Generating Synthetic Populations Using Public Use Microdata

October 25th, 2005

Bill Wheaton
Manager,
Geospatial Science & Technology Program
wdw@rti.org

Bernadette Chasteen
GIS Analyst/Programmer
Geospatial Science & Technology Program
bmc@rti.org
Acknowledgments

- This work funded under the Models of Infectious Disease Agent Study (MIDAS) for the National Institute of General Medical Sciences (NIGMS)
- RTI wishes to thank Irene Eckstrand and the MIDAS Steering Committee for funding and support.

- Prior Research and Techniques:
  - TranSims: Transportation Analysis Simulation System.
    - http://transims.tsasa.lanl.gov
MIDAS Objectives and Overview

- Research Groups Develop Agent-based Models
  - Johns Hopkins (and affiliates)
  - Emory University (team moving to Washington)
  - VBI
  - New Research Groups to be Added

- RTI Provides:
  - Computing Infrastructure
    - 2 64-Node Linux Clusters
  - Data
    - Geospatial Data
  - Data to Support All Modelers
Agent-Based Models

- **Stochastic Microsimulation Models**
  - Large, complex, quantitative, dynamic, stochastic, behavioral, spatial models

- **Microsimulation**
  - Based on individual or unit-level data, not aggregated data
  - Public Use Microdata (PUMS)
  - Issue: Confidentiality

- **How Agent-Based Models work**
  - Agents
  - Agent interactions (Social Networks)
    - Family
    - School/Work
    - Neighborhood
    - Random

- **Example: Schelling**
Micro (Individual) vs. Macro (Aggregate) Data

- Macro/Aggregate Data:
  - Census counts by geographic area
  - State, County, Census Tract, Block Group
  - *Does not provide information on family structure!*

- Micro/Individual Data:
  - Individual or Household-level data
  - *Family Structure Maintained!!*
“Microsimulation methodologies aim at building large-scale data sets on the attributes of individuals or households and on the attributes of individual firms or organizations and at analyzing policy impacts on these micro-units through the simulation of economic, demographic and social processes.

“If we do not have a micro data base on individuals and households then there is a necessity to simulate one”


Thus the Idea: Produce a national, geospatially-explicit synthetic population for the United States.
Creating a Synthetic Population: Data Inputs and Techniques

- **Block-group Level Demographics**
  - SF3

- **Public Use Microdata (PUMS)**
  - Actual Census long-form records (from U.S. Bureau of the Census)
  - Household and individual level data
  - Family structure maintained
  - 5% Sample within Public Use Microdata Areas (PUMAs)
  - PUMAs contain at least 100,000 persons

- **Household Locations**
  - Randomly generated w/in block groups

- **Iterative Proportional Fitting (IPF)**
  - Uses conditional probabilities to fill out a synthetic population that matches SF3 counts based on PUMS samples.
Geographical Context

- Counties
- Census Tracts
- Block Groups
- Public Use Microdata Areas (PUMAs)
- Households
- “Clone” particular records of the 5% PUMS sample (red outlines) to match census counts at block group level (black outlines)
### PUMS Examples

#### Household Record

<table>
<thead>
<tr>
<th>SERIALNO</th>
<th>PUMA5</th>
<th>HWEIGHT</th>
<th>PERSONS</th>
<th>VACSTAT</th>
<th>UNITTYPE</th>
<th>BEDRMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>103</td>
<td>00905</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>455</td>
<td>01000</td>
<td>31</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>457</td>
<td>00904</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>.</td>
</tr>
<tr>
<td>715</td>
<td>01000</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Person Record

<table>
<thead>
<tr>
<th>Serialno</th>
<th>Pnum</th>
<th>Pweight</th>
<th>Sex</th>
<th>Age</th>
<th>Race1</th>
<th>Earns</th>
</tr>
</thead>
<tbody>
<tr>
<td>103</td>
<td>1</td>
<td>22</td>
<td>2</td>
<td>38</td>
<td>1</td>
<td>0083000</td>
</tr>
<tr>
<td>455</td>
<td>1</td>
<td>34</td>
<td>1</td>
<td>48</td>
<td>1</td>
<td>0034000</td>
</tr>
<tr>
<td>455</td>
<td>2</td>
<td>38</td>
<td>2</td>
<td>45</td>
<td>1</td>
<td>0043000</td>
</tr>
<tr>
<td>457</td>
<td>1</td>
<td>22</td>
<td>1</td>
<td>41</td>
<td>2</td>
<td>0000000</td>
</tr>
</tbody>
</table>

www.rti.org
Iterative Proportional Fitting (IPF)

- Conditional Probabilities
  - Once calculated, used to choose PUMS records.

- Before IPF

<table>
<thead>
<tr>
<th>Household Attributes</th>
<th>Economically Active</th>
<th>Employee</th>
<th>Self-Employed</th>
<th>On Welfare</th>
<th>Un-Employed</th>
<th>Row Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWD</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
<td><strong>0.825000</strong></td>
</tr>
<tr>
<td>Married</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
<td><strong>0.175000</strong></td>
</tr>
<tr>
<td>Column Constraint</td>
<td>0.054054</td>
<td>0.689189</td>
<td>0.027027</td>
<td>0.027027</td>
<td>0.202703</td>
<td>1</td>
</tr>
</tbody>
</table>

- After IPF

<table>
<thead>
<tr>
<th>Household Attributes</th>
<th>Economically Active</th>
<th>Employee</th>
<th>Self-Employed</th>
<th>On Welfare</th>
<th>Un-Employed</th>
<th>Row Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWD</td>
<td>0.044595</td>
<td>0.568581</td>
<td>0.022297</td>
<td>0.022297</td>
<td>0.167230</td>
<td><strong>0.825000</strong></td>
</tr>
<tr>
<td>Married</td>
<td>0.009459</td>
<td>0.120608</td>
<td>0.004730</td>
<td>0.004730</td>
<td>0.035473</td>
<td><strong>0.175000</strong></td>
</tr>
<tr>
<td>Column Constraint</td>
<td>0.054054</td>
<td>0.689189</td>
<td>0.027027</td>
<td>0.027027</td>
<td>0.202703</td>
<td>1</td>
</tr>
</tbody>
</table>

*Examples from Norman working paper.*
Processing Status

- Database Design: Done
- Generating 97,000,000 Household Locations: Done
- Input Data Collection: Done
  - Nationwide SF3 Data
  - Nationwide PUMS Data
- Use of TranSims Code to Run IPF and Generate Synthetic Population: About to Start
Database Design
Expected Results

- Households
  - X,Y coordinates
  - Household attributes

- Persons
  - Individual attributes (age, sex, etc.)

- Family Structures Maintained

- Closely Matches Census Counts at Block Group Levels
Conclusion

- Input Data, Statistical Tools, and Software Needed to Build a National Synthetic Population Exist
  - SF3, PUMS, Census Geography, Census Counts, IPF
- Once Created, the U.S. Synthetic Population will support modeling, not just for infectious disease, but for many other fields as well.
- Questions/Comments?