SPEW: Synthetic Populations and Ecosystems of the World

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

Over the past two decades, agent based modeling (ABM) has gained significance. Because of improvements in technology, larger environments can be modeled, in terms of number of both number of agents and features of said agents Add citation. Examples of ABMs can be found in in cival engineering [**?**], finance Add citations, and especially epidemiology [**?**], [**?**]. In particular, ABMs, for instance, allow epidemiologists to model the spread of disease and also to simulate disesease prevention strategies as was done in the software, FRED [**?**].

As input, ABMs rely on pre-specified *agents* or microdata which represent individual persons or objects with a given set of characteristics. Generally, these agents represent diverse populations. As ABMs necesitate 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*.

Ultimately, ABM modelers desire to create useful models that adequately reflect reality and have worthwile insight with regards to decision making. As such, there is a demand for high quality agents for use in ABMs. Expanding on the work of Wheaton et al. [**?**], we originally intended to focus on person-related synthetic ecosystems within the United States. However, with disease outbreaks such as Ebola and Zika, we extended our agent building to affected countries such as Sierra Leone and Brazil and eventually the world at large.

While creating these international agents, challenges quickly arose due to data availability. In response to these challenges, we develop a flexible, modular program, Synthetic Populations and Ecosystems of the World (SPEW), geared toward generating specific ecosystems for users. At its core, SPEW creates ecosystems by consolidating three data sources:

1. Population counts,
2. Geography, and
3. Microdata.

However, the novelty of SPEW lies in the control the user has over the program. SPEW is modular. It can incorporate schools and workplaces if available, the same with churches or hospitals. SPEW is flexible: SPEW allows for many methods of sampling and we include uniform sampling, iterative proportional fitting, and a feature based matching method in order to create synthetic individuals. SPEW is parallelizable: we break geographies into granular regions and run the program on each of these regions. Finally, SPEW is visualizable. We incorporate diagnostic summaries in the form of static reports along with interactive maps of generated populations.

To date, SPEW has synthesized over 3.5 billion individuals from over 50 countries worldwide. More importantly, SPEW gives the users a tool to generate ecosystems according to their needs along with summary information and diagnostics.

The rest of the paper is organized as follows. In section 2, we discuss previous work on generating synthetic populations. In section 3, we explain the necessary data, supplementary data, challenges about integrating data from many sources, and how we overcame such challenges. In section 4, we explain the features of SPEW, including computational, statistical, and graphical components. In section 5, we discuss the synthesization of three different countries, all with different features to emphasize. Finally, we summarize our report in sectoin 6. Contained in the appendices A and B are links to our code and data included in our synthesis.

2 Prior Work

The first working ABMs can be traced back to the the late 1960s and 70s with Conway's Game of Life [**?**], along with Schelling's segregation simulation [**?**]. 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 ([**?**] and [**?**]), logistics [**?**], civil science [**?**], [**?**] 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 [**?**] 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) [**?**] 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. [**?**] 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 5.2.1). 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 [**?**] and [**?**] 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 [**?**], 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

At its core, SPEW relies on three primary source of data: population counts, geography, and microdata. These three elements form the skeleton of any SPEW generated ecosystem. With this in mind, it should come as no surprise a majority of the time spent generating these ecosystems was devoted to collecting data and making sure it was integrated together.

Perhaps the most challenging feature in developing SPEW is collecting and integrating numerous sources of data to create a harmonized synthetic ecosystem. There are two key steps to assembling the necessary data for generating our synthetic ecosystems: Collection and Integration. By collection, we mean the identification and download of the raw data sources. By integration, we refer to the process of making sure that these raw data-sources are aligned with one another. For example, integrating our population counts with our geographies means that we are making sure the region names/identifiers contained within our shapefiles match with the names of our population counts. As one can imagine, the ease with which we could collect and integrate our data-sources together varied by country. In order to make our efforts reproducible, we have made all of the code we used to download, and integrate the data available online, at the following addresss:

`https://github.com/leerichardson/spew_olympus`

We have included this for two main purposes. The first, is that it makes things easier for use when we need to track down issues that arise when generating our synthetic ecosystems. The second reason is that, if desired, users of our synthetic ecosystems will be able to understand down every decision we made, and the journey from turning our raw data-sources into our synthetic populations.

3.1 Counts

Counts refers to the population counts of a region. At the very least, SPEW needs the number of individuals per region. However, more data can be incorporated such as household totals, gender totals, income totals, and race totals. Counts are generally the easiest piece of data to find. For instance, the United States, `census.gov` cite provides practically all the count total data one could need down to the tract level, which consists of about 4,000? persons. For context, there are approximately 88,000 tracts currently in the United States. add THE census graphic.

