CONTRIBUTI ISTAT

n. 4/2007

La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS

M. Scannapieco, D. Zardetto e G. Barcaroli

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Le collane esistenti presso l'ISTAT - *Rivista di Statistica Ufficiale*, *Contributi ISTAT* e *Documenti ISTAT* - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell' ISTAT, del SISTAN, o da studiosi esterni.

La *Rivista di Statistica Ufficiale* accoglie lavori che hanno come oggetto la misurazione dei fenomeni economici, sociali, demografici e ambientali, la costruzione di sistemi informativi e di indicatori, le questioni di natura metodologica, tecnologica o istituzionale connesse al funzionamento dei sistemi statistici e al perseguimento dei fini della statistica ufficiale.

I lavori pubblicati in *Contributi ISTAT* sono diffusi allo scopo di stimolare il dibattito intorno ai risultati preliminari di ricerca in corso.

I *Documenti ISTAT* forniscono indicazioni su linee, progressi e miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.

Il Comitato di redazione esamina le proposte di lavori da pubblicare nelle tre collane sopra indicate. Quelli pubblicati nei Contributi ISTAT e nei Documenti ISTAT sono valutati preventivamente dai dirigenti dell'Istituto, mentre i lavori pubblicati nella Rivista di Statistica Ufficiale sono subordinati al giudizio di referee esterni.

ISTAT – Produzione libraria e centro stampa

CONTRIBUTI ISTAT

n. 4/2007

La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS

M. Scannapieco(*), D. Zardetto(**) e G. Barcaroli(**)

(*) ISTAT - Direzione Centrale per le tecnologie ed il supporto metodologico
(**) ISTAT - Servizio Metodologie, tecnologie e software per la produzione dell'informazione statistica

Contributi e Documenti Istat 2007

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

Produzione libraria e centro stampa:
Carla Pecorario
Via Tuscolana, 1788 - 00173 Roma

1 Introduzione 5
2 R ed il Package Survey 9
2.1 Cenni su R 9
2.2 Il Package Survey: la Struttura e le Principali Funzioni 10
3 La Calibrazione 12
3.1 Formulazione del Problema 12
3.2 La Calibrazione nel Package Survey 14
3.3 La Calibrazione in GENESEES 16
3.4 La Calibrazione nel Package Survey ed in GENESEES: Analogie e Differenze 16
4 Survey/R e GENESEES/SAS: Confronto Sperimentale tra le Funzioni di Calibrazione 18
4.1 Fasi della Sperimentazione 18
4.2 I Data Set dell'Indagine Forze di Lavoro e l'Ambiente di Sperimentazione 19
4.2.1 Cenni sull'Indagine Forze di Lavoro 19
4.2.2 Ambiente di Sperimentazione 21
4.3 Risultati della Sperimentazione: Test di Efficacia e di Efficienza 21
4.3.1 Utilizzo delle Funzioni di Survey: Problemi e Soluzioni 21
4.3.2 Test di Efficacia e di Efficienza: Survey/R e GENESEES/SAS 25
5 Conclusioni 31
6 Bibliografia 32

1 Introduzione

Da tempo all'interno dell'Istituto Nazionale di Statistica è in corso una riflessione di carattere generale sull'utilizzo di strumenti software di tipo *open source*, prodotti e resi disponibili a titolo gratuito da una libera comunità di sviluppatori, contrapposti a quelli proprietari disponibili sul mercato a titolo oneroso.
Tale riflessione non poteva non riguardare un sistema importante quale SAS (Statistical Applications System), presente in Istituto dai primi anni '80 e utilizzato, non solo, e a questo punto forse non tanto, a fini di analisi e ricerca statistica, ma in larga parte all'interno dei processi di produzione dell'informazione statistica, come sistema di elaborazione *tout court*.
Al di là della possibile concorrenza rispetto a SAS di prodotti *open* che potrebbero consentire risparmi di spesa anche notevoli, un ulteriore elemento ha spinto ad investigare seriamente tale possibilità. Negli ultimi due anni l'Istat si è posto il problema della oggettiva pericolosità derivante dalla dipendenza da un prodotto, quale SAS, che non è possibile acquistare, ma solo "noleggiare": se per ventura da un giorno all'altro tale prodotto non fosse più disponibile per l'Istituto, un certo numero di settori di produzione si troverebbero nell'impossibilità di assicurare l'uscita delle informazioni statistiche di competenza. Per parte sua il CNIPA, già nel parere trasmesso nel 2004 in occasione del rinnovo delle licenze fornite dalla SAS Institute, premeva affinché l'Istat effettuasse una sperimentazione "*di prodotti software alternativi, anche di tipo open source, al fine di eliminare, o di contenere, la dipendenza dei sistemi – e di conseguenza dell'attività istituzionale dell'Istituto - da un unico fornitore e da un'unica tecnologia proprietaria*". Nel parere del 2005, il CNIPA, ribadiva quanto sopra ed invitava l'Istat a procedere in tempi brevi a sperimentare tali prodotti.

Nell'impostare l'attività di studio e sperimentazione, sono state considerate le classiche fasi che caratterizzano il processo di produzione delle indagini campionarie:

1. disegno del campione e selezione delle unità campionarie;
2. acquisizione dei dati (intesa come data entry);
3. trattamento dei dati (controllo e correzione; integrazione);
4. tabulazione;
5. data warehousing;
6. calcolo delle stime campionarie e dei relativi errori;
7. analisi dei dati.

e per ognuna di esse sono state considerate le possibili alternative a SAS, open source o di mercato.

1. Per quanto riguarda il *disegno del campione*, si tratta di un'attività che normalmente è svolta una tantum, in occasione del lancio di una nuova indagine campionaria o per la ristrutturazione di una esistente, oppure periodicamente per il miglioramento della efficienza del campione. L'uso del SAS è attualmente preponderante, sia attraverso programmi ad hoc che mediante l'utilizzo di software generalizzato (BETHEL-MAUSS). L'alternativa è nell'uso di software *open source* come R, che offre potenzialmente tutte le funzionalità necessarie.

2. Negli applicativi di *data entry*, l'uso di SAS/AF con SCL (Screen Control Language) è abbastanza diffuso, laddove si utilizzino data set sequenziali. L'alternativa può essere costituita, permanendo l'ottica sequenziale, dall'utilizzo di BLAISE; oppure, presupponendo una migrazione verso una logica relazionale, degli strumenti ORACLE per l'acquisizione (come SQLForms), o comunque di linguaggi come PL/SQL. Non è da escludere anche l'utilizzo di sistemi *open source* attualmente in fase di sperimentazione, come PHPSurveyor in ambiente MySQL.
3. Per quanto riguarda la fase di *trattamento dei dati* (controllo e correzione degli errori ed integrazione delle mancate risposte parziali), occorre distinguere tra procedure (o segmenti di procedure) che fanno uso di sistemi generalizzati, e procedure (o segmenti di esse) che fanno uso di programmi di tipo deterministico, generalmente sviluppati con programmi ad hoc. I sistemi generalizzati che in questo settore fanno uso di SAS sono CONCORD/SCIA e GEIS/BANFF, operanti in ambiente Windows. E' da sottolineare però che di entrambi esistono versioni non Windows (Linux) che non richiedono SAS. E' quindi possibile pensare a migrazioni dalle prime alle seconde. Oppure si può pensare di mantenere l'utilizzo delle versioni Windows (più ricche di funzioni e maneggevoli) nelle attività di sviluppo e di test (di fascia (b), come riportato più avanti), utilizzando le versioni Linux per la produzione. Per quanto riguarda invece i programmi di tipo deterministico, in molte situazioni questi sono sviluppati in SAS. In questo caso, qualsiasi linguaggio procedurale (Java, C, PHP) può costituire una valida alternativa.
4. Nelle attività di reporting, il SAS è utilizzato nella fase di tabulazione essenzialmente per la PROC TABULATE e la PROC FREQ. Qui l'alternativa naturale è costituita da Business Objects e/o da ORACLE (o comunque un DBMS relazionale anche *open source*).
5. Il SAS non è attualmente utilizzato per lo sviluppo di *data warehouses* su microdati o macrodati: dopo l'esperienza effettuata nel Censimento Intermedio di Industria e Servizi (1996), l'Istituto ha adottato altri strumenti (Business Objects e, in prospettiva, il toolkit ISTAR);
6. Per quanto riguarda il *calcolo delle stime campionarie e dei relativi errori*, le procedure di stima per le indagini campionarie fanno normalmente uso di sistemi generalizzati standard (GENESEES o versioni precedenti come SGCS), sviluppati in SAS. Per questi sistemi occorre prevedere lo sviluppo con altri strumenti e linguaggi di sistemi che assicurino le stesse funzionalità. Si può pensare, a tal fine, di utilizzare R che possiede un linguaggio particolarmente potente per il trattamento degli array, unitamente a Java o PHP per le necessarie interfacce.
7. Intendiamo per *analisi dei dati* sia l'attività condotta da ricercatori tematici volta ad approfondire la conoscenza dell'area da investigare, che quella svolta da metodologi e tecnologi ai fini della valutazione ed ottimizzazione dei processi di produzione. Ad esempio, INSIGHT di SAS è utilizzato per studiare la struttura dei dati e calibrare i parametri di procedure per la ricerca ed il trattamento degli *outlier*, o l'individuazione di errori sistematici. Le stesse attività possono essere svolte da R, che permette di utilizzare, senza necessità di scrivere codice mediante l'interfaccia "R Commander", una serie di pacchetti per la produzione di statistiche mono e multivariate, per la stima di modelli e per la produzione di grafici, pur senza raggiungere la sofisticatezza di INSIGHT. Accanto ad R sono da considerare una serie di strumenti *open source* per il data mining, come WEKA, recentemente acquisito ed in fase di sperimentazione, YALE e KNIME.

Una volta considerate le fasi di cui sopra, per individuare quella di maggior interesse per la sperimentazione, è opportuno raggrupparle a seconda dei diversi *livelli di dipendenza* che l'uso di SAS può produrre in ognuna di esse:

a. le attività correnti di produzione istituzionale delle informazioni statistiche periodicamente diffuse dall'Istituto: fasi 2, 3, 4, 5 (per la parte riguardante il datawarehousing per utenti esterni) e 6 (per la parte riguardante il calcolo delle stime);
b. le attività di disegno, sviluppo e test direttamente finalizzate alle prime, svolte principalmente in occasione dell'impianto di nuove indagini o della ristrutturazione di quelle esistenti: fase 1, 5 (per la parte riguardante il datawarehousing per utenti interni) e 6 (per la parte riguardante il calcolo degli errori campionari);
c. le attività parallele di studio e analisi, solo indirettamente finalizzate a quelle di produzione, svolte in particolare dai ricercatori dell'Istituto: fase 7.

E' chiaro che la dipendenza maggiormente rischiosa ed indesiderabile dell'Istat da SAS è quella nella fascia (a): per questo tipo di attività, infatti, una improvvisa indisponibilità produrrebbe blocchi di produzione inaccettabili. Chiaramente, anche le attività nelle altre fasce sono importanti, ma l'Istituto dispone per queste di maggiori margini di manovra e la relativa dipendenza risulta meno rischiosa.

Per la nostra attività di sperimentazione è stata individuata la fase 6 (*calcolo delle stime campionarie e dei relativi errori*) per i seguenti motivi:

- in tale fase, SAS è presente in praticamente tutte le indagini campionarie;
- il calcolo delle stime campionarie ricade all'interno della fascia (a);
- R è già stato individuato come naturale alternativa a SAS.

Abbiamo indicato R come alternativa a SAS perché a nostro parere non avrebbe senso investire pesantemente in attività di migrazione per andare verso un altro sistema proprietario (ad esempio SPSS), solo per un eventuale minor costo. R è un sistema *open source* in continuo sviluppo grazie ad una numerosa comunità di statistici ed informatici, che per il momento però soffre dei seguenti limiti:

- difficoltà a trattare elevate moli di dati;
- mancanza di adeguate interfacce che ne facilitino l'utilizzo per l'utente non programmatore.

E' prevedibile però, dato il continuo sforzo di potenziamento di R compiuto dalla comunità degli sviluppatori, che in un futuro non lontano tali limiti possano essere superati.

Come terreno concreto di sperimentazione, è stata scelta la fase di calcolo delle stime campionarie e dei relativi errori all'interno dell'indagine sulle Forze di Lavoro, per i seguenti motivi essenziali:

a. in tale indagine la fase di calcolo delle stime e degli errori campionari è effettuata mediante utilizzo, indiretto, di GENESEES (GENEralised Sampling Estimates and Errors in Surveys), un software generalizzato sviluppato in SAS, ed utilizzato nella gran parte delle indagini campionarie Istat: ne consegue che qualora la

sperimentazione avesse avuto esito positivo, sarebbe già stato affrontato uno degli elementi di maggior criticità nei processi di produzione;

b. in tale indagine le elaborazioni, effettuate mediante utilizzo dello stimatore di regressione generalizzata, per quanto riguarda il calcolo delle stime, e con il metodo della linearizzazione in serie di Taylor, per quanto riguarda il calcolo dei relativi errori campionari, sono caratterizzate da notevole complessità computazionale, sia per il particolare disegno adottato, di tipo misto (ad uno e a due stadi), che per l'alto numero di vincoli posti nella fase di riponderazione, e rappresentano quindi da questo punto di vista un adeguato banco di prova.

E' da rilevare come sarebbe stato proibitivo, e non avrebbe avuto senso, ai fini della sola sperimentazione, sviluppare ex novo in R gli stessi moduli contenuti in GENESEES: ci si è quindi avvalsi della disponibilità di uno specifico package di R, il package *survey*, che permette di calcolare sia le stime che gli errori campionari secondo le modalità richieste dall'indagine sulle Forze di Lavoro, con un minimo intervento di programmazione.
La sperimentazione effettuata si è rivelata particolarmente utile. In primo luogo, ha dimostrato la praticabilità dell'utilizzo di R in questa specifica fase del processo di produzione, comune a tutte le indagini campionarie correntemente portate avanti dall'Istat. Inoltre, ha permesso di approfondire notevolmente la conoscenza dello strumento, sia dal punto di vista del linguaggio di programmazione, che dell'operatività, sia in ambiente Windows che in quello Linux.
Il presente contributo è così articolato:

- Nella Sezione 2 si illustra in modo sintetico il software R, e si introduce il package survey, utilizzato nella sperimentazione.
- Nella Sezione 3 è fornito un quadro generale sul problema della calibrazione e sui meccanismi con cui è risolto in GENESEES e nel package survey.
- Nella Sezione 4 si procede ad una valutazione comparata di efficacia ed efficienza relativamente ai due sistemi posti a confronto mediante la sperimentazione: survey/R e GENESEES/SAS.
- Infine, nella Sezione 5 si presentano le conclusioni e si accenna alla fase successiva della sperimentazione che riguarderà il calcolo delle stime e degli errori campionari.

