Generating Synthetic Populations for Social Modeling

Tutorial at AAMAS 2017, São Paulo, Brazil

Samarth Swarup & Madhav V. Marathe,
Network Simulation Science and Advanced Computing Division,
Biocomplexity Institute and Initiative,
University of Virginia.
{swarup, marathe}@virginia.edu
Technical report NDSSL-TR-17-073

With contributions by: Abhijin Adiga, Richard Beckman, Jiangzhuo Chen, Youngyun Chungbaek, Stephen Eubank, Kristian Lum, Achla Marathe, Henning Mortveit, Yihui Ren, Srinivasan Venkatramanan

Special thanks to Marta C. González, MIT, for graciously sharing her work and recent slides
Acknowledgements

• Thanks to members of the Network Dynamics and Simulation Science Laboratory, BI and Discovery Analytics Center (DAC), both at Virginia Tech and our collaborators at SUNY Albany, UMD, Harvard and others.

• Support
  – National Science Foundation: HSD grant SES-0729441, NSF PetaApps grant OCI-0904844, NSF NetSE grant CNS-1011769, NSF SDCI grant OCI-1032677,
  – Defense Threat Reduction Agency grant HDTRA1-11-1-0016, DTRA CNIMS contract HDTRA1-11-D-0016-0001,
  – National Institute of Health Midas grant 2U01GM070694-09,
  – Intelligence Advanced Research Projects Activity (IARPA) via the US Department of Interior (DoI) National Business Center (NBC): D12PC000337.

The US government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US government.
Slides for tutorial at AAMAS, São Paulo, Brazil, May 9, 2017

The current version of the slides are available at: http://people.virginia.edu/~ss7rs/synthetic_population_tutorial_2

Suggested citation:
GOALS OF THE TUTORIAL
Objectives

• Overview the state of the art in synthetic populations
  – Discuss how to synthesize populations, networks and more broadly information
  – Discuss some of the challenges in the field
  – Illustrate how advances in computation and data are useful in advancing the field.
  – Provide audience an interesting topical area to work in.
• Describe open problems and challenges in the area
• Does not claim to be extensive. Due to time constraints some topics will not be covered
  – E.g. Detailed discrete choice theory, details of applications where these populations have been useful.
  – References are provided for additional reading
Section

POPULATION SYNTHESIS
Goal: To generate a population of agents with realistic demographic attributes

Input: • Distributions over demographics (marginal distributions),
• A sample of census records

Method: Iterative Proportional Fitting (IPF)
Geographical resolution

Standard hierarchy of US Census geographic entities

- NATION
  - REGIONS
    - DIVISIONS
      - STATES
        - Counties
          - Voting Districts
          - Traffic Analysis Zones
          - County Subdivisions
          - Subminor Civil Divisions
          - Census Tracts
            - Block Groups
              - Census Blocks
              - Places
              - Public Use Microdata Areas
              - State Legislative Districts
              - Urban Growth Areas
              - Core Based Statistical Areas
              - Urban Areas
              - AIANNH Areas* (American Indian, Alaska Native, Native Hawaiian Areas)
    - ZIP Code Tabulation Areas
    - School Districts
    - Congressional Districts
Geographical resolution

Standard hierarchy of US Census geographic entities
Geographical resolution

- **Block Groups (BGs)** are statistical divisions of census tracts, are generally defined to contain between 600 and 3,000 people.
  - The US Census provides various demographic distributions at the BG level.
Geographical resolution

- **Public Use Microdata Areas** (PUMAs) are statistical geographic areas defined for the dissemination of Public Use Microdata Sample (PUMS) data.
  - This is a 5% sample of the Census records.
  - A PUMA contains at least 100,000 people.
  - PUMAs are built on Census tracts and counties.
Generating a base population

- The Census gives marginal information about some variables at household level for each block group.
- Variables used (e.g.):
  - Householder’s age
  - Household income
  - Household size
- What we need:

<table>
<thead>
<tr>
<th>Hsize</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;74</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>0</td>
<td>121</td>
<td>214</td>
<td>46</td>
<td>46</td>
<td>36</td>
<td>25</td>
</tr>
</tbody>
</table>

For census tract 1, block group 2 of Los Alamos county, NM
Generating a base population

- Use PUMS data (5% sample data)
  - A PUMA can contain multiple census block groups.
  - Gives detailed information about household and person demographics.

<table>
<thead>
<tr>
<th>Householder’s age</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;74</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsize 1</td>
<td>2</td>
<td>11</td>
<td>9</td>
<td>3</td>
<td>26</td>
<td>64</td>
<td>42</td>
<td>157</td>
</tr>
<tr>
<td>Hsize 2</td>
<td>11</td>
<td>108</td>
<td>122</td>
<td>48</td>
<td>80</td>
<td>61</td>
<td>18</td>
<td>448</td>
</tr>
<tr>
<td>Hsize 3</td>
<td>28</td>
<td>135</td>
<td>274</td>
<td>156</td>
<td>85</td>
<td>22</td>
<td>6</td>
<td>706</td>
</tr>
<tr>
<td>Hsize &gt;3</td>
<td>0</td>
<td>3</td>
<td>65</td>
<td>76</td>
<td>40</td>
<td>10</td>
<td>3</td>
<td>197</td>
</tr>
<tr>
<td>Total Hsize</td>
<td>41</td>
<td>257</td>
<td>470</td>
<td>283</td>
<td>231</td>
<td>157</td>
<td>69</td>
<td>1508</td>
</tr>
</tbody>
</table>

For PUMA containing census tract 1, block group 2 of Los Alamos county, NM
Generating a base population

- Use Iterative Proportional Fitting (IPF) Algorithm.
- Uses block group marginal information and PUMA data.
- Generates joint distribution for each block group in given PUMA.

<table>
<thead>
<tr>
<th>Householder’s age</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;74</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.003</td>
<td>0.141</td>
<td>0.061</td>
<td>0.020</td>
<td>0.047</td>
<td>0.063</td>
<td>0.000</td>
<td>0.336</td>
</tr>
<tr>
<td>3</td>
<td>0.009</td>
<td>0.228</td>
<td>0.178</td>
<td>0.086</td>
<td>0.065</td>
<td>0.030</td>
<td>0.000</td>
<td>0.594</td>
</tr>
<tr>
<td>&gt;3</td>
<td>0.000</td>
<td>0.003</td>
<td>0.022</td>
<td>0.022</td>
<td>0.016</td>
<td>0.007</td>
<td>0.000</td>
<td>0.069</td>
</tr>
<tr>
<td>Total</td>
<td>0.011</td>
<td>0.372</td>
<td>0.261</td>
<td>0.128</td>
<td>0.128</td>
<td>0.100</td>
<td>0.040</td>
<td></td>
</tr>
</tbody>
</table>

For census tract 1, block group 2 of Los Alamos county, NM

- Sample the required number households from PUMS data from the same category.
How Iterative Proportional Fitting works

There exists a universe of N (here 10,000) individuals that can be represented by a 2-way contingency table. For simplicity, let’s say that there are 2 hair colors and 2 eye colors.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>2000</td>
<td>1500</td>
<td>3500</td>
</tr>
<tr>
<td>Black Hair</td>
<td>500</td>
<td>6000</td>
<td>6500</td>
</tr>
<tr>
<td>Totals</td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>
The Set-Up

There exists a universe of $N = 10,000$ individuals that can be represented by a 2-way contingency table. For simplicity, let’s say that there are 2 hair colors and 2 eye colors.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>?</td>
<td>?</td>
<td>3500</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>?</td>
<td>?</td>
<td>6500</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

But, all we know are the marginal totals...
The Set-Up

The goal is to fill in as best we can the missing cells, while maintaining the interaction structure between hair and eye color.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond  Hair</td>
<td>?</td>
<td>?</td>
<td>3500</td>
</tr>
<tr>
<td>Black Hair</td>
<td>?</td>
<td>?</td>
<td>6500</td>
</tr>
<tr>
<td>Totals</td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

But, all we know are the marginal totals...
A naïve approach

Assume hair and eye color are independent...

