STRUMENTI R PER LA STATISTICA UFFICIALE: L'ESPERIENZA DELL'ISTAT

di Paolo Righi (parighi@istat.it)

Il progetto "Stat2015" che impegnerà l'Istat nei prossimi anni ha come principale obiettivo l'innovazione dei processi e di prodotti nelle fasi della produzione statistica (cfr. NewsStat n. 4/2012). Per la fase di elaborazione dei dati il progetto prevede la realizzazione di un sistema industrializzato che sfrutterà metodologie consolidate e strumenti generalizzati ad alto livello di automazione. In tale ambito la disponibilità di software statistico costituisce un elemento fondamentale ai fini del buon esito del progetto.

LO SVILUPPO DI PACKAGE R

Negli anni recenti ricercatori statistici ed esperti informatici dell'Istituto hanno sviluppato numerosi strumenti IT che rispettano appieno le esigenze di un sistema industrializzato. Di questi, gran parte sono implementati con il linguaggio R, ponendo l'Istat tra i primi posti come contributore per numero di package resi disponibili in rete per la statistica ufficiale e le metodologie di indagine (http://cran.r-project.org/web/views/OfficialStatistics.html). L'impiego di R offre alcuni vantaggi: è un sistema open che la comunità scientifica può migliorare secondo uno spirito di ricerca e collaborazione, e permette di sviluppare sistemi software liberamente fruibili e completamente interoperabili, incentivando la diffusione di buone pratiche caratterizzate da elevati standard metodologici.

Gli strumenti R sviluppati nell'Istituto coprono diversi aspetti della produzione dei dati e possono essere scaricati dalle pagine dell'OTS, l'Osservatorio Tecnologico per Software generalizzati (www.istat.it/it/strumenti/metodi-e-software/softaware) o dal repository del Cran (http://cran.r-project.org). A breve tutti i software saranno comunque disponibili sul sito dell'OTS.

DISEGNO DEL CAMPIONE E TRATTAMENTO DEGLI ERRORI DI MISURA

SamplingStrata (http://cran.r-project.org/web/packages/SamplingStrata/index.html) è il package che permette di determinare la stratificazione ottima di un frame. È implementato un algoritmo genetico per la definizione degli strati, il metodo di Bethel-Chromy per l'allocazione campionaria negli strati e alcune funzioni per analizzare i risultati ottenuti.

MAUSS-R è uno strumento per la definizione del piano di campionamento. Specificati gli obiettivi e i vincoli operativi dell'indagine (costi della rilevazione, precisione desiderata delle stime, tipi di domini di studio, stratificazione), il software determina la dimensione campionaria minima. Il software è flessibile e facile da gestire.

SeleMix (http://cran.r-project.org/web/packages/SeleMix/index.html) è il package che implementa i principali metodi basati sui modelli per classi latenti per individuare errori potenzialmente influenti sulle stime campionarie (editing selettivo), permettendo di ottimizzare le risorse per la revisione interattiva dei dati rilevati, limitandole ai soli casi in cui il beneficio atteso è elevato.

PRODUZIONE DI STIME E MISURE DELLA VARIABILITÀ

ReGenesees è il software che gestisce tutte principali fasi della stima campionaria (dal calcolo dei pesi alla reportistica finale ed analisi della precisione delle stime) per le principali strategie campionarie complesse (cfr. NewsStat n. 4/2012). Dispone di una GUI potente, intuitiva e versatile, che rende amichevole e semplice l'uso anche ad utenti che non siano esperti di R. Recentemente, il package è stato oggetto di una valutazione indipendente effettuata dall'Istituto nazionale di statistica britannico (ONS). Lo studio ha comparato l'efficienza di ReGenesees e del software GES di Statistics Canada, rilevando come ReGenesees replichi correttamente i risultati di GES, garantendo un significativo incremento di efficienza con un dimezzamento in media i tempi di elaborazione.

EVER è il package che effettua il calcolo delle stime e degli errori di campionamento mediante il metodo Delete A Group Jackknife, una variante computazionalmente efficiente del tradizionale metodo Jackknife stratificato. Tale metodo permette di gestire stimatori non analitici ad elevata complessità che sono di particolare rilievo nella statistica ufficiale (ad esempio, gli indicatori di povertà relativa e gli indici di Laeken). Il package può integrarsi con ReGenesees.

INTEGRAZIONE DEI DATI

StatMatch (http://cran.r-project.org/web/packages/StatMatch/index.html) è il package che fornisce alcune funzioni R per effettuare il matching statistico, ovvero l'integrazione di due fonti di dati, quali indagini campionarie complesse, che condividono un certo numero di variabili comuni, riferite alla stessa popolazione obiettivo. Alcune funzioni possono anche essere utilizzate per imputare le mancate risposte parziali mediante metodi *hot-deck*. RELAIS, inoltre, è un toolkit per il Record Linkage. L'obiettivo del software è individuare i record che si riferiscono ad uno stesso oggetto del mondo reale, anche qualora lo rappresentino in modi diversi e provengano da sorgenti dati diverse. Il toolkit sfrutta le potenzialità di R per implementare il metodo di Fellegi-Sunter.