblici. Le amministrazioni pubbliche hanno con ciò assunto un ruolo molto importante, orientando a fini scientifici i dati amministrativi.

² Con la promulgazione della l. 675/96 "Tutela delle persone e di altri soggetti rispetto al trattamento dei dati personali", la possibilità di far colloquiare archivi diversi attraverso l'introduzione di codici identificativi non sembrerebbe praticabile, a meno del consenso degli interessati. Se, infatti, la legge esclude la necessità del consenso nel caso in cui il trattamento sia finalizzato a scopi di ricerca scientifica o di statistica, sancisce anche la necessità dell'anonimato (art.12) ed il codice identificativo dovrebbe contrastare con questa esigenza. A

realizzabile solo a partire dalla creazione di una duplice consapevolezza:

- a) del ruolo informativo associato all'attività istituzionale, che deve portare le pubbliche amministrazioni a curare maggiormente i propri archivi, registri, elenchi, perché dal loro grado di aggiornamento, completezza ed affidabilità dipenderà la bontà dei dati da essi ricavabili³;
- b) dello sforzo metodologico degli statistici, che devono imparare a utilizzare fonti per loro inusuali, e predisporre metodi e tecniche di analisi specifiche per il nuovo tipo di dati con cui ci si trova ad operare. Si tratta di informazioni finalizzate alla gestione, al perseguimento di specifici fini istituzionali, e non primariamente alla conoscenza, che dunque abbisognano di preliminari verifiche, aggiustamenti, in seguito a valutazioni della loro adeguatezza, completezza, qualità e accuratezza (Filippucci, 1997).

D'altra parte, l'utilizzo di queste fonti tende a divenire indispensabile per il costo crescente delle indagini e la riduzione dei fondi destinati alla ricerca, e visto anche l'aumento generalizzato (osservato un po' ovunque) delle non risposte a rilevazioni progettate *ad hoc*.

In Italia la produzione scientifica di carattere demografico ha solitamente privilegiato l'utilizzo dei risultati forniti, oltre che, naturalmente, dal censimento della popolazione (limitatamente alle caratteristiche strutturali della popolazione e delle famiglie e con timide esperienze di sfruttamento per analisi indirette del movimento migratorio e naturale) e dalle tradizionali rilevazioni dello Stato Civile (per il movimento naturale e per lo studio della nuzialità) e dell'Anagrafe (per il movimento migratorio e ancora per analisi strutturali), dalle indagini campionarie, che permettono quegli approfondimenti che le rilevazioni totali non possono consentire. Anche da questo punto di vista emerge, però, una sottoutilizzazione del materiale disponibile, che viene sfruttato spesso solo per produzioni di *routine*. Oggi sembra che le difficoltà normative di accesso ai microdati siano in parte superate, ma il loro costo risulta proibitivo per il singolo ricercatore o per l'Istituto Universitario, mentre, da parte loro, i privati sono più interessati ad archivi integrati, che forniscono cioè indicatori di diversa

dell'anonimato (art.12) ed il codice identificativo dovrebbe contrastare con questa esigenza. A tale proposito il Governo è impegnato dalla l. 676/96 "Delega al Governo in materia di tutela delle persone e di altri soggetti rispetto al trattamento dei dati personali", ad emanare entro 18 mesi le disposizioni integrative per l'applicazione della l. 675/96, con specifico riferimento (art.1d) all'individuazione dei presupposti per l'attribuzione di un numero di identificazione personale, per il suo trattamento e per il trattamento delle informazioni ad esso connesse, per il collegamento con altri dati.

³ L'orientamento del Programma Statistico Nazionale è inoltre volto al trasferimento della rilevazione stessa dall'ISTAT alle singole amministrazioni, a partire dai Ministeri.

natura (ad esempio SEAT, 1995; Dicosoft, 1996).

Se, dunque, le stesse fonti di natura peculiarmente demografica, non sembrano adeguatamente utilizzate, la domanda informativa può trovare altresì nuove risposte nello sfruttamento di archivi, elenchi, registri tenuti da pubbliche amministrazioni, da agenzie o organizzazioni private, pur nel rispetto della normativa sulla tutela dei dati personali.

2. Vi sono Paesi con una lunga esperienza di sfruttamento di fonti amministrative a fini demografici, utilizzate per sopperire alla mancanza di specifiche rilevazioni periodiche capaci di "riempire" lo spazio conoscitivo tra due censimenti o, più recentemente, per migliorare la qualità e la quantità dell'informazione già disponibile.

