Working Title on Simulation of Synthetic Ecosystems

Shannon K. Gallagher, Lee F. Richardson, Samuel L. Ventura, and William F. Eddy
Carnegie Mellon University Department of Statistics

Abstract

This is not yet abstract but will become so with the passage of time.

1 Introduction

With the increased ability in computing in the past two decades, agent based modeling (ABM) has gained significance in civic engineering ADD CITATIONS [2], finance ADD CITATIONS, and especially epidemiology ADD citations (FRED, Porco's latest, find others). In particular, agent based models allow epidemiologists to model the spread of disease and also simulate disease prevention strategies as in FRED Paper citation.

ABM as input relies on pre-specified *agents* or microdata which represent individual objects or individuals with a given set of characteristics. Generally, the agents represent diverse populations. As ABM necesitates that agents interact with one another (and possibly their environment), agents with a richer set of qualities are preferred. We call the agents together with their environment a *synthetic ecosystem*.

For instance we have a table which represents a family in the United States. Finish example. they interact at work, school, home, have race/age/etc.

Ultimately, ABM modelers desire to create useful models that are reflect reality, and have value guiding decision making. As such, we need to create accurate microdata as input to these models. Expanding the work of Wheaton et al. in [11], we intended to focus on syntheic ecosystems within the United States. However, when the Ebola epidemic broke out during the summer of 2014, we extended our model to affected countries in Western Africa, such as Sierra Leone, Liberia, Mali, and more. Challenges quickly arose that did not occur in dealing with the United States due to the quality and type of data available.

In response to these challenges, we develop a flexible, modular proram, Synthetic Populations and Ecosystems of the World (SPEW), geared toward generating specific ecosystems for users. At its core, SPEW creates ecosystem's by consolidating three data sources

1. Population totals
2. Sample microdata
3. Regional geography

along with additional sources of data such as workplaces and schools. As compared to previous instations of synthetic ecosystems, SPEW, is flexible enough to incorporate any human poplation given the availability of data. We currently have generated Western Africa, the United States, and hopefully we have more to create more than (impressively large number) of synthetic people.

Accurate populations have different meanings for different users and thus SPEW is designed to create ecosystems 'on the fly' with variables of interested included for the user. The user can select the region and depth of geographical hierarchy of interest along with traits of the individuals such as age, income, race, or simply the region totals. SPEW in turn finds the most appropriate data and selects the stastical method to best match the user's preferences.

Although we can generate the world on a supercomputer over a few days for ourselves, agent based models are a heavy computational burden, and we provide code for individuals who can easily create moderate sized populations (∼10 million individuals) for their use cases.

Wrapper paragraph.

The rest of the paper is organized as follows. In Section **??**, we discuss the evolution of synthetic populations for ABMs. In Section 3, we describe in the detail the challenges and variety of data we incorporate into SPEW, first focusing on the United States and then the rest of the world. In Secton 4, we describe how we orient our synthetic ecosystems to the user's goals and how we utilize parallel computing. In Section 5, we discuss the main results and how we verify the accuracy of our synthetic ecosystems. Finally, in Section 6, we summarize SPEW's capabilities and describe what we plan to include in future iterations of the program.

2 Prior Work

The first working ABMs can be traced back to the the late 1960s and 70s with Conway's Game of Life [1], along with Schelling's segregation simulation [10]. The first model being an agent based model with deterministic decision rules and the latter probalistic. In both cases, the actual agents are very simple representing agents with one or two qualitites.

As technology progressed, so has the work with ABMs, which can be found in epidemiology ([5] and [8]), logistics [7], civil science [2], [9] and more. Most of these applications focused more on the outputs of the ABMs rather than the inputs or agents.

For our purposes, the biggest development came in 1996. Beckman et. al [2] were particularly interested in creating accurate agents for modeling traffic simulation in Chicago, and they incorporated Deming and Stephan's Iterative Proportional Fitting Procedure (IPFP) [4] as a way of matching population demographics which tables representing their marginal distributions. They utilized the TRANSIMS look up acronym software which still exists today. The IPFP is a way to find the Maximum Likelihood Estimator (MLE) for cells of a contingency table given the marginal totals for certain variables. Using this technique to first create a contingency table from existing marginal totals and sample microdata, Beckman devised sampling weights in which to create full and accurate synthetic ecosystems.

