Synthetic Populations and Ecosystems of the World

MIDAS Informatics Services Group

February 21, 2016

Contents

1 Introduction

The purpose of this document is to provide the specific details on how we generated our synthetic ecosystems. Our hope is that it can serve as both reference guide and a way to answer commonly asked questions. As the methodology and data sources change over time, this document will be continuously updated to reflect these changes.

The code used to generate our synthetic ecosystems is in the R package, Synthtic Populations and Ecosystems of the World (SPEW). The purpose of SPEW is to abstract the process of moving from multiple data sources to final synthetic ecosystems. For example, although we used different data sources to generate synthetic ecosystems for the United States and Canada, both of the final ecosystems were generated using SPEW. See figure 1 for a description of the high level functionality of SPEW.

The creation of SPEW enhances our synthetic ecosystems in many ways. For example, because SPEW requires data in a standardized format, we know precisely how to tranform raw data into usable inputs to SPEW. This greatly reduces the amount of time needed to move from raw data to synthetic ecosystem. SPEW also makes our ecosystems more reliable; Since we use SPEW for every ecosystem, we only need to fix bugs and add features in one place, and these propogate to all of our ecosystems. Finally, the modularity of SPEW gives us a straighforward way to add new parts to the ecosystem (eg: restaraunts, hospitals, etc...) as new data becomes avaialable.

Aside from striving for reproducibility, providing the source code allows users to look into the exact details, add functionality and create ecosystems of their own. Our software is publically accesible on Github at:

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

2 Data Sources

2.1 Collecting the Data

In this section, we describe both the data sources used, and how we collected them. For terminology purposes, let's say that each synthetic ecosystem corresponds to a location. Note that this location not constrained geographically,

only by the availability of data. To generate a synthetic ecosystem for a particular location, we require three sources of data:

1. **Population counts**: The number of people in a location. This can be split into regions within a location, eg: The number of people per state within the United States

2. **Geographies**: A map of the location, and it's regions (if any). Note that this is typically a shapefile, in which polygons correspond to regions, and they are mapped accoriding to latitude and longitude.

3. **Microdata**: Individual records. This is typically a data-set in which each row is an individual, and each column is a characteristics of that individual. Some example columns are age, sex, and income.

While include more data (eg: schools and workplaces for the United States) when available, all of our ecosystems include these three sources.

Collecting and formatting all of our data sources was a non-trivial process. Because of this, we have included all of the code used to transform raw data into valid inputs on Github `https://github.com/leerichardson/spew_olympus`.

Recording the collection and integration of our data sources helps in many ways. For one, when questions about the data arise we can quickly trace it back to the original source. More importantly, it makes the process of moving from raw data to final ecosystem more transparent. Many decisions about the data are made along the way, and it's important for eventual users to understand exactly how their data was impacted from beginning to end.

2.2 United States

Here we detail the data sources used for the United States synthetic ecosystem. Similar to the other ecosystems, the United States is made available at `http://data.olympus.psc.edu/syneco/west/north_america/united_states/`. Note that at this link, we also include meta data, such as the geographies used here `lookup/` directory. In addition to population counts, geographies, and microdata, out United States ecosystem includes schools and workplaces. Schools and workplaces are an important features for agent based models, as it gives information on where individual agents spend their time (and spead disease).

Finally, note that our ecosystems for the United States are created at the tract level, as this was the lowest level of geography available. Users can combine these tracts to generate larger ecosystems as desired, using our geographic lookup table mentioned in the previous paragraph.

2.2.1 Population Counts

American Community Survey Summary Tables (2006-2010)

- Available at: `https://www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html`

- Total number of households by Tract

2.2.2 Geographies

US Census Topologically Integrated Geographic Encoding and Referencing (TIGER) Shapefiles (2010)

- Available at `https://www.census.gov/geo/maps-data/data/tiger.html`
- Geographies at the Census tract level

2.2.3 Microdata

1-Year American Community Survey (2013)

- Available at: `http://www2.census.gov/acs2013_1yr/pums/`
- Corresponds to 2010 defined Census geography
- Both household and people populations
- See appendix for variables used

2.2.4 Schools

National Center for Education Statistics School Data (2011-2013)

- Available at: `http://nces.ed.gov/ccd/elsi/tableGenerator.aspx`
- Public Schools (2013) have latitude/longitude information. Private schools (2011) only have county level information.
- More information at: `http://data.olympus.psc.edu/syneco/west/north_america/united_states/schools/`

2.2.5 Workplaces

ESRI Workplace Data (2009)

- Available with a license from ESRI
- ID, employee counts, and county of different businesses in the US
- More information at `http://data.olympus.psc.edu/syneco/west/north_america/united_states/workplaces/`

2.3 Canada

The Canadian synthetic ecosystem is made up of population counts, geographies, and microdata. Data from Statistics Canada is similar to data from the United States, since the lowest level of geography available is also at the tract level. Finally, we point out that we needed to obtain special permissions from the Canadian government in order to use the microdata defined below.