Gli USA, che non hanno mai avuto un registro completo della popolazione, ma che dispongono di un'ampia gamma di registri parziali⁴, hanno imparato a far tesoro delle informazioni in essi contenute, derivandone stime della popolazione e delle forze di lavoro, tavole di mortalità per specifiche sottopopolazioni in base a cui ricavare stime di mortalità per la popolazione complessiva. In particolare si sono sfruttati a fini demografici i dati dei registri del sistema di sicurezza sociale, relativi a disoccupazione, pensionamento, malattia, assistenza pubblica e assegni familiari. Anche basi di dati di natura diversa sono state utilizzate per ricavare stime di popolazione: elenchi di contribuenti, indirizzari cittadini, elenchi di aderenti alle diverse chiese, elenchi telefonici, registrazioni di licenze edilizie e di demolizione, elenchi di abbonati ai servizi pubblici (luce, acqua, gas). Stime della composizione per sesso, età e razza sono state costruite a partire dalle registrazioni del servizio militare (Shryock, Siegel e ass., 1980). Per la stima intercensuaria della popolazione ultrasessantacinquenne sono state utilizzate le liste degli iscritti ai servizi sanitari; per la stima degli emigrati, i pagamenti assicurativi effettuati da individui che abitano all'estero; il saldo migratorio è stato ricavato da dati fiscali, iscrizioni scolastiche e licenze di guida. Negli anni Ottanta vari archivi sono stati utilizzati per sopperire alle sottoenumerazioni censuarie e stimare il livello di copertura del censimento. Un *post enumeration program* era stato approntato per correggere il dato censuario mediante *dual* o *triple-system estimates*, basate oltre che sulle risultanze censuarie, su due rilevazioni, la *Current Population Survey* e i *Tax returns* (Herriot, 1983).

La Danimarca ha basato il Censimento 1981 su dati amministrativi. Ciò

⁴ I registri parziali riguardano, per specifici scopi amministrativi, solo i soggetti direttamente interessati ad un determinato programma (ad es. liste elettorali, liste di leva, lavoratori o pensionati di una particolare categoria, ecc.). Alcuni di tali registri sono continuamente tenuti e aggiornati, altri sono occasionali e sono attivati per particolari contingenze.

è stato possibile a partire da una importante riforma (1968) che ha prodotto l'implementazione di un sistema elettronico di elaborazione e la creazione del CPR (*Central Population Register*), un registro nazionale della popolazione danese su supporto magnetico, e dall'introduzione del Numero Personale, necessario all'operatività del registro. L'introduzione dello stesso numero negli altri settori della Pubblica Amministrazione rende possibile il collegamento tra le fonti: alle informazioni riguardanti nascite e morti, ad esempio, sono associate le informazioni fornite dalle autorità sanitarie. Se il numero personale assicura il collegamento con gli altri registri di tipo individuale, l'indicazione dei legami familiari permette l'automatico collegamento tra i componenti della famiglia, inoltre la presenza dell'indirizzo della *household* permette il collegamento col Registro Centrale delle Abitazioni. Negli anni Cinquanta fu anche creato un registro degli affari da cui estrarre i campioni per i diversi fenomeni economici; negli anni Settanta fu stabilito un registro nazionale delle costruzioni e delle abitazioni. Attraverso i tre principali registri gli individui sono dunque collegati al loro lavoro e alla loro abitazione, oltre che agli altri componenti della famiglia; vi sono poi altri registri e *Systematic Data Modules* che contengono altre informazioni riferite a settori specifici. In particolare sono fonte di informazione statistica i dati raccolti dal sistema fiscale, da quello pensionistico e dal sistema degli assegni familiari. In Danimarca le informazioni sull'occupazione della popolazione economicamente attiva sono desunte dalle dichiarazioni fiscali annuali, ma vi sono problemi circa la qualità dei dati così raccolti (mancata trasmissione da parte delle autorità locali, dichiarazioni poco affidabili dell'occupazione da parte dei cittadini), che fa nascere l'esigenza di integrare le informazioni a partire da una fonte supplementare (Thygesen, 1983).

La Norvegia ha una lunga esperienza di uso di registri e nel corso del tempo ha realizzato un sistema coordinato di *records* amministrativi, e ciò anche grazie ad una specifica normativa per lo scambio di informazioni tra le diverse agenzie. Lo scopo primario di tale attività non era tanto il miglioramento della produzione statistica, ma piuttosto l'aumento dell'efficienza dell'amministrazione. *Statistics Norway* ha giocato un ruolo primario nell'opera di coordinamento, facendo sì che le registrazioni amministrative diventassero anche un importante tassello nella produzione dei dati statistici e impegnandosi a garantire il controllo della loro qualità. L'interesse attuale è quello di realizzare combinazioni di informazioni tra registri e indagini: la soluzione ottimale rimane quella del collegamento a livello di singola unità statistica, ma utili combinazioni possono venir realizzate anche a livelli superiori di aggregazione. Gli studi più recenti si concentrano proprio sulla presentazione e sulla comparazione dei risultati