Wheaton et al. [11] extended Beckman's program to generate synthetic ecosystems of the entire United States matching on the variables: number of children, household income ($), household size, household population, and vehicles available, disseminating the data at a county level and using marginal totals at a block group level (see Figure 3.1). Their synthetic ecosystem population totals are based off the 2010 Decennial US Census. In addition to the four variables that were matched on, Wheaton incorporated schools and workplaces for which the individuals of the synthetic ecosystem would attend. These synthetic ecosystems were designed specifically for ABMs and both [5] and [8] incorporate them in their models. Limiting capabilities of the Wheaton population include which agent qualities to match on and adherence to the 2010 Decennial Census numbers. In addition to the household and individual populations, Wheaton produced a separate group quarters population including assisted living facilities, prisons, dorms, etc.

While our specific purpose is to create synthetic populations for ABMs, it should be noted that there is lot's of research done creating synthetic populations for privacy purposes. A Bayesian approach to population generation is implemented by Hu, Reiter, and Wang [6], which creates completely synthetic data, rather than sampling multiple copies from microdata as in in the IPF or naive sampling. However, it should be noted that Hu et. al's population is generated with the aim of privacy and not necessarily for the purpose of input to use in ABMs. Hu's populations are designed for communities with the order of magnitude of about 10^4 individuals and it is currently unclear how household populations can be combined with individual populations.

3 Data

A difficult challenge in creating SPEW is consolidating data from a variety of sources. As mentioned above, the necessary data ingredients include population totals, sample microdata, and regional geography. The three necessary Fortunately, for the most part, relevant data exists, but the sources vary in what we call their *harmonization*, ie: how easily the different sources of data sync up with one another. We begin with an example of well synchronized data for the United States and then generalize our approach for the rest of the world.

3.1 United States

Nationwide data is available from the US Census for all three of necessary data sources, and since they all originate from the same organization, the data is harmonized to a high degree. We have a detailed description fo the data in Appendix A.

For population totals, we have both household and individual counts available from the American Community Survey (ACS) Summary Files (SF). These counts are available at the block group level, a census unit consisting of about 100 double check households. However, we work at the tract level which is the union of of census block groups and consists of about 4,000 people per tract. The advantage of using tracts over block groups is they are less variable with the passage of time than block groups and some conditional tables of block groups are suppressed by the Census for privacy reasons.

In addition to providing marginal counts, the Census provides sample microdata or Public Use Micro Samples (PUMS) of actual de-identified individuals from 5%? of the population. Due to privacy reasons, the locations of the agents in the PUMS are only available at the Public Use Micro Area (PUMA) level.

As illustrated in Figure 3.1, there is no direct relationship between PUMAs and counties, which is usually a desired input for ABM. This discrepancy between the data highlights the challenge of synthesizing data, even in a highly harmonized place like the United States.

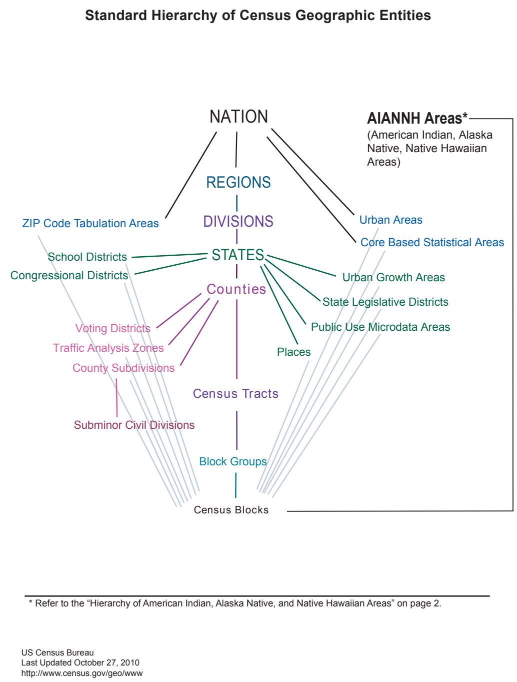

Figure 1: From `census.gov`. Geographical hierarchy of US regions. Of note, we see that PUMAs and counties do not have a nested relationship, an issue which we handle by using the largest common geography of these two: the census tract.

Along with the counts and microdata, we also have to include locations for our synthetic agents by incorporating regional geographies. Borders are dynamic, especially as we move down the geographical hierarchy, which adds a final challenge to consolidating our data sources for use in SPEW. For the United States, we use the Census Topologically Integrated Geographic Encoding and Referencing (TIGER) products for the different borders which allow us to assign locations our synthetic agents.