Although other governments likely provide census counts a la the United States, a useful site we found for population counts is `geohive.com` cite, which is a compendium of population totals for nearly all countries in the world. In addition, geohive population totals are generally available at the state or county equivalent. However, only total individual population counts are available as opposed to individual and household counts as in the United States.

Due to the different nature of available data, we developed SPEW with the data as a proxy for the config file. A prevailing theme of SPEW with regards to data is if we have more features available, then include them. If not, then make do with what we have.

3.2 Geography

In addition to counts, SPEW requires geographic information as an ecosystem is incomplete without supplying its agents a physical location. SPEW assigns each individual a latitude and longitude which is inferred from the input geography. A point location is more useful than a regional location (state, county, tract) because interactions with the environment. For instance with a point location, we can assign children to nearby schools within a county and

their parents to a nearby workplace. In this way, we enable ABM modelers to create richer agent-environment interactions.

The geography's input is generally as a shapefile (`.shp`). Contained within a shapefile is polygon, line, and point data. Also associated with a shapefile are files with such extensions as `.dbf` and `.prj` which attach data and labels to the shapes and map projections, respectively. A diagram would be useful here In this way, SPEW can sample points from the different shapes which allow for different densities of populations.

Again, the United States census cite provides US boundary data down to a tract level. For other countries, we found two useful sites, IPUMS-I and GADM citations, which provide shapefiles of the world at different levels of granularity.

3.3 Microdata

While counts and geographical data are vital to SPEW, the heart and soul of the synthetic ecosystems are generated from sample microdata. This microdata is what differentiates SPEW's output from being points on a map into rich set of individuals with diverse characteristics.

A natural question the reader may ponder is if SPEW needs microdata to make microdata, why just not use the original microdata. The answer is that the input microdata to SPEW is usually only a small sample of the total population. In addition, we have yet to see microdata with physical locations attached. Finally, the microdata may not emphasize the proper variables, and when extrapolating to a full population size, some features of the data may be incorrect unless synthesized in a deliberate way.

Still the microdata is very important, especially with comparisons to marginal totals of features. In this way we are using real data as the basis of our populations rather than fabricated data.

Again, the United States provides the microdata, called Public Use Micro Samples (PUMS) cite. PUMS are available at the PUMA level, which are geographic units defined the census, each approximately containing 100,000 persons. Revealing its namesake, IPUMS-I also hosts many collections of microdata of different countries in the world.

3.4 Supplementary Data

Once the three primary data sources are accounted for, the user can turn their attention to other sources of data. For instance, we have added modules to add schools and workplaces for the United States, from the National Center for Education and Statistics and ESRI, respectively. Other data sources include places of worship and hospitals, and airports which we all found from OpenStreetMap cite. Perhaps, the user wants to add data about smoker status to the population. We provide an interface for that. This is definitely a TODO item. There is effectively no limit to the data one can add to the population, and we include an interface for adding different features to our people.

3.5 Harmonization of Data Sources

3.6 SPEW is made to handle different data

4 SPEW

4.1 SPEW overview

Who created SPEW? The text below makes it sound like it was not the authors

Good for a draft, prolly not needed in final version.

In the previous section, we've described the data necessary in order to generate synthetic ecosystems. In particular, the required data sources are population counts, geographies, and microdata. In this section, we describe how we moved from the three datasources to synthetic ecosystems.

These two sentences say the same thing.

A key development in generating our synthetic ecosystems has been the creation of the R Package, SPEW (Synthetic Populations and Ecosystems of the World). The purpose of SPEW is to provide a general engine for generation of synthetic ecosystems. The need for a general program arose when we realized that once we had collected and integrated the data sources for a particular country, the process of moving from data to ecosystem was the same. The creation of SPEW has enhanced our capabilities of generating synthetic ecosystems in two main ways. First, because SPEW expects data in a particular format, we now know precisely what colections of data-souces must look like before they can be turned into a synthetic ecosystem. This is helpful because it gives us a clear goalpost to aim for while collecting and integrating data. Second, by abstracting the process of moving from data to ecosystem, we were able to generate all of our ecosystems with SPEW. This not only makes our ecosystems more reliable, but it gives us a straightforward path to add functionality in terms of new methods, modules, etc.. down the road.

Same as what?

New paragraph

Again, sentences say the same thing.

Note that we have made the source code for SPEW available online at the following address. Aside from striving for reproducibility, providing the source code allows users to look into the exact details, as well as add functionality and eventually create ecosystems of their own.

`https://github.com/leerichardson/spew`

Can't they already create ecosystems with SPEW?

