

IL NUOVO SOFTWARE BASATO SU R PER IL CALCOLO DELLE STIME E DEGLI ERRORI CAMPIONARI

di Diego Zardetto (zardetto@istat.it)

■ Nel mese di dicembre 2011 l'Istat ha reso disponibile la versione 1.0 del sistema ReGenesees (*R evolved Generalised software for sampling estimates and errors in surveys*). Il software può essere scaricato dal sito web dell'Istat e dal *repository* per il software *open source* della Commissione europea JOINUP.

CORRETTEZZA, SICUREZZA E RIPRODUCIBILITÀ

Le attività di calibrazione dei pesi campionari, il calcolo delle stime di interesse e la valutazione dei relativi errori di campionamento costituiscono uno snodo cruciale del processo di produzione dell'informazione statistica ufficiale.

Si tratta di attività complesse, la cui corretta esecuzione esige una buona conoscenza della teoria statistica sottostante, la piena consapevolezza dei piani di campionamento dell'indagine e una discreta "sensibilità" per i dati e la loro interpretazione dal punto di vista dei fenomeni indagati.

Queste qualità non possono essere integralmente surrogate da un software, per quanto sofisticato e potente esso possa essere. Tutti i principali istituti di statistica concordano, tuttavia, nel ritenere che la disponibilità di sistemi software altamente evoluti e la definizione di protocolli standard per il loro utilizzo siano elementi imprescindibili per garantire la correttezza, la sicurezza e la piena riproducibilità dei processi di produzione dell'informazione statistica. Questa visione strategica ha spinto, negli anni passati, l'Istat e altri INS a investire nello sviluppo in-house di software dedicati alla fase di calibrazione, calcolo delle

stime e degli errori campionari. La medesima visione strategica spinge, oggi, gli stessi istituti a rinnovare ed arricchire la loro offerta di sistemi software. Il sistema ReGenesees è il frutto più recente dello sforzo profuso dall'Istat in questa direzione.

REGENESEES IN UN GUSCIO DI NOCE

ReGenesees è un sistema software basato su R per l'analisi *design-based* e *model-assisted* di indagini campionarie complesse. Il nome del sistema intende enfatizzare la continuità dell'offerta di strumenti Istat per la calibrazione ed il calcolo delle stime e degli errori (di qui il richiamo al suo predecessore SAS: Genesees) sottolineando, al contempo, l'evoluzione e l'arricchimento di tale offerta attraverso R. ReGenesees, tuttavia, è il frutto di un progetto nuovo e indipendente: risponde, infatti, a un radicale cambiamento di logica applicativa che, oltre a consentire un più agevole e sicuro utilizzo del software, garantisce un notevole ampliamento della scelta di stimatori rispetto ai quali calcolare le stime e gli errori campionari. In estrema sintesi:

■ l'interazione dell'utente con ReGenesees avviene ad elevato livello di astrazione. Questo nuovo paradigma riduce sensibilmente il carico di lavoro sull'utente, incrementando l'usabilità del sistema e la sua robustezza rispetto a possibili errori. Nello specifico: l'utente non ha più bisogno di preprocessare i dati di indagine ricorrendo a programmi ad hoc: si limita a fornirli al software così come sono, corredandoli di metadati che descrivono in modo

simbolico il disegno di campionamento ed il modello di calibrazione; è il sistema stesso a trasformare i dati di indagine nelle complesse strutture necessarie a risolvere il problema di calibrazione e a calcolare stime ed errori;

■ oltre a totali, medie e frequenze, ReGenesees consente di calcolare stime ed errori campionari rispetto a quantili, rapporti, e, più in generale, rispetto a qualsiasi stimatore complesso, purché esprimibile come funzione differenziabile di stimatori di Horvitz-Thompson o di Calibrazione. Si noti che tali stimatori complessi possono essere definiti in modo del tutto libero: l'utente deve solo fornire al sistema l'espressione simbolica dello stimatore come funzione matematica. ReGenesees, infatti, è in grado di linearizzare automaticamente gli stimatori complessi, consentendone così la stima della varianza campionaria a costo zero per l'utente.

Entrambe le innovazioni fanno leva su un punto di forza specifico del linguaggio R: la sua capacità di elaborare informazioni simboliche. Sviluppare le medesime funzionalità in SAS sarebbe stato semplicemente impossibile. Ecco dunque, materializzato in ReGenesees, un esempio di ciò che abbiamo chiamato "opportunità del cambiamento tecnologico".

Il cambiamento tecnologico, d'altra parte, comporta sempre sfide e prezzi da pagare. La sfida più insidiosa affrontata dal progetto ReGenesees è stata quella di dimostrare che un sistema interamente sviluppato in R fosse davvero in grado di gestire efficientemente le ingenti moli di dati caratteristiche delle indagini campionarie che l'Istat conduce su vasta scala.

Oggi, in virtù delle evidenze empiriche e dei risultati riproducibili accumulati durante una estesa ed accurata fase di test e validazione, siamo certi che la sfida sia stata superata.