3.2 The World

Although the US has easily accessible, high quality data available, especially for the 3 primary ingredients for our synthetic population, this is generally not the case. However, there are sources of international data which work for our purposes, some of which are harmonized across countries, and others which come from statistical agencies.

For international population totals, we use `geohive.com`. Geohive has the equivalent of level 2 geography, which are the equivalent of states for nearly every country in the world. The levels represent the granularity of the regions with a larger level being more granular than the previous. We have an example of different levels in Table **??**. For some countries, we have Level 3 geography available, which would be the equivalent of counties in the US.

The counts, in comparison to the US, represent population totals only. This presents a challenge for us because we sample from households PUMS, which in turn generate the people. There are many solutions to this issue, and one we employ entails finding the household average for each country and using that to find the number of households per region. In general, there is a tradeoff in balancing the correct populations of people and households, but this tradeoff can be mitigated using more advanced sampling techniques such as mean matching, or the Iterative Proportional Fitting (IPF) algorithm. Again, this just emphasizes the importance of the user's objectives. We can design a population to accurately reflect the variables the user needs for her research.

Add table of levels of geography

For PUMS data we have available IPUMS-I [3], which are PUMS for many countries in the world, although many are unavailable. In the case where we cannot find microdata for a country, we use a neighboring country, where again 'neighboring' is defined in the terms of the user's goals. If the user desires a country that is an economic neighbor, than we use those PUMS. Our program is flexible as to meet the needs of the users.

Finally, our primary source for international shapefiles comes from the Global Administrative Areas (GADM) site this project. In general, these files provide us with a baseline for how to split a country into its subregions. We use this baseline to guide how we match the population counts from geohive, and employ record linkage techniques when there are discrepancies.

4 Methods

4.1 Approach

At a high level, our program can be divided into three steps which is visualized in Figure Make figure:

1. Data Formatting
2. Ecosystem Generation (Parallel Step)
3. Dissemination

Each step emphasizes a different issue which we expound upon in the below.

4.1.1 Data Formatting

This is the most manual labor-intensive step of our program. The issue lies within the fact that we, more often than not, aggregate data from unharmonized sources. In particular, we need to match geographical IDs to one another, find the relationship between the marginal totals and the available microdata, and find geographies that are syncrhonous with our current data. The task is made difficult because each of these pieces can be quite dynamic through time.

Geographical IDs are easily matched in the case where we use data from the same source and year, such as the US Census. The US Census geographical hierarchy assigns unique ID numbers to each tract through an 11 digit ID, `SSCCCTTTTTT`, which represents a 2 digit state number, a 3 digit county number, and a 6 digit tract number.

For most countries, however, geographies do not have a unique identification number, and we are forced to match by the character names, a feat made challenging particularly for non-English speaking countries. For example, this would happen in an international country when the level 2 geography in a GADM shapefile doesn't match exactly with the level 2 geography from the geohive counts. As a result, we use a small form of Record Linkage to create similarity scores of region names from different sources. We print these scores in our log files so we can always check double check our work, and have a record of what was done.

Our formatting is done in a pre-processing step with the help of a configuration file in which we maintain region specific instructions for our program. This step syncronizes all of the data together across sources, and set's the table for the actual sampling and generation of microdata.

4.1.2 Ecosystem Generation

This step is the 'workhorse' of our program and is designed to work in parallel. The data formatting step puts our data in such a form that the proper microdata, population totals, and geography is loaded at a granular level (about the size of a county). This means that we now know how many people to sample in each region, where to put them, and which portion of the microdata to sample from. Our process then samples household microdata (from a selection of sampling techniques), samples longitude and latitude coordinates from the correponding geographic region, combines the household populations with individuals, and incorporates other agent data requested, such as school and workplace assignments.

4.1.3 Dissemination

Our program works by splitting a country into mutually exclusive regions, the union of which adds up to the entire country. From this, we can generate a synthetic population for each one of these subregions, using minimal data as an input. From minimalist perspective, all we need is the number of housegolds in the subregion, the appropriate microdata, and the latitude and longitude bounding box. We generate a synthetic population for each subregion, and organize them as requested by the user.

For example, in the united states we generate a unique synthetic population for each tract. In each tract, we sample from the microdata corresponding to whichever PUMA the tract is located inside, and sample the location of the household from the TIGER file corresponding to the apropriate tract. We organize these tracts into subdirectories organized by county. Thus, the default United States synthetic population has a subdirectory for each state, each county within the state, and a synthetic population file for each tract.