We estimate the marginal probabilities as...

\[
\begin{align*}
&\text{Pr(Blue)} = 0.35 & &\text{Pr(Blond)} = 0.25 \\
&\text{Pr(Brown)} = 0.65 & &\text{Pr(Black)} = 0.75 \\
\end{align*}
\]

Under the assumption of independence, the expected number of people in each of the cells should be...

\[
\begin{align*}
\text{Blue Eyes} & \times \text{Brown Eyes} \times 10000 \\
\text{Blond Hair} & : 0.35 \times 0.25 \times 10000 = 875 & 0.65 \times 0.25 \times 10000 = 1625 & 3500 \\
\text{Black Hair} & : 0.35 \times 0.75 \times 10000 = 2625 & 0.65 \times 0.75 \times 10000 = 4875 & 6500 \\
\text{Totals} & : 2500 & 7500 & 10000 \\
\end{align*}
\]
A naïve approach

The marginal totals still match, but the dependence structure is totally changed.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>.35 x .25 x 10000 = 875</td>
<td>.65 x .25 x 10000 = 1625</td>
<td>3500</td>
</tr>
<tr>
<td>Black Hair</td>
<td>.35 x .75 x 10000 = 2625</td>
<td>.65 x .75 x 10000 = 4875</td>
<td>6500</td>
</tr>
<tr>
<td>Totals</td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

Table estimated under independence assumption from marginal totals

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>2000</td>
<td>1500</td>
<td>3500</td>
</tr>
<tr>
<td>Black Hair</td>
<td>500</td>
<td>6000</td>
<td>6500</td>
</tr>
<tr>
<td>Totals</td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

Original table

The combination of blue eyes and blond hair is estimated to be much less common than it is in the real population because hair and eye color are dependent, i.e. \( \Pr(\text{Blue Eyes and Blond Hair}) \neq \Pr(\text{Blue Eyes}) \Pr(\text{Blond Hair}) \)
The Set-Up

We obtain a sub-sample of size $n=100$ and obtain the following counts...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>18</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td>Black Hair</td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td>Totals</td>
<td>21</td>
<td>79</td>
<td>100</td>
</tr>
</tbody>
</table>
The Set-Up

The challenge is to maintain the dependence structure of the sub-sample while matching it to the whole population’s marginal totals.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>18</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td>Black Hair</td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td>Totals</td>
<td>21</td>
<td>79</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>m11</td>
<td>m12</td>
<td>3500</td>
</tr>
<tr>
<td>Black Hair</td>
<td>m21</td>
<td>m22</td>
<td>6500</td>
</tr>
<tr>
<td>Totals</td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

What does this mean?
Intuition

**Minimum discrimination information principle (Kullback)**

Suppose we have a prior distribution over some variables, $p(x)$. If we want to revise this distribution based on some new observations, we should choose a new distribution, $p'(x)$, that matches the new observations, while minimizing $D(p' || p)$.

$$I(p; \pi) = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \ln \frac{p_{ij}}{\pi_{ij}}, \text{ where } \pi_{ij} = \frac{n_{ij}}{n}$$

In other words, we are trying to find a joint distribution that matches the marginals, while staying as close to the sample distribution as we can.
The IPF Algorithm

- Here we work in terms of cell probabilities instead of counts.
  - Let \( m_{ij} = np_{ij} \)
  - The marginal probabilities are then fixed.
    \[
    \frac{N_i}{N} = p_i = \sum_j p_{ij} \quad \frac{N_j}{N} = p_j = \sum_i p_{ij}
    \]

- Initialize \( p_{0ij} = n_{ij}/n \)
- For \( t \geq 1 \)
  - Set \( p_{ij}^{(2t-1)} = p_{ij}^{(2t-2)} \frac{p_i}{p_i^{(2t-2)}} \) and \( p_{ij}^{(2t)} = p_{ij}^{(2t-1)} \frac{p_j}{p_j^{(2t-1)}} \)
IPF on our example data

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>18</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>21</td>
<td>89</td>
<td>100</td>
</tr>
</tbody>
</table>

Subject to $p_1 = .35$, $p_2 = .65$, $p_1 = .25$, $p_2 = .75$
IPF on our example data

Initialize

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>18/100 =</td>
<td>17/100=</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>.18</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>3/100=</td>
<td>62/100=</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>.03</td>
<td>.63</td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>.21</td>
<td>.79</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to \( p1. = .35, p2. = .65, p1 = .25, p2 = .75 \)
### IPF on our example data

Adjust rows to match given proportion... factor for this adjustment are .35/.34 and .65/.66...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>.18x.35/.3 4 = .185</td>
<td>.16x .35/.34 = .165</td>
<td>.35</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>.03x .65/.66 = .030</td>
<td>.63x .65/.66 = .620</td>
<td>.65</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>.215</td>
<td>.785</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75
IPF on our example data

Adjust columns to match given proportion... factor for this adjustment are .25/.215 and .75/.785...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>.185x</td>
<td>.165x</td>
<td>.373</td>
</tr>
<tr>
<td></td>
<td>.25/.215 =</td>
<td>.75/.785 =</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.215</td>
<td>.158</td>
<td></td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>.030x</td>
<td>.620x</td>
<td>.627</td>
</tr>
<tr>
<td></td>
<td>.25/.215 =</td>
<td>.75/.785 =</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.035</td>
<td>.592</td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>.25</td>
<td>.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75
IPF on our example data

Adjust rows to match given proportion... factor for this adjustment are .35/.373 and .65/.627 ...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>.215x</td>
<td>.158x</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>.35/.373 =.202</td>
<td>.35/.373 =.148</td>
<td></td>
</tr>
<tr>
<td>Black Hair</td>
<td>.035x</td>
<td>.592x</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td>.65/.627 =.036</td>
<td>.65/.627 =.614</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>.238</td>
<td>.762</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to p1. = .35, p2. = .65, p.1 = .25, p.2 = .75
IPF on our example data

Adjust columns to match given proportion... factor for this adjustment are .25/.238 and .75/.762 ...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>.202x</td>
<td>.148x</td>
<td>.358</td>
</tr>
<tr>
<td></td>
<td>.25/.238</td>
<td>.75/.762</td>
<td></td>
</tr>
<tr>
<td></td>
<td>= .212</td>
<td>= .146</td>
<td></td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>.036x</td>
<td>.614 x</td>
<td>.642</td>
</tr>
<tr>
<td></td>
<td>.25/.238</td>
<td>.75/.762</td>
<td></td>
</tr>
<tr>
<td></td>
<td>= .038</td>
<td>= .604</td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>.25</td>
<td>.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75
IPF on our example data

Adjust rows to match given proportion... factor for this adjustment are \( \frac{.35}{.358} \) and \( \frac{.65}{.642} \) ...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>( .212 \times ) ( \frac{.35}{.358} = .207 )</td>
<td>( .146 \times ) ( \frac{.35}{.358} = .143 )</td>
<td>( .35 )</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>( .038 \times ) ( \frac{.65}{.642} = .038 )</td>
<td>( .604 \times ) ( \frac{.65}{.642} = .612 )</td>
<td>( .65 )</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>( .244 )</td>
<td>( .755 )</td>
<td>( 1.00 )</td>
</tr>
</tbody>
</table>

Subject to \( p1. = .35 \), \( p2. = .65 \), \( p1 = .25 \), \( p2 = .75 \)
## IPF on our example data

Adjust columns to match given proportion... factor for this adjustment are .25/.244 and .75/.755 ...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>.210</td>
<td>.142</td>
<td>.352</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>.040</td>
<td>.608</td>
<td>.648</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>.25</td>
<td>.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75
IPF on our example data

Adjust rows to match given proportion... factor for this adjustment are .35/.352 and .65/.648 ...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>.209</td>
<td>.141</td>
<td>.350</td>
</tr>
<tr>
<td>Black Hair</td>
<td>.040</td>
<td>.609</td>
<td>.649</td>
</tr>
<tr>
<td>Totals</td>
<td>.249</td>
<td>.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

All changes were less than .001 from last iteration, so we’ve reached our stopping criterion. Done!
Notice that the marginal probabilities are very close to what we’ve constrained them to be. If we’d continued, we could get even closer.
Generating a base population

<table>
<thead>
<tr>
<th>Householder’s age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsize</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>&gt;3</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

For census tract 1, block group 2 of Los Alamos county, NM

- Once we have a joint distribution, we sample from it repeatedly and choose matching PUMS records to create the synthetic population.
Other approaches to population synthesis

- Combinatorial optimization
  - Estimating the micro-population by stochastically reweighting the given micro-data
  - Randomly allocate individuals from the micro-data to each zone, and iteratively replace to improve goodness-of-fit

- Generalized Regression Weighting (GREGWT), Simulated Annealing (GenSA)

- They can be considered to be broadly equivalent, as both involve some form of reweighting

Relationship between IPF and CO

Examples

- **TRANSIMS** - [https://www.fhwa.dot.gov/planning/tmip/resources/transims/](https://www.fhwa.dot.gov/planning/tmip/resources/transims/)
  - Used for accurate and sensitive travel forecasts for transportation planning and emission analysis

  - To simulate metropolitan real estate markets and study the impact of land use policies

- **EUROMOD** - [https://www.euromod.ac.uk/](https://www.euromod.ac.uk/)
  - EU based microsimulation to calculate effects of taxes and benefits on incomes and work incentives

  - Canada based Longitudinal population health microsimulation model to rationally compare competing health intervention alternatives

  - To understand the potential outcomes of public policy changes such as welfare reform, tax reform, and national health care reform.
Some useful references

• Ireland, C. T. and Kullback, S. “Contingency Tables with Given Marginals.” *Biometrika* 55 (1968), 179-188.
Section

MULTI-LEVEL POPULATION SYNTHESIS
**Goal:** To generate a population of individuals and households with realistic demographic attributes

**Input:**
- Distributions over demographics (marginal distributions),
- A sample of census records

**Method:** Iterative Proportional Updating (IPU)
Motivation

• We may have marginals at both household and person level.

