“Metodi per l’integrazione tra la base dati Health Search e l’indagine Istat sulle condizioni di salute”

Marco Di Zio – Istat

La comunicazione ha come obiettivo la presentazione del contesto metodologico e dei metodi utilizzati per l’integrazione dei dati Health-Search e l’indagine Istat sulle condizioni di salute. Il contenuto può essere suddiviso in tre blocchi. Si inizia illustrando il contenuto informativo del problema, mettendone in risalto gli aspetti principali nell’ambito di una procedura di integrazione statistica. La non presenza di unità in comune nelle due fonti di dati porta all’utilizzo di tecniche sviluppate nell’ambito dello statistical matching (abbinamento statistico).

Nello statistical matching l’obiettivo è quello di analizzare il comportamento di variabili osservate distintamente nei due data set (Y quelle osservate in un data set e Z quelle nell’altro) sfruttando l’informazione contenuta nelle variabili X osservate in entrambe le fonti di dati.

La parte centrale della presentazione è dedicata alla presentazione del contesto metodologico dello statistical matching ed all’illustrazione delle tecniche sviluppate in questo ambito. In particolare l’attenzione viene posta sul problema dell’assunzione dell’indipendenza condizionata (IC) su cui si basano le usuali procedure e sull’illustrazione di tecniche per l’analisi dell’incertezza che permettono di fare inferenza non basata sull’ipotesi IC.

L’indipendenza delle variabili non osservate congiuntamente condizionatamente all’osservazione delle variabili comuni è una ipotesi forte, in essa si sta assumendo che la conoscenza della X sia sufficiente a spiegare il comportamento delle Y e Z. Il problema è che questa ipotesi non può essere verificata con i dati a disposizione perché i valori assunti dalle variabili Y e Z non sono mai osservati congiuntamente.

L’analisi dell’incertezza non si basa su tale assunzione, l’inferenza prodotta tramite questo approccio però non produce stime puntuali del parametro di interesse ma solo degli intervalli i cui estremi sono dati dai valori che il parametro può assumere condizionatamente ai dati osservati. Nel caso in cui le variabili X, Y e Z siano variabili categoriali, un parametro di interesse può essere la frequenza di ogni singola cella della relativa tabella di contingenza. In questo caso l’analisi dell’incertezza restituirà un minimo e massimo per la frequenza di ogni singola cella. Tali limiti possono essere calcolati analiticamente tramite la disuguaglianza di Fréchet.

Nel problema di integrazione analizzato c’è però un ulteriore complicazione. Nelle variabili comuni si potrebbe osservare un disallineamento (differenza delle distribuzioni non causata solo da variabilità campionaria) dovuto al fatto che nell’indagine Istat i rispondenti sono gli individui, mentre in HS sono i medici. Nel primo caso una componente di soggettività nella dichiarazione di alcune patologie potrebbe essere un fattore importante da tenere in considerazione.

A tal fine sono state sviluppate delle tecniche che estendono i metodi utilizzati nello statistical matching. In questa prima applicazione, è stato fatto uso di un modello utilizzato nell’ambito dei problemi che trattano variabili misclassificate. In tale modello si ipotizza che se l’individuo non è affetto da patologia, con certezza l’individuo ci fornisce l’informazione corretta, ovvero non essere affetto da malattia. Viceversa, se l’individuo è affetto da malattia, la probabilità che questo ci dia l’informazione corretta è pari ad una probabilità incognita p strettamente minore di 1. Nelle slide viene illustrato come stimare la probabilità p di corretta classificazione nel caso di variabili dicotomiche. Questa probabilità viene poi utilizzata per

prevedere nel data set Istat il valore della variabili affette da una componente soggettiva. Una volta predetto questo valore, il matching viene condotto sotto l'ipotesi di indipendenza condizionata.

Viene anche effettuata l'analisi dell'incertezza. In quest'ultimo caso non è però possibile utilizzare metodi analitici per il calcolo degli estremi degli intervalli che descrivono i valori plausibili per il parametro oggetto di interesse. Per questo motivo, grazie alla collaborazione con l'Università La Sapienza di Roma, sono stati sviluppati degli algoritmi ad-hoc.