IL PACKAGE "STATMATCH" PER LO STATISTICAL MATCHING E L'IMPUTAZIONE IN R

di Marcello D'Orazio (madorazi@istat.it)

Ad inizio di dicembre è stata rilasciata la versione 1.2.0 del package "StatMatch" per l'ambiente R (http://CRAN.R-project.org/package=StatMatch) e nei prossimi mesi il software sarà disponibile anche nelle pagine web del rinnovato Osservatorio Tecnologico dei Software dell'Istituto Nazionale di Statistica.

StatMatch (Statistical Matching) è un package open source, che contiene funzioni per l'integrazione di dati attraverso tecniche di abbinamento statistico (statistical matching o data fusion). Nel corso degli anni le funzionalità sono state arricchite e migliorate anche grazie alle attività del progetto "ESSnet on Data Integration" coordinato dall'Istat (http://www.cros-portal.eu/content/data-integration-1).

L'ABBINAMENTO STATISTICO PER INTEGRARE LE FONTI DATI

Le tecniche di abbinamento statistico si propongono di integrare fonti dati riferite alla medesima popolazione (es. indagini campionarie) che condividono alcune informazioni (variabili comuni), con l'obiettivo di studiare le relazioni tra variabili non osservate congiuntamente nella stessa indagine; ad esempio, in ambito sociale l'integrazione dei dati sul reddito degli individui con i dati sui consumi, permette di approfondire la relazione tra redditi-consumi. Lo statistical matching può essere anche utilizzato per creare un data set omnicomprensivo su cui condurre studi di microsimulazione.

L'abbinamento statistico può essere finalizzato alla stima di parametri di modelli matematico-statistici (correlazione, regressione, ecc.) o di tabelle in cui si incrociano variabili non osservate congiuntamente. In alternativa si può voler creare un data set "sintetico" in cui siano disponibili anche tutte le variabili di interesse. Si parla di data set sintetico perché solo una parte delle variabili è stata effettivamente osservata mentre un'altra (variabili non disponibili nella stessa fonte dati) è stata ricostruita utilizzando lo statistical matching. Numerose delle tecniche di abbinamento derivano dall'imputazione dei valori mancanti di una indagine. Per maggiori dettagli sulle tecniche di abbinamento statistico si veda la monografia "Statistical matching: Theory and practice" (M. D'Orazio, M. Di Zio e M. Scanu, 2006).

PRINCIPALI FUNZIONALITÀ DI STATMATCH

Il package StatMatch contiene funzioni per condurre statistical matching e funzioni di supporto. Diverse funzioni implementano metodi di tipo non parametrico: si tratta di procedure di imputazione "hot deck" dei valori delle variabili mancanti nella fonte dati considerata come "ricevente". È possibile utilizzare il metodo del donatore di distanza minima (nearest neighbour hot deck): per ogni unità ricevente si cerca nella seconda fonte dati il donatore più vicino in base ad una distanza calcolata sulle variabili comuni alle fonti dati. Sono disponibili numerosi tipi di distanze, semplici o complesse. La selezione del donatore più vicino può essere vincolata in modo da evitare che una unità possa essere utilizzata più volte come donatore. Ciò comporta un maggiore sforzo computazionale ma, in genere, fornisce una distribuzione delle variabili imputate più vicina a quella osservata nella fonte dati donatrice. Sono state implementate anche diverse forme di scelta casuale del donatore (random hot deck) con possibilità di utilizzare anche dei pesi delle unità.

In presenza di variabili continue è possibile utilizzare dei metodi misti che, per creare il data set sintetico, combinano l'uso di modelli di regressione con il donatore di distanza minima.

L'abbinamento statistico di dati di indagini campionarie complesse può essere condotto attraverso metodi che combinano opportuni modelli di regressione e tecniche di ripondera-zione dei pesi campionari.

Infine alcune funzioni di StatMatch permettono lo studio dell'incertezza relativa alla probabilità che si verifichi uno o più eventi dati dalla combinazione di modalità delle variabili non osservate congiuntamente. Questo approccio non conduce ad una stima puntuale della probabilità di interesse quanto piuttosto ad un intervallo di valori plausibili. Tanto più stretto è l'intervallo tanto minore è l'incertezza relativamente all'evento in questione.