

# Public Use Samples (PUS)

## Campioni dei censimenti svizzeri della popolazione ad uso pubblico

### Introduzione

Il censimento della popolazione, condotto ogni dieci anni dal 1850, è un pilastro importante della statistica pubblica svizzera. Fornisce dati demografici di base e informazioni fondamentali, paragonabili in termini di spazio e di tempo, sull'insieme degli abitanti, delle economie domestiche, dei posti di lavoro e delle abitazioni.

Tuttavia, l'accesso alla raccolta di dati del censimento è limitato dalle dimensioni dello stesso - anche con i mezzi informatici più moderni, un file contenente i dati di diversi milioni di persone è difficile da gestire - e la tutela del carattere confidenziale dei dati rimane un problema essenziale.

Per porre rimedio a queste difficoltà, l'Ufficio federale di statistica, in collaborazione con il Consiglio politico delle scienze sociali (CPS) dell'Accademia svizzera delle scienze

umane e sociali, ha deciso di realizzare dei campioni di dati individuali ad uso pubblico, i cosiddetti "Public Use Samples" (PUS). In realtà, in questo contesto l'UST non compie un lavoro da pioniere. Infatti, numerosi uffici di statistica nazionali mettono a disposizione dei ricercatori dei PUS, vale a dire dei campioni dei censimenti nazionali. Basta menzionare "il campione al 5%" dell'istituto francese INSEE o i campioni dell'"U.S. Census Bureau", senza scordare gli esempi dei Paesi Bassi o del Canada.

Lo scopo di questi PUS è di consentire a tutte le persone interessate - ricercatori, studenti, scienziati - di analizzare i dati del censimento nell'ambito di un "modello ridotto", in particolare a scopo di insegnamento e di ricerca. Mostrando la ricchezza di informazioni esistente, facendo luce sulle possibilità di utilizzazione e di analisi offerte da un cen-



**Dominique Spahn  
e Monique Graf,  
Ufficio federale di statistica**

simento, i PUS dovrebbero consentire di abbassare le barriere di accesso ai dati del censimento. Tanto più che questi PUS, composti da dati estratti a caso rappresentanti il 5% della popolazione, possono essere diramati senza restrizioni preventive, dato che sono state adottate misure specifiche per assicurarne l'anonimato.

### Tre Public Use Samples: censimenti della popolazione del 1970, 1980 e 1990

Finora sono stati realizzati tre "Public Use Samples", disponibili a partire dal mese di dicembre 2001<sup>1</sup>. Primi di una serie, forniscono i dati dei censimenti del 1970, 1980 e 1990. È inoltre prevista la creazione di uno o più PUS contenenti dati del censimento 2000.

I dati sono incentrati sulle persone e sono corredati di informazioni sulle economie domestiche, sulle abitazioni e sugli edifici. I campioni al 5%, rappresentativi della popolazione, comprendono la maggior parte

foto Ti-press



<sup>1</sup> Questo lavoro è frutto della collaborazione di autori che lavorano, rispettivamente, presso il Servizio svizzero di informazione e archiviazione dei dati per le scienze sociali (SIDOS), presso Consult AG Berna e presso l'Ufficio federale di statistica.



delle informazioni fornite dal censimento, tuttavia in forma meno dettagliata.

I dati sulle economie domestiche, sulle abitazioni e sugli edifici forniscono informazioni contestuali sulle condizioni di vita degli individui, ma non possono essere usati in modo isolato poiché sono il risultato di un

campione di persone. È probabile che l'UST produca in futuro altri campioni rappresentativi delle economie domestiche o delle abitazioni, paragonabili a quanto ha fatto ad esempio l'ufficio di statistica canadese con il suo *File di microdati a grande diffusione (FMGD) delle economie domestiche e delle abitazioni o il suo FMGD sulle famiglie*.

I dati dei tre PUS sono pubblicati sotto forma di CD-ROM. Quest'ultimo è corredato di un opuscolo, in francese e tedesco, comprendente una breve introduzione ai PUS. Nel CD-ROM sono inclusi:

- il testo dell'opuscolo stesso;
- i 3 PUS, vale a dire i file comprendenti i campioni di dati dei tre anni di censimento;
- un file tecnico per la lettura dei dati ASCII, nonché i programmi di conversione SAS e SPSS;
- la descrizione delle variabili dei PUS;
- i questionari originali (individuali, economie domestiche, edifici e abitazioni) per ognuno dei 3 anni;
- gli elenchi dei codici;
- varie altre informazioni.

