Generation of realistic consumer populations for electrical demand simulation in the context of the smart grid
Introduction

- Agent-based simulations for energy systems require as an input both:
  - the structure of the system (a network or graph) and
  - the attributes of the agents (or nodes)

- Moreover, the agents are positioned over the network according to their characteristics

- Generation of synthetic, networked populations involve two steps:
  - generate a population of agents (that represent energy consumers) with realistic attributes
  - position them over a network according to their attributes
Use case

- In a previous work, authors proposed a model of energy consumption that reconstructs the load curve from the bottom up [3,6]

- Parameters of this model include a list of the devices of each household, and the corresponding usage of these appliances (the « synthetic population »)

- Data:
  - we previously conducted a survey on 769 households [3]. Data includes household usage rates, sociodemographics and lifestyle issues.
  - Objective: extrapolate « smartly » this sample to large synthetic population sizes to be used into simulations

- This case is representative of most data-driven agent-based simulations in which data is rare and costly
First approach:
Hand-made generation of attributes
(without network)
Main approaches to population generation

- Survey data
- Real population(s)
- Aggregate data
- Resizing
- Fusion
- Random generation
- Synthetic populations
Aggregating data

- We rejected population resizing in order to avoid errors caused by the low size of our sample.
  - for instance, this ~700 sample contains a household with 1 adult and 12 children
  - extrapolated to 10,000 households => 15 of such an household, which is actually rare in the actual population.

- We aggregate data by detecting correlations, based on:
  - **common sense**: «the number of showers may probably depends on the number of people in the household or the number of children»
  - **statistical analysis**: «does the income determines the number of showers ?»
  - **number of samples in the cross analysis**: too few correlations, reject the correlation
<table>
<thead>
<tr>
<th>nrshowers_class</th>
<th>npers</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Count</td>
<td>12</td>
<td>22</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>% within npers</td>
<td>4,2%</td>
<td>4,5%</td>
<td>.7%</td>
<td>4,2%</td>
<td>2,6%</td>
<td>.0%</td>
</tr>
<tr>
<td>0_to_4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>108</td>
<td>137</td>
<td>15</td>
<td>18</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>% within npers</td>
<td>37,6%</td>
<td>28,0%</td>
<td>10,6%</td>
<td>12,7%</td>
<td>5,1%</td>
<td>16,7%</td>
</tr>
<tr>
<td>10_to_14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>18</td>
<td>188</td>
<td>41</td>
<td>25</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>% within npers</td>
<td>6,3%</td>
<td>38,4%</td>
<td>28,9%</td>
<td>17,6%</td>
<td>25,6%</td>
<td>25,0%</td>
</tr>
<tr>
<td>15_to_19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>1</td>
<td>12</td>
<td>28</td>
<td>15</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>% within npers</td>
<td>.3%</td>
<td>2,4%</td>
<td>19,7%</td>
<td>10,6%</td>
<td>17,9%</td>
<td>.0%</td>
</tr>
<tr>
<td>20_to_24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>0</td>
<td>3</td>
<td>26</td>
<td>25</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>% within npers</td>
<td>.0%</td>
<td>.6%</td>
<td>18,3%</td>
<td>17,6%</td>
<td>15,4%</td>
<td>8,3%</td>
</tr>
<tr>
<td>25_to_29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>23</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>% within npers</td>
<td>.0%</td>
<td>.0%</td>
<td>1,4%</td>
<td>16,2%</td>
<td>7,7%</td>
<td>8,3%</td>
</tr>
<tr>
<td>30_to_34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>% within npers</td>
<td>.0%</td>
<td>.2%</td>
<td>.0%</td>
<td>2,8%</td>
<td>2,6%</td>
<td>8,3%</td>
</tr>
<tr>
<td>35_to_39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>% within npers</td>
<td>.0%</td>
<td>.0%</td>
<td>.0%</td>
<td>.0%</td>
<td>15,4%</td>
<td>.0%</td>
</tr>
<tr>
<td>40_more</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>% within npers</td>
<td>.0%</td>
<td>.0%</td>
<td>.0%</td>
<td>.7%</td>
<td>2,6%</td>
<td>8,3%</td>
</tr>
<tr>
<td>5_to_9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>148</td>
<td>127</td>
<td>29</td>
<td>25</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>% within npers</td>
<td>51,6%</td>
<td>25,9%</td>
<td>20,4%</td>
<td>17,6%</td>
<td>5,1%</td>
<td>25,0%</td>
</tr>
<tr>
<td>Total</td>
<td>287</td>
<td>490</td>
<td>142</td>
<td>142</td>
<td>39</td>
<td>12</td>
</tr>
<tr>
<td>% within npers</td>
<td>100,0%</td>
<td>100,0%</td>
<td>100,0%</td>
<td>100,0%</td>
<td>100,0%</td>
<td>100,0%</td>
</tr>
</tbody>
</table>