forniti da metodi e tecniche volti a risolvere il problema del collegamento tra le diverse fonti. Nel lavoro presentato alla Conferenza di Reykjavik, Thomsen e Kleive Holmoy (1996) hanno illustrato alcuni metodi di combinazione (ancora a livello individuale) tra informazioni tratte da registri e da indagini. L'esperienza norvegese mostra che l'uso di fonti supplementari incrementa comunque sempre l'efficienza delle stime, qualunque sia il loro metodo di costruzione: per la Norvegia ciò è attribuibile alla buona qualità dei registri. I progetti per il futuro prevedono in questo Paese la combinazione di informazioni tra fonti le cui unità differiscano dall'una all'altra. In questo caso, mentre alcune informazioni potranno essere associate a livello individuale, altre dovranno venir combinate a livelli superiori.

La Svezia, che ha in gran parte sostituito la rilevazione censuaria con i registri, fonda la sua conoscenza del mercato del lavoro sull'integrazione di tre registri (della popolazione totale, dei redditi, delle imprese): questo sistema integrato, denominato SORE (*Statistics on Regional Employment*) può a sua volta trovar beneficio dal collegamento con un'altra rilevazione svedese sul mercato del lavoro, questa volta di tipo campionario (*Labour Force Survey*, LFS). La diversa natura delle informazioni raccolte (tipicamente strutturale quelle del SORE, più attenta alle caratteristiche qualitative del mercato del lavoro la LFS) porta a ritenere particolarmente vantaggiosa la combinazione delle due fonti. Tra i benefici che si possono trarre dall'uso congiunto, non va poi trascurato l'aspetto relativo al reciproco controllo che si attiva tra le fonti utilizzate (Carlsson, 1996).

Nel 1979 *Statistics Canada* avviò una ricerca sulla possibilità di sviluppare statistiche sociali su piccole aree da dati amministrativi che fossero comparabili con quelle desumibili dai censimenti e da indagini campionarie sulle famiglie, concernenti i temi della mobilità territoriale, del reddito, della composizione familiare e delle forze di lavoro. A questo fine sono state utilizzate in modo congiunto quattro fonti di cui una principale, il *Personal Income Tax File* che copre circa il 90% della popolazione, e tre secondarie che si riferiscono a settori di popolazione in parte esclusi dal file principale. L'*Health and Welfare Old Security Files*, che riguarda programmi di assistenza per la popolazione sopra i 64 anni; il *Family Allowance File* che interessa tutti i minori di 15 anni e il 95% dei ragazzi sotto i 18 anni e si basa sui pagamenti di assegni familiari ai genitori; l'*Unemployment Insurance Files*, che completa l'informazione per i titolari di redditi bassi che sono esclusi dal pagamento delle tasse e dunque dal file principale. Il collegamento tra le quattro fonti è assicurato dalla comune presenza di un codice identificativo individuale, il *Social Insurance Number (SIN)*: ciò permette di integrare le fonti ed ampliare la copertura

dell'universo, oltre che rendere possibile l'estrazione di un campione longitudinale (del 10%). Il programma, che si prefiggeva analisi di popolazione su aree di ridotte dimensioni, ha portato alla creazione di una *Divisione per lo studio delle piccole aree e dei dati amministrativi (SAADD)* che ha posto particolare attenzione alla messa a punto della metodologia di stima dei movimenti migratori annuali tra *Census Divisions* a partire dai dati di fonte fiscale, in assenza di una fonte demografica diretta per la rilevazione di tali eventi. Si tratta di una procedura consolidata (l'esperienza è ormai ventennale) della quale pare opportuno dare conto, sia pure in maniera sintetica, nella scheda riportata in Appendice.