• The earlier IPF method is applied just at the household level.
  – Households are sampled from the joint distribution and copied over into the synthetic population
  – This may result in a discrepancy in the distributions at the person level
## Multi-level constraints

<table>
<thead>
<tr>
<th>Hsize</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;74</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.003</td>
<td>0.141</td>
<td>0.061</td>
<td>0.020</td>
<td>0.047</td>
<td>0.063</td>
<td>0.000</td>
<td>0.336</td>
</tr>
<tr>
<td>3</td>
<td>0.009</td>
<td>0.228</td>
<td>0.178</td>
<td>0.086</td>
<td>0.065</td>
<td>0.030</td>
<td>0.000</td>
<td>0.594</td>
</tr>
<tr>
<td>&gt;3</td>
<td>0.000</td>
<td>0.003</td>
<td>0.022</td>
<td>0.022</td>
<td>0.016</td>
<td>0.007</td>
<td>0.000</td>
<td>0.069</td>
</tr>
<tr>
<td>Total</td>
<td>0.011</td>
<td>0.372</td>
<td>0.261</td>
<td>0.128</td>
<td>0.128</td>
<td>0.100</td>
<td>0.040</td>
<td></td>
</tr>
</tbody>
</table>

### Blue Eyes vs. Brown Eyes

<table>
<thead>
<tr>
<th>Blond Hair</th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.209</td>
<td>.141</td>
<td>.350</td>
</tr>
<tr>
<td>Black Hair</td>
<td>.040</td>
<td>.609</td>
<td>.649</td>
</tr>
<tr>
<td>Totals</td>
<td>.249</td>
<td>.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Simplified example

Household-level IPF results:

<table>
<thead>
<tr>
<th></th>
<th>Income &lt; $50K</th>
<th>Income &gt;= $50K</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 person</td>
<td>0.25</td>
<td>0.15</td>
<td>0.4</td>
</tr>
<tr>
<td>&gt;= 2 people</td>
<td>0.1</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Totals</td>
<td>0.35</td>
<td>0.65</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Individual-level IPF results:

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>0.2</td>
<td>0.15</td>
<td>0.35</td>
</tr>
<tr>
<td>Black Hair</td>
<td>0.05</td>
<td>0.6</td>
<td>0.65</td>
</tr>
<tr>
<td>Totals</td>
<td>0.25</td>
<td>0.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

If we sample households using the household-level joint distribution, the synthetic population will not match the individual-level joint distribution.
Iterative Proportional Updating

• Do IPF separately for the household-level constraints and the individual-level constraints.
• The household records corresponding to each cell of the joint distribution will contain different numbers of people of each type.
• Adjust the sampling weights of the household records so that the distribution over individuals in the synthetic population more closely matches the results of the individual-level IPF.

## Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>HH Size</th>
<th>HH Income</th>
<th>Blond, blue</th>
<th>Blond, brown</th>
<th>Black, blue</th>
<th>Black, brown</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>40,000</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>45,000</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>35,000</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>42,000</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>80,000</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>60,000</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1.3</td>
</tr>
</tbody>
</table>
## Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>&lt;50K, 1</th>
<th>&lt;50K, &gt;=2</th>
<th>&gt;=50K, 1</th>
<th>&gt;=50K, &gt;=2</th>
<th>Blond, blue</th>
<th>Blond, brown</th>
<th>Black, blue</th>
<th>Black, brown</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Total number of households = 2000
Total number of people = 4000
## Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>HH Type</th>
<th>Person type</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;50K, 1</td>
<td>&lt;50K, &gt;=2</td>
<td>&gt;=50K, 1</td>
<td>&gt;=50K, &gt;=2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\rho = \frac{c_k}{\sum_q d_{s_{qk}, k} \times w_{s_{qk}}} = \frac{0.25 \times 2000}{1 \times 1 + 1 \times 1} = \frac{500}{2} = 250
\]
## Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>HH Type</th>
<th>Person type</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$&lt;50K$, 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$&gt;=50K, &gt;2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>250</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

\[
\rho = \sum_q d_{sq_k,k} \times w_{sq_k} = \frac{0.1 \times 2000}{1 + 1 \times 1} = \frac{200}{2} = 100
\]
## Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>HH Type</th>
<th>Person type</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;50K, 1</td>
<td>Blond, blue</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>&gt;=50K, 1</td>
<td>Blond, brown</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[
\rho = \frac{c_k}{\sum_q d_{s_qk,k} \times w_{s_qk}} = \frac{0.2 \times 4000}{1 \times 250 + 1 \times 100 + 2 \times 500} = \frac{800}{1350} = 0.593
\]
Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>HH Type</th>
<th>Person type</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;50K, 1</td>
<td>&lt;50K, 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;=50K, 1</td>
<td>&gt;=50K, 1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\rho = \frac{c_k}{\sum_q d_{s_{ak},k} \times w_{sak}} = \frac{0.15 \times 4000}{1 \times 250 + 1 \times 100 + 1 \times 500} = \frac{600}{850} = 0.706
\]
Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>HH Type</th>
<th>Person type</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;50K, 1</td>
<td>&lt;50K, &gt;=2</td>
<td>&gt;=50K, 1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\rho = \sum_{q} \frac{c_k}{d_{s_{q,k}} \times w_{s_{q,k}}} = \frac{0.05 \times 4000}{1 \times 70.6 + 1 \times 296.5} = \frac{200}{367.1} = 0.545
\]
### Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>HH Type</th>
<th>Person type</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;50K, 1</td>
<td>&lt;50K, &gt;=2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\rho = \frac{\sum_{q} c_k}{\sum_{q} d_{sqk}k \times w_{sqk}} = \frac{0.6 \times 4000}{1 \times 59.3 + 1 \times 161.6 + 2 \times 303} = \frac{2400}{826.9} = 2.9
\]
Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>&lt;50K, 1</th>
<th>&lt;50K, &gt;=2</th>
<th>&gt;=50K, 1</th>
<th>&gt;=50K, &gt;=2</th>
<th>Blond, blue</th>
<th>Blond, brown</th>
<th>Black, blue</th>
<th>Black, brown</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

$$\delta = \frac{\sum_j \sum_i |d_{i,j} w_i - c_j|}{\sum_j c_j}$$
Simplified example

