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

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

`http://pplapi.com/123.csv`	Retrieve an agent by its number - in this case, 123. Receive the results as a CSV file.
`http://pplapi.com/123.xml`	Retrieve an agent, receiving results as XML.
`http://pplapi.com/random.csv`	Retrieve one agent at random.
`http://pplapi.com/country/ca/random.csv`	Retrieve one agent at random from the country indicated by its Internet Top Level Domain (TLD) - in this case, the country is Canada.
`http://pplapi.com/batch/5/sample.csv`	Retrieve a random sample of 5 agents.
`http://pplapi.com/batch/5/country/ca/sample.csv`	Retrieve a random sample of 5 agents from the indicated country.
`http://pplapi.com/metrics.csv`	Retrieve metrics about *pplapi.com* itself.

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* ≈ 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.

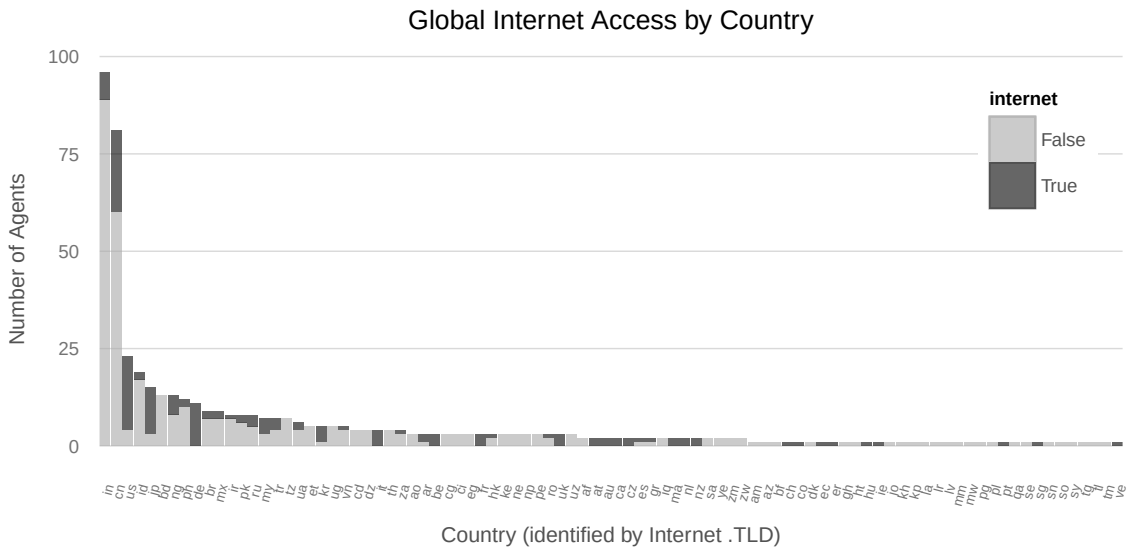

Fig. 1. In this plot, $n = 500$ agents are selected at random from the synthetic world population. To acquire data for this plot, the URL `http://pplapi.com/batch/500/sample.csv` was queried. A stacked histogram displays the total number of individuals selected from each country. Each country is then 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.

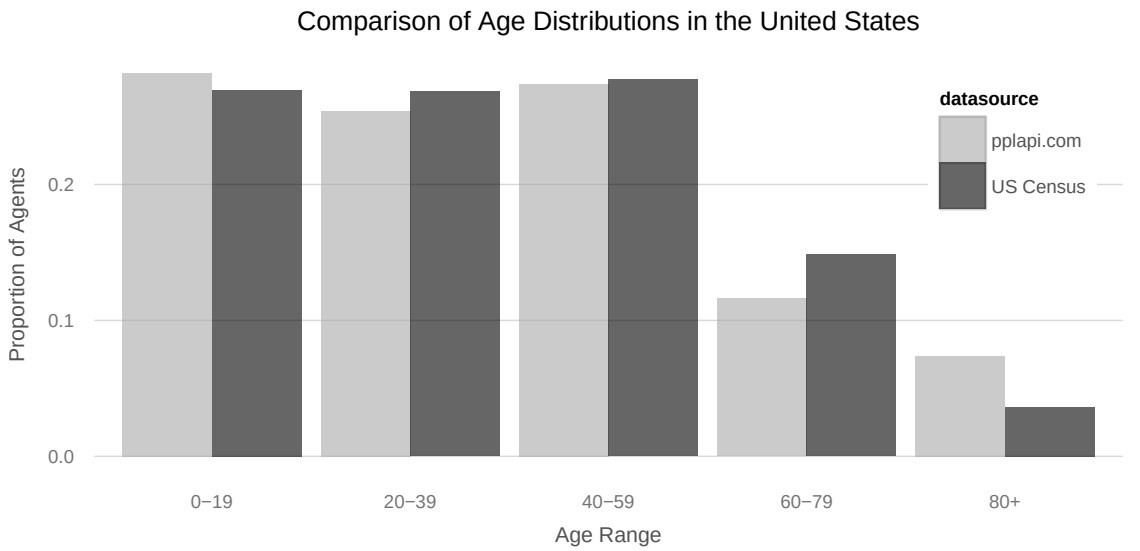

Fig. 2. $n = 500$ agents were selected at random from the United States, and were then grouped by age. To acquire data for this plot, the URL `http://pplapi.com/batch/500/country/us/sample.csv` was queried. Comparable age groups were aggregated from the United States 2010 Census [Howden and Meyer 2010]. Age ranges from *pplapi.com* and the US Census are then compared side-by-side according to proportion. From this visualization, it is clear that *pplapi.com* underestimates the 60-79 range and overestimates the 80+ range. In each comparison, *pplapi.com* deviates from the US Census by less than 4%.