4.1.4 User Flexibility

A feature of our program is the flexibility the user has to generate synthetic ecosystems that best fit her needs.

4.2 Computing

Something which can be deduced from the dissemenation section is that our method for generating synthetic populations is well suited for parallel computing. In particular, once we split our large region into subregions, we can use a single node to quickly generate a synthetic population. In particular, we use the Olympus computing cluster, hosted by the Pittsburgh Supercomputing Center. This cluster has 23 compute nodes, each containing 64 total computing cores. Thus, we we split our computation across 1536 computing cores while utilizing the entire cluster, greatly reducing the time it takes to complete an entire region.

For example, the United states is split into 74,000 tracts. The time to generate an individual tract is two minutes. If we split out our computations across the 1500 Olympus cores, it will take us around $(74000/1500) \times 2 =\approx 100$ minutes to generate the entire United States. In the future, it may be more computationally feasible to

5 Results and Vetting

How are we validating or population? Add a section on Automated checks and diagnostics

6 Conclusions and Future Work

A Data List

1. 2006-2010 5-year ACS PUMS
   - Available at: `http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t`
   - Corresponds to 2000 defined Census geography
   - Household and People populations
   - For detailed information see: `http://www.census.gov/acs/www/data_documentation/documentation_main/`

   (a) `pums_h.csv`
   - The variables correspond to different household attrributes, about 80 of which are weights.

   (b) `pums_p.csv`
   - People population subset of the PUMS
   - The variables correspond to different pepole attributes, around 90 of which are weights.
2. US Census TIGER Shapefiles– 2010
   - Available at `https://www.census.gov/geo/maps-data/data/tiger.html`
   - Geographical boundaries of different census regions. Currently have block group level, which is the most fine unit diseminated by the Census.
3. National Center for Education Statistics School Data
   - Available at: `http://nces.ed.gov/ccd/elsi/tableGenerator.aspx`
   - Can find school data for given year and region.
   - Variables include enrollment information, latitude and longitude coordinates, and other useful variables.
   - Both public and private school data available
4. ESRI workplace data

References

[1] Andrew Adamatzky. *Game of Life Cellular Automata*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[2] R.J. Beckman, K.A. Baggerly, and M.D. McKay. Creating synthetic baseline populations. *Transportation Research Part A*, 30(6):415–429, 1996.

[3] Minnesota Population Center. Integrated public use microdata series, international: Version 6.3, 2014. [Machine-readable database].

[4] W. Edwards Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):pp. 427–444, 1940.

[5] Grefenstette JJ, Brown ST, Rosenfeld R, Depasse J, Stone NT, Cooley PC, Wheaton WD, Fyshe A, Galloway DD, Sriram A, Guclu H, Abraham T, and Burke DS. Fred (a framework for reconstructing epidemic diseases): An open-source software system for modeling infectious diseases and control strategies using census-based populations., October 2013. PubMed PMID: 24103508.

[6] Jingchen Hu, JeromeP. Reiter, and Quanli Wang. Disclosure risk evaluation for fully synthetic categorical data. In Josep Domingo-Ferrer, editor, *Privacy in Statistical Databases*, volume 8744 of *Lecture Notes in Computer Science*, pages 185–199. Springer International Publishing, 2014.

[7] Fu-ren Lin and Shyh-ming Lin. Enhancing the supply chain performance by integrating simulated and physical agents into organizational information systems. *Journal of Artificial Societies and Social Simulation*, 9(4):1, 2006.

[8] Fengchen Liu, WayneTA Enanoria, Jennifer Zipprich, Seth Blumberg, Kathleen Harriman, SarahF Ackley, WilliamD Wheaton, JustineL Allpress, and TravisC Porco. The role of vaccination coverage, individual behaviors, and the public health response in the control of measles epidemics: an agent-based simulation for california. *BMC Public Health*, 15(1), 2015.

[9] David L. Sallach and Charles M. Machal. Introduction: The simulation of social agents. *Social Science Computer Review*, 19:245–248, Fall 2001.

[10] Thomas Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 1971.

[11] William D. Wheaton, James C. Cajka, Bernadette M. Chasteen, Diane K. Wagener, Philip C. Cooley, Laxminarayana Ganapathi, Douglas J. Roberts, Justine L. Allpress, and James C. Cajka. Rti press synthesized population databases: A us geospatial database for agent-based models, 2009.