SPEW
Synthetic Populations and Ecosystems of the World

Version 1.2.0

January 15, 2017

Shannon Gallagher
William F Eddy
Lee Richardson
Sam Ventura

MIDAS Informatics Services Group

Department of Statistics, Carnegie Mellon University

Contents

1 Introduction

The SPEW framework generates synthetic ecosystems for using different data sources and methodologies. A detailed description of the SPEW framework is given in Gallagher et al. (2016), an article which will be referenced frequently. In contrast, this document provides details on the latest release of SPEW synthetic ecosystems, version 1.2.0. The goal is to explain the details of the released ecosytems and software so users can quickly understand and get started. SPEW 1.2.0 data files and software are distributed simentaneously. So, SPEW 1.2.0 refers to both data files and the software that generated it.

There are two main improvements to SPEW 1.2.0. First, United States synthetic ecosystems use an Iterative Proportional Fitting (IPF) based methodology for sampling population characteristics. Second, each synthetic ecosystem is paired with a diagnostic report, summarizing the contents of each ecosystem.

In total, SPEW 1.2.0 contains 126 synthetic ecosystems, described in Table 1. The 126 ecosystems are categorized into three groups, by their the input data: United States, IPUMS, and Canadianda. The data sources used for each ecosystem group are described in Section 2. Next, the methodologies used for generating synthetic ecosystems are descibed in Section 3. Finally, the structure of the data files are described in Section 4.

Table 1: The three data groups for current SPEW 1.2.0 synthetic ecosystems: U.S., IPUMS, and Canada. Count gives the total number of ecosystems in this group, level gives the size of each ecosystem, and region level gives the lowest region level generated

Group	Count	Level	Region Level
United States	52	State	Tract
IPUMS	73	Country	Admin. Level 1
Canada	1	Country	Tract

2 Data Sources

This section describes the data sources used for generating SPEW ecosystems. Each SPEW 1.2.0 ecosystem requires three inputs:

1. Population counts
2. Geography
3. Population Characteristics

These data sources are described in detail in Gallagher et al. (2016). This section details the data used for each of the three data groups: United States, IPUMS, and Canada.

2.1 United States

United States synthetic ecosystems are the most detailed. A synthetic ecosystem is produced for each tract, containing approximately $1000 - 2000$ households and $3000 - 7000$ agents. In addition to the three required data sources, the United states using road-level data for spatial sampling, summary file data for Iterative Proportional Fitting, and school and workplace data for environmental components.

2.1.1 Population Counts

American Community Survey Summary Tables (2006-2010) U.S. Census Bureau (2010b)

- Available at: `https://www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html`
- Total number of households by Tract

2.1.2 Geographies

US Census Topologically Integrated Geographic Encoding and Referencing (TIGER) Shapefiles (2010) U.S. Census Bureau (2010a)

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

2.1.3 Microdata

1-Year American Community Survey (2013) U.S. Census Bureau (2013)

- Available at: `http://www2.census.gov/acs2013_1yr/pums/`
- Corresponds to 2010 census geography
- Both household and people populations

2.1.4 Summary files

5-Year American Community Survey Summary files U.S. Census Bureau (2010b)

- Available at: `http://www.census.gov/programs-surveys/acs/data/summary-file.html#.html`
- Corresponds to 2010 census geography
- Population characteristic totals at the tract level

2.1.5 Schools

National Center for Education Statistics School Data (2013) of Education. Institute of Education Sciences. National Center for Education Statistics (2015)

- 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.

2.1.6 Workplaces

ESRI Workplace Data (2009) Esri (2009)

- Available with a license from ESRI
- ID, employee counts, and county of different businesses in the US

2.2 IPUMS

IPUMS synthetic ecosystems only use the three required data sources. Microdata and Geographies come from IPUMS, and population counts come from Geohive.

2.2.1 Population Counts

Geohive GeoHive (2016)

- 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.