<table>
<thead>
<tr>
<th>HH ID</th>
<th>HH Type</th>
<th>Person type</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;50K, 1</td>
<td>&lt;50K, &gt;=2</td>
<td>&gt;=50K, 1</td>
<td>&gt;=50K, &gt;=2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.35</td>
<td>0.05</td>
<td>1</td>
<td>0.347</td>
</tr>
</tbody>
</table>

\[ \delta = 0.5131 \]
Example

Table 1. An Example of the Iterative Proportional Updating (IPU) Algorithm

<table>
<thead>
<tr>
<th>Household ID</th>
<th>Weights</th>
<th>Household Type 1</th>
<th>Household Type 2</th>
<th>Person Type 1</th>
<th>Person Type 2</th>
<th>Person Type 3</th>
<th>Weights 1</th>
<th>Weights 2</th>
<th>Weights 3</th>
<th>Weights 4</th>
<th>Weights 5</th>
<th>Final Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11.67</td>
<td>11.67</td>
<td>9.51</td>
<td>8.05</td>
<td>12.37</td>
<td>1.36</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>11.67</td>
<td>11.67</td>
<td>9.51</td>
<td>9.51</td>
<td>14.61</td>
<td>25.66</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>11.67</td>
<td>11.67</td>
<td>9.51</td>
<td>8.05</td>
<td>8.05</td>
<td>7.98</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1.00</td>
<td>13.00</td>
<td>10.59</td>
<td>10.59</td>
<td>16.28</td>
<td>27.79</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1.00</td>
<td>13.00</td>
<td>10.59</td>
<td>10.59</td>
<td>16.28</td>
<td>18.45</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1.00</td>
<td>13.00</td>
<td>10.59</td>
<td>8.97</td>
<td>8.97</td>
<td>8.64</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1.00</td>
<td>13.00</td>
<td>10.59</td>
<td>8.97</td>
<td>13.78</td>
<td>1.47</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1.00</td>
<td>13.00</td>
<td>10.59</td>
<td>8.97</td>
<td>8.97</td>
<td>8.64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weighted Sum 1</th>
<th>3.00</th>
<th>5.00</th>
<th>9.00</th>
<th>7.00</th>
<th>7.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constraints</td>
<td>35.00</td>
<td>65.00</td>
<td>91.00</td>
<td>65.00</td>
<td>104.00</td>
</tr>
<tr>
<td>$\delta_h$</td>
<td>0.9143</td>
<td>0.9231</td>
<td>0.9011</td>
<td>0.8923</td>
<td>0.9327</td>
</tr>
<tr>
<td>Weighted Sum 1</td>
<td>35.00</td>
<td>5.00</td>
<td>51.67</td>
<td>28.33</td>
<td>28.33</td>
</tr>
<tr>
<td>Weighted Sum 2</td>
<td>35.00</td>
<td>65.00</td>
<td>111.67</td>
<td>88.33</td>
<td>88.33</td>
</tr>
<tr>
<td>Weighted Sum 3</td>
<td>28.52</td>
<td>55.38</td>
<td>91.00</td>
<td>76.80</td>
<td>74.39</td>
</tr>
<tr>
<td>Weighted Sum 4</td>
<td>25.60</td>
<td>48.50</td>
<td>80.11</td>
<td>65.00</td>
<td>67.68</td>
</tr>
<tr>
<td>Weighted Sum 5</td>
<td>35.02</td>
<td>64.90</td>
<td>104.84</td>
<td>85.94</td>
<td>104.00</td>
</tr>
<tr>
<td>$\delta_h$</td>
<td>0.0006</td>
<td>0.0015</td>
<td>0.1521</td>
<td>0.3222</td>
<td>0.0000</td>
</tr>
<tr>
<td>Final Weighted Sum</td>
<td>35.00</td>
<td>65.00</td>
<td>91.00</td>
<td>65.00</td>
<td>104.00</td>
</tr>
</tbody>
</table>

Some useful references


Section

POPULATION EVOLUTION
**Goal:** To evolve a population of individuals and households with realistic demographic attributes over time

**Input:**
- An existing synthetic population,
- Various data on probabilities of life events

**Method:** Logistic regression; direct lookup in probability tables
Modeling with life events: FPOP

- FPOP models six kinds of life events:
  - aging,
  - mortality,
  - birth,
  - marriage/union formation
  - marriage/union dissolution,
  - migration

Modeling with life events: FPOP

Modeling with life events: FPOP

Modeling with life events: UrbanSim

- Economic and demographic transition models
- Household and employment mobility models
- Household and employment location models
- Real estate development model
- Land price model

Modeling with life events
Namazi-Rad et al. (2014)

• Generalized Malthusian model of population growth:

\[
P^{(t+1)} = P^{(t)} + \sum_{i=1}^{w} \left[ F_i^{(t)} \times \varphi\left( F_i^{(t)} \right) \right] \\
+ \sum_{j=1}^{w} \left[ M_j^{(t)} \times \varphi\left( M_j^{(t)} \right) \right] + \sum_{i=1}^{w} \sum_{j=1}^{w} \left[ C_{(i,j)}^{(t)} \times \varphi\left( C_{(i,j)}^{(t)} \right) \right]
\]

• Growth rates for males, females, and couples are derived from data.

Modeling with life events
Namazi-Rad et al. (2014)

- Household transitions:
  - Death rate: $D'_{ijkl} = [D^{(M1)}_i, D^{(F1)}_j, D^{(F2)}_k, D^{(M2)}_l]$, 
  - Marriage rate: $\Gamma'_{ij} = [\Gamma^{(1)}_{ij}, \Gamma^{(2)}_{ij}]$, 
  - Birth rate: $B'_{ij} = [B^C_{ij}]$, 
  - Divorce rate: $\Lambda'_{ij} = [\Lambda^C_{ij}]$, and 
  - Leaving the parental home: $\Lambda'_{kl} = [\Lambda^{(F)}_k, \Lambda^{(M)}_l]$. 

Modeling with life events
Namazi-Rad et al. (2014)

Some useful references

  – https://udst.github.io/urbansim/


Section

ACTIVITY ASSIGNMENT
**Goal:** To assign a realistic daily activity sequence to each agent

**Input:**
- A synthetic population of agents with demographics,
- A household activity survey

**Methods:**
- Classification and Regression Trees
- Fitted Values Method
Problem

• Given a population of agents, we want to assign them realistic activity sequences that
  – Obey the within-household dependence structure of the given data
  – Are consistent with the individual demographic characteristics of each agent.

• Why is this important?
  – In transportation modeling, this will help determine which locations people go to.
  – In epidemiology, this will determine contact patterns.
The National Household Travel Survey

The NHTS/NPTS serves as the nation's inventory of daily travel. Data is collected on daily trips taken in a 24-hour period, and includes:

- purpose of the trip (work, shopping, etc.);
- means of transportation used (car, bus, subway, walk, etc.);
- how long the trip took, i.e., travel time;
- time of day when the trip took place;
- day of week when the trip took place; and
- if a private vehicle trip:
  - number of people in the vehicle, i.e., vehicle occupancy;
  - driver characteristics (age, sex, worker status, education level, etc.); and
  - vehicle attributes (make, model, model year, amount of miles driven in a year).

These data are collected for:

- all trips,
- all modes,
- all purposes,
- all trip lengths, and
- all areas of the country, urban and rural.

http://nhts.ornl.gov
Activity Selection

What do I want to do?

Where do I want to go?
## Notation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{ij}$</td>
<td>Synthetic person $i$ in synthetic household $j$.</td>
</tr>
<tr>
<td>$H_j$</td>
<td>Synthetic household $j$.</td>
</tr>
<tr>
<td>$P^<em>_i{j}</em>$</td>
<td>Survey person $i^<em>$ in survey household $j^</em>$.</td>
</tr>
<tr>
<td>$H^<em>_j{^</em>}$</td>
<td>Survey household $j^*$.</td>
</tr>
<tr>
<td>$X_{ij}$</td>
<td>Personal and household demographics of the $i^{th}$ synthetic person in the $j^{th}$ synthetic household.</td>
</tr>
<tr>
<td></td>
<td>Example: $X = {\text{age, sex, highest income in household}}$</td>
</tr>
<tr>
<td>$X^<em>_i{j}</em>$</td>
<td>Personal and household demographics of the $i^{*th}$ person in the $j^{*th}$ survey household.</td>
</tr>
<tr>
<td>$A^<em>_i{j}</em>$</td>
<td>Activity sequence of the $i^{*th}$ person in the $j^{*th}$ survey household.</td>
</tr>
<tr>
<td>$Y^*_i{j}^*k$</td>
<td>Number of minutes the $i^{*th}$ person in the $j^{*th}$ survey household spends performing activity-type $k$ (home, work, school, shopping, ...)</td>
</tr>
</tbody>
</table>
Use survey households in CART to break into leaf nodes based on total time spent doing each of several activities.
CART-based re-sampling

Split 1 [ex: age of oldest member]

Split 2a [ex: income of oldest member]

Split 2b [ex: number of household members]

Use tree made from survey individuals to find pool of survey households with which to match synthetic households
How to assign individual activity schedules?

1) Select at random from the households in the leaf node

2) Assign the schedules according to age (i.e. oldest member of synthetic household is assigned the schedule of survey household’s oldest member).
Fitted Values Method: general idea

We want to match survey households to synthetic households based on some measure of similarity.