2 R ed il Package Survey

Questa sezione introduce brevemente le caratteristiche generali di R (Sezione 2.1). Per una trattazione più ampia dell'ambiente R si veda [R-intro,2006]; per approfondimenti sul linguaggio R si rimanda a [R-lang,2006] [S-Programming,2000].
La Sezione 2.2 fornisce una prima panoramica della struttura del package survey e delle funzionalità da esso rese disponibili.

2.1 Cenni su R

La natura di R è correttamente, seppur sinteticamente, descritta dalla definizione che la schermata di avvio del software presenta all'utente: "*R – A Language and Environment*".
R è un *ambiente* in quanto rende disponibile, ed agevolmente fruibile anche ad utenti non programmatori, una suite integrata di strumenti per la manipolazione, l'analisi e la visualizzazione dei dati. R è un *linguaggio* poiché fornisce gli utenti programmatori di un potente ed elegante linguaggio di programmazione general purpose, mediante il quale è possibile estendere ed arricchire le funzionalità base del software.
Il sistema può essere schematizzato come una struttura a tre livelli. La base è costituita dal programma C che implementa l'interprete del linguaggio R. Il secondo livello include le librerie (package) della distribuzione standard del software, nelle quali sono organizzati i programmi R che realizzano funzioni di utilità generale (algoritmi di ordinamento, di ottimizzazione, di algebra lineare, definizioni di classi e metodi, etc.). Lo strato più esterno è rappresentato da circa 1.000 (ad oggi) package aggiuntivi, dedicati alla risoluzione di problemi specifici.
Come è stato anticipato, la sperimentazione oggetto di questo lavoro è volta a stabilire se, nel processo di produzione dei dati, R possa configurarsi come una valida alternativa a software di elaborazione commerciali. In questa prospettiva lo studio degli strumenti in dotazione ai singoli package, pur doveroso, non appare sufficiente. Emerge chiaramente la necessità di estendere l'indagine alle potenzialità di R come linguaggio di programmazione *tout court*. A questo scopo ne riportiamo, senza pretesa di completezza, alcune caratteristiche rilevanti:

- Il linguaggio R è un "dialetto" del linguaggio S.

S è stato sviluppato per oltre 20 anni (a partire dalla fine degli anni '70) da John Chambers e collaboratori presso i Bell labs. Dal 1993 è disponibile un software commerciale, S-Plus, che utilizza con licenza esclusiva il linguaggio S. Conviene, a tale proposito, chiarire come R non sia una replica o una versione gratuita di S-Plus. Al contrario, il linguaggio R costituisce una implementazione indipendente (*open source*) di S, dal quale, infatti, differisce in aspetti sottili ma rilevanti (ad esempio le regole di visibilità degli identificatori).

- R è un linguaggio di alto livello, di notevole potere espressivo, orientato principalmente al calcolo matematico.

Da un punto di vista formale tutti i linguaggi di programmazione di uso comune sono Turing-completi: in essi è possibile esprimere esattamente lo stesso insieme di algoritmi. In pratica, d'altra parte, la semplicità di programmazione e l'economia di codice (le dimensioni fondamentali di ciò che qui intendiamo per "espressività") non sono

equivalenti in tutti i linguaggi. R, soprattutto in virtù della natura intrinsecamente vettoriale delle sue funzionalità primitive, si rivela particolarmente elegante e potente nella programmazione orientata al calcolo matematico. Ciò ne spiega il successo e la diffusione in ambito statistico.

- R è un linguaggio interpretato.

Questa caratteristica penalizza in modo particolare l'efficienza dei programmi R che fanno ricorso a cicli: sebbene più veloci degli analoghi S-PLUS, essi risultano 2 ordini di grandezza più lenti di compilati C o Fortran. R, d'altra parte, dispone di metodi di vettorizzazione davvero efficienti e flessibili, che consentono, nella maggior parte dei casi, di evitare le strutture di ciclo.
Il fatto che il sistema R sia costruito sulla base di un linguaggio interpretato presenta anche aspetti positivi. Il principale, almeno dal punto di vista di un ricercatore statistico, è certamente la possibilità di lavorare con R in modalità interattiva, circostanza che rende l'analisi esplorativa dei dati davvero semplice ed agile. Altro vantaggio, percepibile soprattutto dall'utente programmatore, è costituito dalla ricca dotazione di strumenti avanzati per il debug. In R è, a titolo di esempio, possibile interrompere in un punto arbitrario il flusso di esecuzione di un programma, avere completo accesso allo stato corrente della memoria, alterare il valore degli oggetti in essa presenti ed effettuare le elaborazioni desiderate, far ripartire l'esecuzione del programma step-by-step o in modo non controllato, ispezionare la call-stack, etc.

- In R la gestione della memoria secondaria non è ottimizzata.

Al contrario del SAS, R *non* dispone di un sofisticato sistema di gestione dello swapping su memoria secondaria: lo delega totalmente al sistema operativo. Per conseguenza, se un processo R usa più memoria della RAM disponibile, tipicamente diventa molto lento. Inoltre una sessione R su piattaforma Windows a 32 bit non può ottenere dal sistema operativo più di 3GB di memoria allocabile.

- R è un linguaggio funzionale.

Il costrutto base del linguaggio è la funzione: una unità di programma che associa ad un insieme di valori per gli argomenti un unico valore di ritorno. In R le funzioni sono first-class objects: sono, cioè, trattate alla stregua delle strutture dati ordinarie (di tipo integer, real, character, etc.) comuni a tutti i linguaggi di programmazione. In particolare, in R esiste un tipo di dato i cui valori sono funzioni (cioè le funzioni possono comparire a destra di una istruzione di assegnazione) ed oggetti funzione possono essere argomento o valore di ritorno di altre funzioni.
Vale la pena specificare come, sebbene aderisca al paradigma funzionale, R disponga sia delle istruzioni tipiche dei linguaggi imperativi (assegnazione ed istruzioni condizionali e di ciclo), sia di strumenti in grado di "simulare" uno stile di programmazione object-oriented (i metodi S3 e S4).

2.2 Il Package Survey: la Struttura e le Principali Funzioni

Il package di R *survey* (la cui home page è: http://faculty.washington.edu/tlumley/survey/) consente di effettuare analisi di dati risultanti da indagini campionarie, anche caratterizzate da disegni di campionamento complessi. L'autore del package, Thomas

Lumley, è professore associato dell'Università di Washington e membro del "Core Development Team" di R; nell'ambito di quest'ultima attività, il package survey costituisce il suo principale progetto. La versione attuale del package è la 3.6 e contiene poco più di 6.000 linee di codice R[1].
La filosofia di implementazione del package ben supporta i principi di incapsulamento ed interfacciamento esplicito propri di una buona progettazione del software. L'incapsulamento è realizzato mediante l'introduzione di "oggetti" di tipo specifico, su cui le analisi statistiche devono essere condotte. In tal modo, gli utilizzatori degli oggetti possono accedere solo a quanto è loro necessario, mentre non hanno alcuna visione dei dettagli non utili. L'interfacciamento esplicito è favorito, tra gli altri elementi, dalla scelta di specificare le variabili di interesse per l'analisi mediante "formule" fornite come parametri alle funzioni, insieme ad ulteriori informazioni di interesse per il disegno.

In particolare, il package si può pensare strutturato in due principali insiemi di funzioni:

- Funzioni di creazione e manipolazione degli oggetti specifici del package. Tali oggetti sono costituiti dall'unione dei dati campionari rilevati, rappresentati come *data frame*[2], e dei metadati che descrivono la strategia di campionamento adottata ed hanno lo scopo di abilitare e guidare le successive analisi e computazioni. Tra le funzioni che svolgono il ruolo di "costruttori" di oggetti, le principali sono la funzione `svydesign` e la funzione `svrepdesign` che creano, rispettivamente, oggetti di classe `survey.design` e `svyrep.design`. La differenza sostanziale fra le due classi risiede nella tecnica di calcolo degli errori standard associati agli stimatori costruiti sui relativi oggetti: per la classe `survey.design` è utilizzato il metodo della linearizzazione di Taylor, per la classe `svyrep.design` sono supportati diversi metodi di replicazione del campione (ad esempio Jackknife, bootstrap, BRR). Poiché gli oggetti delle due classi sono corredati dei metadati di disegno, la loro manipolazione necessita di una gestione specifica per i metadati. Ad esempio, per disegni di campionamento complessi, il calcolo dell'errore standard di uno stimatore in una sottopopolazione non dipende solo dai dati che costituiscono la sottopopolazione stessa. Il package gestisce correttamente, ed in maniera trasparente per l'utente, aspetti di questo tipo ed a tal fine prevede la definizione di metodi per l'accesso (`"["`), la selezione di sottopopolazioni (`subset`), la modifica di valori (`update`), e la gestione dei valori nulli (`na.action`).
- Funzioni di elaborazione ed analisi statistica che operano sugli oggetti specifici del package. Il package rende disponibili, fra le altre, funzioni per il calcolo dei pesi di calibrazione (`calibrate`, `rake` e `postStratify`), per la stima di medie (`svymean`), totali (`svytotal`), rapporti (`svyratio`), quantili (`svyquantile`) e dei relativi errori standard e design effect, funzioni per la stima della varianza campionaria (`svrVar` e

[1] Nella Sezione 2.1 è stato fatto cenno alla "economia di codice" caratteristica della programmazione R. A tale proposito vale la pena osservare come il software *VPLX* (Variance Estimation for Complex Samples) del U.S Census Boureau, che implementa un *sottoinsieme* delle funzionalità rese disponibili dal package survey di R, contiene circa 250.000 linee di codice Fortran.
[2] Il data frame è la struttura dati di R deputata a rappresentare le "matrici dei dati" (in cui ad ogni riga corrisponde una osservazione e ad ogni colonna una variabile).

`svyrecvar`), per la tabulazione (`svytable`) e per la rappresentazione grafica (`svyplot`).

Il package survey è realizzato secondo il sistema ad oggetti di R noto come S3 System. Il sistema ad oggetti S3 prevede la definizione di molte delle funzioni del package come metodi (`S3method`) che operano su oggetti del package. In virtù di tale dispositivo, ed in modo trasparente per l'utente, la richiesta di una generica analisi statistica su un determinato oggetto viene "smistata" in base alla classe dell'oggetto; l'elaborazione che ne consegue è quella specifica della classe identificata. Ad esempio, il codice eseguito a valle dell'invocazione della funzione `svymean` su un oggetto di classe `svyrep.design`, sarà quello appropriato alla metodologia di replicazione del campione. Un'ulteriore caratteristica del package è l'utilizzo dei *namespace*. Il meccanismo dei namespace, recentemente introdotto nei package R, consente di definire cosa è privato, cioè non visibile, e cosa è invece pubblico, cioè visibile ai clienti del package. Le funzioni pubbliche e private del package survey (come per gli altri package R per cui sono utilizzati i namespace) sono definite nel file di testo "NAMESPACE", che è presente nella directory di installazione del package.

3 La Calibrazione

La presente sezione è dedicata alla descrizione generale del problema di calibrazione (Sezione 3.1) ed all'esposizione delle tecniche risolutive adottate da survey/R (Sezione 3.2) e da GENESEES/SAS (Sezione 3.3). La Sezione conclusiva 3.4 riporta una analisi dei punti di contatto e delle differenze tra i due software.

3.1 Formulazione del Problema

Nelle indagini campionarie un importante concetto è il *peso* da attribuire alle unità del campione affinché meglio rappresentino la popolazione di interesse. Un modo comune di descrivere il significato dei pesi è quello di considerare che una unità campionaria che ha peso p rappresenta se stessa ed altre $p-1$ unità della popolazione (non selezionate nel campione). Nel calcolo della stima dei parametri della popolazione di interesse (ad esempio medie o totali di variabili rilevate), i valori osservati per ciascuna unità campionaria contribuiscono, pertanto, secondo il peso associato all'unità stessa.
I cosiddetti pesi *diretti* sono definiti come inversi delle probabilità di inclusione, vale a dire le probabilità note con le quali le unità sono selezionate sulla base di uno specifico disegno di campionamento.
Nel caso dello stimatore di *Horvitz-Thompson* [Horvitz-and-Thompson,1952], i pesi utilizzati sono semplicemente i pesi diretti. Indicando con s il campione, con d_k il peso diretto della k-esima unità campionaria e con y_k il valore su essa osservato per una variabile di interesse y, l'espressione dello stimatore di Horvitz-Thompson del totale di y sulla popolazione è la seguente:

$$\hat{Y}_{HT} = \sum_{k \in s} d_k y_k \qquad (1)$$

Lo stimatore di Horvitz-Thompson è uno stimatore corretto. Nondimeno, l'accuratezza delle stime che esso produce, in termini di errore quadratico medio, può essere migliorata mediante l'utilizzo di stimatori alternativi. Gli stimatori di *calibrazione* (o di *riponderazione*), solo asintoticamente corretti, garantiscono tale risultato in virtù di un notevole guadagno di efficienza [Särndal-et-al,1992] [Singh-and-Mohl-1996]. Essi richiedono, tuttavia, un sistema di calcolo dei pesi più complesso. Per la precisione, i pesi *finali*[3] si ottengono risolvendo un problema di minimo vincolato, in cui la funzione da minimizzare è una funzione di distanza G tra i pesi diretti d_k e i pesi finali w_k, e i vincoli sono rappresentati da condizioni di uguaglianza delle stime campionarie di *variabili ausiliarie* con i rispettivi *totali noti* della popolazione (desunti da fonti esterne all'indagine):

$$\begin{cases} \min \sum_{k \in s} G(d_k, w_k) \\ \sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X} \end{cases} \qquad (2)$$

Nell'equazione precedente è stato indicato con $\mathbf{X}$ il vettore dei totali noti e con $\mathbf{x}_k$ il vettore delle variabili ausiliare osservate sulla k-esima unità campionaria.
Al problema formulato nell'equazione (2) vengono, molto spesso, aggiunti ulteriori vincoli, detti *vincoli di range*, del tipo seguente:

$$L \le \frac{w_k}{d_k} \le U \quad k \in s \qquad \text{con} \qquad 0 \le L \le 1 \le U \qquad (3)$$

Attraverso il sistema di disequazioni (3) è, infatti, possibile eliminare il rischio che i pesi finali, soluzione del problema (2), presentino valori patologici (ad esempio, negativi) o indesiderati (ad esempio, enormi rispetto ai pesi diretti).
Il complesso delle equazioni (2) e (3) è comunemente designato come *problema di calibrazione* [Vanderhoeft-2001]; l'espressione del corrispondente stimatore di calibrazione è formalmente analoga alla (1):

$$\hat{Y}_{CAL} = \sum_{k \in s} w_k y_k \qquad (4)$$

Lo stimatore di *regressione generalizzata* può essere riguardato come un caso particolare della famiglia degli stimatori di calibrazione: quello ottenuto utilizzando la distanza *euclidea*[4] [Särndal-et-al,1992]. E' possibile definire diversi stimatori di calibrazione in corrispondenza di diverse funzioni distanza, e la scelta dipende dalle esigenze della specifica indagine.