Per le statistiche francesi il problema è quello di sostituire l'assenza della registrazione continua, assicurata altrove dal registro della popolazione, con stime intercensuarie annuali. Una metodologia messa a punto dall'INSEE permette di usare l'insieme delle fonti esistenti per reciproche correzioni e per produrre stime di popolazioni locali. In particolare in Francia non viene rilevato il movimento migratorio, per cui si rende necessaria la costruzione di informazioni circa il movimento della popolazione a livelli territoriali inferiori a quello nazionale. I registri amministrativi sono di ausilio per la ripartizione del movimento complessivo tra le diverse aree geografiche. La metodologia fin qui adottata produceva le stime richieste attraverso un procedimento a tappe, e per ognuna di esse veniva utilizzata una fonte diversa. Dal 1996 si è invece passati al metodo *multi-sources* in quanto una stima basata sulla considerazione congiunta di più fonti risente in misura più ridotta delle distorsioni prodotte dai difetti, dalle disomogeneità territoriali, dalla variabilità temporale, dagli effetti delle modificazioni normative, ecc. L'utilizzo di più fonti implica dapprima un procedimento di comparazione e, successivamente, una sintesi dei diversi indicatori. Queste operazioni non sono immediate: ad esempio per la stima di una popolazione locale il numero degli scolari non è direttamente comparabile con quello delle abitazioni o con quello degli elettori. Alcune fonti forniscono indicatori diretti del fenomeno che si intende misurare: ad esempio il saldo migratorio può essere direttamente ricavato da una sottopopolazione (ad es. quella degli scolari), conoscendo la proporzione che la lega al gruppo complessivo (popolazione complessiva di una certa area). Altre fonti (ad es. l'elenco degli abbonati alla rete elettrica) non forniscono direttamente stime della popolazione, ma occorre applicare dei correttivi (riferiti ad esempio all'ampiezza delle famiglie) (Jacod, 1996).

Il problema fondamentale del metodo *multi-sources* sta nella sintesi degli indicatori ricavati dalle diverse fonti. L'esperienza francese indica che la media semplice risente troppo dei valori estremi, la cui presenza porta a prediligere la mediana; visto però che la qualità dell'informazione fornita

dalle diverse fonti è variabile, sembra preferibile una ponderazione dei vari indicatori, con pesi desunti dall'analisi dell'efficienza delle diverse stime.

Statistics Norway ha sperimentato la sintesi delle informazioni provenienti da più fonti, sia attraverso il metodo della ponderazione, sia applicando un metodo multivariato di analisi della regressione (*calibration*), in cui il set di variabili indipendenti è formato da indicatori (tra loro poco correlati) ricavati da registri e riferiti alla popolazione complessiva, mentre le variabili dipendenti sono osservate solo parzialmente (nel caso specifico provengono da un'indagine campionaria).

3. Queste metodologie appaiono di grande interesse e se ne potrebbe prospettare l'applicazione, per il nostro Paese, per il collegamento tra registro della popolazione e indagini campionarie quali la Multiscopo, l'Indagine sui Consumi delle famiglie e quella sulle Forze di Lavoro: l'obiettivo potrebbe essere quello di ottenere stime, a livello regionale e provinciale, delle variabili, non presenti nel registro, che le indagini forniscono solo per grandi ripartizioni territoriali.

In Italia l'interesse per l'integrazione delle fonti o comunque per lo sfruttamento di fonti "non usuali" comincia del resto a farsi strada. Righi e Terra Abrami (1993) hanno recentemente svolto interessanti riflessioni circa la possibilità di uno "sfruttamento integrato delle indagini e delle rilevazioni ufficiali" (anche se sempre di natura strettamente demografica), in vista della costituzione, nell'ambito del SISTAN di un *Sistema Informativo Demografico*: auspicano il collegamento tra censimento, anagrafe e stato civile mediante l'introduzione nelle schede di nascita, di morte e di migrazione dello stesso codice identificativo del censimento. Una esperienza in questa direzione è stata quella condotta in tema di mortalità differenziale, attraverso il collegamento tra schede di morte e dati censuari: il progetto originario prevedeva un approccio longitudinale che né l'indagine pilota, né la seconda indagine, hanno però potuto realizzare. Altri ricercatori dell'ISTAT hanno iniziato a valutare le metodologie necessarie all'implementazione di collegamenti tra le fonti, orientandosi principalmente verso il *matching* esatto piuttosto che su quello statistico⁵. Analisi e sfruttamento di fonti non tradizionali sono state effettuati per fenomeni per cui le fonti usuali si mostrano più deficitarie, primo tra tutti lo studio dell'immigrazione straniera per il quale il ricorso a basi di dati di natura amministrativa (tenute dai vari Ministeri competenti nelle diverse materie)

⁵ Per *matching* "esatto" si intende il collegamento tra informazioni riferite alla stessa unità statistica; per *matching* "statistico" il collegamento tra informazioni di unità ritenute simili in base a qualche criterio.

risulta imprescindibile⁶, tuttavia i diversi archivi sono stati utilizzati singolarmente, senza passare alla loro integrazione. In altri casi il ricorso a fonti non usuali è derivato dalla necessità di reperire informazioni di natura non strettamente demografica (principalmente di tipo economico) ritenute esplicative o, quantomeno, connesse a fenomeni demografici. Angeli e altri (1994) e Pasquini e Samoggia (1996) hanno utilizzato i dati dell'*Indagine sui bilanci delle famiglie nel 1991* della Banca d'Italia per studiare le relazioni tra struttura familiare, ruoli femminili e reddito; Zaccarin (1994) la *Rilevazione delle Forze di Lavoro* per una analisi longitudinale (su dati aggregati) dell'esperienza lavorativa, familiare, riproduttiva di alcune coorti femminili; Clerici (1994, 1996) un'anagrafe automatizzata per ricostruire tratti di biografie familiari e l'*Indagine I Consumi delle Famiglie* (ISTAT), per lo studio delle relazioni tra tipologie di consumo e stadio del ciclo di vita attraversato dalle famiglie (Clerici, 1997).