**Step 1:** Select a survey household based on the similarity between the it and the synthetic household (probabilistic, minimum distance, etc.)

**Step 2:** Find the survey individual in the selected survey household that is most similar to each synthetic individual

**Step 3:** Assign schedules accordingly

<table>
<thead>
<tr>
<th>Schedule:</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00am     wake up</td>
</tr>
<tr>
<td>9:00am     travel to work</td>
</tr>
<tr>
<td>9:15am     arrive at work</td>
</tr>
<tr>
<td>1:00pm     go to lunch</td>
</tr>
<tr>
<td>2:00pm     return to work</td>
</tr>
<tr>
<td>6:00pm     travel to store</td>
</tr>
<tr>
<td>6:25pm     shopping</td>
</tr>
<tr>
<td>7:00pm     travel home</td>
</tr>
<tr>
<td>7:30pm     home</td>
</tr>
</tbody>
</table>

Synthetic  Survey
Define:

\[ D_{HH}(H_j, H_j^*) = \max_i \{ D_{PH}(P_{ij}, H_j^*) \} \]  

\[ D_{PH}(P_{ij}, H_j^*) = \min_{i^*} \{ D(P_{ij}, P_{i^* j^*}) \} \]

This is the [asymmetric] Hausdorff distance between synthetic household \( j \) and survey household \( j^* \).

Algorithmically, for each person in synthetic household \( j \), find the person in survey household \( j^* \) that most closely resembles them—this is that person’s distance from the survey household (represented by Eq. (2)). The distance between the households \( j \) and \( j^* \) is the farthest any of the members of household \( j \) is from \( j^* \).
Similarity measures between survey and synthetic households

\[ D(1, 1^*) = 1.2 \]
\[ D(2, 1^*) = 2.7 \]
\[ D(1, 2^*) = 0.5 \]
\[ D(2, 2^*) = 3.7 \]
\[ D(2, 3^*) = 4.0 \]
\[ D(1, 3^*) = 3.8 \]
Similarity measures between survey and synthetic households

\[
D_{PH}(1, H_{j*}) = \min(1.2, 0.5, 3.8) = 0.5
\]
\[
D_{PH}(2, H_{j*}) = \min(2.7, 3.7, 4.0) = 2.7
\]

\[
D_{HH}(H_j, H_{j*}) = \max(D_{PH}(1, H_{j*}), D_{PH}(2, H_{j*})) = 2.7
\]
What about D?

Euclidean distance between covariate vectors?

$$\| \mathbf{X}_{ij} - \mathbf{X}_{i* j*} \|$$

Mahalanobis distance between covariate vectors?

$$(\mathbf{X}_{ij} - \mathbf{X}_{i* j*}) S^{-1} (\mathbf{X}_{ij} - \mathbf{X}_{i* j*})$$
What about D?

We use...

\[ D(P_{ij}, P_{ij}^*) = \| \hat{Y}_{ij} - \hat{Y}_{ij}^* \| \]

where

\[ \hat{Y}_{ij} = \{ \hat{f}_k(X_{ij}) : k = 1 \ldots K \} \]

i.e. the set of fitted values from some model, \( f_k \), for each of the activity times. For example, if \( f_k \) is just a regression function*, for each activity type, \( k \), we’d have

\[ \hat{Y}_{ij} = X_{ij} \hat{\beta}_k \]

Where \( \hat{\beta}_k \) is vector of estimated coefficients from the model

\[ \hat{Y}_{ij}^* = X_{ij}^* \hat{\beta}_k + \epsilon_{ij} \]

*We don’t have to use linear regression. Something like CART, random forest, etc. could also be used here for \( f_k \).
Results: Work

Smoothed Template
Average Work Hours

VSP

Average Hours for People with Activity of WORK

Fitted-Values Match

Average Hours for People with Activity of WORK
Results: College

Smoothed Template
Persons with College Activity

VSP
Persons with Activity of COLLEGE

Fitted-Values Match
Persons with Activity of COLLEGE

Proportion vs Age

Female
Male
Some useful references


ASSIGNING NEW ATTRIBUTES/ACTIVITIES TO SYNTHETIC AGENTS
Augment the list of attributes in synthetic population

- The synthetic population has a set of available features but additional features may be desirable to expand its application to other areas of interest.
- For example to study obesity, tobacco addiction, activity on social media etc.

<table>
<thead>
<tr>
<th>Available Features</th>
<th>Example Desirable Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Personal income</td>
</tr>
<tr>
<td>Gender</td>
<td>Have a computer/cell phone</td>
</tr>
<tr>
<td>Marital status</td>
<td>Level of physical activity</td>
</tr>
<tr>
<td>Employment status</td>
<td>Number of virtual friends</td>
</tr>
<tr>
<td>Household size</td>
<td>Time spent on social media</td>
</tr>
<tr>
<td>Household income</td>
<td>Have a credit card</td>
</tr>
<tr>
<td>Education level</td>
<td>Tobacco use</td>
</tr>
<tr>
<td>Number of vehicles</td>
<td>Gym membership</td>
</tr>
<tr>
<td>Religion</td>
<td>Body mass index</td>
</tr>
<tr>
<td>Race</td>
<td>Have a bank account</td>
</tr>
</tbody>
</table>
Classification and regression tree (CART)

- Predictive modeling approach
- CART recursively partitions data into disjoint sets by fitting a simple prediction model within each set.
- The dependent variable is the desired attribute (owns a cell phone) and the independent variables are the available attributes (such as age, gender, income).
- The resulting tree has a flow-chart structure that is easy to interpret.

```
Input: age, gender, occupation, ...

Does the person like computer games

age < 15
Y N

is male?
Y  N

prediction score in each leaf
+2  +0.1  -1

Univ of Virginia
Biocomplexity Institute & Initiative
```
Example application of CART

• Cell phone assignment to synthetic population using CART

• Use National Health Interview Survey (NHIS) data which contains demographic information and cell phone ownership data for households.

• CART model is built using NHIS data. Dependent variable is average number of cell phones among householders and independent variables are demographics e.g. household income, # of workers in the family, family size etc.
Example Application of CART

- The fitted model is applied to the synthetic population to determine the number of cell phones among householders.
- Once the number of cell phones for each family is determined, individuals are assigned cell phone randomly with a higher weight to workers in the family who are between age group 18-50 years.

<table>
<thead>
<tr>
<th>#Devices</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob.</td>
<td>0.35</td>
<td>0.33</td>
<td>0.30</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Device assignment: example
Limitations of CART

• Locally optimal decisions are made at each branch. Does not guarantee globally-optimal decision tree.

• Over-fitting can occur and result in a very complex tree model that does not generalize beyond training dataset.
Assigning energy related activities to individuals

- Goal is to calculate energy demand profiles of households in the synthetic population.

- Currently will only focus on active energy demand calculation.

\[
E_{\text{Active}} \quad (\text{Laundry, Watching TV, Cleaning}) \\
E_{\text{Passive}} \quad (\text{Hot water usage, Space heating})
\]

- Total Energy Demand
  - \( E_{\text{Active}} \)
  - \( E_{\text{Passive}} \)
  - Shared
  - Individual
Active demand

• The root node represents the complete set of ATUS (American time use survey) respondents
• Contains 24 hour period activity dairies for 13,260 respondents
• The demographic variables are used as the splitting variables namely
  – Age, Employment Status, Household Size, School Enrollment Status, Number of Children and Martial Status
• We use CART to construct the binary decision tree
• At each stage, the algorithm tries to split the node into two groups based on the best possible splitting variable
Active energy demand

- Each household member in the synthetic population is assigned a leaf node based on his/her demographic variables.
- From the matching demographic variables, we select a ATUS survey respondent at random and assign the activity pattern of the ATUS survey respondent to the synthetic population household member.
- For each energy activity, a list of appliances are associated to calculated the energy used.
Overall view

Synthetic population - Demographic information

Extrapolating characteristics from survey data on to synthetic data

Statistical models for activity sequence generation

Synthetic household level activity sequence generation

Activity-appliance association and demand profile generation

Appliance energy rating from EIA

Energy demand profile

ATUS Survey - Individual's daily activity schedule

EIA-RECS Survey - Building characteristics and appliance information

Statistical models for activity sequence generation
Some useful references


Goal: To assign a geographical location for each activity for each agent

Input:

• A synthetic population of agents with demographics and daily activity sequences
• Geographical data on roads, residence types, business locations, school locations, and other points of interest

Methods:

