NBA Game outcome prediction

Lee Richardson, Daren Wang, Xiaofeng Yu, Chi Zhang

Carnegie Mellon University

Objectives

Introduction

midas.jpg

Data

The following publically available data sources were utilzedi:

- **U.S. Census Products**
  - **ACS SF** (Summary Files)- Counts for marginal variables per Census Tract. Available in 5-year, 3-year, and 1-year format.
  - **ACS PUMS** (Public Use Micro Samples)- De-identified individual records available in a 5-year, 3-year, and 1-year format. Subset of actual records from corresponding ACS SF.
  - **TIGER** (Topologically Integrated Geographic Encoding and Referencing) Products- Census tract maps for each state. Used for 2000 and 2010 census-defined tract maps.
  - **County-PUMA Relationship** table. Allows switching between different geographies of PUMA and county. Available for 2000 and 2010.

Geography

- Must reconcile geographies defined by different entities
  - e.g. ACS SF are available at tract level
  - PUMS data's base level is the PUMA

censusgeochart.png

Figure 1: From U.S. Census. Geographical hierarchy diagram.

Conclusion

Using our adapted methods, we are able to reproduce the results of RTI. These produced microdata can be used in epidemiology models used for the study of the spread of disease.

Future Work

References

Acknowledgements

This research was made possible under NIH Grant MIDAS. We would also like to thank Carnegie Mellon University's Department of Statistics and the SURE 2014 proram as well as Dr. Bill Eddy for his guidance and support.

Contact Information

- Web: portal.isg.pitt.edu/midas/home.dob
- Email: sgallagh@andrew.cmu.edu

Results

As a result of our procedure for generating microdata, we are able to reproduce the results of Allegheny County as done by a previous group, the Research Triangle Institute (RTI) [**?**]. We produce micro-data that is accurate for each tract, per the U.S. Census data.

Method

For each Census tract in the U.S., we generated micro-data in the following manner:

1. Find aggregate counts for householder age, householder race, household size, and household income from 2007-2011 ACS SF
2. Implement IPFP procedure, similar to that in the one of TransSim by Beckman, Baggerly, and McKay [**?**]
3. Sample from 2007-2011 PUMS records from corresponding PUMA to generate synthetic data
4. Sample from appropriate spatial polygons to assign a latitude and longitude coordinate, as shown in Figure 1

Sampling Example

allegheny_sample.pdf

Figure 2: Example of how coordinates were selected for synthetic households in Allegheny County, PA. Between 5 and 10 locations were sampled per tract.