2.3.1 Population Counts

Statistics Canada Census Profile (2011)

- Available at: https://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/details/download-telecharger/comprehensive/comp-csv-tab-dwnld-tlchrgr.cfm?Lang=E specifying the Census Tracts option.
- Total population for every tract

2.3.2 Geographies

Statistics Canada Boundary File (2011)

- Available at: https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2011-eng.cfm specifying the English, ArcGIS, and Census tract option.
- Geographies at the Census Tract level

2.3.3 Microdata

Public Use Microdata File (2011)

- Obtained with special permissions from Statistics Canada
- Variables defined in the appendix

2.4 IPUMS

For many countries, we used a combination of the three data sources defined below. This was possible because each data source had information for various countries, which allowed us to combine them together to form ecosystems. In particular, the Minnesota Population Center's International Public Microdata Sample (IPUMS) provided two of these three data sources, so we refer to this category as IPUMS.

Below, we reference administrative level geographies. By this, we mean the way countries are divided into regions according to the International Organization for Standardization http://www.iso.org/iso/home.html.

2.4.1 Population Counts

Geohive

- Available at: http://www.geohive.com/
- Compiles population statistics from various statistical agencies throughout the world (The list can be seen here http://www.geohive.com/earth/statorgz.aspx)
- Population counts from over 150 countries at various administrative levels.

2.4.2 Geography

IPUMS Shapefiles

- Available at: `https://international.ipums.org/international/`
- Shapefiles corresponding to IPUMS microdata.
- Available at administrative level 1

2.4.3 Microdata

International Public Use Microdata Sample (IPUMS)

- Available at: `https://international.ipums.org/international/`
- Microdata from 82 different countries
- See the appendix for the variables used

3 Methods

In this section, we describe how SPEW generates synthetic ecosystems. Recall that before SPEW can run, the data sources must be converted into a specific format. This format is made up of two separate requirements. First, the data must be converted to use a unified naming system. Second, each data source must have an identical set of geographies. For example, Canada uses three different sources: Population counts, geographies, and microdata. Each one of these sources defines it's own geographic division of Canada. Before SPEW can generate the Canadian ecosystem, we must make sure the the geographic divisions are linked to and identical with eachother. We have included a series of diagnostic tests on our SPEW github page that interested users can run to verify SPEW will run for their particular inputs.

SPEW generates synthetic ecosystems for locations specified by the user. For example, we can generate ecosystems for continents, countries, states, provinces, tracts, etc., as long as the required data is available. See algorithm 1 for a high level overview of how SPEW moves from raw data to synthetic ecosystems. First, SPEW verifies that the data is valied. Next, we split the location into mutually exclusive regions (eg: Census Tracts), the sum of which equals the entire location. Once the location is split, a synthetic ecosystem is generated for each region. For example, Pennsylvalnia is split into tracts, and we generate a synthetic ecosystem for each tract. The idea here is to create synthetic ecosystems at the lowest granular level possible, in order to have more information rich ecosystems. Synthetic ecosystems are generated according to 1, and below we detail how each one of these steps is carried out.

Finally, we must point out that beyond locations, another piece of geographical data we use is the Public Use Microdata Area (PUMA). At a high level, we use PUMA's to obtain more accurate samples. To understand PUMA's, it's

important to note that when we generate synthetic ecosystems, we are sampling records from Microdata. Note that we have microdata for entire locations, and occasionally this microdata contains a variable which indicates a more specific region. If a PUMA is available, we use it to subset our microdata before generating ecosystems for particular regions. Note that a PUMA is always larger than a region.