4. Le esperienze fin qui illustrate hanno mostrato diversi livelli di utilizzo degli archivi amministrativi e diversi criteri e finalità dell'integrazione tra le fonti:

Primo livello (uso non integrato di fonti "non usuali"):

- a) uso di registri (parziali) per stime intercensuali di caratteri strutturali o di fenomeni di movimento;
- b) sfruttamento di indagini non demografiche per studi demo-economici.

Secondo livello (integrazione tra più fonti per analisi trasversali):

- a) collegamento tra registri e censimento per controlli di qualità (copertura);
- b) collegamento tra registri e indagini per stime intercensuarie;
- c) collegamento tra archivi per sostituire il censimento.

Terzo livello (sfruttamento di fonti, integrate o meno, per analisi longitudinali):

- a) estrazione di campioni longitudinali da archivi storici;
- b) estrazione di campioni longitudinali da *linkage* tra censimenti successivi;
- c) estrazione di campioni longitudinali da *linkage* tra archivi e censimenti.

Inoltre, a ciascuno dei livelli considerati, l'interesse può riguardare l'intero territorio nazionale o realtà locali.

⁶ Vedi, soprattutto, ISTAT (1989).

Questa pur parziale, ma significativa, ricognizione può portare a riflettere sulle possibili applicazioni che le diverse strategie di ricerca possono avere nel nostro Paese. L'uso dei registri appare particolarmente utile in una situazione, quale quella italiana, caratterizzata dall'assenza di rilevazioni campionarie longitudinali. L'unica esperienza italiana in questo senso è la *Rilevazione delle Forze di Lavoro*, che attraverso il meccanismo della rotazione vede (ma solo parzialmente) reintervistato il suo campione. Tutte le altre sono indagini trasversali che sfruttano, anche in maniera ridotta, le domande retrospettive.

L'Italia, diversamente dalla maggioranza dei paesi analizzati, ha la fortuna di disporre di un registro generale della popolazione residente (a cui una recente normativa aggiunge gli stranieri con permesso di soggiorno), che appare però ampiamente sottoutilizzato rispetto alle sue potenzialità⁷. La qualità dell'informazione anagrafica, principalmente la sua copertura, è messa in discussione: sicuramente i problemi esistono e vanno affrontati con maggiore decisione se, tra l'altro, si vuole continuare a reputare affidabili i risultati delle principali indagini nazionali i cui campioni sono generalmente estratti, mediante complesse operazioni di stratificazione, proprio dalle anagrafi comunali.

Una strada possibile per il miglioramento delle conoscenze sulla reale consistenza della popolazione che dimora stabilmente su un dato territorio potrebbe ad esempio essere quella dell'integrazione tra anagrafe e archivio degli assistiti dal Sistema Sanitario Nazionale. E' infatti noto come sia possibile richiedere, con certe modalità, l'iscrizione ad una USL diversa da quella di residenza. Ciò avviene nei casi in cui è conveniente per gli individui mantenere, per vari motivi, la residenza anagrafica in luogo diverso da quello di effettiva dimora. Poiché è interesse di tali soggetti fruire dell'assistenza sanitaria nel luogo di effettiva dimora, si può ritenere la fonte sanitaria più veritiera ed aggiornata di quella anagrafica. Pertanto un incrocio delle informazioni (limitatamente alla dimora del soggetto sul territorio o, se affidabili, ad altre caratteristiche demografiche), non può che produrre un miglioramento della qualità dell'informazione. Un uso analogo potrebbe essere fatto degli elenchi di iscritti alla scuola dell'obbligo, in cui possono risultare inseriti alunni residenti anagraficamente in comune diverso da quello della scuola.

Anche nel caso di buoni livelli di copertura esistono comunque, per l'Anagrafe, problemi di qualità dell'informazione riferita alle varie caratteristiche individuali: è noto come alcuni caratteri presenti nella scheda individuale non siano affidabili (ad es. titolo di studio, professione). Un

⁷ Per un esempio di sfruttamento di un'Anagrafe informatizzata, vedi Clerici (1994, 1996).

altro, ma non ultimo, aspetto da considerare è che ancora un 30% delle anagrafi non è stata automatizzata e che in Italia l'Anagrafe è un registro locale, non centralizzato⁸. Tutto ciò non deve far perdere di vista le ampie possibilità di sfruttamento che si aprono in prospettiva, che però debbono indirizzarsi verso un uso integrato di questo archivio "di base", su cui innestare, attraverso procedure di collegamento a vari livelli, le informazioni ulteriori desumibili da fonti "integrative", di natura ancora amministrativa o da indagini campionarie.