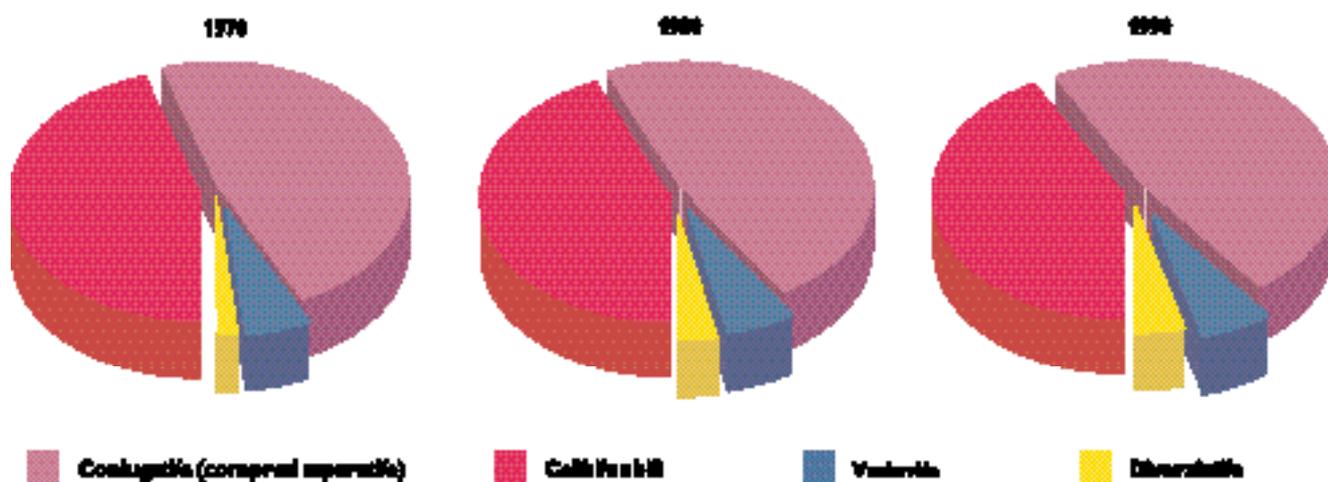
Al fine di facilitare la comprensione dei

dati messi a disposizione, si è dedicata particolare attenzione alla descrizione delle variabili, descrizione che figura sul CD-ROM. Per ognuna di queste variabili, oltre al suo nome e alla sua etichetta, viene fornita pertanto una breve introduzione che illustra il contenuto della variabile, un'indicazione sulla popolazione di riferimento coinvolta (ad es. popolazione residente o popolazione residente occupata), nonché informazioni sulle particolarità proprie a ogni anno di censimento. Inoltre, per ognuno degli anni 1970, 1980 e 1990, le frequenze della variabile sono presentate sotto forma di tabella e di grafico.

### Procedura di campionamento

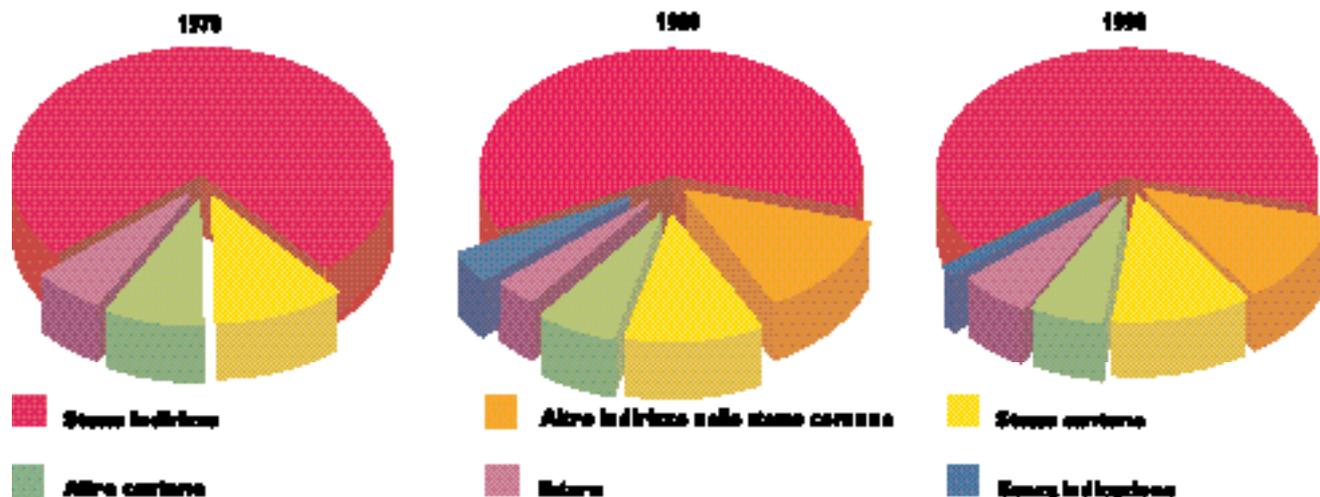
I campioni sono stati estratti indipendentemente l'uno dall'altro sulla scorta dei dati dei censimenti degli anni 1970, 1980 e 1990. La procedura adottata è quella del campionamento causale semplice senza reimmissione e ogni campione comprende il 5% della popolazione di riferimento. In altri termini, qualsiasi persona aveva le stesse probabilità di far parte del campione in ogni anno di censimento.