---

[3] Nella pratica Istat ci si riferisce comunemente ai pesi finali con la locuzione "pesi di riporto all'universo".
[4] Con un piccolo abuso di linguaggio, che costituisce una consuetudine in letteratura, la funzione di distanza euclidea sarà nel seguito indicata anche come "lineare": ad essere lineare è in realtà la *funzione di calibrazione* (l'inversa della derivata prima) ad essa associata.

3.2 La Calibrazione nel Package Survey

Come accennato nella Sezione 2.2, il package survey rende disponibili diverse funzioni per il calcolo dei pesi di calibrazione. Alla funzione `calibrate` è, in particolare, demandata la soluzione del problema di calibrazione nella sua formulazione più generale[5]. Prima di poter invocare la `calibrate`, è necessario creare un oggetto specifico mediante una delle funzioni "costruttore" – `svydesign` e `svyrepdesign` – introdotte nella Sezione 2.2. Poiché l'approfondimento dei metodi di replicazione del campione esula dalle finalità della sperimentazione effettuata, ci limiteremo, in quanto segue, a descrivere i principali parametri di input della funzione `svydesign`. Tali parametri sono tipicamente specificati in termini di *formule* (ottenibili utilizzando l'operatore infisso `~`) riferite alle colonne del data frame dei dati campionari, ed includono:

- `data`: identifica il data frame che contiene i dati campionari.
- `ids`: individua gli identificatori delle unità campionarie ed è, quindi, un parametro obbligatorio. E' possibile specificare un disegno di campionamento a più stadi semplicemente utilizzando una formula in cui compaiano gli identificatori delle unità selezionate nei singoli stadi. Ad esempio la scrittura `ids=~id_PSU+id_SSU` dichiara un campionamento a due stadi in cui le unità di primo stadio sono identificate dalla variabile `id_PSU` e quelle di secondo stadio dalla variabile `id_SSU`.
- `strata`: identifica le variabili di stratificazione. Il parametro è opzionale: per default il campione è ipotizzato non stratificato. In modo del tutto analogo a quanto illustrato per `ids`, è possibile definire disegni di campionamento a più stadi con stratificazione: basterà far comparire, nella formula legata al parametro `strata`, le variabili che identificano gli strati in ciascuno stadio.
- `weights`: specifica i pesi diretti, ed è anch'esso opzionale.
- `probs`: indica le probabilità di inclusione nel campione e, dunque, costituisce una alternativa ai pesi diretti.
- `fpc`: il parametro opzionale `fpc` (finite population correction) consente di ricostruire il fattore di correzione che occorre includere nel calcolo della varianza quando le unità selezionate nel campione siano una frazione apprezzabile della popolazione obiettivo. Può essere specificato sia come totale della popolazione in ogni strato, sia come frazione della popolazione campionata in ogni strato. Ad esempio, con un campionamento di 100 unità da uno strato di dimensione 500, `fpc` può valere 500, oppure 100/500=0,2. Se `fpc` è specificato, ma non lo è il parametro `probs` (o `weights`), le probabilità di campionamento sono calcolate a partire dai valori di `fpc`, assumendo una strategia di campionamento casuale semplice entro ogni strato. Nel caso di campionamento a più stadi, in fpc dovrà comparire il numero complessivo di unità di ciascuno stadio (o la corrispondente frazione di campionamento).

L'oggetto di classe `survey.design` creato dall'invocazione di `svydesign` contiene sia i dati che i metadati di disegno ed è uno dei parametri di input della funzione `calibrate`.

---

[5] Contrariamente alle funzioni `rake` e `postStratify`, che gestiscono unicamente variabili ausiliarie di tipo *qualitativo*, la `calibrate` consente l'utilizzo di variabili ausiliarie sia qualitative che *quantitative* (eventualmente continue).

La `calibrate` implementa le funzioni di distanza lineare, lineare troncata, raking, raking troncata, logit (=logaritmica) e logit troncata. Per tutte le distanze il problema di minimo vincolato è risolto con il metodo dei moltiplicatori di Lagrange, utilizzando (con l'unica eccezione della lineare non troncata) l'algoritmo di Newton-Raphson [Deville-et-al, 1993].
I principali parametri formali della funzione sono di seguito elencati e commentati[6]:

- `design`: rappresenta l'oggetto di classe `survey.design` su cui il processo di calibrazione deve operare.
- `formula`: specifica simbolicamente, ed in modo univoco, il modello di calibrazione[7] che si intende utilizzare. Il modello di calibrazione identifica le variabili ausiliarie ed esplicita la struttura dei vincoli. Ad esempio la scrittura `formula=~(`X_i`+`X_j`):C` definisce il problema di calibrazione in cui i vincoli vengono posti sui totali delle variabili ausiliarie (quantitative) X_i ed X_j all'interno delle sottopopolazioni identificate dalle diverse modalità della variabile di classificazione (qualitativa) `C`.
- `population`: specifica il vettore dei totali noti per le variabili ausiliarie secondo il modello di calibrazione definito da `formula`. Se il disegno di campionamento prevede grappoli, il parametro attuale di `population` dovrà essere la lista dei vettori dei totali per ogni grappolo.
- `bounds`: consente di includere nel problema di calibrazione *vincoli di range* sui pesi finali. Il parametro attuale di `bounds` deve essere un vettore numerico a due componenti. Ad esempio la scrittura `bounds=c(L,U)` equivale al sistema di disequazioni:

  $$L \le \frac{w_k}{d_k} \le U \qquad k \in s$$

  Il valore di default è `c(-Inf,Inf)` e corrisponde ad assenza di *vincoli di range*.
- `calfun`: il parametro attuale corrispondente a `calfun` deve essere una stringa che specifica la funzione di distanza. I tre valori possibili sono `"linear"`, `"raking"`, e `"logit"`.
- `maxit`: determina il numero massimo di iterazioni dell'algoritmo di Newton-Raphson cui è demandata la soluzione del problema di calibrazione. Il valore di default del parametro `maxit` è 50.
- `epsilon`: individua la massima differenza tollerata fra una stima campionaria ed il corrispondente totale noto. Il valore di default del parametro `epsilon` è 10^{-7}. I valori di `epsilon` e `maxit` costituiscono, dunque, i criteri di convergenza ed arresto dell'algoritmo iterativo alla base della funzione `calibrate`.

L'output della `calibrate` è ancora un oggetto di classe `survey.design`, ma contenente i pesi finali (accessibili, ad esempio, mediante la funzione `weights`).

---

[6] I parametri sono illustrati con particolare riferimento all'utilizzo della `calibrate` nell'ambito della sperimentazione. Sono possibili utilizzi più generali, per i quali si rimanda al manuale del package survey [Manuale-Survey,2007].

[7] Nel linguaggio dello stimatore di regressione generalizzata è il modello di regressione lineare soggiacente al problema di calibrazione. Si veda [Wilkinson-and-Rogers,1973].

3.3 La Calibrazione in GENESEES

L'interazione con il software generalizzato GENESEES avviene mediante un ambiente a finestre, che consente la specifica interattiva dei parametri di input alle varie funzionalità rese disponibili dal software. In particolare, il processo di calibrazione in GENESEES è realizzato dalla *funzione di riponderazione*; i principali passi necessari all'esecuzione di tale funzione sono[8]:

- Selezione del data set SAS contenente i dati campionari.
- Selezione del data set SAS contenente i totali noti.
- Per il data set dei dati campionari sono specificati i seguenti parametri di input: (i) popolazioni pianificate usate per lo stimatore, (ii) codice identificativo dell'unità campionaria, (iii) variabili ausiliarie, (iv) peso diretto, (v) peso distanza[9] (opzionale) e (vi) funzione di distanza. Il parametro (i) identifica una partizione della popolazione obiettivo in sottopopolazioni per le quali siano noti i totali delle variabili ausiliarie specificate in (iii)[10]. Le funzioni di distanza supportate dal software GENESEES sono: la lineare e la lineare troncata, la logaritmica e la logaritmica troncata, la funzione di Hellinger, la funzione di minima entropia, e la funzione Chi-quadrato.
- Per il data set dei totali noti sono specificati: (i) popolazioni pianificate usate per lo stimatore, e (ii) totali noti delle variabili ausiliarie.

I pesi finali calcolati da GENESEES sono resi accessibili all'utente attraverso appositi data set SAS, oltre ad esser riportati in una colonna supplementare, di nome *coeffin*, del data set dei dati campionari. In aggiunta, GENESEES produce alcuni data set per la gestione di condizioni di errore eventualmente rilevate sugli input, nonché tavole di sintesi contenenti statistiche relative alle distribuzioni dei pesi diretti e dei pesi finali ed alle stime dirette e finali dei totali delle variabili ausiliarie.

3.4 La Calibrazione nel Package Survey ed in GENESEES: Analogie e Differenze

Prima di illustrare il confronto sperimentale tra la calibrazione realizzata tramite le funzioni del package survey e la calibrazione in GENESEES, descriviamo in questa sezione aspetti condivisi e differenze tra i due software.

Come accennato nelle precedenti sezioni la funzione di riponderazione di GENESEES "corrisponde" alla funzione `calibrate` del package survey.

---

[8] Come per l'illustrazione della `calibrate`, anche nel caso dell'illustrazione del software GENESEES ci si atterrà all'utilizzo che di tale software è stato fatto nel contesto della sperimentazione condotta. Per una trattazione generale si rimanda al manuale di GENESEES [Manuale-GENESEES,2005].

[9] Il peso distanza è un fattore correttivo del peso diretto, ad esso non correlato. Nella maggioranza delle applicazioni si assume che i pesi distanza siano pari ad 1. L'utilizzo di pesi distanza variabili può essere appropriato per risolvere particolari problemi di sottocopertura [Manuale-GENESEES,2005].

[10] In realtà, qualora il disegno di campionamento sia *stratificato*, occorre che le sottopopolazioni della partizione siano ottenibili aggregando uno o più strati: di qui l'espressione "popolazioni pianificate". Si noti che le informazioni fornite dai parametri (i) e (iii) equivalgono a quelle contenute nel parametro `formula` della funzione `calibrate` del package survey di R.

Una prima differenza riguarda il progetto delle due funzioni. Nel caso di GENESEES, tutti i parametri necessari al processo di calibrazione sono forniti in input alla funzione in un'unica soluzione. Nel caso della `calibrate`, al contrario, è dapprima necessario costruire un oggetto `survey.design`, integrando i dati di indagine con le informazioni relative al disegno di campionamento; solo a questo punto è possibile invocare la `calibrate` fornendole in input l'oggetto costruito. Con riferimento alla qualità della progettazione, è preferibile la soluzione di R che isola due funzionalità indipendenti favorendo la coesione dei moduli funzionali realizzati.
Come si evince dalle Sezioni 3.2 e 3.3, il numero di funzioni di distanza implementate da GENESEES è superiore a quello attualmente supportato dalla `calibrate`. Tuttavia, la `calibrate` rende disponibili tutte le funzioni di distanza effettivamente utilizzate nelle indagini concrete su larga scala.
Come la funzione di riponderazione di GENESEES, anche la `calibrate` risolve il problema di minimo vincolato con il metodo dei moltiplicatori di Lagrange, utilizzando (con l'unica eccezione della lineare non troncata) l'algoritmo di Newton-Raphson. Nondimeno i due software differiscono nei criteri di arresto e convergenza adottati. In GENESEES il parametro che controlla il numero massimo di iterazioni effettuate dall'algoritmo è incorporato nel codice e *non* è, dunque, modificabile dall'utente. Il valore prescelto è pari a 30 iterazioni. Il parametro `maxit` della calibrate (cfr. Sezione 3.2) può, invece, essere alterato liberamente e presenta un valore di default di 50 iterazioni.
Più interessante è la differenza fra i criteri di convergenza. In GENESEES il processo iterativo è interrotto quando la massima differenza relativa (in valore assoluto) fra i moltiplicatori di Lagrange calcolati nell'iterazione corrente ed in quella precedente scende al di sotto di una certa soglia. Tale valore di soglia, fissato a 10^{-8}, è, ancora una volta, incorporato nel codice e *non* può essere alterato dall'utente. Semplificando, si può dire che l'elaborazione in GENESEES viene arrestata quando i pesi "finali"[11], calcolati nelle successive iterazioni, smettono di cambiare apprezzabilmente. Dunque, contrariamente a quanto accade per la funzione `calibrate` (cfr. Sezione 3.2), il criterio di convergenza di GENESEES non garantisce, di per sé, che vi sia accordo (entro una data tolleranza) fra le stime campionaria calcolate con i pesi finali ed i corrispondenti totali noti[12].
La differenza di maggior rilievo fra la funzione `calibrate` di R e GENESEES risiede nel trattamento dei vincoli di range (cfr. Sezione 3.1). Nel caso della `calibrate` l'utente è chiamato a fornire, mediante il parametro formale `bounds`, il limite inferiore (L) e superiore (U) dell'intervallo di variazione *effettivamente* accessibile ai rapporti fra pesi finali e diretti. Il punto debole di questa soluzione risiede nel fatto che l'utente, non essendo supportato dal software nella scelta, può specificare valori di L ed U che rendono *incompatibile* il problema di calibrazione. In tal caso la funzione `calibrate`, che è ovviamente destinata a non convergere, spreca tempo di elaborazione, in quanto effettua comunque `maxit` iterazioni. In più, ad elaborazione conclusa, la funzione `calibrate` si

[11] Infatti i pesi "finali", calcolati ad una data iterazione, dipendono (attraverso la *funzione di calibrazione*, cfr. nota 4) dai moltiplicatori di Lagrange (oltre che, ovviamente, dai pesi "diretti").
[12] In GENESEES è, in ogni caso, possibile analizzare *a posteriori* la qualità dell'accordo fra le stime ed i totali noti mediante un apposito report prodotto in output.

limita ad informare l'utente della mancata convergenza, senza indicarne la causa (vale a dire la scelta "patologica" dei `bounds`).
La soluzione offerta da GENESEES, più robusta ed efficiente, è certamente preferibile. Il software calcola preliminarmente, ed in modo del tutto trasparente all'utente, il limite inferiore (L_{max}) e superiore (U_{min}) del *minimo intervallo teorico* in cui è possibile vincolare i rapporti fra pesi finali e diretti se si richiede che il problema di calibrazione ammetta soluzione. Solo a questo punto l'utente è chiamato a specificare due moltiplicatori ($\alpha_L \leq 1$ e $\alpha_U \geq 1$, con default $\alpha_L=0,5$ e $\alpha_U=1,5$) che fungono da coefficienti di espansione dell'intervallo minimo teorico. Con questo accorgimento GENESEES garantisce che l'intervallo effettivo scelto dall'utente (delimitato da $L=\alpha_L \cdot L_{max}$ e $U=\alpha_U \cdot U_{min}$) contenga l'intervallo minimo teorico e conduca, pertanto, ad un problema sempre *compatibile*.