La possibilità di archiviare periodicamente (annualmente) i dati correnti apre inoltre la strada alla ricostruzione, almeno per intervalli (annuali), dell'evolversi delle condizioni personali, familiari, abitative, residenziali, occupazionali, di reddito, di salute, ecc. Disponendo di una serie abbastanza consistente di *files* storici in cui siano presenti quei caratteri che permettono il collegamento dei *records* tra le diverse basi di dati, è possibile operare ricostruzioni longitudinali di tratti o di intere biografie. Pensiamo ad esempio al collegamento tra anagrafe e dichiarazioni dei redditi delle persone fisiche: alle informazioni anagrafiche sulla composizione della famiglia e sulle sue trasformazioni potrebbero venir associate (per coloro che presentano la dichiarazione dei redditi), informazioni circa il reddito da lavoro, la proprietà di terreni e fabbricati (ma soprattutto dell'abitazione) ed il loro cambiamento.

L'accesso ai registri è spesso negato in nome della riservatezza, ma non vi è motivo di vietare una selezione campionaria di tali informazioni, privata dei riferimenti che permettano di risalire agli individui selezionati. Il collegamento tra registri o tra fonti diverse, necessariamente realizzato mediante codici personali o set di caratteri che permettono l'identificazione dei soggetti, può passare attraverso procedure di campionamento o attribuzione di codici identificativi fittizi che solo l'amministrazione rimarrebbe in grado di decodificare. Anche l'associazione tra archivi diversi potrebbe portare alla costruzione di basi di dati temporanee, che rimangono memorizzate per il tempo necessario all'applicazione di procedure di estrazione campionaria e successivamente cancellate.

Un progetto di integrazione tra fonti per migliorarne la qualità, seguito dall'estrazione annuale di un campione longitudinale, creerebbe la possibilità di applicare quell'approccio biografico allo studio della popolazione che può condurre, ad esempio, a fare miglior luce sui processi decisionali che hanno prodotto le attuali congiunture di invecchiamento, di denatalità, di deurbanizzazione che molto bene riusciamo ormai a descrivere ma che faticiamo ancora a comprendere nella loro genesi (Bonaguidi, Santini, 1993).

⁸ Il progetto Istatel prevede la trasmissione in via telematica dei dati sulla dinamica demografica dei comuni all'Istat.

BIBLIOGRAFIA

- ANGELI A., COCCHI D., PASQUINI L., SAMOGGIA A. (1994), "Struttura delle famiglie e reddito: primi risultati di un'analisi su dati trasversali", *Statistica*, 4.
- BONAGUIDI A., SANTINI A. (1993), "Informazione statistica e ricostruzione dei processi demografici", ISTAT, *Quaderni di Ricerca. Famiglia e Società*, 4.
- CALOSSO G. (1997), *Gli aspetti giuridici dell'impiego dei dati individuali*, Bozza provvisoria, presentata al Workshop SIS-ISTAT "L'utilizzo dei dati amministrativi e fiscali a fini statistici: quali prospettive per la ricerca in Italia?", Roma, 16 giugno 1997.
- CARLSSON F. (1996), *Use of Labour Force Survey (LFS) to Improve the Quality of Statistics, Based on Administrative Registers*, IAOS, Statistics Iceland, 5th Independent Conference, Reykjavik, 2-5 July 1996.
- CIUCCI L., RACIOPPI F. (a cura di) (1994), *Studi di popolazione. Nuovi approcci per la descrizione e l'interpretazione*, Dipartimento di Scienze Demografiche, Università degli Studi di Roma "La Sapienza", Roma.
- CLERICI R. (1994), "Lo sfruttamento di un archivio anagrafico automatizzato: una coorte di matrimoni nel primo decennio di vita", in CIUCCI L., RACIOPPI F. (1994) *Studi di popolazione. Nuovi approcci per la descrizione e l'interpretazione*, Dipartimento di Scienze Demografiche, Università degli Studi di Roma "La Sapienza", Roma, 317-339.
- CLERICI R. (1996), "Le famiglie e la città: le scelte residenziali delle famiglie dei coniugati nel 1982", in MICHELI G.A., CLERICI R. (a cura di), *Morfologia e mobilità urbana*, Franco Angeli, Milano, 17-113.
- CLERICI R. (1997), "Nuovi modelli familiari e consumi privati", in MICHELI G.A., RIVELLINI G. (a cura di), *Popolazione e Mercato*, Franco Angeli, Milano.
- DICOSOFT s.r.l. (1996), *Atlante statistico informatizzato* (materiale pubblicitario), Roma, s.d. (ma 1996).
- FILIPPUCCI C. (1997), "Utilizzo di dati amministrativi: prospettive e problemi", *Giornale del SISTAN*, 6.
- HERRIOT R.A. (1983), *The Use of Administrative Records in Social and Demographic Statistics*, ISI, Madrid.
- ISTAT (1989), "Analisi delle fonti statistiche per la misura dell'immigrazione straniera in Italia: esame e proposte", *Note e Relazioni*, 6.
- ISTAT (1990), "La mortalità differenziale secondo alcuni fattori socio economici", *Note e Relazioni*, 2.
- ISTAT (1995), "Gli stranieri in Italia, fonti statistiche", *Note e Relazioni*, 1.