• Gravity model
• Trip chaining model
• Radiation model
Problem

• Given a population of agents with demographics and daily activity sequences, we want to assign a location for each activity
  – Activity types should be consistent with location types
  – It should result in sensible travel patterns
Home location assignment

- Data used:
  - Household structure (type of the building, capacity) i.e. single family household, duplex, apartment etc.
  - Street data from NAVTEQ/HERE, i.e. name, type of the road/street, length and other geometry info
- Housing unit (home location) is assigned to a link of given category with probability proportional to its length.
Home location assignment

- Synthetic households are assigned home locations with probability proportional to home location weights (home location type – building capacity)
- Output: located base population
Assigning locations for all other activities

- Step 1: Use a discrete choice model to generate all work or school locations.
- Step 2: Use trip-chaining discrete choice models to generate locations for other activities.
Step 1: work & school locations

- A work/school location, \( L \), is chosen with probability,

\[
p(L) = \frac{e^{a(L) + b_m t(H, L)}}{\sum e^{a(L') + b_m t(H, L')}}
\]

\( a(L) \) : Attractor weight
\( b_m \) : Travel mode coefficient
\( t(H, L) \) : Travel time

For computational reasons, only locations within a radius \( r_{ab} \) of the home location are considered.
Step 2: other activity locations

• To generate locations for other activities, we use a logistic multinomial choice chaining model.

\[ p(L_1) = \frac{e^{b_{m1}t(L,L_1) + a(L_1) + b_{m2}t(L_1,H)}}{\sum e^{b_{m1}t(L,L'_1) + a(L'_1) + b_{m2}t(L'_1,H)}} \]

• This means that the location choice for a non-anchor activity takes into account the locations of the previous and next anchor activities.
Step 2: other activity locations

• To generate locations for other activities, we use a logistic multinomial choice chaining model.
Question:
How many people travel from \( a \) to \( b \)?

1. Gravity Model
2. Radiation Model
Gravity Models:
In Analogy with Newton’s Law of Gravity

The Gravity Model is a fitting based model with a form of:

\[ \phi_{ab} = C \frac{m_a^\alpha n_b^\beta}{f(r_{ab})} \]

- \( m_a \) - source population size
- \( n_b \) - destination population size
- \( r_{ab} \) - distance between \( a \) and \( b \) (Euclidian distance).
- \( f(r) \) - deterrence function, typically either \( f(r) = r^\gamma \) or \( f(r) = e^{\delta r} \)

\( \alpha, \beta, \gamma, \delta \) are the fitting parameters
Gravity Models: In Analogy with Newton’s Law of Gravity

Originally, introduced by Zipf, G. K. (1946) as a hypothesis with the form:

\[ \phi_{ab} = \frac{m_a n_b}{r_{ab}^\gamma} \]

Zipf, G. K. The \( P_1P_2/D \) hypothesis: on the intercity movement of persons. Am. Sociol. Rev. 11, 677–686 (1946)

Railway, \( \gamma = 1 \)  
Highway, \( \gamma = 1.25 \)  
Airway, \( \gamma = ? \)
Gravity Models: 
In Analogy with Newton’s Law of Gravity

Original 1-parameter fitting form is not satisfying.

Ex.1. Global cargo ship movements:

\[ \phi_{ab} = \frac{\alpha_a \beta_b m_a n_b}{f(r_{ab})} \]
\[ f(r_{ab}) = r_{ab}^\gamma e^{\delta r_{ab}} \]


Ex.2. Airway Traffic:

\[ \phi_{ab} = C \frac{m_a^\alpha n_b^\beta}{f(r_{ab})} \]
\[ f(r_{ab}) = e^{\delta r_{ab}} \]


*Requires two sets of parameters, for \( r_{ab} > 300 \ km \) and \( r_{ab} < 300 \ km \)
Gravity Models: In Analogy with Newton’s Law of Gravity

Some Limitations of Gravity Models:

• Lacking theoretical guidance or a rigorous derivation
• Parameters having no physical meaning
• Requiring data to fit. Once the system is changed, previous parameters are no longer valid.

“With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.” — John von Neumann

*Nature* 427, 297 (22 January 2004) | doi:10.1038/427297a
Radiation Model

Simini F, et. al. (2012) Nature 484: 96-100
“A universal model for mobility and migration patterns”

\[ \phi_{ab} = \zeta m_a p_{ab} \]

- \( \zeta \) – fraction of traveling people. An overall scaling factor for all locations.

- \( p_{ab} \) – probability for a person from location \( a \) (population \( m_a \)) to travel to location \( b \) (population \( n_b \))

\[ p_{ab} = \frac{m_a n_b}{(s_{ab} + m_a)(s_{ab} + m_a + n_b)} \]

- \( s_{ab} \) – total population of nodes within the disk of radius \( r_{ab} \), except \( m_a \).

Where does this equation come from?
A person living at location \( a \), is looking for a job with salary or benefits expectation of at least \( z \).

The number of job offers at a location is proportional to the population of that location.

Values of offers are denoted by \( z \).

Let’s assume that

\[
\text{This person only takes the job that satisfies his expectation } ( z’ \geq z) \text{ the closest to } a. 
\]

and \( z’ \) are drawn from the same distribution.

What **is the probability** that a person with salary expectation \( z \), living at location \( a \) will take a job at location \( b \) and **not** anywhere closer?
What is the probability that a person with salary expectation, living at location, will take a job at location and not anywhere closer?

Simini et al. use an analogy with radiation emission and absorption processes to formalize this problem.

- The source, $a$, is assumed to be emitting particles with absorption thresholds, $z_x^a$.
- The locations around $a$ are assumed to absorb particles with probabilities, $z_x^b$.
- A particle is absorbed by the closest location whose absorbance is higher than its absorption threshold.

From this, they work out the probability of one emission/absorption event between any two locations, and thereby obtain an analytical expression for the flux between them.

$$p_{ab} = \frac{m_an_b}{(s_{ab} + m_a)(s_{ab} + m_a + n_b)}$$

Note that this turns out to be independent of the distribution, $p(z)$, of the emission and absorption thresholds.
The Generalized Radiation Model

Ren Y., et. al. (2014) Nature Comm. 5: 5347 “Predicting commuter flows in spatial networks using a radiation model based on temporal ranges.” DOI:10.1038/ncomms6347

The problem: \( s_{ab} \) is the population within \( r_{ab} \). Here \( r_{ab} \) acts as a cost measure.

Notice: there is no coupling between the mobility flux and the network.

However:

People don’t estimate costs based on Euclidean distance.

They have to travel on paths of the network.

Hence, they will estimate costs on the network.

This couples the radiation model with the network.

Therefore, we need to redefine the area population \( s_{ab} \).
The Generalized Radiation Model

Here, $c_{ab}$ is a general travel cost on the network.
Some useful references

- Yingxiang Yang, Shan Jiang, Daniele Veneziano, Shounak Athavale, Marta C. Gonzalez, “TimeGeo: a spatiotemporal framework for modeling urban mobility without surveys.” PNAS.
NETWORK CONSTRUCTION
Constructing synthetic information and networks: an overview

**Synthetic Agents**

**Agents**
- People
- Computer, mobile devices, etc.
- Hosts (plant models)

**Places of interaction**
- work
- residence
- school
- trading facility

**Data**
- Demographics
- Activity surveys
- Infrastructure
- GIS
- Trade
- Timeline, surveillance data

**Interactions**

**Data driven**
- Activities of people
- Social & professional relationships
- Transportation
- Commodity flow
- Long distance travel

**Procedure driven**
- Behavior
- Network protocols

**Structural properties**
- Network measures
  - degree distribution
  - number of components
  - diameter
  - spectral radius
- Sensitivity to structural uncertainty.

**Dynamical properties**
- Simulations
- Sensitivity to model choice and parameters

**Analyses**
- V&V
- UQ

---

- Social networks
- Friend networks
- Disease networks
- Addiction networks
- Radio networks
- HIV
- Flu
- Ebola

---

- Biocomplexity Institute & Initiative
Example: An induced contact network

**SOCIAL CONTACT NETWORK**
Using co-occupancy at residence and activity locations, we can infer contacts and their durations. From this we infer the social contact network.

**PEOPLE-LOCATION NETWORK**
The computational model allows us to assign people to locations with durations of visit and through this determine their contacts and interactions.
Relational, dynamic social contact networks

People (8 million)

Vertex attributes:
- age
- household size
- gender
- income
- ...