4.1.1 How SPEW Works

Never lead with a figure

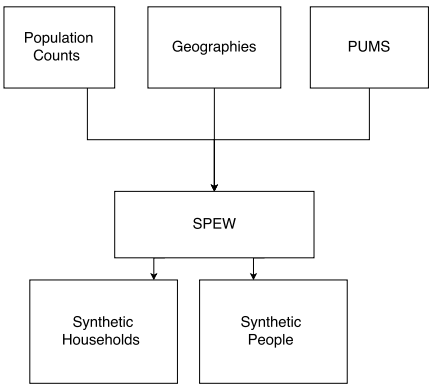

Figure 1: This diagram shows what spew does as a high level: Takes formatted population counts, geographices, and microdata and outputs synthetic ecosystems

you already said this.

At a high level, spew performs the function of taking our three integrated data sources, and outputs a synthetic ecosystem, see **??** for a demonstration. More specifically, spew works by splitting a location into mutually exclusive regions, the union of which adds up to the entire location. From this, we generate a synthetic population for each one of these regions.

I'm expecting to lean about spew here not data format.

It's important to point out that our PUMS data usually contains a variable which corresponds to a specific region within the synthetic population. This variable is usually a superset of many smaller regions, and we refer to it as the `puma_id`. Thus, for each region we typically subset the PUMS data to contain only data from the corresponding `puma_id`. This leads to synthetic ecosystems which are more representative of the marginal distributions of each tract.

For example, in the United States we generate a unique synthetic population for each tract. In this case, we can think of each tract as one of our mutually exclusive regions. Note that each tract is contained within a Public Use Microdata Area (PUMA), and the United States PUMS data has a variable indicating which PUMA each record is located within. Thus, for each tract we subset the PUMS data to contain all samples from the particular PUMA the tract is located in. Once we have the correct PUMS data, we sample the appropriate number of households for the particular tract. Next, we sample the location of each household using the Tiger shapefile, which serves the role of our geographies. Note that before we can generate the populations, we verify that all of the regions in the geography files match up with all of the regions in the population counts file. Finally, we organize these tracts into subdirectories organized by PUMA. Thus, the default United States synthetic population has a subdirectory for each state, each PUMA within the state, and a synthetic population

Para does two things, and the most important is buried.

input : Population counts, geographies, microdata, other ...
1. Verify each data source has necessary components ;
2. Verify data sources align with one another ;

for *Every Region* **do**
1. Sample Households ;
2. Sample Locations ;
3. Attach People to Households ;
4. Add other data as desired (eg: schools, workplaces, etc...) ;

end
output: Synthetic Households, People, etc...

An algorithm is like a piece of math, and needs to be incorporated into the sentence around it (use the name "Algorithm 1" in a sentence, e.g.).

As we discussed in class, you may want to move this up higher in the document so that it provides a sort of "automatic outline" for what you want to write about -- that way you are less likely to forget details, e.g.

Algorithm 1: Pseudocode for generating Synthetic Ecosystems with SPEW

It's important to point out that the pseudocode in the above algorithm is fairly general. In particular, one could use any method they wanted to sample households, locations, and even people within households. Also note that while the three required data-sources needed to generate the synthetic households and people, there is in principle no type of data, be it schools, workplaces, hospitals, mosquitoes, etc.., that we could not include into this framework.

sentences don't scan

This generality of SPEW is by design, and we think that it is one of our best features. Because we are dealing with many heterogenous data-sources, and can not predict the future types of ecosystems which will be requested by Agent Based Modelers, we strove to create an engine in which it would be easy to implement new features and requests as desired. For instance, we imagine an interested party could simply give us an algorithm and data to assign agents to schools, and we could then easily incorporate and run this through spew.

Right now, spew uses very rudimentary methods for these different steps. In particular, instead of using IPF as in [**?**], we employ simple random sampling. This means that we are simply re-sampling the pre-existing records until we have enough for a synthetic ecosystem. Along these sample lines, to sample locations we are simply uniformly sampling over each particular region, without focus on the locational features within the ecosystem. While these

which different steps?

where is this defined?

basic techniques allowed us to set up the initial framework and get the first round of populations released, there is much room to advance the sophistication of the methods we use. One could imagine using density estimation to sample the households, and attribute based sampling to location them. While our current populations lack sophisticated methods, we believe this gives us lots of room for growth in future iterations of the program.

4.2 Computational

My strong sense is that you've just written down some stream of consciousness thoughts here without organizing them into a sequence of points that will stick with the reader.

By design, generating synthetic populations is a very computationally intensive task. Even in just generating one ecosystem, when there are millions of people with many characteristics to be generated, the input/output load is quite expansive. Fortunately, we had access to the Olympus Cluster, hosted by the Pittsburgh Supercomputing Center.