- JACOD M. (1996), *Estimations locales de population multi-sources*, IAOS, Statistics Iceland, 5th Independent Conference, Reykjavik, 2-5 July 1996.
- LEYES J.M. (1980), *Developing Administrative Data: Research and Development and Prototypes Underway in Statistics Canada*, Statistics Canada.
- MONTGOMERY K.J., STANDISH L.D. (1993), *Report on the Methodology of Migration Data from Taxation Records*, Small Area and Administrative Data Division, Statistic Canada.
- NORRIS D., LEYES J., KOPUSTAS N. (1981), *The Development of Social Data from Administrative Records: the Canadian Experience*, American Statistical Association Meeting, Detroit.
- NORRIS D., STANDISH L. (1983), *A Technical Report on the Development of Migration Data from Taxation Records*, Administrative Data Development Staff, Statistics Canada.
- PASQUINI L., SAMOGGIA A. (1996), "Ruoli della donna e reddito: una prima analisi", in SIS, *Atti della XXXVIII Riunione Scientifica*, Maggioli Editore.
- RIGHI A., TERRA ABRAMI V. (1993), "Verso lo sfruttamento integrato delle indagini e delle rilevazioni ufficiali a fini demografici", ISTAT, *Quaderni di Ricerca. Famiglia e Società*, 1.
- SEAT (1995), *Gli strumenti del geomarketing*, SEAT Divisione STET S.p.A., Pomezia (Roma), s.d. (ma 1995).
- SHRYOCK H.S., SIEGEL J.S. & ass. (1980), *The Methods and Materials of Demography*, U.S. Bureau of the Census, U.S. Government Printing Office, Washington.
- SISTAN-ISTAT (1994a), *Il sistema statistico nazionale*, Roma.
- SISTAN-ISTAT (1994b), *Rapporto Annuale. L'attività e le risorse della statistica ufficiale. 1993*, Roma.
- SISTAN-ISTAT (1997), *Programma Statistico Nazionale (triennio 1997-99)*, Roma, s.d. (ma 1997).
- THYGESEN L. (1983), *Methodological Problems Connected with a Socio-demographic Statistical System Based on Administrative Records*, ISI, Madrid.
- THOMSEN I., KLEIVE HOLMOY A.M. (1996), *Combining Data from Surveys and Administrative Record System. The Norwegian Experience*, IAOS, Statistics Iceland, 5th Independent Conference, Reykjavik, 2-5 July, 1996.
- ZACCARIN S. (1994), "Generazioni di donne a confronto: un tentativo di analisi longitudinale da dati aggregati", in CIUCCI L., RACIOPPI F. *Studi di popolazione. Nuovi approcci per la descrizione e l'interpretazione*, Dipartimento di Scienze Demografiche, Università degli Studi di Roma "La Sapienza", Roma, 341-357.
- ZAJCZYK F. (1991), *La conoscenza sociale del territorio. Fonti e qualità dei dati*, Collana di Sociologia, Franco Angeli, Milano.

APPENDICE

**STATISTICS CANADA:
STIMA DEI FLUSSI MIGRATORI DA FONTE FISCALE**

Dal 1976 *Statistics Canada* ha approntato un metodo di stima dei movimenti migratori tra le divisioni territoriali censuarie a partire dalla fonte fiscale. Con riferimento a ciascun anno, e alle divisioni di censimento, vengono sistematicamente prodotte quattro tavole di dati relative a:

- a) immigrazioni, emigrazioni e saldo migratorio totale;
- b) immigrazioni, emigrazioni e saldo migratorio per classe di età;
- c) immigrazioni, emigrazioni e saldo migratorio totale per tipo di spostamento (intraprovinciale, interprovinciale, internazionale) e sesso;
- d) principali flussi migratori per origine e destinazione.