REFERENCES

Dimitris Ballas, Graham Clarke, Danny Dorling, Heather Eyre, Bethan Thomas, and David Rossiter. 2005. SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place* 11, 1 (Jan. 2005), 13–34. DOI:http://dx.doi.org/10.1002/psp.351

Christopher L. Barrett, Richard J. Beckman, Maleq Khan, V. S. Anil Kumar, Madhav V. Marathe, Paula E. Stretz, Tridib Dutta, and Bryan Lewis. 2009. Generation and Analysis of Large Synthetic Social Contact Networks. In *Winter Simulation Conference (WSC '09)*. Winter Simulation Conference, Austin, Texas, 1003–1014. http://dl.acm.org/citation.cfm?id=1995456.1995598

Johan Barthelemy and Philippe L. Toint. 2013. Synthetic population generation without a sample. *Transportation Science* 47, 2 (May 2013), 266.

Riccardo Boero and Flaminio Squazzoni. 2005. Does empirical embeddedness matter? Methodological issues on agent-based models for analytical social science. *Journal of Artificial Societies and Social Simulation* 8, 4 (2005). http://jasss.soc.surrey.ac.uk/8/4/6.html

C.I.A. 2014. *The World Factbook*. Technical Report.

Joshua M. Epstein and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press.

Roy T. Fielding and Richard N. Taylor. 2002. Principled Design of the Modern Web Architecture. *ACM Trans. Internet Technol.* 2, 2 (May 2002), 115–150. DOI:http://dx.doi.org/10.1145/514183.514185

LM Howden and Julie Meyer. 2010. *Age and Sex Composition: 2010*. Technical Report. U.S. Department of Commerce, Economics and Statistics Administration. U.S. Census Bureau.

Sean Luke, Claudio Cioffi-Revilla, Liviu Panait, and Keith Sullivan. 2004. Mason: A new multi-agent simulation toolkit. In *Proceedings of the 2004 swarmfest workshop*, Vol. 8. 44. http://cobweb.cs.uga.edu/~maria/pads/papers/mason-SwarmFest04.pdf

David Masad and Jacqueline Kazil. 2015. Mesa: An Agent-Based Modeling Framework. (2015). http://conference.scipy.org/proceedings/scipy2015/pdfs/jacqueline_kazil.pdf

MathWorks. 1998. The MATLAB User Guide. *Inc., Natick, MA* 5 (1998), 333.

Ian Dennis Miller. 2016. Flask-Diamond. (Jan. 2016). http://flask-diamond.org/

Mohammad-Reza Namazi-Rad, Payam Mokhtarian, and Pascal Perez. 2014. Generating a Dynamic Synthetic Population Using an Age-Structured Two-Sex Model for Household Dynamics. *PLoS ONE* 9, 4 (April 2014), e94761. DOI:http://dx.doi.org/10.1371/journal.pone.0094761

Cynthia Nikolai and Gregory Madey. 2009. Tools of the trade: A survey of various agent based modeling platforms. *Journal of Artificial Societies and Social Simulation* 12, 2 (2009), 2. http://jasss.soc.surrey.ac.uk/12/2/2.html

Python-Software-Foundation. 2010. Python Language. (July 2010). http://python.org

Rand-Corporation. 1955. *A million random digits with 100,000 normal deviates*. Minnesota Historical Society.

RDevelopment-Core-Team. 2008. R: A language and environment for statistical computing. *R Foundation Statistical Computing* (2008).

Thomas C. Schelling. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1, 2 (1971), 143–186. http://www.tandfonline.com/doi/abs/10.1080/0022250X.1971.9989794

Michael Schierl. 2014. FactbookXML. (Dec. 2014). http://jmatchparser.sourceforge.net/factbook/

John. Von Neumann. 1966. *Theory of self-reproducing automata*. University of Illinois Press, Urbana.

William D. Wheaton, James C. Cajka, Bernadette M. Chasteen, Diane K. Wagener, Philip C. Cooley, Laxminarayana Ganapathi, Douglas J. Roberts, and Justine L. Allpress. 2009. Synthesized population databases: A US geospatial database for agent-based models. *Methods report (RTI Press)* 2009, 10 (2009), 905. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875687/

Uri Wilensky. 1999. NetLogo. (1999).

Pei-jun Ye, Xiao Wang, Cheng Chen, Yue-tong Lin, and Fei-yue Wang. 2016. Hybrid Agent Modeling in Population Simulation: Current Approaches and Future Directions. *Journal of Artificial Societies and Social Simulation* 19, 1 (2016), 12. http://jasss.soc.surrey.ac.uk/19/1/12.html