Generating a Synthetic U.S. Population- MIDAS

Shannon Gallagher

Carnegie Mellon University, Dept. of Statistics

Objectives

- Generate **micro-data** which simulates U.S. population
  - Households and Person Records
- Reconcile hierarchies of U.S. geography
- Utilize **IPFP** to create populations
  - Accurately reflect U.S. population
- Assign synthetic households appropriate locations
  - Latitude and Longitude
- Reproduce and improve upon results of a previous group, RTI

Introduction

MIDAS (Models of Infectious Disease Agent Study)

- Develops computational models of the interaction of infectious agents
- Models run on synthetic **micro-data**
  - Household records

House ID	Geo ID	Race	Income	Age	Size	Lat.	Long.
11815531	0103003	2	$29,300	50	3	40.43826	-79.9835

  - Person records

House ID	Geo ID	Age	Sex	Race	SP Order	Relate	School ID	Work ID
11815531	0103003	50	2	2	1	0	NA	513987696
11815531	0103003	19	1	2	2	2	NA	NA
11815531	0103003	18	2	2	3	2	450086847	NA

Generating micro-data for the U.S

- Utilize **U.S. Census Products**
  - American Community Survey (ACS) data
  - TIGER Map Products
- Implement **IPFP** (Iterative Proportional Fitting Procedure) and assign locations

MIDAS

Funded by the National Institutes of Health

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

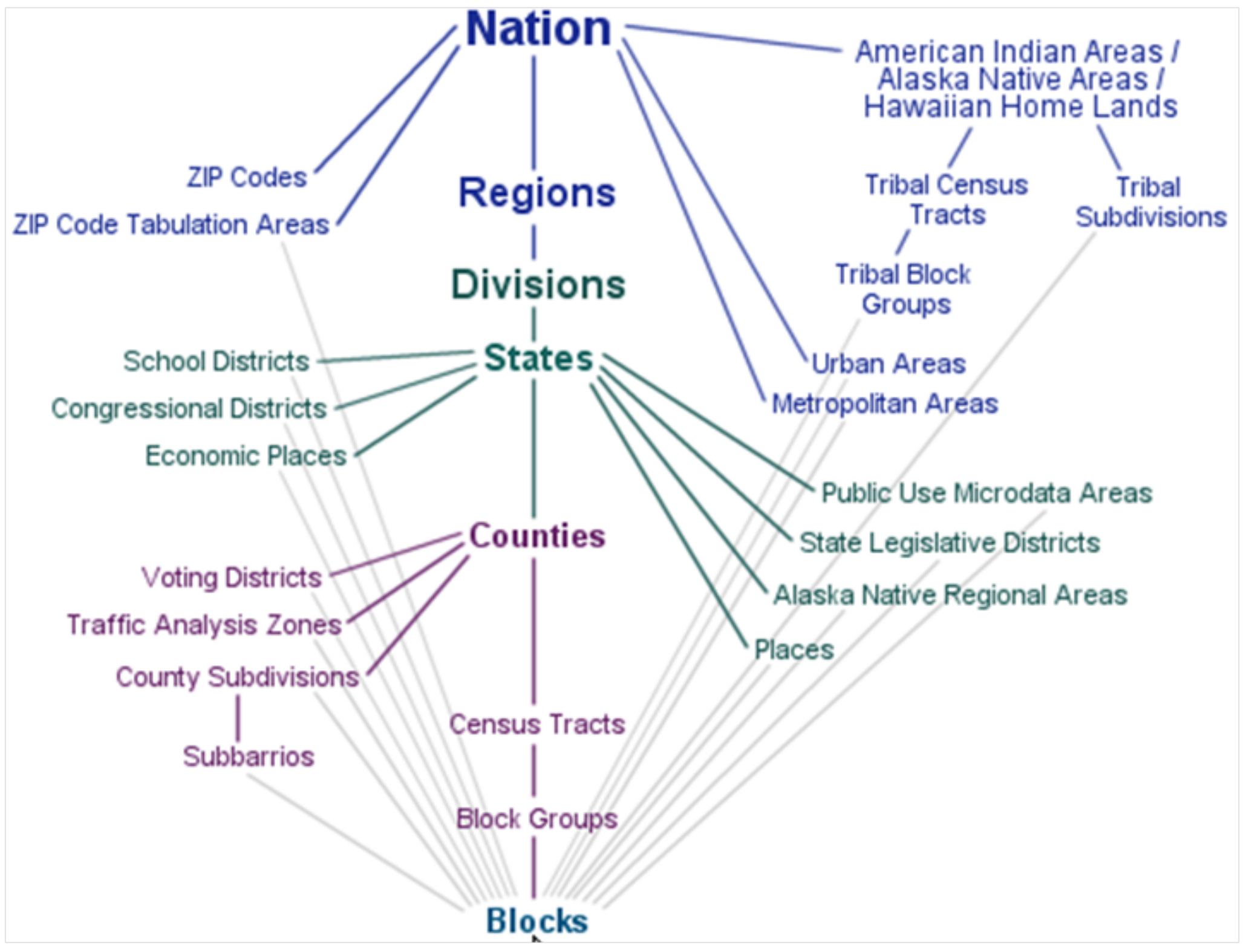

Figure 1: From U.S. Census. Geographical hierarchy diagram.