La messa a punto della metodologia di stima è derivata dalla limitata disponibilità di informazioni dirette sul fenomeno migratorio, soprattutto per aggregati territoriali di piccole dimensioni. La fonte principale per lo studio delle migrazioni è infatti il censimento, che contiene un quesito retrospettivo circa la residenza di 5 anni prima, di cui sono però noti i limiti.

È dunque sorta la necessità di ricorrere a fonti indirette. La scelta è caduta sulla fonte fiscale (*Revenue Canada Master File* da cui vengono estratti gli *individual income tax records*). Tale fonte contiene, oltre ad un codice identificativo personale per ciascun dichiarante, l'indirizzo (da cui trarre il codice postale), il sesso, l'anno di nascita, lo stato civile, il reddito dichiarato, un codice indicativo di eventuali movimenti internazionali. Mediante il confronto tra due successive dichiarazioni dei redditi si deducono lo spostamento e le sue caratteristiche.

Il procedimento di stima dei flussi migratori è articolato in quattro principali fasi:

- 1) attribuzione di un codice geografico (*census division*) al record individuale del contribuente, a partire dal codice postale tratto dall'indirizzo o dal recapito postale;
- 2) stima delle persone a carico (soprattutto figli e coniuge);
- 3) identificazione dei migrati per sesso e classe di età;
- 4) aggiustamento per la completa copertura della popolazione.

Nel 1989 una riforma del sistema di detrazioni portò a modificare il metodo di stima delle persone a carico (ovvero della fase 2: questo passaggio è cruciale in quanto alle persone a carico il metodo di stima fa assumere lo stesso comportamento migratorio del dichiarante a cui sono associati). Fino alla riforma la stima era realizzata a partire da un campione di dichiaranti ai

quali venivano attribuite delle persone a carico a seconda dell'età, sesso e stato civile del dichiarante e alle detrazioni da lui effettuate, e successivamente imputando a ciascun dichiarante le persone a carico. Queste venivano poi classificate per sesso e per età in base alle distribuzioni censuarie. La procedura, secondo *Statistics Canada*, copriva approssimativamente il 90-95% della popolazione totale e, d'altra parte, erano presenti doppi conteggi valutati attorno al 7%.

In occasione della riforma fiscale del 1989 fu approntato un nuovo sistema di stima. Esso si basa sulla ricostruzione delle famiglie: a ciascun dichiarante vengono associati (se presenti nel file) o imputati (se non presenti) coniuge e figli utilizzando informazioni quali lo stato civile (ad ogni coniugato viene associato il coniuge se dichiarante o attribuito se non presente nel file), il numero del sistema di sicurezza sociale (SIN), l'indirizzo, le detrazioni, i crediti fiscali per i figli, gli assegni familiari, le esenzioni addizionali, ecc. Dal file delle famiglie si ricava un file di individui e su questo si applicano le successive fasi di stima delle migrazioni.

Vi sono diversi aspetti interessanti nella procedura di definizione delle caratteristiche dei figli a carico: di costoro si desume l'esistenza per la presenza di detrazioni o di assegni o di crediti fiscali, ma non si conosce né il sesso, né sempre è nota l'età. L'informazione mancante è dunque imputata con metodo probabilistico a partire dai dati di censimento. Analogamente viene imputata l'età dei coniugi dei dichiaranti. In tal modo si ricostruisce la struttura per età e sesso della parte di popolazione interessata direttamente (dichiaranti) o indirettamente (persone a carico) alle pratiche fiscali. In questa ricostruzione non vengono considerati altri componenti della famiglia (non a carico e non dichiaranti, in particolare anziani che non presentano dichiarazione dei redditi), ma è stato valutato che il 97% della popolazione (1990-91) era contribuente o a carico di un contribuente. Il confronto tra popolazione stimata e censita permette una valutazione del metodo di stima e fornisce gli elementi per la copertura totale della popolazione.

Dei dichiaranti si conosce, tramite il confronto di due successive dichiarazioni, la mobilità: la stessa mobilità è imputata ai familiari. Per la popolazione non coperta si assume un comportamento migratorio simile a quello riscontrato tra i dichiaranti.

Molte sono evidentemente le assunzioni e le ipotesi che regolano l'imputazione dei dati non direttamente noti, ma le verifiche condotte in termini comparativi con i risultati censuari sembrano confermare la validità di un metodo molto ingegnoso, che certo presenta dei limiti e delle carenze, ma che pure ci si presenta come uno stimolo a sfruttare meglio ed in modo innovativo la ricchezza informativa che anche il nostro Paese inizia a valorizzare.