## A Persone residenti in Svizzera secondo lo stato civile nel 1970, 1980 e 1990



<sup>2</sup> European Strategic Programme for Research and Development in Information Technologies/Statistical Disclosure Control.

## B Persone residenti in Svizzera secondo il luogo di residenza cinque anni prima nel 1970, 1980 e 1990



### Protezione dei dati

La creazione di un campione ad uso pubblico contenente dei dati individuali pone dei problemi particolari in termini di protezione dei dati. È stato pertanto necessario dedicare particolare attenzione a questa problematica. Infatti, è indispensabile accertarsi che le persone incluse nel campione non possano essere identificate in base a un sottoinsieme di variabili. Per conseguire questo obiettivo, l'UST ha richiesto la collaborazione di esperti di matematica statistica. Per i lavori di anonimizzazione legati alla protezione dei microdati, gli autori si sono ispirati alle ricerche effettuate dal gruppo di esperti del progetto europeo ESPRIT/SDC<sup>2</sup>.

La prima forma di protezione dell'anonimato delle persone incluse nel file è costituita dal tempo trascorso dalla raccolta dei dati, ossia 11 anni per il censimento più recente. Infatti, è molto più difficile identificare una persona sulla scorta di informazioni risalenti a 11 anni prima, di quanto non sia sulla scorta di informazioni recenti. Il rischio della divulgazione va inoltre commisurato con l'interesse di un'eventuale intrusione. Nel caso di dati così vecchi, sembra ridursi la motivazione per una divulgazione di dati personali,

come sembra ridursi la capacità di dimostrare che una persona particolare sia identificata, vale a dire a dimostrare che, per una combinazione particolare di variabili, essa sia stata unica nella popolazione 11 anni prima.

L'identificazione di una persona non può che basarsi sulla combinazione di informazioni fornite da variabili diverse, dato che ogni identificatore diretto viene naturalmente soppresso. Il promulgatore di dati deve pertanto compiere una scelta strategica: può sopprimere una parte delle variabili per evitare delle combinazioni identificanti, oppure mantenere l'insieme delle variabili presentando tuttavia l'insieme di queste variabili in forma meno dettagliata (ossia riducendo il numero di modalità). I PUS sono stati concepiti conformemente alla seconda possibilità, nell'intento di mantenere la maggior parte delle informazioni sociologiche disponibili e di fornire un'immagine completa, seppure meno dettagliata, dei censimenti.

Non tutte le variabili hanno lo stesso potere identificante, per cui il livello di dettaglio non è stato ridotto in modo uniforme. Sono state impiegate procedure diverse. Per motivi attinenti alla stessa protezione dei dati, non possiamo comunicare quali procedure

sono state adottate esattamente, né in quale misura.

### Conclusione

È incontestabile che i limiti posti alla realizzazione di un PUS, in particolare per quanto concerne la protezione dei dati, limitano determinate analisi. A seconda dell'argomento di ricerca, rimarrà necessario utilizzare il file di dati completo, ciò che implica la sottoscrizione di un contratto di protezione dei dati. Ciò nonostante, i PUS sono destinati a diventare un componente della gamma di strumenti offerti dall'UFS per l'analisi dei dati dei censimenti. Che si tratti ad esempio di convalidare i dati di indagini o di controllarne i campioni, di arricchire i dati raccolti nell'ambito di altre ricerche, o di formare gli studenti di statistica, si ritiene che il ricorso ai PUS diventerà irrinunciabile.

### Riferimenti

UST, Public Use Samples (PUS) - Echantillons à usage public des recensements suisses de la population, Neuchâtel 2001.  
N. di ordinazione: 001-9094, prezzo (CD-ROM incluso): 25 franchi. ■