4 Survey/R e GENESEES/SAS: Confronto Sperimentale tra le Funzioni di Calibrazione

Questa sezione descrive la sperimentazione effettuata ed i principali obiettivi che essa ha raggiunto: (i) acquisire familiarità con R ed, in particolare, con il package survey; (ii) testare la possibilità *concreta* – nel caso realistico di una delle maggiori indagini correnti dell'Istat: l'indagine sulle Forze di Lavoro – di utilizzare R nel processo di calibrazione dei dati; (iii) valutare l'efficacia e l'efficienza delle funzioni del package survey mediante un confronto con i risultati ottenuti da GENESEES.
La Sezione 4.1 illustra il metodo e l'articolazione di massima della sperimentazione. La Sezione 4.2 introduce l'indagine sulle Forze di Lavoro, descrive la struttura e le dimensioni dei data set utilizzati e la dotazione software e hardware dell'ambiente informatico di sperimentazione. Infine, la Sezione 4.3 analizza, sia in termini di efficacia che di efficienza, i risultati ottenuti utilizzando il package survey per la calibrazione.

4.1 Fasi della Sperimentazione

La sperimentazione condotta si è articolata in una sequenza temporale di fasi che sono brevemente illustrate nella presente sezione. L'ultima fase ha previsto il confronto sperimentale tra GENESEES/SAS e Survey/R, ed è descritta in dettaglio nella Sezione 4.3.

- Fase 1: Studio del package survey ed individuazione delle funzioni di interesse. Questa fase ha sostanzialmente previsto lo studio della documentazione relativa al package survey. Gli strumenti rilevanti per il calcolo dei pesi di calibrazione sono stati individuati nella funzione `calibrate` e nella funzione `svydesign`, necessaria alla costruzione dell'oggetto `survey.design` input della `calibrate` (si veda la Sezione 2.2).
- Fase 2: Test delle funzioni individuate nella fase 1 su data set "controllati". Questa fase è stata condotta con l'obiettivo di comprendere pienamente il funzionamento della `svydesign` e della `calibrate`, posponendo ogni considerazione relativa alla loro efficienza su data set realistici. Nello specifico, le due funzioni sono state testate su data set di prova, prendendo in esame diverse configurazioni dei parametri di input. Alcune considerazioni interessanti emerse dal test sono riportate nella Sezione 4.3.1. E'

stato, infine, effettuato un primo confronto tra i risultati di GENESEES ed i risultati della `calibrate`, con lo scopo principale di orientare la sperimentazione effettiva che è stata realizzata nella fase 4.

- Fase 3: Identificazione delle analogie e delle differenze tra GENESEES e la `calibrate`. Questa fase ha consentito la sistematizzazione e l'analisi dei punti di contatto e di differenza tra la funzione di riponderazione di GENESEES e la funzione `calibrate` del package survey; l'esito è stato presentato in dettaglio nella precedente Sezione 3.4. Lo studio è stato utile anche per progettare nella maniera più opportuna la sperimentazione condotta nella fase 4. Ad esempio, a causa della diversa gestione dei vincoli di range, si è resa necessaria l'implementazione di un modulo R che "replicasse" il metodo di calcolo dei bounds di GENESEES: solo in questo modo è stato, infatti, possibile dare senso al test di efficacia cui i due software sono stati sottoposti nella fase 4.
- Fase 4: Test di efficacia ed efficienza di GENESEES e `calibrate` sull'indagine Forze di Lavoro. Questa fase è descritta in maniera particolareggiata nelle sezioni successive, ed è stata la più onerosa del processo di sperimentazione. Anticipiamo che i risultati conseguiti sull'indagine sulle Forze di Lavoro – una rilevazione in certo senso "critica" *sia* in termini di dimensioni, *sia* per la complessità del disegno di campionamento e del problema di calibrazione – oltre a dimostrare la possibilità concreta di usare il package survey nella pratica Istat, hanno permesso l'individuazione di criticità e punti di forza generali del linguaggio R.

4.2 I Data Set dell'Indagine Forze di Lavoro e l'Ambiente di Sperimentazione

Questa sezione fornisce alcuni cenni introduttivi sull'indagine Forze Lavoro in generale ed, in particolare, sul disegno di campionamento e sulla struttura del problema di calibrazione (Sezione 4.2.1); per ulteriori dettagli è possibile consultare [FOL-met-norm,2006]. La Sezione 4.2.2 illustra, infine, le proprietà dei data set utilizzati nella sperimentazione e le caratteristiche dell'ambiente in cui essa è stata effettuata.

4.2.1 Cenni sull'Indagine Forze di Lavoro

L'indagine sulle Forze di Lavoro è una delle maggiori rilevazioni campionarie effettuate dall'Istat. Il principale obiettivo dell'indagine è la produzione e la diffusione delle stime ufficiali relative al numero di occupati, di persone in cerca di occupazione e di persone inattive residenti in Italia[13]. La rilevazione è stata svolta senza interruzioni a partire dal 1959. Nel gennaio 2004 è stata completata la transizione (iniziata a gennaio 2003) dalla "tradizionale" versione *trimestrale* dell'indagine, effettuata in una specifica settimana per ciascun trimestre, alla "nuova" versione *continua*, distribuita su tutte le settimane dell'anno.

[13] Per la precisione la popolazione obiettivo dell'indagine si limita ai componenti delle *famiglie* residenti in Italia (esclude, cioè, i membri permanenti delle *convivenze*).

Il disegno di campionamento è del tipo a due stadi con stratificazione delle unità di primo stadio. Le unità di primo stadio sono i comuni e le unità di secondo stadio sono le famiglie anagrafiche. I comuni sono suddivisi in strati omogenei in base all'ampiezza demografica; gli strati così determinati possono essere distinti in due categorie: strati auto-rappresentativi (AR), costituiti da un solo comune che viene incluso nel campione con certezza, e strati non auto-rappresentativi (NAR), contenenti più di un comune. Da ogni strato NAR viene estratto un solo comune con probabilità proporzionale alla dimensione demografica. Da ciascun comune campione viene, infine, selezionato un campione casuale semplice di famiglie e tutti gli individui appartenenti alle famiglie estratte vengono intervistati (campionamento a grappoli). Complessivamente, in un trimestre, l'indagine coinvolge circa 1.200 comuni campione (di cui circa 350 AR) e circa 75.000 famiglie, per un totale di circa 200.000 individui.
Le stime prodotte dall'indagine sulle Forze di Lavoro sono ottenute ricorrendo allo stimatore di calibrazione associato alla distanza logaritmica troncata. Il corrispondente problema di calibrazione viene risolto utilizzando il software GENESEES. Deve essere, a questo punto, sottolineato come i pesi finali vengano calcolati a livello *familiare*: a tutti gli individui di una famiglia viene attribuito lo *stesso* peso finale.
I principali vincoli imposti nel problema di calibrazione sono relativi ai seguenti totali noti:

1. popolazione residente per regione[14], sesso e 14 classi d'età;
2. popolazione residente per provincia, sesso e 5 classi d'età;
3. popolazione residente nei 13 grandi comuni per sesso e 5 classi d'età.

Ulteriori vincoli riguardano il totale dei cittadini stranieri per regione, sesso e nazionalità (UE o non UE), il numero di famiglie per gruppo di rotazione[15] e la popolazione mensile per sesso.
Nella prassi corrente i data set SAS da fornire in input alla funzione riponderazione di GENESEES (si veda la Sezione 3.3) vengono preparati in modo che le "popolazioni pianificate per lo stimatore" siano le 21 popolazioni regionali e le variabili ausiliarie siano numeriche. Il data set dei dati campionari contiene un record per famiglia. Le 182 variabili ausiliarie che in esso compaiono specificano quanti membri di ciascuna famiglia "cadano" in una delle sottopopolazioni di cui sono noti i totali. La struttura dei totali noti 1. e 2., d'altro canto, fa si che non tutte le 182 variabili ausiliarie utilizzate siano linearmente indipendenti[16].

---

[14] Le province autonome di Trento e Bolzano sono trattate separatamente: complessivamente le "regioni" sono, dunque, 21.

[15] Il disegno di rilevazione della nuova indagine continua è del tipo a "panel ruotato". Lo schema di rotazione è del tipo 2 – 2 – 2: ogni famiglia è inclusa nel campione per due rilevazioni successive, esce dal campione nei due trimestri seguenti e, infine, viene reinserita nel campione per altre due rilevazioni. Complessivamente, in ogni trimestre, sono presenti nel campione 4 "gruppi di rotazione" costituiti da famiglie che sono alla 1°, 2°, 3° e 4° intervista.

[16] Infatti la classificazione dell'età relativa al totale noto (2) è ottenuta per aggregazione di categorie della corrispondente classificazione in (1) e, ovviamente, la popolazione di una regione è pari alla somma delle popolazioni delle province che la compongono.

4.2.2 Ambiente di Sperimentazione

Come è stato accennato nelle Sezioni 1 e 4.1, la sperimentazione del package survey di R ed il confronto con GENESEES/SAS sono stati condotti utilizzando i dati rilevati nell'indagine sulle Forze di Lavoro. Fra gli obiettivi programmatici della sperimentazione è stata già menzionata la valutazione delle potenzialità di R nel gestire elaborazioni di elevata complessità computazionale su consistenti moli di dati. Per documentare l'esito di tale studio occorre, in via preliminare, descrivere le dimensioni dei data set utilizzati, nonché la dotazione hardware e software degli elaboratori di cui ci si è avvalsi.
Sia la funzione di riponderazione di GENESEES che la funzione `calibrate` del package survey sono state eseguite sul data set[17] dei record *familiari* dell'indagine sulle Forze di Lavoro. Tale data set, relativo al primo trimestre 2005, è composto da 70.150 righe e 192 colonne (di cui 182 relative alle variabili ausiliarie).
La sperimentazione è stata condotta sia in ambiente PC che in ambiente server. Le caratteristiche rilevanti dei due ambienti sono schematicamente riportate in quanto segue:

- PC
  - sistema operativo Windows XP
  - 760 MB di RAM
  - CPU da 3 GHz
- server
  - sistema operativo Linux
  - 10 GB di RAM
  - 4 CPU da 2 GHz

E' necessario precisare come, nel contesto della sperimentazione, le elaborazioni effettuate sul server abbiano, in realtà, impegnato uno solo dei 4 processori disponibili[18].

4.3 Risultati della Sperimentazione: Test di Efficacia e di Efficienza

In questa sezione descriviamo in dettaglio la sperimentazione effettuata. Dapprima (Sezione 4.3.1) illustriamo i principali problemi incontrati nell'utilizzo delle funzioni del package, accennando alle soluzioni adottate. Successivamente (Sezione 4.3.2) descriviamo i risultati sperimentali del confronto tra Survey/R e GENESEES/SAS sia in termini di efficacia che di efficienza.

4.3.1 Utilizzo delle Funzioni di Survey: Problemi e Soluzioni

Come ogni altro package di R, il package survey è corredato di un manuale on line che ne costituisce la principale fonte di documentazione [Manuale-Survey,2007].

---

[17] In R la struttura dati corrispondente al data set SAS è il *data frame*. Per semplicità, in quanto segue, ci riferiremo sempre a data set, considerando la distinzione sottointesa.
[18] Se i programmi non sono appositamente parallelizzati – ed è questo il caso della nostra sperimentazione – la natura multiprocessore del server interviene *solo* nell'ottimizzare la gestione di elaborazioni concorrenti richieste da utenti diversi.

Documentazione di supporto è disponibile nella pagina web dedicata al package (si veda la Sezione 2.2) a cura dell'autore Thomas Lumley.
L'approccio iniziale con il package è stato *difficoltoso,* soprattutto perché la filosofia di utilizzo, che abbiamo schematicamente delineato nella Sezione 2.2, non è sufficientemente esplicita nelle due fonti sopra citate. Ad esempio, pur essendo relativamente semplici la creazione di un oggetto di classe `survey.design` e la successiva invocazione della funzione `calibrate`, non sono chiaramente indicate le modalità di accesso ai pesi finali, che il package gestisce come *metadati* dell'oggetto calibrato. La documentazione di survey necessita dunque di un'integrazione, in particolare in riferimento agli oggetti specifici del package ed alle funzioni di accesso ai metadati di corredo.
Un secondo problema, ancora ascrivibile ad un difetto di documentazione, è stato riscontrato nell'utilizzo della funzione `calibrate` con distanza *lineare non troncata*, su dati campionari che presentino una situazione di *dipendenza lineare* delle variabili ausiliarie. Come accennato nella Sezione 4.2.1, la collinearità delle variabili ausiliarie interessa anche l'indagine sulle Forze di Lavoro (per la quale, in ogni caso, è prassi consolidata usare la distanza logaritmica troncata). Il punto è che, nel *solo* caso della distanza lineare non troncata, l'algoritmo risolutivo adottato dalla `calibrate` non è quello consueto di Newton-Raphson, di cui è nota l'insensibilità alla dipendenza lineare. Al contrario, nel caso in questione, la `calibrate` ricorre ad una programma Fortran (di cui richiama, fra l'altro, la versione *compilata*) che va in errore in caso di collinearità.
La documentazione del package survey non fa menzione di questo limite della `calibrate`, e, circostanza anche più grave, la funzione `calibrate` non fornisce all'utente alcuna diagnosi dell'errore, quando esso si verifica.
Generalizzando, il codice del package ha richiesto un test accurato (che nel caso specifico è stato effettuato per tutte le funzioni di distanza disponibili) che ha rilevato la presenza di pre-condizioni non esplicitate per l'esecuzione delle funzioni con particolari parametri.
Un'ulteriore difficoltà ha riguardato la creazione di oggetti di classe `survey.design` per data set risultanti da disegni di campionamento complessi, quale quello della rilevazione sulle Forze di Lavoro. Sebbene l'ostacolo in questione si sia manifestato in un contesto – il calcolo degli errori standard delle stime – che non è oggetto del presente documento, vale certamente la pena accennare alla natura del problema: rischi analoghi *possono*, infatti, interessare qualsiasi package di R.
Nello specifico, l'invocazione della funzione `svydesign` sul data set dei dati *individuali* di Forze di Lavoro (circa 200.000 record), con la definizione dell'appropriato disegno di campionamento (si veda la Sezione 4.2.1), ha comportato tempi di esecuzione inaccettabili.
Nella Sezione 2.1 è stato sottolineato che, se un processo R usa più memoria della RAM disponibile, tipicamente diventa molto lento. Poiché i primi passi della sperimentazione sono stati condotti in ambiente PC, è apparso inizialmente naturale imputare la lentezza della funzione `svydesign` alla gestione non ottimizzata della memoria secondaria di R. Questa "falsa pista" è stata immediatamente abbandonata dopo aver studiato il comportamento dell'applicazione in ambiente server. Anche sul server, infatti, i tempi di