The Olympus cluster is made up of 24 nodes, with one serving as the head node and the other 23 nodes serving as compute nodes. Each node has four cite olympus documentation multi core processors, each of which contains 16 compute cores, so each node has 64 compute cores. This means that, in principle, we can run 1536 processes at one time on Olympus.

One other thing that jumps out when inspection reference algorithm, is that spew is what is referred to as an embarrasingly parallel application. By this, we mean that once we have our integrated data, it easy easy to see that by parallelizing each region, we have a straighforward way of generating our synthetic ecosystems in parallel.

4.2.1 Parallelization

4.2.2 Olympus

4.2.3 Generating the World 3.5 billion and counting!

4.3 Statistical (SLV)

4.3.1 Sampling: SRS, IPFP, MMM, Bayes

4.3.2 Other ideas such as variance of pops?

4.4 Graphical (SKG)

4.4.1 Diagnostics Suite

4.4.2 Maps

4.4.3 Interactive Graphics

5 Results

5.1 Overview

5.2 Case Study

5.2.1 US- PUMS - schools and workplace module

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 already highly integrated. We have a detailed description fo the data in Appendix B.

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 PUMS data de-identified individuals from 5%? of the population. This data also comes from the ACS, and in our current iteration we use the one year, 2013 PUMS surveys. Due to privacy reasons, the locations of the individuals in the PUMS are only available at the Public Use Micro Area (PUMA) level.

As illustrated in Figure 5.2.1, there is no direct relationship between PUMAs and counties, and counties are usually the 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.

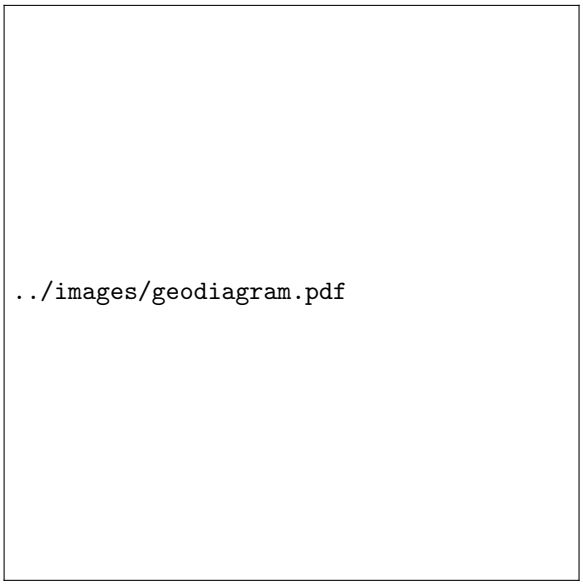

Figure 2: From `census.gov`. Geographical hierarchy of US regions. Of note, we see that PUMAs and counties do not have a nested relationship, an issue which we handle by using the largest common geography of these two: the census tract.

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.

5.2.2 IPUMS

In contrast to the USA, it was more difficult to find harmonized sources of data for other countries. However, we did track down our three required data sources for 83 different countries, largely stemming from the availability of IPUMS data for these countries of interest.

The IPUMS data we have available IPUMS-I [**?**] are simply PUMS data for many countries in the world. In our case, we were able to download PUMS files for 83 countries from the IPUMS website. In these data-sets, the main results such as

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.

5.2.3 CANADA - Custom Synthetic Population

As a final example here, we detail the data we used to create a Canadian Synthetic population. In our view, the Canadian synthetic population represents the way in which we will generate the majority of our ecosystems going forward: Finding data from a location of interest, integrating the sources together, and using SPEW to output a synthetic ecosystem.

In this particular example, we downloaded each of our three data sources from the Statistics Canada website. Once we had the data downloaded, we were able to write a few computer scripts which converted the data into the format suitable for analysis. In particular, we ran the data-set through a series of checks which come with the spew package. Once the data-set made it through these checks, we knew that it was ready for use in SPEW.

The key point in the generation of the Canadian synthetic ecosystem was how quickly we were able to put it together. Specifically, all we needed was links to the three particular data-sources, and after a few hours of munging, we had the synthetic ecosystem. The rapid nature of creating the Canadian Synthetic ecosystem comes largely from the utility of the spew R package we have developed. In particular, we knew exactly how the data needed to look, and once we had it in place we were confident that spew could generate the required ecosystem. While dealing with heterogeneous data-sources always requires some leg-work, we believe our infrastructure has greatly enhanced both the speed and quality with which we can generate synthetic ecosystems.

5.3 In the news...

5.3.1 epimodels.org

5.3.2 Ebola request, Zika

5.3.3 Porco and measles

5.3.4 Hackathon inclusion

5.3.5 MIDAS meeting 2016

6 Discussion and Looking Forward

A Code

B 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