2.2.2 Geography

IPUMS Shapefiles Center (2014)

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

2.2.3 Microdata

Center (2014) 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

2.3 Canada

The Canadian synthetic ecosystem uses custom national data from Statistics Canada producing synthetic ecosystems at the tract level. Only the three required data sources are used

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) Statistics Canada (2011)

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

3 Methods

This section describes the methodology used for SPEW 1.2.0 synthetic ecosystems. SPEW splits a location into **regions**, then generates a synthetic ecosystem for each region. For each region, SPEW samples household characteristics, person characteristics, and assigns locations to agents and households. In addition, SPEW assigns agents to environmental components if the data is available. The SPEW framework is summarized in Algorithm 1.

input : Harmonized counts, geographies, population characteristics and supp. data
for *Every region* **do**
  1. Sample population characteristics of households
  1. Sample population characteristics of agents
  2. Sample locations of agents
  3. Add environmental components (e.g. schools, workplaces, vectors, etc)
  output: Synthetic ecosystem for region
end

Algorithm 1: Process for Generating Synthetic Ecosystems with SPEW

Algorithm 1 is general, and different methodologies can be used for each step. This section details the approaches SPEW 1.2.0 uses for each data group. Table 2 gives the methodology used for each of the three data groups.

Table 2: Methdologies used for the three data groups.

Group	Population Characteristics Sampling	Spatial Sampling	Environmental Components
United States	IPF	Roads	Schools, Workplaces
IPUMS	SRS	Uniform	None
Canada	SRS	Uniform	None

While most sections refer to Gallagher et al. (2016) for details, the software availabe online at:

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

So the exact details are available to look up, if users are interested.

3.1 Sample population characteristics of households

SPEW 1.2.0 uses two different approaches for sampling population characteristics of households:

- Simple Random Sampling
- Iterative Proportional Fitting

Both approaches are explained in Gallagher et al. (2016). In this version, the United States synthetic ecosystems use Iterative Proportional Fitting, and all of the other ecosystems use Simple Random Sampling.

3.2 Sample locations of agents

For SPEW 1.2.0, sampling locations of agents was done in the same for all data groups. Every microdata source contained two files: one for households, and one for people. In addition, an ID variable linked the household and person level microdata. Once households were sampled, the person level population was created by including every person whose household was chosen. If a household was chosen more than once, then the corresponding agents were chosen an equivalent number of times.

3.3 Spatial Sampling of Agents

SPEW 1.2.0 uses two different approaches for spatial sampling of agents:

- Sampling Uniformly Across a Region
- Sampling Uniformly Across Roads

Both approaches are explained in Gallagher et al. (2016). The United States samples uniformly across roads, and all other ecosystems sample uniformly across a region.

3.4 School Assignments

Only the United States synthetic ecosystems include school assignments. School assignments use an adapted gravity model, which depends on school capacity and distane. Like the other methodologies, the description is given in Gallagher et al. (2016).

3.5 Workplace Assignments

Similar to schools, workplaces are only included in United States synthetic ecosystems. The same gravity model framework is used, described in Gallagher et al. (2016).

3.6 Diagnostics

SPEW's current version contains diagnostic reports for each ecosystem. These reports include

- General Info
  - The country (state) name
  - The number of administrative levels available in the country (state)

  - The number of sub-regions
  - A map of population density with real household assignments
- Synthetic Households and Synthetic People
  - Total number in country (state)
  - Graphs of population characteristics per region
  - The population characteristics included in the synthetic ecosystem
- Generation information

These summaries and graphs allow the user to see whether the synthetic ecosystems pass the "eye test." A discussion of diagnostic is contained in Gallagher et al. (2016).

4 Output

SPEW synthetic ecosystems are available online at:

`http://data.olympus.psc.edu/syneco/`.

Synthetic ecosystems are stored in a geographic hierarchy, based on the hierarchy of the United Nations Statistics division. This hierarchy is available at:

`http://unstats.un.org/unsd/methods/m49/m49alpha.htm`).

The lowest level of our geographic hierarchy is a country. Each country has a corresponding ISO3 code, invaluable for matching data accross sources. Sometimes, we have data at lower levels than country. In this case, we extend the geographic hierarchy to include data at lower levels within the country. An example is the United States, where we have data at the state level, so we include a state level in the hierarchy, underneath the US country.