elaborazione risultavano enormi[19], nonostante la macchina possedesse una RAM *superiore* alla memoria complessivamente richiesta dall'applicazione. Questa osservazione ci ha spinto ad "ispezionare" accuratamente il codice della funzione `svydesign`, nell'intento di isolarne le componenti più onerose in termini di tempo di esecuzione. Il contributo dominante al tempo di elaborazione è risultato provenire da `as.fpc`, una funzione privata invocata da `svydesign`. Lo studio del codice di `as.fpc` ci ha, infine, consentito, mediante la modifica di *una sola istruzione* del programma[20], di risolvere definitivamente il problema della patologica inefficienza di `svydesign`. Per la precisione, ricorrendo alla versione ottimizzata della funzione `svydesign` è stato possibile costruire l'oggetto `survey.design` in circa 20 secondi (sia in ambiente PC che server).
Pertanto, come considerazione generale, è necessario rimarcare che il codice reso disponibile dai package di R può *non essere ottimizzato*, e causare quindi un utilizzo delle risorse eccessivo rispetto alle necessità effettive.
I problemi maggiori, però, sono stati riscontrati nell'uso della funzione `calibrate`. In ambiente server, i primi tentativi di invocazione della `calibrate` – sull'oggetto `survey.design` costruito a partire dal data set delle famiglie di Forze di Lavoro – non sono andati a buon fine. Il comportamento dell'applicazione era scoraggiante: dopo esser stata attiva per circa *4 giorni* ed aver compiuto le 50 iterazioni di default, la `calibrate` terminava *senza convergere*. Tale circostanza appariva preoccupante per due ragioni concettualmente distinte: (i) la mancata convergenza dell'algoritmo di Newton-Raphson, nonostante il problema di calibrazione fosse certamente *compatibile* ed il numero di iterazioni effettuate congruo; (ii) l'estrema lentezza dell'applicazione, nonostante la cospicua dotazione di RAM e CPU del server.
Per individuare la causa della mancata convergenza è stato necessario, ancora una volta, procedere ad una analisi puntuale del codice della `calibrate`. Ciò ci ha condotto all'individuazione di un *bug* del programma: un errore commesso nel calcolo della derivata prima della funzione di calibrazione associata alla distanza logaritmica. L'errore in questione degradava a tal punto la velocità di convergenza dell'algoritmo di Newton-Raphson da renderlo, a tutti gli effetti, inutilizzabile.
La correzione del bug ha portato alla scrittura di una nuova funzione `calibrate` che è stata sostituita a quella originariamente fornita dal package survey. Tale sostituzione ha, fra l'altro, richiesto la ricompilazione dei sorgenti del package[21]. Il tempo di esecuzione della nuova `calibrate` su server è risultato di circa 2 ore e mezza, e la convergenza è stata raggiunta in 5 iterazioni. In ultimo, il bug individuato è stato segnalato all'autore di survey che ha provveduto a "fissarlo" nella successiva release del package. In quanto

---

[19] L'esecuzione della funzione `svydesign` su un sottoinsieme del campione di 50.000 record ha richiesto, sul server, circa 41 ore.
[20] Nello specifico si è trattato di convertire il tipo di una variabile da `factor` (il formato in cui R rappresenta le variabili qualitative) a `numeric`. Riguardo il problema dell'impatto delle conversioni di tipo sull'efficienza dei programmi R, è possibile consultare [S-Programming,2000].
[21] L'uso del meccanismo dei namespace nel package ha reso necessario ricompilare i sorgenti. L'alternativa di sostituire la `calibrate` con la `new_calibrate` "caricandola" all'occorrenza si sarebbe rivelata troppo dispendiosa.

segue, a meno che non sia esplicitamente indicato il contrario, ci riferiremo sempre a questa nuova versione corretta della funzione `calibrate`.
Come considerazione generale, è opportuno, dunque, tenere sempre presente la possibilità di funzionamenti *non corretti*, anche nel caso di package giunti a versioni apparentemente "consolidate". In linea con la filosofia open source di R, qualora sia stata individuata la causa di un errato funzionamento, è buona norma fornirne prontamente segnalazione allo sviluppatore del package, in modo che l'intera comunità degli utenti possa beneficiarne.
La correzione del bug, sebbene abbia velocizzato l'algoritmo risolutivo della `calibrate` (diminuendo il numero di iterazioni necessarie per ottenerne la convergenza), non si è rivelata sufficiente a risolvere alla radice il problema della "lentezza" della funzione. Anzi, lo studio del comportamento della `calibrate` in ambiente PC ha confermato in modo evidente come la funzione non fosse dispendiosa solo in termini di cicli CPU, ma anche – e in certo senso più gravemente – in termini di memoria. In effetti, il tentativo di invocare su PC la `calibrate` (con gli stessi parametri utilizzati in ambiente server) ha portato alla saturazione della memoria disponibile, come mostrato in Figura 1.

```
> res=calibrate(design=des,population=totnoti,
calfun="logit",bounds=c(0.5,2.5),formula=formula)
Errore: non posso allocare un vettore di dimensione 2094635 Kb
Inoltre: Warning messages:
1: Reached total allocation of 759Mb: see help(memory.size)
2: Reached total allocation of 759Mb: see help(memory.size)
```

Figura 1: Invocazione della `calibrate` in ambiente PC

La ricerca di una soluzione in grado di garantire, anche in ambiente PC, l'uso della `calibrate`, ci ha condotto a riconsiderare la struttura algebrica del problema di calibrazione. L'esito dell'analisi si è concretizzato nella progettazione di una versione *iterata* della funzione, che indicheremo nel seguito col nome `calibrate_iter`[22]. Per la precisione, in virtù della natura delle variabili ausiliarie e dei vincoli usati nell'indagine sulle Forze di Lavoro, è stato possibile decomporre l'originale problema di calibrazione in sottoproblemi *indipendenti*, ciascuno dei quali relativo ad una sola "popolazione pianificata" (vale a dire ad una regione, si veda la Sezione 4.2.1). I benefici di tale decomposizione del problema riguardano sia la memoria che il carico computazionale, e, quindi, il tempo di elaborazione. In effetti la `calibrate_iter`, oltre ad evitare la saturazione della memoria del PC, si è rivelata particolarmente efficiente, ottenendo il risultato atteso in circa 90 secondi; il tempo di esecuzione registrato in ambiente server è confrontabile con quello osservato in ambiente PC.
Per quanto riguarda eventuali istanze del problema di calibrazione per le quali la soluzione iterata non sia applicabile, è stata effettuata un'analisi dei tempi di esecuzione della `calibrate` al variare di alcune caratteristiche del problema in input. Tale analisi,

[22] Il programma della funzione `calibrate_iter` (che, ovviamente, invoca al suo interno la funzione `calibrate`) è costituito da poco più di 100 linee di codice R.

riportata nella Sezione 4.3.2, può costituire uno strumento utile per valutare la fattibilità di esecuzione della `calibrate` in ambienti PC con risorse di memoria limitate.
In sintesi: sebbene la `calibrate` fornisse (dopo la correzione del bug) una soluzione accettabile in ambiente server, il suo uso concreto su PC ha richiesto uno specifico intervento di programmazione che ha indotto, come benefico effetto collaterale, una drastica riduzione dei tempi di elaborazione anche in ambiente server. Se ne conclude, generalizzando, che in alcuni casi può rivelarsi conveniente (o addirittura obbligatorio) specializzare le funzioni di un package a situazioni ad-hoc, mediante interventi mirati di programmazione.

4.3.2 Test di Efficacia e di Efficienza: Survey/R e GENESEES/SAS

In questa sezione ci concentriamo sui test di efficienza ed efficacia realizzati nel corso della sperimentazione. In particolare, consideriamo i seguenti insiemi di test:

1. Valutazione del tempo di esecuzione della `calibrate` per diverse configurazioni del data set di input e del problema di calibrazione.
2. Confronto dei tempi di esecuzione della `calibrate`, della `calibrate_iter` e della funzione riponderazione di GENESEES.
3. Confronto dell'efficacia dei risultati prodotti da Survey/R rispetto a GENESEES/SAS.

Per comprendere la motivazione dei test dell'insieme (1) è necessario osservare che: (i) benché la soluzione adottata nella `calibrate_iter` (ed in GENESEES) sia praticabile in tutte le maggiori indagini correnti dell'Istat, almeno in astratto non è sempre possibile decomporre il problema di calibrazione in sottoproblemi indipendenti ed operare, poi, su ciascuno di essi separatamente; (ii) non è stato possibile eseguire la `calibrate` in ambiente PC sull'intero data set di Forze di Lavoro (si veda Sezione 4.3.1).
In effetti, i test dell'insieme (1) sono stati concepiti allo scopo di verificare i limiti reali di applicabilità della `calibrate` nel caso di vincoli stringenti sulle risorse di elaborazione (scenario possibile in ambiente PC). Si è inteso, in più, fornire una stima di massima dei tempi di esecuzione della `calibrate` al variare della complessità del problema di calibrazione, sia in ambiente PC che server.
I tempi ed i limiti di esecuzione della `calibrate` sono stati valutati sperimentalmente al variare di tre fattori: il numero di record coinvolti (`r`), il numero di variabili ausiliare (`v`) e il numero di domini di stima[23] (`d`). La selezione di queste tre dimensioni è motivata dall'analisi della complessità computazionale dell'operazione time-critical effettuata dalla `calibrate`, che consiste nel calcolo dell'inversa generalizzata di una matrice. La complessità di questa operazione per matrici di input quadrate di dimensioni `n` è pari ad $O(n^3)$, mentre per matrici rettangolari di dimensione $n \cdot m$ è pari a $O[(n \cdot m)^{3/2}]$. Ebbene, nel caso della `calibrate` la matrice di input (vale a dire la matrice di design del modello

[23] Nella presente sezione ci riferiremo alle "popolazioni pianificate usate per lo stimatore" (introdotte nella Sezione 3.3) con la dizione più concisa "domini di stima". Come osservato nella Sezione 4.2.1, nel caso della rilevazione sulle Forze di Lavoro i domini di stima coincidono con le 21 popolazioni "regionali".

di calibrazione prescelto) ha un numero di righe pari al numero totale dei record del data set (`r`) e un numero di colonne pari al prodotto del numero delle variabili ausiliarie (`v`) per il numero dei domini di stima (`d`).
La `calibrate` è stata eseguita con diversi data set di input, costruiti a partire dal data set delle famiglie di Forze di Lavoro nel modo che segue: (i) sono stati ordinati i domini di stima per numero di record; (ii) i domini così ordinati sono stati aggregati in 6 sottoinsiemi contenenti rispettivamente 1, 6, 15, 18, 20 e 21 domini.
Le prime misure sono state condotte in ambiente server, ed il risultato è mostrato in Figura 2.

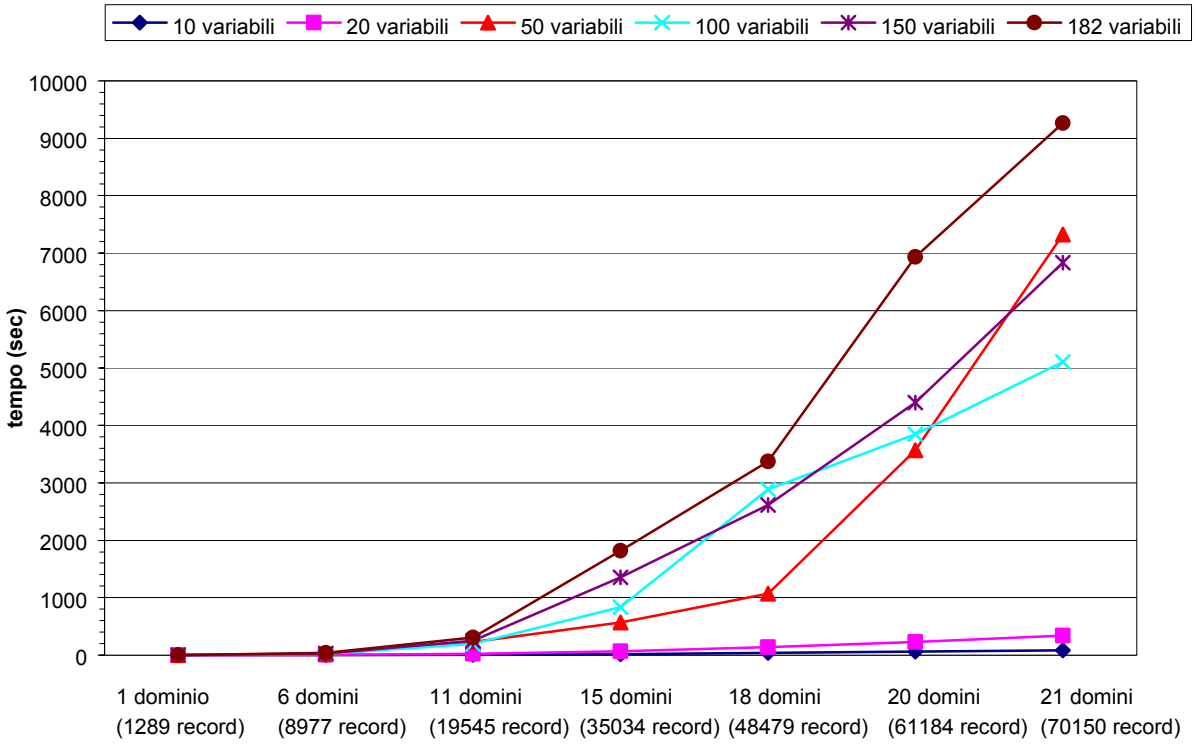

Figura 2: Tempi della `calibrate` in ambiente server

L'asse delle ordinate riporta il tempo di esecuzione in secondi, mentre quello delle ascisse riporta le caratteristiche del data set di input, sia in termini di numero di domini che di numerosità complessiva; ciascuna curva è poi riferita ad un numero diverso di variabili ausiliarie.
E' possibile osservare come l'andamento empirico delle curve, pur rivelando un buon accordo qualitativo con il comportamento atteso in base alla complessità asintotica dell'operazione di calcolo dell'inversa generalizzata, non lo rispecchi integralmente. In effetti l'andamento asintotico non prevede alcuna intersezione fra le curve.
I risultati del medesimo esperimento, condotto questa volta in ambiente PC, sono mostrati in Figura 3.

