A Synthetic World Population for Agent-Based Social Simulation.

IAN DENNIS MILLER and GERALD C. CUPCHIK, University of Toronto

1. INTRODUCTION

pplapi.com (pronounced “people API”) is a web-based data service that provides access to a synthetic world population, \( n = 7,171,922,938 \). Researchers can submit queries to pplapi.com using its Application Programming Interface (API) to obtain samples consisting of synthetic agents drawn from this population. Because researchers do not need to host the synthetic population themselves, pplapi.com reduces start-up costs associated with using synthetic agents in research. pplapi.com provides a canonical namespace for Agent Based Social Simulations, permitting comparisons between models implemented with disparate modeling frameworks and programming languages. pplapi.com follows in the tradition of reference data sets that were historically difficult to compute [Rand-Corporation 1955].

2. BACKGROUND

2.1 Agent-Based Social Simulation

Agent-Based Modeling (ABM) was originally suggested by Von Neumann [1966], but the Schelling [1971] model of segregation is regarded as the first major demonstration of the method. Research by Epstein and Axtell [1996] into artificial societies opened a new wave of investigation into Agent Based Social Simulation (ABSS). General purpose mathematical languages such as MATLAB [MathWorks 1998] and R [RDevelopment-Core-Team 2008] are widely used for ABSS research. ABSS software suites, including NetLogo [Wilensky 1999], MASON [Luke et al. 2004], and Mesa [Masad and Kazil 2015] provide specialized tools for agent research. The software landscape is highly fragmented; Nikolai and Madey [2009] identified more than 50 separate ABM and ABSS toolkits.

2.2 Synthetic Populations

Synthetic populations are useful for research, particularly in the fields of transportation, demography, and public health [Ballas et al. 2005; Wheaton et al. 2009; Namazi-Rad et al. 2014]. A variety of techniques have been explored for generating synthetic populations, including various sampling methods, cross-tabulation methods, iterative fitting methods, and sample-free methods [Barrett et al. 2009; Barthelemy and Toint 2013; Ye et al. 2016]. Many population synthesis algorithms simulate country census entities such as families, households, census tracts, and population centers, potentially limiting the generalizability of the data to a single country or to a particular scientific domain.

3. METHODS

3.1 pplapi.com: synthetic world population as a service

All humans who participate in any modelable social process are drawn from the same world population, \( n = 7,171,922,938 \). Boero and Squazzoni [2005] have argued that agent-based models must be empirically calibrated or else they are at risk of being divorced from reality. On that basis, any Agent-Based Social Simulation that attempts to model human behavior should draw from an agent...
1. Miller and G. Cupchik

Table I. Querying the API. Several example URLs provide an overview of pplapi.com functionality.

<table>
<thead>
<tr>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://pplapi.com/123.csv">http://pplapi.com/123.csv</a></td>
<td>Retrieve an agent by its number - in this case, 123. Receive the results as a CSV file.</td>
</tr>
<tr>
<td><a href="http://pplapi.com/123.xml">http://pplapi.com/123.xml</a></td>
<td>Retrieve an agent, receiving results as XML.</td>
</tr>
<tr>
<td><a href="http://pplapi.com/random.csv">http://pplapi.com/random.csv</a></td>
<td>Retrieve one agent at random.</td>
</tr>
<tr>
<td><a href="http://pplapi.com/country/ca/random.csv">http://pplapi.com/country/ca/random.csv</a></td>
<td>Retrieve one agent at random from the country indicated by its Internet Top Level Domain (TLD) - in this case, the country is Canada.</td>
</tr>
<tr>
<td><a href="http://pplapi.com/batch/5/country/ca/sample.csv">http://pplapi.com/batch/5/country/ca/sample.csv</a></td>
<td>Retrieve a random sample of 5 agents from the indicated country.</td>
</tr>
</tbody>
</table>

population that is empirically grounded within the human population. We believe there is no complete model of humankind suitable for Agent-Based Social Simulation, so we introduce pplapi.com.

3.2 Online API

pplapi.com enables researchers to access a synthetic population of all humans alive in 2014. The data are available through a web-based API that can be accessed using the Hypertext Transfer Protocol (HTTP). Virtually every major programming language supports HTTP, so many existing ABSS toolkits already have HTTP capabilities. The pplapi.com website includes examples for integrating with several research environments, including NetLogo, MASON, and R.

The API is designed according to REST principles [Fielding and Taylor 2002], so the API is operated by downloading specific files (see Table I). As of early 2016, the API provides structured data files in CSV, JSON, TXT, HTML, XML, and LISP format. When the JSON format is used to serialize the entire population space of pplapi.com, it yields a file with size \( \approx 6.8 \) TB.

3.3 Implementation

The web service is implemented in the Python language [Python-Software-Foundation 2010] using the Flask-Diamond project framework [Miller 2016]. We have implemented a sample-free population synthesis algorithm based on country-level global parameter estimates, which are informed by a variety of data sources [C.I.A. 2014; Schierl 2014].

4. RESULTS

To briefly demonstrate the utility of pplapi.com, we asked ourselves: 1) what does the world population distribution look like; and 2) how many of those people have Internet access? Figure 1 depicts a sample of \( n = 500 \) agents selected at random from pplapi.com and plotted by country.

To validate the results generated by pplapi.com, we performed a basic check for convergence between the synthetic population and known population parameters (Figure 2). The initial results are encouraging but a word of caution is in order. Country-level parameters are used during agent synthesis, so research questions that require higher resolution data (e.g. state or county level) will require augmented agents based on additional data sources.

5. CONCLUSION

Altogether, pplapi.com is a universal data platform that enables Agent-Based Social Simulations of the human population.

Collective Intelligence 2016.
In this plot, $n = 500$ agents are selected at random from the synthetic world population. To acquire data for this plot, the URL http://pplapi.com/batch/500/sample.csv was queried. A stacked histogram displays the total number of individuals selected from each country. Each country is then subdivided according to Internet access. Because some countries have larger populations than others, agents from those countries are more likely to be included in this sample.

In each comparison, pplapi.com deviates from the US Census by less than 4%.

Collective Intelligence 2016.
REFERENCES


Collective Intelligence 2016.