IL NUOVO PACCHETTO R SELEMIX PER L'EDITING SELETTIVO

di Ugo Guarnera (guarnera@istat.it)

Da alcuni mesi è disponibile sul repository del CRAN (http://cran.rproject.org) il nuovo pacchetto R SeleMix (Selective Editing via MIXtures) per l'individuazione degli errori influenti nei dati da indagini economiche. Il pacchetto, sviluppato dai ricercatori del servizio metodologico dell'Istat, implementa una metodologia di editing selettivo nata all'interno dell'Istituto e presentata in diversi convegni e seminari scientifici in ambito nazionale ed internazionale.

IL TRATTAMENTO DEGLI ERRORI DI MISURA

Nel contesto degli istituti nazionali di statistica è largamente diffuso il convincimento che una parte rilevante delle risorse impiegate nell'intero processo di produzione dell'informazione statistica debba essere dedicata all'individuazione e al trattamento degli errori di misura. Le attività associate al trattamento degli errori nei dati di indagine sono raggruppabili essenzialmente in due categorie:

1) attività basate su analisi interattiva dei dati (editing interattivo);
2) attività che si avvalgono di tecniche di controllo e correzione automatiche.

La prima categoria si riferisce a quel complesso di operazioni di revisione, compresi il ri-contatto dei rispondenti, l'accurato controllo del questionario ecc., che garantiscono un'elevata affidabilità nel processo di rimozione degli errori, ma che implicano nel contempo costi elevati in termini di tempi e personale specializzato impiegato. I metodi automatici viceversa utilizzano procedure e software in grado di processare grandi moli di dati in tempi ridotti e richiedono essenzialmente solo le risorse necessarie alle attività di manutenzione e "setting" dei parametri. La costante riduzione di fondi destinati alla statistica ufficiale con la conseguente diminuzione delle risorse a disposizione delle strutture di produzione, comportano la necessità di limitare il ricorso all'editing interattivo ai soli casi in cui il beneficio atteso in termini di accuratezza delle stime sia significativo. Questa esigenza ha alimentato negli ultimi anni un attivo filone di ricerca volto alla definizione di metodologie capaci di razionalizzare i criteri di selezione dei dati da destinare ad analisi interattiva (editing selettivo). Nonostante gli sforzi profusi dalla ricerca nel settore tuttavia, la maggior parte delle tecniche in uso sono tuttora di natura euristica e si avvalgono di procedure ad hoc. Di conseguenza i criteri sottostanti la selezione delle unità da sottoporre a revisione interattiva risultano spesso non espliciti o comunque non riconducibili ad una seppur rozza "stima" del rapporto costo/benefici associato alla revisione.

Con il software SeleMix si è voluto proporre uno strumento generalizzato che, basato su ipotesi esplicite circa la distribuzione dei dati e il meccanismo di errore, consenta di affrontare il problema dell'editing selettivo in termini statistico-inferenziali. La modellizzazione utilizzata, coerente con assunzioni distribuzionali standard nel contesto del trattamento di variabili economiche, si è dimostrata appropriata in diverse situazioni di interesse, come evidenziato da numerosi test effettuati su dati simulati e reali. Per queste caratteristiche SeleMix si candida in modo naturale come strumento "standard" dell'Istat per l'editing selettivo.

UN APPROCCIO "MODEL-BASED" ALL'EDITING SELETTIVO

A differenza dalla maggior parte degli strumenti di editing selettivo, SeleMix si basa sulla modellizzazione esplicita dei dati "veri" (cioè non contaminati) e del meccanismo di errore. In particolare il ricorso a un modello di contaminazione a classi latenti consente di catturare la natura "intermittente" degli errori di misura e quindi di attribuire ad ogni osservazione di indagine una probabilità di presenza dell'errore. Quest'ultima può a sua volta essere interpretata come misura di "anomalia" e quindi essere utilizzata nell'ambito dell'analisi degli outlier. Inoltre per ogni valore osservato è fornita una previsione del corrispondente valore "vero" (e quindi dell'errore) basata sulla appropriata distribuzione condizionata. Questa caratteristica consente di associare il numero di unità da revisionare all'accuratezza desiderata per le stime di interesse. È importante inoltre sottolineare che il metodo non richiede la disponibilità simultanea di dati contaminati e "puliti" su cui stimare il modello d'errore. Il package SeleMix è costituito da tre funzioni principali che si occupano rispettivamente di stimare il modello di contaminazione (ml.est), di calcolare le previsioni dei valori veri (pred.y), e di selezionare le unità contenenti errori potenzialmente influenti sulla base di una soglia di accuratezza specificata dall'utente (sel.edit).

Un ultimo cenno va infine dedicato alla possibilità di utilizzare il sofware anche in presenza di dati incompleti. In tal caso le previsioni che corrispondono a valori mancanti possono essere viste come imputazioni "robuste". Informazioni su aspetti tecnici e metodologici sono disponibili sul sito web dell'Istat nelle pagine dedicate all'Osservatorio tecnologico per i software generalizzati.