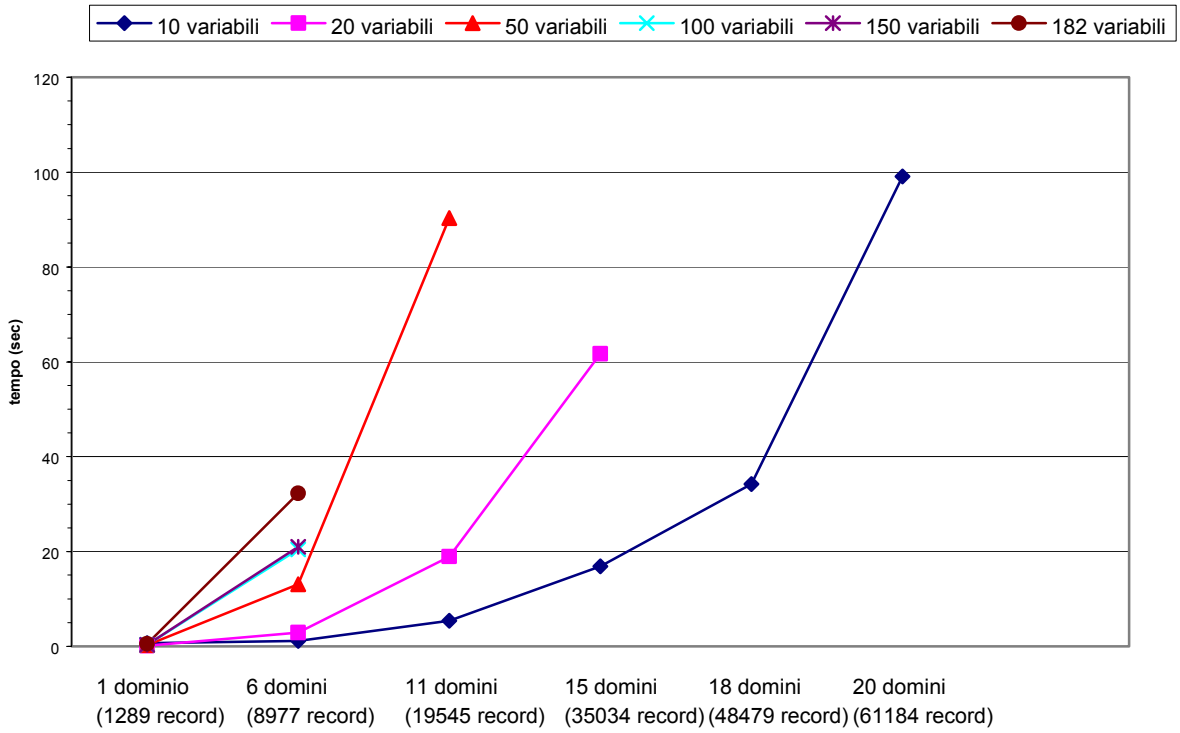

Figura 3: Tempi della `calibrate` in ambiente PC

Si nota immediatamente come in ambiente PC, in virtù del limite imposto dalla RAM disponibile[24], molte computazioni non siano state possibili. Ad esempio, nel caso di 182 variabili ausiliarie, il numero massimo di domini per cui è stato possibile ottenere la conclusione del processo di calibrazione è appena pari a 6, per un totale di 8.977 record. Inoltre, neanche per un numero di variabili pari a 10 si è ottenuto un risultato su tutti i 21 domini di stima.

I test dell'insieme (2) hanno richiesto la misura del tempo di elaborazione necessario a Survey/R ed a GENESEES/SAS per risolvere il problema di calibrazione dell'indagine sulle Forze di Lavoro. Laddove possibile le misure sono state realizzate sia in ambiente server che in ambiente PC.
Nella Sezione 4.3.1, è stato già fatto cenno alle caratteristiche prestazionali (i) della `calibrate` originaria, (ii) della versione derivante dalla correzione di un bug rintracciato nel codice sorgente (`calibrate` corretta), e (iii) della versione iterata (`calibrate_iter`). I corrispondenti tempi di elaborazione sono riportati in maggiore dettaglio nella Figura 4.

[24] R consente di aumentare la memoria allocabile *oltre* il limite costituito dalla RAM della macchina. Nel caso del test in esame, tuttavia, non è stato ritenuto opportuno ricorrere a questa possibilità, poiché, come osservato nella Sezione 2.1, ciò avrebbe comportato un notevole rallentamento delle applicazioni.

Funzione	Ambiente	Tempo
R `calibrate` originaria	PC	–
R `calibrate` originaria	Server	> 4 giorni
R `calibrate` corretta	PC	–
R `calibrate` corretta	Server	155 minuti
R `calibrate_iter`	PC	84 secondi
R `calibrate_iter`	Server	86 secondi

Figura 4: Tempi di esecuzione della `calibrate` originaria, della `calibrate` corretta e della `calibrate_iter`

In ambiente PC, come si vede, la possibilità concreta di usare survey/R per calcolare i pesi finali della rilevazione sulle Forze di Lavoro poggia interamente sulla decomposizione del problema operata dalla `calibrate_iter`. D'altro canto l'effetto benefico della decomposizione non si limita alla gestione della memoria; esso si estende al tempo di elaborazione, come si apprezza confrontando i tempi registrati in ambiente server dalla `calibrate` corretta (155 minuti) e dalla sua versione iterata `calibrate_iter` (86 secondi). La ragione di tale miglioramento può essere compresa ricordando, ancora una volta, l'espressione dalla complessità computazionale dell'inversa generalizzata.
Nel caso della `calibrate`, che opera sulla matrice di design del problema di calibrazione *completo*, la complessità del calcolo può scriversi (con la notazione introdotta in precedenza) nella forma $O[(r \cdot v \cdot d)^{3/2}]$. Per la `calibrate_iter`, che invece risolve il problema di calibrazione separatamente in ciascuno degli `d` domini di stima, sotto l'ipotesi semplificativa che ciascun dominio contenga lo stesso numero di record (pari, dunque, a `r/d`), si ottiene per la complessità la stima $(1/d)^2 \cdot O[(r \cdot v \cdot d)^{3/2}]$. Grosso modo, quindi, ci si deve attendere che la soluzione iterata diminuisca i tempi di esecuzione di un fattore d^2. Nel caso di Forze Lavoro, in cui $d \sim O(10)$, il guadagno atteso è di ordine `100`, in buon accordo con quanto sperimentalmente osservato.

In GENESEES/SAS l'esecuzione della funzione riponderazione, con input costituito dal data set delle famiglie dell'indagine Forze di Lavoro, ha comportato un tempo totale di esecuzione di circa 9 minuti (in ambiente PC). Poiché il processo di calibrazione eseguito da GENESEES è anch'esso iterato, il tempo in esame deve essere paragonato a quello della `calibrate_iter` in ambiente PC. Da tale confronto risulta che la `calibrate_iter` è circa 6 volte più veloce della funzione di riponderazione di GENESEES. La differenza di performance è imputabile a molteplici fattori, che includono certamente la diversa natura dei linguaggi di programmazione utilizzati (R e SAS/IML) e la diversa progettazione del software. In più va ricordato che la funzione di riponderazione di GENESEES, oltre a calcolare i pesi finali del processo di calibrazione, produce anche data set e report di diagnostica (funzionalità, quest'ultima, non prevista dalla `calibrate`). Fra gli altri aspetti vale la pena menzionare il differente criterio di convergenza dell'algoritmo di Newton-

Raphson adottato nei due software (cfr. Sezione 3.4): come si verifica in Figura 5, esso fa si che, dominio per dominio, il numero di iterazioni eseguite da Survey/R sia sempre minore di quello di GENESEES/SAS

Dominio	Iterazioni GENESEES	Iterazioni `calibrate_iter`
1	6	4
2	6	5
3	6	4
5	6	5
6	6	4
7	6	4
8	6	4
9	6	4
10	6	4
11	6	5
12	6	5
13	6	5
14	7	5
15	6	4
16	6	4
17	5	4
18	6	5
19	6	5
20	6	5
41	5	4
42	5	4

Figura 5: Numero di iterazioni per dominio in GENESEES e in `calibrate_iter`

Con riferimento all'insieme di test (3), l'efficacia della calibrazione in Survey/R ed in GENESEES/SAS è stata valutata confrontando i 70.150 pesi finali ottenuti nei due casi.
In Figura 6 i pesi finali calcolati con la `calibrate` (corretta o iterata, in questa sede non fa differenza) sono riportati in grafico in funzione dei pesi calcolati con la funzione di riponderazione di GENESEES. E' immediato constatare come i due software forniscano risultati in ottimo accordo.

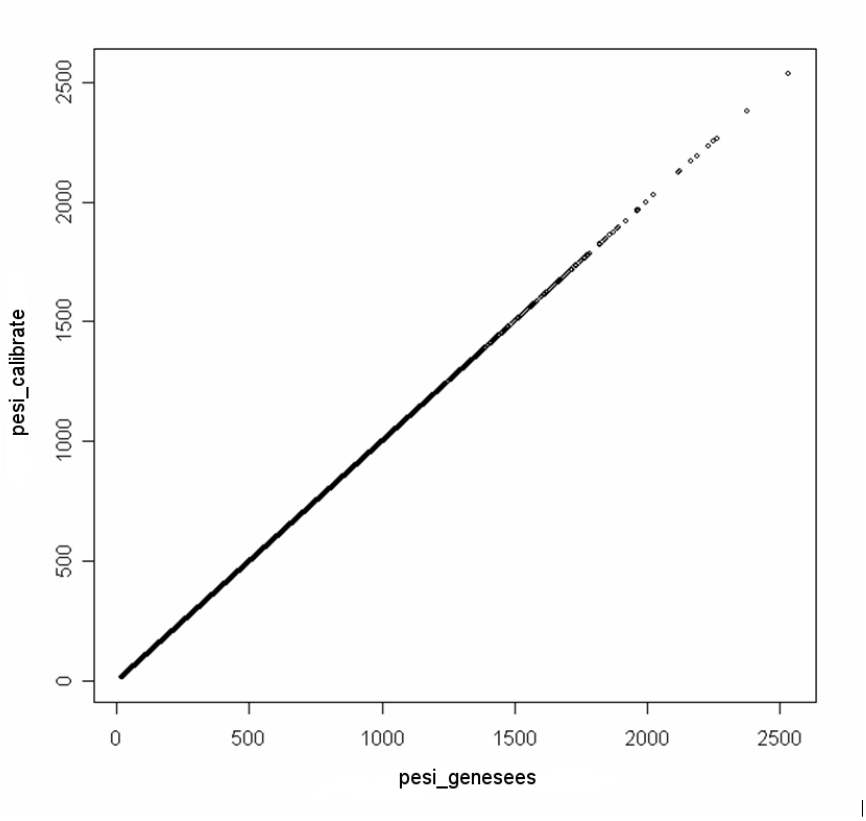

Figura 6. Confronto grafico tra i pesi di GENESEES e della calibrate

Una valutazione più precisa della misura di tale accordo può essere ricavata dallo studio delle distribuzioni delle differenze assolute e relative fra i pesi prodotti in output dai due software. In Figura 7 è riportata la sintesi della distribuzione delle differenza assolute ottenuta usando la funzione `summary` di R.

```
> summary(abs(pesi_genesees-pesi_calibrate))
Min.        Median      1st Qu.     Mean        3rd Qu.     Max.
7.816e-14   1.660e-09   5.548e-09   2.443e-07   3.905e-08   1.090e-04
```

Figura 7: Sintesi della distribuzione delle differenza assolute tra i pesi di GENESEES e della `calibrate`

Come si vede la media delle differenze assolute è di ordine 10^{-7} ed il valore massimo osservato è di ordine 10^{-4}.
Ancor più significativa è, in ogni caso, la valutazione ricavata analizzando la distribuzione delle differenze relative. Il risultato della funzione `summary` di R è riportato in Figura 8.

```
> summary(abs(pesi_genesees-pesi_calibrate)/ pesi_genesees)
Min.        Median      1st Qu.     Mean        3rd Qu.     Max.
8.405e-16   7.460e-12   2.326e-11   5.950e-10   1.189e-10   1.086e-07
```

Figura 8 Sintesi della distribuzione delle differenze relative tra i pesi di GENESEES e della calibrate

La massima differenza relativa fra il peso finale attribuito da GENESEES ad un arbitrario record del data set di Forze Lavoro ed il corrispondente peso calcolato dalla `calibrate` è di ordine 10^{-7}. E', per conseguenza, lecito asserire che i risultati prodotti dai due software sono, a tutti gli effetti pratici, identici.

5 Conclusioni

Nel presente documento sono stati illustrati gli aspetti salienti della sperimentazione condotta su R e, nello specifico, sugli strumenti resi disponibili dal package survey per il calcolo dei pesi finali delle indagini campionarie. Come accennato nell'Introduzione, l'obiettivo strategico della sperimentazione era valutare le potenzialità di R nel gestire elaborazioni di elevata complessità computazionale su consistenti moli di dati. Per questo motivo lo studio del package survey è stato condotto utilizzando i dati dell'indagine sulle Forze di Lavoro, una rilevazione "critica" sia in termini di dimensioni che per la complessità del disegno di campionamento e del problema di calibrazione. Il lavoro è stato concluso dalla comparazione, sia in termini di efficacia che di efficienza, della funzione `calibrate` di survey con la funzione di riponderazione del software generalizzato GENESEES, realizzato in Istat e sviluppato in linguaggio SAS.
I risultati conseguiti dimostrano, a nostro parere, la piena praticabilità dell'utilizzo del package survey nella fase di calibrazione dei dati. I punti principali a favore di questa conclusione sono sinteticamente riportati in quanto segue:

- L'efficienza della `calibrate` è confrontabile, se non addirittura superiore, a quella di GENESEES. Inoltre è possibile eseguire la `calibrate` sia in ambiente PC che in ambiente server. Il prezzo da pagare per questi due risultati è, in termini di interventi mirati di programmazione, decisamente contenuto.
- L'efficacia della `calibrate` è a tutti gli effetti identica a quella ottenuta con GENESEES.
- La copertura delle funzionalità offerte dal package survey rispetto a GENESEES è molto elevata. In ogni caso, la copertura può essere resa completa con un modesto sforzo di programmazione.

La sperimentazione realizzate costituisce, in Istat, uno dei primi esempi di test sistematico di un package R. Essa ha, fra l'altro, richiesto di estendere ed arricchire le funzionalità base del package, o di adattarle ad esigenze specifiche della pratica Istat. Anche da questo punto di vista, l'esperienza fatta ha condotto ad un giudizio positivo, sulla base delle motivazioni di seguito elencate:

- L'accesso al codice sorgente del package è stato fondamentale per personalizzare le soluzioni disponibili.

- La programmazione in R, sebbene di apprendimento non immediato, si è rivelata stimolante ed ha comportato la produzione di codice compatto ed essenziale che ha richiesto tempi di sviluppo contenuti.
- L'incapsulamento, caratteristico dello stile object-oriented con cui è progettato e costruito il package survey, ha comportato un'iniziale difficoltà di comprensione della struttura del package; la difficoltà è stata, comunque, compensata dalla possibilità di effettuare modifiche "locali" sulla base delle esigenze specifiche che si sono presentate nel corso della sperimentazione.