For example, in the United states a PUMA is made up of around 100,000 people, or 30-60 tracts. Recall that for the United States, we generate synthetic ecosystems at the tract level. So before we sample for an individual tract, we subset the microdata to only include data from the PUMA that contains the tract. By doing this, our tract level synthetic ecosystems of the individual regions as opposed to each tract sampling from the entire state data.

input : Population counts, geographies, microdata, other sources
1. Check that each data source has necessary components
2. Check data sources align with one another
for *Every Region* **do**
    1. Sample Households
    2. Sample Locations
    3. Sample People
    4. Assign other data if available (schools, workplaces, etc...)
end
output: Synthetic Ecosystem

Algorithm 1: Pseudocode for generating Synthetic Ecosystems with SPEW

3.1 Sample Households

From algorithm 1, we see that the first action SPEW takes is sampling households. This is because microdata typically comes in two pieces: One for households, and another for people. We sample from households first so that we can link individual people to their respective household, and we want to make sure that households have the correct number of people. Otherwise, we would be unable to generate accurate household information.

We take a straighforward approach to sampling households: uniform sampling. This means is we are simply resampling households at random from the household microdata. For example, lets say that tract x has n households. To create the synthetic households, we simply resample (with replacement) n households from the household microdata. Note that if tract $x \in p$, for some PUMA p, we only sample from households within this particular tract, otherwise we sample from the entire location. Finally, we assign these sampled households to tract x. In future iterations, we will include more sophisticated sampling methods. For example, the American Community Survey publishes summary tables, which contain demographic summary information for each tract. We could weight individual houseolds to make sure that our synthetic ecosystem matches with these summary tables. More generally, we could estimate the

distribution of households, and sample from this density instead. We hope to include these methods in future iterations.

3.2 Sample Locations

After sampling the appropriate number of households, the next step is assigning each household a location. This is where the geographies are used in SPEW, as each region has a polygon associated with it, which specifies the geographic boundaries of the region in a machine readable format. To assign locations within this polygon, we take another straightforward approach: Uniform sampling within the polygon, and assign these locations to the household. Specifically, we use the R function `spsample` from the `sp` package to uniformly assign households to particular locations.

We acknowledge that simply uniform sampling of locations provides limitations.The reason we use it here is that we have not obtained any more granular data, such as population density estimates. Once we have more detailed location data, we will include more sophisticated approaches for location sampling.

3.3 Sample People

Now that we have households and locations assigned, the next step is to sample people for each household. Note that the household and people level microdata always contain an identifier which can be used to link them. Using this variable, to assign people to households we perform a left join operation using the sampled households and people microdata. This means that our synthetic people are simply the people corresponding to the households which were already sampled. We take this approach to ensure the consistency between our household and person ecosystems. Note that the location attached to each person will be identical to the location of their household.

Like our other sampling methods, we believe there is much room to improve here. In particular, we see that that both persons and households are frequently duplicated. To get around this for people, we could estimate the conditional density of people corresponding to household sizes, and sample from this. We hope to include methods like this in future iterations of the program.

3.4 School Assignments

Note that for schools, we only have data corresponding to the United States. For that reason, we will explain our method for assigning schools in the US here. We have access to school data that includes:

- enrollment totals
- latitude and longitude for public schools
- county location for private schools.

In addition, the synthetic people have features for school enrolment (`SCH`), grade level (`SCHG`), and age (`AGEP`).Using these three variables, we can find which people need to be assigned to a public school or private school. Note that our dataset only corresponds here to high schools, so we are not assinging preschool, college, or professional schools.

The algorithm is as follows. For each person, determine whether the person should be sent to school. Use the county of the person to identify the schools (Note that if there are no schools in this particular, country, we use the entire state). Then using `SCH`, determine whether we should use public or private schools. Using the `SCHG` variable, further subset the schools to find those with the proper grade. For private schools, weight the subsetted schools by the enrollment total, assigning more weight to schools with more enrolled students. For public schools, weight the subsetted schools by both the enrollment total and the haversine distance between the person and the schools. Schools with more people are given more weight as are schools that are physically closer to the person. Finally, sample a school with the weights calculated above. In the case where there are no schools in the county, subset the schools only by state.