- Example of crosstabulation using SPSS; note the low number of cases in some cells
Choice of correlations

- The survey data is processed using the SPSS software

- for instance, for computing the number of showers, we detected:
  - \( p(\text{nbshowers} \mid \text{nbpeople}) \): the number of showers mainly depend on the number of people
  - \( p(\text{nbpeople}) \): for generating a population, we will use the frequency distribution of the number of people

- Feedback:
  - on small samples (even for 700 households !), we have to rely only on first-order correlations
  - The use of aggregate statistics facilitates the correction of aberrant cases (by manually correcting some probabilities)
Application

- Algorithm: monte-Carlo sampling:
  - for each household to create
    - randomly select the value of the lifestyle according to \( p(\text{lifestyle}) \),
    - randomly select the value of the number of showers according to \( p(\text{nbshowers} \mid \text{lifestyle}) \)

- Does work, but:
  - we have to encode the generation of the population by ourself; this increase the risk of programming errors or artifacts
  - it provides no solution for the generation of the network according to agents’ attributes
  - => we explore the use of a standalone generic tool for the generation of networked populations [2]
Second approach: use of a generic tool to generate both attributes and a network
The YANG approach

- The YANG approach [2] is devoted to the generation of synthetic networked populations.

- It provides:
  - a meta-model for describing the attributes of individuals and probabilistic generation rules for creating links according to the properties of agents
  - an algorithm for the generation (we ignore the generation of networks in the frame of this study), and an open-source graphical tool for using it
  - a measure of generation errors
Encoding of attributes

- Attributes are encoded as random variables with conditional probabilities into a Bayesian network (same principle with a graphical representation)

- Then YANG uses the same Monte-Carlo principle as ours for generating the population
Generation of networks

- **principle**: households are not positioned randomly over the distribution network; for instance, rural areas often contain bigger households

- YANG enables the description of rules for creating links; these rules describe the probability to create a link given the attributes of individuals

- **First experiment**:
  - we add agents that represent distribution hubs of the network,
  - all the agents are spatialized over 5 fictive spatial areas,
  - households are connected to distribution hubs of their area,
  - distribution hubs are interconnected by a random electric grid
  - household size depends on the spatial location

- Let’s browse an example of generated network
- View of the 5 spatial locations, and of the random distribution grid
- Zoom on this network; household and hubs have the same location as expected
There is now an homogeneity in the size of households per distribution hub
Because the number of showers depends on the household size, it is also different for various hubs.
Discussion and future work
Discussion

- In short:
  - the use of aggregate statistics facilitate the correction of data (reduce the statistical impact of aberrant cases)
  - however, on such a small sample, we can rarely correlate more than two variables
  - the use of a generic tool facilitates prototyping of networks by positionning the agents according to their characteristics

- This opens the way to the reproduction of load curves at the meso level (hubs, spatial areas);

- Next steps:
  - retrieve data on the correlation between household types and their location over a distribution network,
  - reproduce the differences in lifestyles at this local scale (heterogeneity in building isolation, work periods, etc.) and its impact on the load curve
Thanks. questions/comments/critics warmly welcomed!