A questi aspetti positivi, si contrappongono alcuni aspetti negativi, riassumibili nella carenza di documentazione, nella possibilità di errori nel codice disponibile e, talvolta, nella necessità di ottimizzarlo.

Allo stato, è in corso di realizzazione una seconda fase della sperimentazione, che prevede lo studio ed il test del package survey con riferimento alle funzioni di calcolo delle stime e degli errori campionari. Anche in questo caso, è prevista la realizzazione di un estensivo confronto con l'analoga funzionalità resa disponibile dal software GENESEES.

6 Bibliografia

[Deville-et-al,1993] Deville J.C., Särndal C.E. and Sautory O.: "Generalized Raking Procedures in Survey Sampling", Journal of the American Statistical Association, vol.88, 1993.

[FOL-met-norm,2006] "La Rilevazione sulle Forze di Lavoro: Contenuti, Metodologie, Organizzazione", Metodi e Norme, n. 32, Istat, 2006.

[Horvitz-and-Thompson,1952] Horvitz D.G. and Thompson D. J.: "A Generalization of Sampling without Replacement in Finite Universe", Journal of the American Statistical Association, vol. 47, 1952.

[Manuale-GENESEES,2005] GENESEES V. 3.0 – Funzione di Riponderazione, Istat, 2005.

[Manuale-Survey,2007] Lumley, T.: "The survey Package", disponibile on line all'indirizzo: http://cran.r-project.org/doc/packages/survey.pdf

[R-intro,2006] An Introduction to R, disponibile on line all'indirizzo: http://cran.r-project.org/doc/manuals/R-intro.pdf

[R-lang,2006] R Language Definition, disponibile on line all'indirizzo: http://cran.r-project.org/doc/manuals/R-lang.pdf

[S-Programming,2000] Venables W. and Ripley B.D.: "S Programming", Springer, 2000.

[Särndal-et-al,1992] Särndal C.E., Swensson B. and Wretman J.: "Model Assisted Survey Sampling", Springer Verlag, 1992.

[Singh-and-Mohl-1996] Singh A.C. and Mohl C.A. : "Understanding Calibration Estimators in Survey Sampling", Survey Methodology, vol.22, n.2, 1996.

[Vanderhoeft-2001] Vanderhoeft C.: "Generalized Calibration at Statistic Belgium", Statistics Belgium Working Paper n. 3, disponibile on line all'indirizzo: http://www.statbel.fgov.be/studies/paper03_en.asp

[Wilkinson-and-Rogers,1973] Wilkinson G.N. and Rogers C.E. : "Symbolic Description of Factorial Models for Analysis of Variance", Journal of the Royal Statistical Society, series C (Applied Statistics), Vol. 22, p.181-191, 1973.

Contributi ISTAT(*)

1/2002 - Francesca Biancani, Andrea Carone, Rita Pistacchio e Giuseppina Ruocco - *Analisi delle imprese individuali*

2/2002 - Massimiliano Borgese .-.*Proposte metodologiche per un progetto d'indagine sul trasporto aereo alla luce della recente normativa comunitaria sul settore*

3/2002 - Nadia Di Veroli e Roberta Rizzi - *Proposta di classificazione dei rapporti di lavoro subordinato e delle attività di lavoro autonomo: analisi del quadro normativo*

4/2002 – Roberto Gismondi – *Uno stimatore ottimale in presenza di non risposte*

5/2002 – Maria Anna Pennucci – *Le strategie europee per l'occupazione dal Libro bianco di Delors al Consiglio Europeo di Cardiff*

1/2003 – Giovanni Maria Merola – *Safety Rules in Statistical Disclosure Control for Tabular Data*

2/2003 – Fabio Bacchini, Pietro Gennari e Roberto Iannaccone – *A new index of production for the construction sector based on input data*

3/2003 – Fulvia Cerroni e Enrica Morganti – *La metodologia e il potenziale informativo dell'archivio sui gruppi di impresa: primi risultati*

4/2003 – Sara Mastrovita e Isabella Siciliani – *Effetti dei trasferimenti sociali sulla distribuzione del reddito nei Paesi dell'Unione europea: un'analisi dal Panel europeo sule famiglie*

5/2003 – Patrizia Cella, Giuseppe Garofalo, Adriano Paggiaro, Nicola Torelli e Caterina Viviano – *Demografia d'impresa: l'utilizzo di tecniche di abbinamento per l'analisi della continuità*

6/2003 – Enrico Grande e Orietta Luzi – *Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in Istat*

7/2003 – Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Indagine sperimentale sui posti di lavoro vacanti*

8/2003 – Mario Adua – *L'agricoltura di montagna: le aziende delle donne, caratteristiche agricole e socio-rurali*

9/2003 – Franco Mostacci e Roberto Sabbatini – *L'euro ha creato inflazione? Changeover e arrotondamenti dei prezzi al consumo in Italia nel 2002*

10/2003 – Leonello Tronti – *Problemi e prospettive di riforma del sistema pensionistico*

11/2003 – Roberto Gismondi – *Tecniche di stima e condizioni di coerenza per indagini infraannuali ripetute nel tempo*

12/2003 – Antonio Frenda – *Analisi delle legislazioni e delle prassi contabili relative ai gruppi di imprese nei paesi dell'Unione Europea*

1/2004 – Marcello D'Orazio, Marco Di Zio e Mauro Scanu – *Statistical Matching and the Likelihood Principle: Uncertainty and Logical Constraints*

2/2004 – Giovanna Brancato – *Metodologie e stime dell'errore di risposta. Una sperimentazione di reintervista telefonica*

3/2004 – Franco Mostacci, Giuseppina Natale e Elisabetta Pugliese – *Gli indici dei prezzi al consumo per sub popolazioni*

4/2004 – Leonello Tronti – *Una proposta di metodo: osservazioni e raccomandazioni sulla definizione e la classificazione di alcune variabili attinenti al mercato del lavoro*

5/2004 – Ugo Guarnera – *Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi: il softaware* Quis

6/2004 – Patrizia Giaquinto, Marco Landriscina e Daniela Pagliuca – *La nuova funzione di analisi dei modelli implementata in Genesees v. 3.0*

7/2004 – Roberto Di Giuseppe, Patrizia Giaquinto e Daniela Pagliuca – *MAUSS (Multivariate Allocation of Units in Sampling Surveys):*

un software generalizzato per risolvere il problema dell' allocazione campionaria nelle indagini Istat

8/2004 – Ennio Fortunato e Liana Verzicco – *Problemi di rilevazione e integrazione della condizione professionale nelle indagini sociali dell'Istat*

9/2004 – Claudio Pauselli e Claudia Rinaldelli – *La valutazione dell'errore di campionamento delle stime di povertà relativa secondo la tecnica Replicazioni Bilanciate Ripetute*

10/2004 – Eugenio Arcidiacono, Marina Briolini, Paolo Giuberti, Marco Ricci, Giovanni Sacchini e Giorgia Telloli – *Procedimenti giudiziari, reati, indagati e vittime in Emilia-Romagna nel 2002: un'analisi territoriale sulla base dei procedimenti iscritti nel sistema informativo Re.Ge.*

11/2004 – Enrico Grande e Orietta Luzi **–** *Regression trees in the context of imputation of item non-response: an experimental application on business data*

12/2004 – Luisa Frova e Marilena Pappagallo **–** *Procedura di now-cast dei dati di mortalità per causa*

13/2004 – Giorgio DellaRocca, Marco Di Zio, Orietta Luzi, Emanuela Scavalli e Giorgia Simeoni **–** *IDEA (Indices for Data Editing Assessment): sistema per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori SIDI*

14/2004 **–** Monica Pace, Silvia Bruzzone, Luisa Frova e Marilena Pappagallo **–** *Review of the existing information about death certification practices, certificate structures and training tools for certification of causes of death in Europe*

15/2004 – Elisa Berntsen – *Modello Unico di Dichiarazione ambientale: una fonte amministrativa per l'Archivio delle Unità Locali di Asia*

16/2004 – Salvatore F. Allegra e Alessandro La Rocca – *Sintetizzare misure elementari: una sperimentazione di alcuni criteri per la definizione di un indice composto*

17/2004 – Francesca R. Pogelli – *Un'applicazione del modello "Country Product Dummy" per un'analisi territoriale dei prezzi*

18/2004 – Antonia Manzari – *Valutazione comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi*

19/2004 – Claudio Pauselli – *Intensità di povertà relativa: stima dell'errore di campionamento e sua valutazione temporale*

20/2004 – Maria Dimitri, Ersilia Di Pietro, Alessandra Nuccitelli e Evelina Paluzzi – *Sperimentazione di una metodologia per il controllo della qualità di dati anagrafici*

21/2004 – Tiziana Pichiorri, Anna M. Sgamba e Valerio Papale – *Un modello di ottimizzazione per l'imputazione delle mancate risposte statistiche nell'indagine sui trasporti marittimi dell'Istat*

(*) ultimi cinque anni

22/2004 – Diego Bellisai, Piero D. Falorsi, Annalisa Lucarelli, Maria A. Pennucci e Leonello G. Tronti – *Indagine pilota sulle retribuzioni di fatto nel pubblico impiego*

23/2004 – Lidia Brondi – *La riorganizzazione del sistema idrico: quadro normativo, delimitazione degli ambiti territoriali ottimali e analisi statistica delle loro caratteristiche strutturali*

24/2004 – Roberto Gismondi e Laura De Sandro – *Provisional Estimation of the Italian Monthly Retail Trade Index*

25/2004 – Annamaria Urbano, Claudia Brunini e Alessandra Chessa – *I minori in stato di abbandono: analisi del fenomeno e studio di una nuova prospettiva d'indagine*

26/2004 – Paola Anzini e Anna Ciammola – *La destagionalizzazione degli indici della produzione industriale: un confronto tra approccio diretto e indiretto*

27/2004 – Alessandro La Rocca – *Analisi della struttura settoriale dell'occupazione regionale: 8° Censimento dell'industria e dei servizi 2001 7° Censimento dell'industria e dei servizi 1991*

28/2004 – Vincenzo Spinelli e Massimiliano Tancioni – *I Trattamenti Monetari non Pensionistici: approccio computazionale e risultati della sperimentazione sugli archivi INPS-DM10*

29/2004 – Paolo Consolini – *L'indagine sperimentale sull'archivio fiscale modd.770 anno 1999: analisi della qualità del dato e stime campionarie*

1/2005 – Fabrizio M. Arosio – *La stampa periodica e l'informazione on-line: risultati dell'indagine pilota sui quotidiani on-line*

2/2005 – Marco Di Zio, Ugo Guarnera e Orietta Luzi – *Improving the effectiveness of a probabilistic editing strategy for business data*

3/2005 – Diego Moretti e Claudia Rinaldelli – *EU-SILC complex indicators: the implementation of variance estimation*

4/2005 – Fabio Bacchini, Roberto Iannaccone e Edoardo Otranto – *L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione*

5/2005 – Marco Broccoli – *Analisi della criminalità a livello comunale: metodologie innovative*

6/2005 – Claudia De Vitiis, Loredana Di Consiglio e Stefano Falorsi – *Studio del disegno campionario per la nuova rilevazione continua sulle Forze di Lavoro*

7/2005 – Edoardo Otranto e Roberto Iannaccone – *Continuous Time Models to Extract a Signal in Presence of Irregular Surveys*

8/2005 – Cosima Mero e Adriano Pareto – *Analisi e sintesi degli indicatori di qualità dell'attività di rilevazione nelle indagini campionarie sulle famiglie*

9/2005 – Filippo Oropallo – *Enterprise microsimulation models and data challenges*

10/2005 – Marcello D' Orazio, Marco Di Zio e Mauro Scanu – *A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study*

11/2005 – Stefania Macchia, Manuela Murgia, Loredana Mazza, Giorgia Simeoni, Francesca Di Patrizio, Valentino Parisi, Roberto Petrillo e Paola Ungaro – *Una soluzione per la rilevazione e codifica della Professione nelle indagini CATI*

12/2005 – Piero D. Falorsi, Monica Scannapieco, Antonia Boggia e Antonio Pavone – *Principi Guida per il Miglioramento della Qualità dei Dati Toponomastici nella Pubblica Amministrazione*

13/2005 – Ciro Baldi, Francesca Ceccato, Silvia Pacini e Donatella Tuzi – *La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento*

14/2005 – Stefano De Francisci, Giuseppe Sindoni e Leonardo Tininini – *Da Winci/MD: un sistema per data warehouse statistici sul Web*

15/2005 – Gerardo Gallo e Evelina Palazzi – *I cittadini italiani naturalizzati: l'analisi dei dati censuari del 2001, con un confronto tra immigrati di prima e seconda generazione*

16/2005 – Saverio Gazzelloni, Mario Albisinni, Lorenzo Bagatta, Claudio Ceccarelli, Luciana Quattrociocchi, Rita Ranaldi e Antonio Toma – *La nuova rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*

17/2005 – Maria Carla Congia – *Il lavoro degli extracomunitari nelle imprese italiane e la regolarizzazione del 2002. Prime evidenze empiriche dai dati INPS*

18/2005 – Giovanni Bottazzi, Patrizia Cella, Giuseppe Garofalo, Paolo Misso, Mariano Porcu e Marianna Tosi – *Indagine pilota sulla nuova imprenditorialità nella Regione Sardegna. Relazione Conclusiva*

19/2005 – Fabrizio Martire e Donatella Zindato – *Le famiglie straniere: analisi dei dati censuari del 2001 sui cittadini stranieri residenti*

20/2005 – Ennio Fortunato – *Il Sistema di Indicatori Territoriali: percorso di progetto, prospettive di sviluppo e integrazione con i processi di produzione statistica*

21/2005 – Antonella Baldassarini e Danilo Birardi – *I conti economici trimestrali: un approccio alla stima dell'input di lavoro*

22/2005 – Francesco Rizzo, Dario Camol e Laura Vignola – *Uso di XML e WEB Services per l'integrazione di sistemi informativi statistici attraverso lo standard SDMX*

1/2006 – Ennio Fortunato – *L'analisi integrata delle esigenze informative dell'utenza Istat: Il contributo del Sistema di Indicatori Territoriali*

2/2006 – Francesco Altarocca – *I design pattern nella progettazione di software per il supporto alla statistica ufficiale*

3/2006 – Roberta Palmieri – *Le migranti straniere: una lettura di genere dei dati dell'osservatorio interistituzionale sull'immigrazione in provincia di Macerata*

4/2006 – Raffaella Amato, Silvia Bruzzone, Valentina Delmonte e Lidia Fagiolo – *Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage*