In this way, students will not cross county borders when being assigned a school. The enrollment totals and distances are used to assign probabilities to various schools, as opposed to hard restrictions. The process is summarized in Algorithm 2.

input : synthetic people, schools data
for *Every person* **do**
  Determine whether child should be sent to school **if** *no school* **then**
    school ID ← NA
  else
    if *coordinates of school exist* **then**
      weight schools based on distance from child and enrollment totals
    else
      weight schools based on enrollment totals
    end
    sample school from county (use state if there are no schools in a county) based on weights school ID ← sampled school ID
  end
end
output: School IDs

Algorithm 2: Pseudo code for generating schools

3.5 Workplace Assignments

Workplaces are assigned im a similar way as schools. The synthetic people have a feature employment status recode (`ESR`), which tells us if the person is working. The workplace data we have contains the number of employees as well as the state and county of the workplace. We first subset the workplaces to match

the state and county of the person in question. Then we weight the workplaces by the number of employees, with more employees given more weight. Finally, we sample a workplace for each person. The pseudo code for this assignment is shown in Algorithm 3.

```
input : synthetic people, workplace data
for Every person do
    Determine whether person is in workforce
    if no employment then
        work ID ← NA
    else
        weight workplaces in county based on number of employees
        sample a workplace based on weights
        workplace ID ← sampled workplace ID
    end
end
output: Workplace IDs
```

Algorithm 3: Pseudo code for generating workplaces

4 Output

In this section, we will describe the output of SPEW. First, we describe the organization of the data. Next, we explain the other files that come along with our populations, including diagnostic plots and geographies.

Our synthetic ecosystems can be found online at the following web address, `http://data.olympus.psc.edu/syneco/`. Our ecosystems are organized in a geographic hierarchy: The ordering goes: Hemisphere, continent, country, and State/Province. Users should be able to navigate this geographic hierarchy intuitively to find the synthetic ecosystem of interest.

4.1 File Structure

Recall that in the methods section, we introduced three geographic levels SPEW uses to generate syntehtic ecosystems: location, PUMA, and region. We use these three pieces of information to organize our output files geographically. We have a subdirectory corresponding to each location, and a subdirectory for each puma within that geography. Within this PUMA directory, the region-level synthetic ecosystems are stored. For example, let's consider Alabama, which has a US State ID of 01. Using this ID, we can locate the `http://data.olympus.psc.edu/syneco/west/north_america/united_states/01/` directory, within the United States folder. Note that this subdirectory corresponds to the geography of Alabama. Clicking on this, and we see subdirectories corresponding to PUMAS within Alabama, let's take puma ID 100 for example. If we click on this, within the `http://data.olympus.psc.edu/syneco/west/north_america/united_states/01/100/` link, we see various .csv files, starting with

either `people` or `household`. These files correspond to the synthetic households and people level populations of various tracts, within PUMA 100.

4.2 Diagnostics

We are building a set of automatic diagnostics for synthetic ecosystems, which will be available soon.

4.3 Lookup geographies

We are adding an output to include the final geographies used for generation the synthetic ecosystem for particular locations.

A Codebook

A.1 United States: American Community Survey

See http://www2.census.gov/programs-surveys/acs/tech_docs/code_lists/2010_ACS_Code_Lists.pdf

A.2 Canada: Public Use Microdata File

See http://data.olympus.psc.edu/syneco/west/north_america/canada/docs/pums/2011%20PUMF_FMGD/Hierarchical%20file/English/Documentation%20and%20user%20guide/2011%20NHS%20Hierarchical%20PUMF%20User%20Guide.pdf

A.3 IPUMS: International Public Use Microdata Sample

See https://usa.ipums.org/usa/resources/codebooks/DataDict0610.pdf

B Acknowledgements

This work was supported by the Models of Infectious Disease Agency Study (MIDAS) from the National Institute of General Medical Sciences (NIGMS), grant number NIH 1 U24 GM110707-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the National Institutes of Health (NIH).

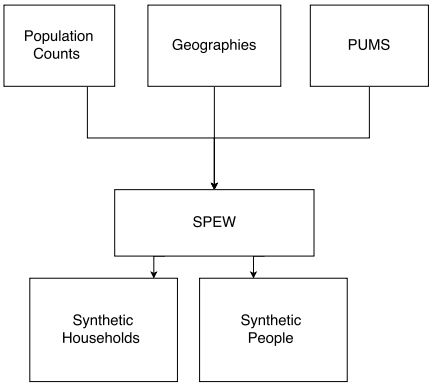

Figure 1: This figure shows the high level functionality of SPEW. The main point to emphasize is that SPEW takes multiple data sources for a location of interest, and converts these into final synthetic ecosystems