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) [2]. 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 [1]
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

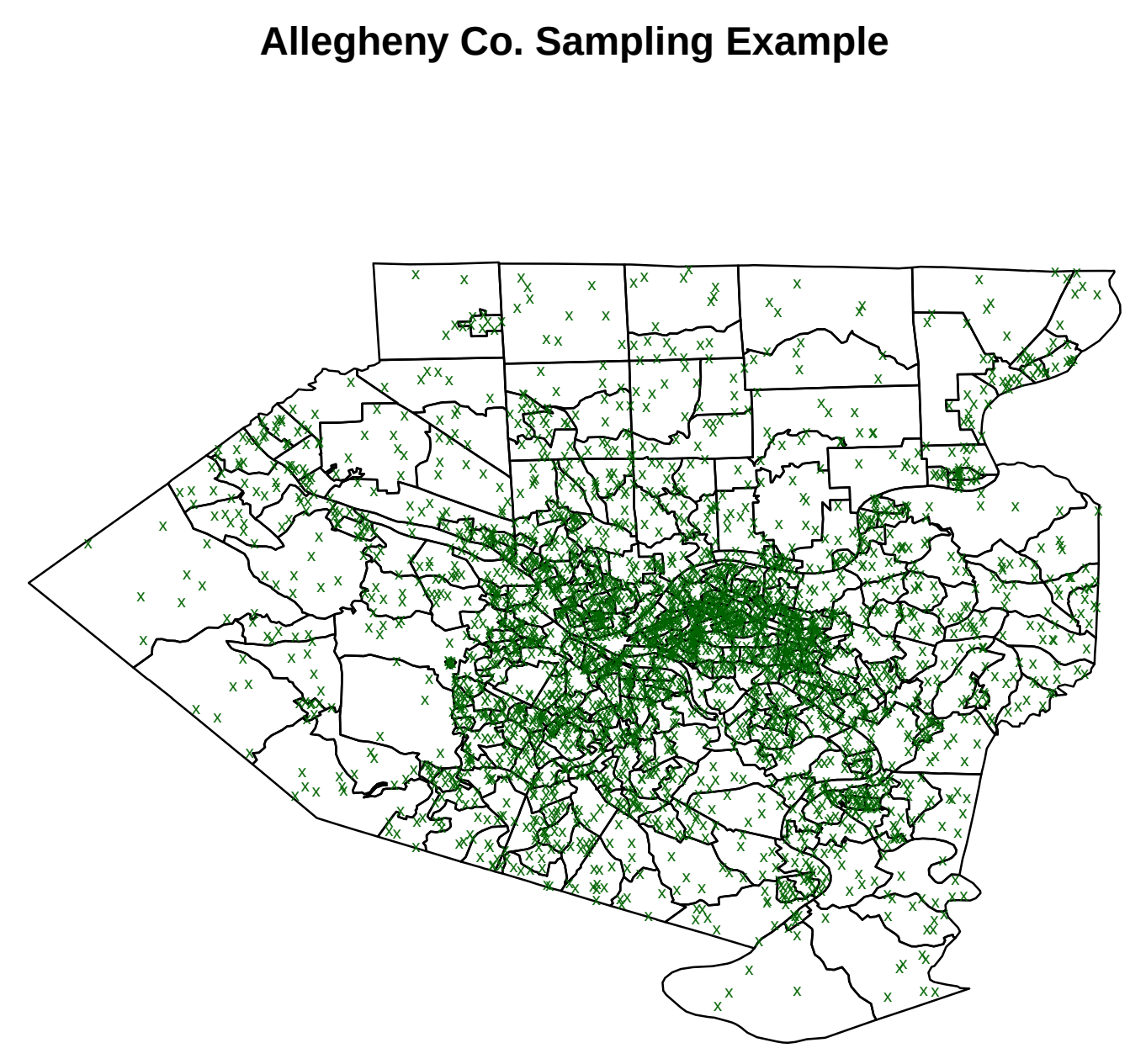

Figure 2: Example of how coordinates were selected for synthetic households in Allegheny County, PA. Between 5 and 10 locations were sampled per tract.

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

- Include school and workplace data for individual synthetic records
- Generate group quarters, such as university dorms, prisons, and military bases, and group quarters residents
- Allow the user to specify the year, perhaps to study past epidemics

References

[1] Baggerly Beckman and McKay.
"Creating synthetic baseline populations".
Transportation Research Part A: Policy and Practice **30**(6): 415-429. http://www.sciencedirect.com/science/article/B6VG7-3VWT9NR-2/2/829cf6e7e3efc66bcb969d9709c6a6e3, 1996.

[2] W.D. Wheaton.
"U.S. Synthetic Population 2010: Quick Start Guide".
RTI International. Retrieved from
https://www.epimodels.org/midasdocs/SynthPop/2010_synth_pop_ver1_quickstart.pdf, 2014.

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

CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900