5/2006 – Alessandro La Rocca – *Fuzzy clustering: la logica, i metodi*

6/2006 – Raffaella Cascioli – *Integrazione dei dati micro dalla Rilevazione delle Forze di Lavoro e dagli archivi amministrativi INPS: risultati di una sperimentazione sui dati campione di 4 province*

7/2006 – Gianluca Brogi, Salvatore Cusimano, Giuseppina del Vicario, Giuseppe Garofalo e Orietta Patacchia – *La realizzazione di Asia Agricoltura tramite l'utilizzo di dati amministrativi: il contenuto delle fonti e i risultati del processo di integrazione*

8/2006 – Simonetta Cozzi – *La distribuzione commerciale in Italia: caratteristiche strutturali e tendenze evolutive*

9/2006 – Giovanni Seri – *A graphical framework to evaluate risk assessment and information loss at individual level*

10/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Feasibility studies for the coverage of public institutions in sections N and O*

11/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Quarterly labour cost index in public education*

12/2006 – Silvia Montagna, Patrizia Collesi, Florinda Damiani, Danila Fulgenzio, Maria Francesca Loporcaro e Giorgia Simeoni – *Nuove esperienze di rilevazione della Customer Satisfaction*

13/2006 – Lucia Coppola e Giovanni Seri – *Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC*

14/2006 – Lidia Brondi – *L'utilizzazione delle surveys per la stima del valore monetario del danno ambientale: il metodo della valutazione contingente*

15/2006 – Carlo Boselli – *Le piccole imprese leggere esportatrici e non esportatrici: differenze di struttura e di comportamento*

16/2006 – Carlo De Greogorio – *Il nuovo impianto della rilevazione centralizzata del prezzo dei medicinali con obbligo di prescrizione*

1/2007 – Paolo Roberti, Maria Grazia Calza, Filippo Oropallo e Stefania Rossetti – *Knowledge Databases to Support Policy Impact Analysis: the EuroKy-PIA Project*

2/2007 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, e Marina Sorrentino – *Production of job vacancy statistics: coverage*

3/2007 – Carlo Lucarelli e Giampiero Ricci – *Working times and working schedules: the framework emerging from the new Italian lfs in a gender perspective*

4/2007 – Monica Scannapieco, Diego Zardetto e Giulio Barcaroli – *La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS*

Documenti ISTAT(*)

1/2002 – Paolo Consolini e Rita De Carli - *Le prestazioni sociali monetarie non pensionistiche: unità di analisi, fonti e rappresentazione statistica dei dati*

2/2002 – Stefania Macchia - *Sperimentazione, implementazione e gestione dell'ambiente di codifica automatica della classificazione delle Attività economiche*

3/2002 – Maria De Lucia - *Applicabilità della disciplina in materia di festività nel pubblico impiego*

4/2002 – Roberto Gismondi, Massimo Marciani e Mauro Giorgetti - *The italian contribution towards the implementantion of an european transport information system: main results of the MESUDEMO project*

5/2002 – Olimpio Cianfarani e Sauro Angeletti - *Misure di risultato e indicatori di processo: l'esperienza progettuale dell'Istat*

6/2002 – Riccardo Carbini e Valerio De Santis – *Programma statistico nazionale: specifiche e note metodologiche per la compilazione delle schede identificative dei progetti*

7/2002 – Maria De Lucia – *Il CCNL del personale dirigente dell'area 1 e la valutazione delle prestazioni dei dirigenti*

8/2002 – Giuseppe Garofalo e Enrica Morganti – *Gruppo di lavoro per la progettazione di un archivio statistico sui gruppi d'impresa*

1/2003 – Francesca Ceccato, Massimiliano Tancioni e Donatella Tuzi – *MODSIM-P:Il nuovo modello dinamico di previsione della spesa pensionistica*

2/2003 – Anna Pia Mirto – *Definizioni e classificazioni delle strutture ricettive nelle rilevazioni statistiche ufficiali sull'offerta turistica*

3/2003 – Simona Spirito – *Le prestazioni assistenziali monetarie non pensionistiche*

4/2003 – Maria De Lucia – *Approfondimenti di alcune tematiche inerenti la gestione del personale*

5/2003 – Rosalia Coniglio, Marialuisa Cugno, Maria Filmeno e Alberto Vitalini – *Mappatura della criminalità nel distretto di Milano*

6/2003 – Maria Letizia D'Autilia – *I provvedimenti di riforma della pubblica amministrazione per l'identificazione delle "Amministrazioni pubbliche" secondo il Sec95: analisi istituzionale e organizzativa per l'anno 2000*

7/2003 – Francesca Gallo, Pierpaolo Massoli, Sara Mastrovita, Roberto Merluzzi, Claudio Pauselli, Isabella Siciliani e Alessandra Sorrentino – *La procedura di controllo e correzione dei dati Panel Europeo sulle famiglie*

8/2003 – Cinzia Castagnaro, Martina Lo Conte, Stefania Macchia e Manuela Murgia – *Una soluzione in-house per le indagini CATI: il caso della Indagine Campionaria sulle Nascite*

9/2003 – Anna Pia Maria Mirto e Norina Salamone – *La classificazione delle strutture ricettive turistiche nella normativa delle regioni italiane*

10/2003 – Roberto Gismondi e Anna Pia Maria Mirto – *Le fonti statistiche per l'analisi della congiuntura turistica: il mosaico italiano*

11/2003 – Loredana Di Consiglio e Stefano Falorsi – *Alcuni aspetti metodologici relativi al disegno dell'indagine di copertura del Censimento Generale della Popolazione 2001*

12/2003 – Roberto Gismondi e Anna Rita Giorgi – *Struttura e dinamica evolutiva del comparto commerciale al dettaglio: le tendenze recenti e gli effetti della riforma "Bersani"*

13/2003 – Donatella Cangialosi e Rosario Milazzo – *Fabbisogni formativi degli Uffici comunali di statistica: indagine rapida in Sicilia*

14/2003 – Agostino Buratti e Giovanni Salzano – *Il sistema automatizzato integrato per la gestione delle rilevazioni dei documenti di bilancio degli enti locali*

1/2004 – Giovanna Brancato e Giorgia Simeoni – *Tesauri del Sistema Informativo di Documentazione delle Indagini (SIDI)*

2/2004 – Corrado Peperoni – *Indagine sui bilanci consuntivi degli Enti previdenziali: rilevazione, gestione e procedure di controllo dei dati*

3/2004 – Marzia Angelucci, Giovanna Brancato, Dario Camol, Alessio Cardacino, Sandra Maresca e Concetta Pellegrini – *Il sistema ASIMET per la gestione delle Note Metodologiche dell'Annuario Statistico Italiano*

4/2004 – Francesca Gallo, Sara Mastrovita, Isabella Siciliani e Giovanni Battista Arcieri – *Il processo di produzione dell'Indagine ECHP*

5/2004 – Natale Renato Fazio e Carmela Pascucci – *Gli operatori non identificati nelle statistiche del commercio con l'estero: metodologia di identificazione nelle spedizioni "groupage" e miglioramento nella qualità dei dati*

6/2004 – Diego Moretti e Claudia Rinaldelli – *Una valutazione dettagliata dell'errore campionario della spesa media mensile familiare*

7/2004 – Franco Mostacci – *Aspetti Teorico-pratici per la Costruzione di Indici dei Prezzi al Consumo*

8/2004 – Maria Frustaci – *Glossario economico-statistico multilingua*

9/2004 – Giovanni Seri e Maurizio Lucarelli – *"Il Laboratorio per l'analisi dei dati elementari (ADELE): monitoraggio dell'attività dal 1999 al 2004"*

10/2004 – Alessandra Nuccitelli, Francesco Bosio e Luciano Fioriti – *L'applicazione RECLINK per il record linkage: metodologia implementata e linee guida per la sua utilizzazione*

1/2005 – Francesco Cuccia, Simone De Angelis, Antonio Laureti Palma, Stefania Macchia, Simona Mastroluca e Domenico Perrone – *La codifica delle variabili testuali nel 14° Censimento Generale della Popolazione*

2/2005 – Marina Peci – *La statistica per i Comuni: sviluppo e prospettive del progetto Sisco.T (Servizio Informativo Statistico Comunale. Tavole)*

3/2005 – Massimiliano Renzetti e Annamaria Urbano – *Sistema Informativo sulla Giustizia: strumenti di gestione e manutenzione*

4/2005 – Marco Broccoli, Roberto Di Giuseppe e Daniela Pagliuca – *Progettazione di una procedura informatica generalizzata per la sperimentazione del metodo Microstrat di coordinamento della selezione delle imprese soggette a rilevazioni nella realtà Istat*

5/2005 – Mauro Albani e Francesca Pagliara – *La ristrutturazione della rilevazione Istat sulla criminalità minorile*

6/2005 – Francesco Altarocca e Gaetano Sberno – *Progettazione e sviluppo di un "Catalogo dei File Grezzi con meta-dati di base" (CFG) in tecnologia Web*

* ultimi cinque anni

7/2005 – Salvatore F. Allegra e Barbara Baldazzi – *Data editing and quality of daily diaries in the Italian Time Use Survey*

8/2005 – Alessandra Capobianchi – *Alcune esperienze in ambito internazionale per l'accesso ai dati elementari*

9/2005 – Francesco Rizzo, Laura Vignola, Dario Camol e Mauro Bianchi – *Il progetto "banca dati della diffusione congiunturale"*

10/2005 – Ennio Fortunato e Nadia Mignolli – *I sistemi informativi Istat per la diffusione via web*

11/2005 – Ennio Fortunato e Nadia Mignolli – *Sistemi di indicatori per l'attività di governo: l'offerta informativa dell'Istat*

12/2005 – Carlo De Gregorio e Stefania Fatello – *L'indice dei prezzi al consumo dei testi scolastici nel 2004*

13/2005 – Francesco Rizzo e Laura Vignola – *RSS: uno standard per diffondere informazioni*

14/2005 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Launching and implementing the job vacancy statistics*

15/2005 – Stefano De Francisci, Massimiliano Renzetti, Giuseppe Sindoni e Leonardo Tininini – *La modellazione dei processi nel Sistema Informativo Generalizzato di Diffusione dell'ISTAT*

16/2005 – Ennio Fortunato e Nadia Mignolli – *Verso il Sistema di Indicatori Territoriali: rilevazione e analisi della produzione Istat*

17/2005 – Raffaella Cianchetta e Daniela Pagliuca – *Soluzioni Open Source per il software generalizzato in Istat: il caso di PHPSurveyor*

18/2005 – Gianluca Giuliani e Barbara Boschetto – *Gli indicatori di qualità dell'Indagine continua sulle Forze di Lavoro dell'Istat*

19/2005 – Rossana Balestrino, Franco Garritano, Carlo Cipriano e Luciano Fanfoni – *Metodi e aspetti tecnologici di raccolta dei dati sulle imprese*

1/2006 – Roberta Roncati – *www.istat.it (versione 3.0) Il nuovo piano di navigazione*

2/2006 – Maura Seri e Annamaria Urbano – *Sistema Informativo Territoriale sulla Giustizia: la sezione sui confronti internazionali*

3/2006 – Giovanna Brancato, Riccardo Carbini e Concetta Pellegrini – *SIQual: il sistema informativo sulla qualità per gli utenti esterni*

4/2006 – Concetta Pellegrini – *Soluzioni tecnologiche a supporto dello sviluppo di sistemi informativi sulla qualità: l'esperienza SIDI*

5/2006 – Maurizio Lucarelli – *Una valutazione critica dei modelli di accesso remoto nella comunicazione di informazione statistica*

6/2006 – Natale Renato Fazio – *La ricostruzione storica delle statistiche del commercio con l'estero per gli anni 1970-1990*

7/2006 – Emilia D'Acunto – *L'evoluzione delle statistiche ufficiali sugli indici dei prezzi al consumo*

8/2006 – Ugo Guarnera, Orietta Luzi e Stefano Salvi – *Indagine struttura e produzioni delle aziende agricole: la nuova procedura di controllo e correzione automatica per le variabili su superfici aziendali e consistenza degli allevamenti*

9/2006 – Maurizio Lucarelli – *La regionalizzazione del Laboratorio ADELE: un'ipotesi di sistema distribuito per l'accesso ai dati elementari*

10/2006 – Alessandra Bugio, Claudia De Vitiis, Stefano Falorsi, Lidia Gargiulo, Emilio Gianicolo e Alessandro Pallara – *La stima di indicatori per domini sub-regionali con i dati dell'indagine: condizioni di salute e ricorso ai servizi sanitari*

11/2006 – Sonia Vittozzi, Paola Giacchè, Achille Zuchegna, Piero Crivelli, Patrizia Collesi, Valerio Tiberi, Alexia Sasso, Maurizio Bonsignori, Giuseppe Stassi e Giovanni A. Barbieri – *Progetto di articolazione della produzione editoriale in collane e settori*

12/2006 – Alessandra Coli, Francesca Tartamella, Giuseppe Sacco, Ivan Faiella, Marcello D'Orazio, Marco Di Zio, Mauro Scanu, Isabella Siciliani, Sara Colombini e Alessandra Masi – *La costruzione di un Archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane*

13/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Intrastat*

14/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Extrastat*

15/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: comparazione tra rilevazione Intrastat ed Extrastat*

16/2006 – Fabio M. Rapiti – *Short term statistics quality Reporting: the LCI National Quality Report 2004*

17/2006 – Giampiero Siesto, Franco Branchi, Cristina Casciano, Tiziana Di Francescantonio, Piero Demetrio Falorsi, Salvatore Filiberti, Gianfranco Marsigliesi, Umberto Sansone, Ennio Santi, Roberto Sanzo e Alessandro Zeli – *Valutazione delle possibilità di uso di dati fiscali a supporto della rilevazione PMI*

18/2006 – Mauro Albani – *La nuova procedura per il trattamento dei dati dell'indagine Istat sulla criminalità*

19/2006 – Alessandra Capobianchi – *Review dei sistemi di accesso remoto: schematizzazione e analisi comparativa*

20/2006 – Francesco Altarocca – *Gli strumenti informatici nella raccolta dei dati di indagini statistiche: il caso della Rilevazione sperimentale delle tecnologie informatiche e della comunicazione nelle Pubbliche Amministrazioni locali*

1/2007 – Giuseppe Stassi – *La politica editoriale dell'Istat nel periodo 1996-2004: collane, settori, modalità di diffusione*

2/2007 – Daniela Ichim – *Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment*

3/2007 – Ugo Guarnera, Orietta Luzi e Irene Tommasi – *La nuova procedura di controllo e correzione degli errori e delle mancate risposte parziali nell'indagine sui Risultati Economici delle Aziende Agricole (REA)*

4/2007 – Vincenzo Spinelli – *Processo di Acquisizione e Trattamento Informatico degli Archivi relativi al Modello di Dichiarazione 770*