

Un data warehouse per il Comune di Milano

Gian Paolo Almirante

Un progetto, sviluppato all'Università Milano-Bicocca su software SAS, nato per integrare dati anagrafici e dichiarazioni dei redditi, fornisce uno spaccato preciso della società milanese.

Per far fronte alla crescente domanda di informazioni da parte della Pubblica Amministrazione in relazione alle politiche di servizio il Settore Statistica del Comune di Milano ha finanziato un progetto, sviluppato dal Dipartimento di Statistica dell'Università Milano-Bicocca (Unimib), chiamato AMeRiCA (Anagrafe Milanese e Redditi Individuali con Archivio). Il progetto, basato su un data warehouse SAS, si propone di integrare i dati anagrafici con le dichiarazioni dei redditi delle persone fisiche. "Oltre all'innovatività dal punto di vista statistico-informatico – dichiara **Biancamaria Zavanella**, Facoltà di Scienze Statistiche, Dipartimento di Statistica, Statistica Economica - Unimib – uno degli aspetti fondamentali del progetto è che mai sino ad ora, eccezion fatta per il censimento, è stato possibile rappresentare in modo così preciso e dettagliato lo spaccato della società milanese. Il progetto AMeRiCA presenta anche un'ulteriore potenzialità: la struttura del data warehouse permette infatti di integrare i dati già esistenti con ulteriori fonti aumentando così le dimensioni di analisi".

Come illustra – **Mario Mezzanzanica**, Facoltà di Scienze Statistiche, Organizzazione dei Sistemi Informativi Aziendali-Unimib – il progetto si è articolato su varie fasi, ad ognuna delle quali corrisponde una serie di procedure il cui scopo è migliorare la qualità dell'informazione contenuta nei dati e riorganizzarne la struttura.

Le fonti utilizzate sono due: l'Anagrafe del Comune di Milano e i modelli per la dichiarazione dei redditi delle persone fisiche; i dati anagrafici contengono sia i cittadini residenti, identificati come 'attivi', sia i cittadini iscritti alla lista Aire (Anagrafe degli Italiani Residenti all'Estero). La prima difficoltà incontrata deriva dal fatto che Anagrafe e Ministero delle Finanze fanno uso di classificazioni diverse per uno stesso tipo di dato o utilizzano entrambe classificazioni non adeguate alla successiva fase di analisi. Le principali classificazioni che sono state quindi rielaborate e integrate in: nazionalità, ricondotta ad un unico standard prevedendo l'aggregazione in alcune macroaree; residenze, con diverse tipologie di aggregazioni in

modo da permettere analisi più o meno dettagliate dal punto di vista territoriale, evidenziando gli indirizzi rappresentanti eventuali convivenze e le situazioni indicanti casi anomali come ad esempio i 'senza fissa dimora'; i rapporti di parentela, semplificati e riclassificati per permettere una ricostruzione più agevole delle famiglie; l'età, aggregate in differenti classi e infine le famiglie, classificate seguendo gli standard attualmente in uso, prevedendo la distinzione tra nuclei famigliari e distinguendoli in base alla numerosità e alla tipologia dei componenti.

Una volta acquisiti, i dati hanno subito una prima fase di 'pulizia' dal punto di vista sintattico, riguardante cioè la correttezza formale. Al termine del processo di pulizia sintattica è stata operata anche una pulizia di tipo semantico, eliminando ad esempio i dati palesemente errati dovuti ad errori nella compilazione dei moduli di dichiarazione dei redditi o alla presenza di moduli diversi ed incompatibili all'interno del medesimo anno di imposta.

Un volta acquisiti e puliti, i dati sono stati riorganizzati seguen-