OnTheMap™: Synthetic Data Protection

Local Employment Dynamics — All About Jobs

Background

OnTheMap is the first partially synthetic data released by the Census Bureau. The web-based[1] application using this data was first released in 2006. It is updated annually and now in its third version. The next version is scheduled for release in December 2009. A unique feature of the synthetic data approach is that OnTheMap provides detailed data about where workers live, where workers work, and the combination of home and work location pairs at the census block level, while confidentiality is still strictly protected.

How Synthetic Data Work

Imagine the United States to be a giant jigsaw puzzle made of 8 million unique, disjoint pieces also known as census blocks in OnTheMap. The boundaries for these pieces may change from year to year; so will the content of these pieces.

However, once the content for each census block has been estimated, the count results can be aggregated easily for standard higher geographies such as tracts, counties, cities, or any layer that can be created as aggregates of census blocks.

Each of the census blocks may have none or some workers living or working (or both) in it. Each count of workers by categories of age, earnings, and industry, as well as their aggregate, may vary from year to year.

Unlike censuses or surveys from which confidential micro-data are tabulated into statistical tables, OnTheMap employs a calibrated Bayesian[2] modeling approach to produce its synthetic data on workers in residential area by census block, from which statistical results are tabulated.

In particular, we used the data from the 2000 Census Transportation Planning Package[3] and built a priori probability distributions of the count of workers residing in a census block, conditional on their characteristics on age, earnings, industry, and place of work. In addition, the parameters in the prior distributions were controlled by infusing additional noise for confidentiality protection.

The observed confidential data in 2002, the first year such micro-data were used by OnTheMap, form the likelihood function for the year 2002. The combination of the 2000 prior distribution and the 2002 likelihood function yielded the posterior predictive distribution for the year 2002. This probability distribution was then used to synthesize workers in residential areas for OnTheMap in 2002.

Thus, random draws from the posterior predictive distributions preserve the statistical properties and analytical validity of the underlying confidential data. In addition, there are no actual worker's data involved in the tabulation of statistical results by home and work location pairs for OnTheMap.

---

[1] Available at http://lehd.did.census.gov on May 26, 2009.
[2] Little, Roderick (2006). "Calibrated Bayes: A Bayes/Frequentist Roadmap." The American Statistician,60,213-223.
[3] Available at http://www.fhwa.dot.gov/ctpp/ on May 26, 2009.

The first draw from the posterior predictive distribution in 2002 then became the prior distribution for workers in 2003. This prior distribution was combined with the 2003 observed confidential data to yield the posterior predictive distribution for 2003, from which synthetic workers were randomly generated for OnTheMap in 2003.

This approach was repeated iteratively for subsequent years to complete one implicate for the nation, which was then implemented in OnTheMap. The approach is analogous to contingency table modeling with a Multinomial-Dirichlet natural conjugate prior/posterior family.

Notes and Remarks

1. Synthetic data modeling is substantially more complex than simply anonymizing home and work location pairs.

2. No actual worker data are released in the synthetic data modeling approach. Therefore, re-identification, if it occurs, applies only to a synthetic worker, not an actual worker. A synthetic worker has no actual personal information to reveal or to re-identify.

3. Implicates are expected to vary due to probabilistic fluctuations, but they should retain similar statistical properties and analytical validity as designed by the synthetic data modeling approach. Release of multiple implicates can reduce the degree of protection offered by the synthetic data modeling approach.