Locations (1 million)

Vertex attributes:
- \([x,y,z]\)
- land use
- ...

Edge attributes:
- activity type: shop, work, school
- \([\text{start time 1, end time 1}]\)
- \([\text{start time 2, end time 2}]\)
- ...

Locations

Vertex attributes:
Various projections of person-location temporal social visitation network
Refining subnetworks

• Refine a subpopulation and the associated subnetwork in an existing synthetic population/network when more detailed data on the subpopulation becomes available.
  – Special locations: schools, office work/college campuses, hospitals, military bases, hotels, etc.
  – Additional data: mobility data from surveys, sensors, phones, fitbits, etc.
Subnetwork construction

• Identify subpopulation in the whole synthetic population
• Construct subnetwork as follows:
  – Activity data: can replace existing activities occurring at the location in each individual’s daily activity sequence
    • Class registration and schedule data in a high school describes how students move between rooms (sublocations) during the day.
    • Activity vignettes for different cohorts describe activities of military personnel and civilians on post.
    • Patient admission data and hospital work log data describe how patients and health care workers move between rooms (sublocations).
    • Activity survey data for slum subpopulations in Delhi describes the daily activities of people living in slums.
      \textit{Note that these individuals’ activities outside of the location can be kept.}
  – Connection data: is sometimes directly available.
    • Collected by close proximity sensors carried by participants: e.g. Salathe et al. 2010 describes such a data set from a high school.
Embedding a subnetwork

(a) identifying and mapping
(b) removing school edges
(c) embedding
Subpopulation and network inference: Survey and administrative information versus digital traces

• Subpopulation and network attributes can be obtained via various sources:
  – Subjective surveys, administrative data, digital traces from sensors, phones, call records.

• Each method has its pros and cons:
  – E.g. Call data records are not the best for synthesizing networks for specific locations; data although collected by phone companies is not available below the resolution of cell towers at best.
Example: Synthesizing high school networks and embedding them

<table>
<thead>
<tr>
<th>no. of students</th>
<th>school 1</th>
<th>school 2</th>
<th>school 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of students in grade [9,10,11,12] respectively</td>
<td>[293,266,298,259]</td>
<td>[311,231,220,196]</td>
<td>[107,115,105,98]</td>
</tr>
<tr>
<td>no. of classrooms</td>
<td>78</td>
<td>66</td>
<td>45</td>
</tr>
<tr>
<td>no. of classes per day</td>
<td>9</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>duration of classes (minutes)</td>
<td>45</td>
<td>90</td>
<td>50-56</td>
</tr>
</tbody>
</table>

Table 1: Statistic of real class schedules (data from registration information).

Hybrid approach that combines digital traces with administrative and survey information:

- Administrative information obtained from schools and anonymized
- Digital traces collected by RFID sensors and provide in-class networks
Challenges

• Uncertainty and misalignment in subpopulation identification
  – Uncertainty: Real (anonymized) data of high schools (in NRV): Multiple ways to match real and synthetic high school students
  – Misalignment: In the whole synthetic population, the number of people identified through activity locations may not equal the actual size of subpopulation.
    • Identified > actual: need random selection
    • Identified < actual: need expansion (relaxation of identification criteria)

• Adequacy of the refinement: not all details matter.
Do details matter?

Embedding refined NRV high school subnetworks in NRV synthetic network did not significantly affect the network structure of the whole network.

---

**Figure 2:**

- **(a)** Degree Distribution for NRV Networks
- **(b)** Contact Duration Distribution for NRV Networks
- **(c)** Weighted-Degree Distribution for NRV Networks
- **(d)** Clustering Coefficient Distribution for NRV Networks
- **(e)** Shortest Path Distribution for NRV Networks
- **(f)** Conductance Distribution for NRV Networks

---

**Figure 3:**

- **(a)** Readers a sense about the variability of simulation readings.
- **(b)** Displays the fraction of contacts, fraction of nodes, and conductance distribution for NRV networks.
Do details matter?

But the refinement significantly affects the infectious disease dynamics on the whole network, even though the subpopulation is only a small fraction (1.7%) of the whole population.

Epicurves for population in schools (School 1,2 and 3)

Epicurves of NRV Networks: class schedule comparison
### Comparing subnetwork refinement

<table>
<thead>
<tr>
<th>type</th>
<th>whole population</th>
<th>subpopulation</th>
<th>extra detailed data</th>
</tr>
</thead>
<tbody>
<tr>
<td>high school</td>
<td>New River Valley, VA</td>
<td>students attending three high schools</td>
<td>class registration data of each student</td>
</tr>
<tr>
<td>hospital</td>
<td>Virginia</td>
<td>patients and health care workers in a hospital</td>
<td>electronic medical records and direct observations of healthcare workers</td>
</tr>
<tr>
<td>military base</td>
<td>several metropolitan regions in US</td>
<td>people residing and working at military bases</td>
<td>demographic distributions and activity vignettes (templates) of different cohorts</td>
</tr>
<tr>
<td>slum</td>
<td>Delhi, India</td>
<td>people residing in slum areas</td>
<td>demographic and activity survey on households in slum areas, Delhi</td>
</tr>
</tbody>
</table>
# Comparing subnetwork refinement

<table>
<thead>
<tr>
<th>type</th>
<th>subpopulation identification</th>
<th>subnetwork construction</th>
<th>subnetwork embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>high school</td>
<td>individuals in synthetic population who have school type activities in the three schools</td>
<td>usual method is used to combine activity sequences with sublocation model to derive co-location contacts; we have studies different sublocation models</td>
<td>replace existing in-school contacts with new contacts</td>
</tr>
<tr>
<td>hospital</td>
<td>patients can be identified in an epidemic simulation; healthcare workers are those having work type activities in this hospital</td>
<td>activities are created for each healthcare worker, patient, and visitor, from which a people-location network is generated</td>
<td>N/A [not done yet]</td>
</tr>
<tr>
<td>military base</td>
<td>synthetic households whose home locations are within [not enough] or near the bases</td>
<td>on-base activity sequences are assign to individuals in subpopulation by cohort; usual method is used to combine activity sequences with sublocation model to derive co-location contacts</td>
<td>replace existing on-base contacts with new contacts</td>
</tr>
<tr>
<td>slum</td>
<td>synthetic households whose home locations are within the slum areas [defined by polygons]; some living in the same wards are also selected and moved into slum areas</td>
<td>activity templates are derived from survey data; activity sequence is assigned to each individuals in subpopulation according demographics; usual method is used to combine activity sequences with sublocation model to derive co-location contacts</td>
<td>replace existing edges of individuals in slum subpopulation with new edges</td>
</tr>
</tbody>
</table>
Extensions

• Other networks we have constructed
  – Large Geographical areas:
    • Military base: Administrative data available; daily activities are different; mixture of civilians and military personnel; work and residential quarters
    • Slums: high density living spaces, data is not trivial to obtain; within-house network is not easy; can span large geographical area
  – Small set of Buildings
    • Hospitals: detailed data obtained by shadowing the personnel; patients in rooms for long periods; staff moves in and out at short time scales
    • Office Building: Detailed maps available, people reside in their space for long periods; relatively low levels of mobility
Some useful references


Section

POPULATION SYNTHESIS FOR DATA-POOR REGIONS
New data sources and/or classes

- WorldPop.org Population density
- ADCW 7.2-7.3 NDSSL-customized database from ADCi
- US Census Bureau Auto-harvested age/gender distributions
- European Social Survey Household composition
- ILO Labor force statistics
- The World Bank School enrollment rates
- OECD Daycare and preschool rates
- ORNL LandScan Population Counts
- ESRI Geographic Boundary Shapefiles
- OSM Road networks
- Additional data for specific target countries
Methodology Overview

- For regions with limited available data
- We model people with attributes, household compositions, home- and activity locations, human activities and where they occur, and all contacts between people on a typical day.
- Activity types: Home; Work; School; Other
Model Input Data Sets

- Joint age/gender distributions (U.S. Census)
- Household size distribution
- School distributions by type
- Workplace size distributions
- Main daytime activity type by age and location type
Model Inputs and Parameters

- Population density estimates (LandScan; WorldPop)
- Activity templates
- Adult age limit (Eric – what is the list of features governed)
- Maximal commute distance (default is 80km)
- Target commute distance (currently 6km for mean value; exponent in gravity model for distance term)
Population augmentation: activities & locations

- The constructed populations have basic demographic information, properties and relations.

