

IN ARRIVO FS4, NUOVO SOFTWARE OPEN SOURCE PER IL CAMPIONAMENTO

di Raffaella Cianchetta (cianchet@istat.it)

■ Il package open source FS4 (*First Stage Stratification and Selection in Sampling*), attualmente in versione prototipale, è un software generalizzato per la stratificazione e selezione delle unità di primo stadio in disegni di campionamento a due o più stadi, sviluppato completamente in linguaggio R e dotato di un'interfaccia utente di tipo grafico (GUI). La versione 1.0 del package sarà a breve disponibile sul sito web dell'Istituto nazionale di statistica, nelle pagine dedicate all'Osservatorio Tecnologico per i Software Generalizzati.

CARATTERISTICHE

Il software effettua, all'interno di ciascun dominio di stima, il calcolo di una soglia dimensionale per una data misura di ampiezza. Le unità di primo stadio (PSU, *Primary Sampling Unit*) con la misura di ampiezza superiore alla soglia sono identificate come autorappresentative (SR, *Self Representative*) e ciascuna di esse forma uno strato a sé. Le restanti PSU non autorappresentative (NSR, *Non Self Representative*) sono ordinate per la misura di ampiezza e suddivise in strati di dimensione approssimativamente costante alla soglia corretta e con PSU aventi dimensioni il più possibile omogenee. La selezione di un campione a dimensione fissa di PSU per ciascuno strato è realizzata con il metodo Sampford, implementato nella funzione `UPsampford` (selezione con probabilità proporzionale alla dimensione e senza reinserimento) del package R "sampling". FS4 determina, per ogni PSU, la dimensione del campione di unità finali sulla base di dimensioni pianificate in ingresso.

La flessibilità del software è rappresentata dalla possibilità fornita all'utente di:

- scegliere la misura d'ampiezza per definire la stratificazione e la probabilità di inclusione per il campionamento PPS (*Probability Proportional to Size*);
- inserire in input una singola o plurima (variabile tra domini) dimensione pianificata del campione di unità finali da osservare in ogni PSU;
- lanciare la procedura in due passi separati, consentendo di visionare un primo output e sulla base di questo modificare i parametri di input.

Il package dispone di una GUI user-friendly, che ne consente l'utilizzo anche ad utenti non conoscitori del linguaggio R, implementata in Tcl/Tk tramite il package R "tcltk".

L'ORGANIZZAZIONE DELLE MASCHERE

Nella barra dei menu sono indicate le funzioni disponibili distinte per macrofunzionalità, ad eccezione della funzione principale di stratificazione. La sua selezione rende visibile una maschera nella quale l'utente ha la possibilità di scegliere due dataframe tra quelli presenti all'interno del workspace: quello della popolazione di PSU e quello di allocazione. La scelta di ciascun dataframe rende visualizzabile l'elenco delle variabili che lo compongono, alcune delle quali saranno selezionate obbligatoriamente per popolare la maschera. I parametri di input obbligatori sono rappresentati dal numero di PSU campione da selezionare in ciascuno strato e dalla dimensione singola o plurima pianificata del campione di

unità finali da osservare in ogni PSU. Il parametro concernente il lancio parziale della procedura è impostato su "Yes". I campi obbligatori di output sono relativi ai nomi che l'utente dovrà assegnare ai quattro dataframe di output resi disponibili: i primi due a livello di dominio, il terzo ed il quarto rispettivamente a livello di strato e di PSU. Il primo output visualizza l'elenco delle PSU campione SR e NSR prima della stratificazione dell'insieme PSU NSR, mentre nel secondo si aggiunge la dimensione del campione delle unità finali distinto per PSU SR e NSR. Il terzo output evidenzia, per ogni strato, il numero di PSU campione ed il numero di PSU totali. Il quarto output fornisce, per ogni PSU, alcune informazioni quali, ad esempio, le probabilità d'inclusione e la frazione di campionamento. È possibile visualizzare e rimuovere i dataframe di output e tutti quelli appartenenti al workspace grazie a specifiche funzioni del menu "Tools". L'importazione di file e l'esportazione di dataframe, nei formati ".txt" e ".csv", sono disponibili mediante funzionalità del menu "Dati".