- For specific scenarios and analyses, one generally needs to introduce additional, scenario-specific properties. Examples:
  - Funeral attendance (West Africa; Ebola)
  - Funeral locations (West Africa; Ebola)
  - Military sub-population with activities (National guard force readiness)
  - Integration of mosquito prevalence map and risk maps (vector-borne diseases: Malaria; Zika; Dengue)
Population augmentation: long distance travel and transportation of commodities

- For pandemics and many other phenomena, it is essential to capture the impact of travel (people) and transportation (commodities)
- Populations and models must consider long distance domestic and international T&T
- Challenge: a myriad of existing data source; new ones come online: how to ingest and integrate?
- From a systems perspective, we need a framework consisting of a *harmonizing database*, flow modeling, and translations/augmentation
- Purpose: support quick, unified handling of broad ranges of data sources when augmenting for T&T
Population augmentation: long distance travel and transportation of commodities

Commodity flow

Fresh Tomatoes Export Volume 2013 (Data source: FAOSTAT)
Nodes size proportional to import volume
Edges: > 1000 tons
Population augmentation: long distance travel and transportation of commodities

- Framework naturally supports most patch models
- A radiation-based model can be used for capturing commuter flow on road networks
VALIDATION, VERIFICATION AND SENSITIVITY ANALYSIS
V&V and UQ

• Validation based on
  – Independently collected data: surveys, surveillance information, etc.
  – Expert opinion

• Uncertainty quantification
  – How does uncertainty in the network structure affect its properties?
  – How good is a network based on sampled data?
  – E.g. robustness of rankings based on k-core [PKDD’13]
Criteria for measuring the utility of network modification

- Along the three versions of methodologies, we
  - Add New information [residential networks]
  - Replace raw data with higher resolution data [MMI]
  - Correct wrong data [activity survey]
- We expect to see a higher “quality” network through the modifications
  - What does “quality” mean?
  - How do we compare two networks?
- Metrics of data quality
  - **Accuracy**: How close the measured value is to the actual value in the real world
  - **Precision**: How consistent the simulation outcomes are across replicates.
  - **Fidelity**: The number of featural components included in the model
  - **Resolution**: The granularity of the features represented
  - **Model Purpose**: Affected Epidemics
How can we compare two synthetic networks?

• No single metric which serves to “validate” a system dynamics model (Jay Forrester)

• A number of metrics, organized as *multiple levels/parts of the synthetic network*
  – Entity level: the population, built infrastructure and their layout
  – Collective level: validate against aggregate statistics.
  – Network level: structural properties
  – Epidemic dynamics level: policy effects
Analysis and comparison of three synthetic networks for Delhi NCR

• Different population demographic structure
  – Individual level: V1 matches census data better
  – Household level: V2/V3 matches survey data
Analysis and comparison of three synthetic networks for Delhi NCR

• Location layout
  – The three versions are progressively of higher “realism” in disaggregated structure of location
Analysis and comparison of three synthetic networks for Delhi NCR

• Different activity patterns
  – In the figure are distributions of 7 most frequent human mobility motifs in V2 and V3 (red dots represent home).
Analysis and comparison of three synthetic networks for Delhi NCR
Analysis and comparison of three synthetic networks for Delhi NCR

- Delhi-V1 has a group of people with very low degrees and high clustering coefficients.
- The abnormal peaks disappear in Delhi-V2/Delhi-V3 due to the modeling of residential contacts.
- Residential contacts modeling increases the model fidelity.
Analysis and comparison of three synthetic networks for Delhi NCR

- The distribution of unlabeled graphlets in three networks are very similar.
Analysis and comparison of three synthetic networks for Delhi NCR

• The distribution of labeled graphlets in three networks differs significantly.
Disease Spread in a Social Network

• **Within-host disease model**: SEIR
  
  ![SEIR Model Diagram](image)
  
  - S: Susceptible
  - E: Exposed (Incubating)
  - I: Infectious
  - R: Recovered

• **Between-host disease model**: 
  - probabilistic transmissions along edges of social contact network
  - from infectious people to susceptible people

• **Vaccination Strategies:**
  - implemented at the beginning of epidemic;
  - vaccinate only a selected subpopulation;
  - vaccine stockpile is 10% of the whole population;
  - compliance rate 67%
Analysis and comparison of three synthetic networks for Delhi NCR: Intervention Comparison: the Impact of Age-Mixing Patterns to Epidemics
SENSITIVITY ANALYSIS VIA LOCAL PERTURBATION
Network Sensitivity and importance of details

• A common assumption in the current literature: local properties of a network (such as degree distribution) are often adequate to determine dynamical properties of the system abstracted by the network.

• We show that local properties such as the degree distribution or assortativity might not be sufficient to characterize the global dynamics of reaction-diffusion processes in realistic social networks.

• **Define:** Network Sensitivity (NS) as the change in the dynamical outcome as a function of network structure and propose a systematic way to study NS.

• **Method:** By a simple edge switch operations, we generate a Markov chain of graphs with fixed structural constraints (e.g. degrees) and study disease dynamics on these families of graphs.
Consider the space of all graphs

There are $2^{O(N^2)}$ graphs on $N$ vertices. When the vertices are distinctly labeled, many of these graphs are topologically distinct.
Some of these graphs have a particular degree distribution.

Proofs about properties of scale-free graphs (e.g. “no epidemic transition in a scale-free graph”) describe mean outcomes across the hexagonal region, sometimes variances.
Some of these might represent synthetic networks

The set of synthetic networks may overlap the set of scale free graphs, but may occupy a “small region”. Analytical proofs that weigh each network in this region “equally” therefore are not applicable since most of the weight in an average is not from realistic social networks.
Create “more random” graphs by switching edge endpoints
Constrained edge switch keep local structural invariants
Different switching rules keep different structural constraints.

This choice maintains degree distribution (actually it keeps the degree of all the nodes unchanged).
Different switching rules keep different structural constraints

This choice maintains assortativity by color
Different switching rules keep different structural constraints.

No choice maintains assortativity by degree and color.
Different switching rules keep different structural constraints

Any choice maintains clustering distribution invariant
Edge switch is one step in a Markov chain.

Our Markov chain mixes quickly (empirically), sampling from all graphs with the constrained property.
The chain’s starting point is important.

Similar studies have gone the other direction, introducing structure into random graphs. Harder since one does not know the minimal structural features that need to be maintained.
Simple approach for deciding which details matter

• Given
  – a representational detail in the system interaction generator
  – A labeled graph abstraction $G$ of the interaction structure
  – A dynamical process $P$ over the graph

• Let $S(G)$ be a edge-switch operation that holds certain structural constraints invariant and let $S(G)$ denote the iterated application (walk on the Markov chain). If $P$ changes “sufficiently” as a result, then the features held invariant are not sufficient descriptors of the network model.
Switching operations

Preserves degree distribution

Switch operations

Assortativity preserving Switching

- **Age assortativity (AAS):** switch edges if end points have similar age
- **Degree assortativity (DAS):** switch edges if end points have similar degrees
Edge switch changes structural properties

Switch rate: fraction of the edges switched

Structure of the networks changes significantly with switch rate.
Effect of switching on disease dynamics

**Attack rate**: fraction of population that gets infected

- Disease spreads quickly in the switched networks
Effect of switching on intervention policies

Efficacy of various interventions: School Closure (SC), Vaccination (VC), Work Closure (WC), and Social Distancing (SD)

- Order of efficacy changes in the switched networks

In original network:
SC > VC > WC > SD

In the shuffled network:
WC > SD > VC > SC

“A > B” means intervention A is more effective than B.
Section

CONCLUSION
Paths forward

- Synthetic populations provide a data-driven framework for microsimulation and agent-based modeling.
- Applications are numerous and growing:
  - Infectious and chronic disease epidemiology
  - Disaster response
  - Infrastructure modeling
  - Land use
  - Regional economic modeling
  - Environmental epidemiology
  - Refugees, ...
- Each application brings new methodological challenges
- There is also a need for integrated software platforms for data management, provenance, version tracking, quality evaluation, and more.
THANK YOU!

Samarth Swarup
swarup@virginia.edu
http://people.virginia.edu/~ss7rs

Madhav V. Marathe
marathe@virginia.edu
https://biocomplexity.virginia.edu/madhav-marathe

Useful websites:
NSSAC: https://biocomplexity.virginia.edu/nssac
Biocomplexity Institute: http://biocomplexity.virginia.edu