The hierarchy is as follows:

1. Region
2. Sub-region
3. Country
4. Lower level data (if available)

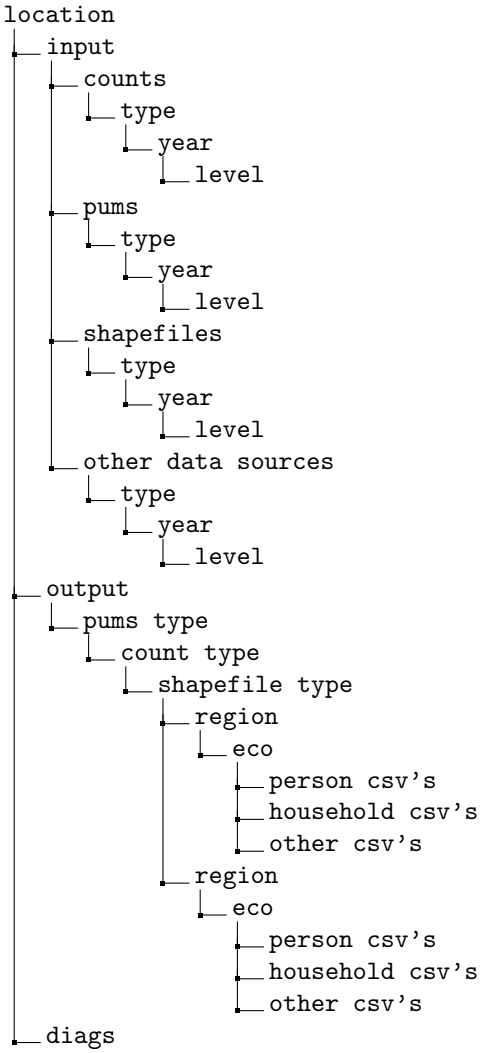
Below are example file-paths for China and Califoria, within the hierarchy:

- `spew_1.2.0/asia/eastern_asia/chn`
- `spew_1.2.0/americas/northern_america/usa/06`

4.1 Directory Structure

Each synthetic ecosystem is contained within its own directory. In this directory, the input data goes in the `input/` folder, and the output data in the `output/` folder. The `input/` directory organizes input data by type (counts, shapefiles, etc), year, and geographic level. The geographic levels won't necesarily match accross different data types, as they come from entirely different sources. Finally, the `diags/` folder contains an html file with the diahnostics report.

The specific directory structure is



In addition, since Canada required lots of manual preparation, we have included a `prep/` folder, with the custom scripts to prepare the data.

References

Center, M. P. (2014). Integrated public use microdata series, international: Version 6.3. [Machine-readable database].

Esri (2009). US business data throughout this paper originated from Esri. Esri Business Data is the intellectual property of Esri and are used herein under license. Copyright Esri. All rights reserved. For more information about Esri software, please visit `www.esri.com`.

Gallagher, S., Richardson, L., Ventura, S. L., and Eddy, W. F. (2016). SPEW: Synthetic Populations and Ecosystems of the World. *Submitted.*

GeoHive (2016). `http://www.geohive.com/`.

of Education. Institute of Education Sciences. National Center for Education Statistics, U. D. (2015). Elsi tablegenerator. Electronic.

Statistics Canada (2011). Census of Canada, 2011, Families File (public-use microdata file). CD. All computations, use and interpretation of these data re entirely those of the author.

U.S. Census Bureau (2010a). 2010 TIGER/LineShapefiles (machine-readable data files). Electronic.

U.S. Census Bureau (2010b). American Community Survey Summary Files 2006-2010. Electronic.

U.S. Census Bureau (2013). American Community Survey Public Use Microdata Sample 2013. Electronic.

A Codebook

SPEW 1.2.0 ecosystems use a different source of microdata for each data group. This section gives the variables used for each data group and points to the relevant documentation.

A.1 United States

United States ecosystems use the American Community Survey, and release the following variables. THe codebook for this version of the ACS is located at:

`https://usa.ipums.org/usa/resources/codebooks/DataDict2013.pdf`

A.2 Canada

The Canada codebook is online at:

`http://data.olympus.psc.edu/syneco/spew_1.2.0/americas/northern_america/can/input/prep/Individual%20file/English/Documentation%20and%20user%20guide/2011%20NHS%20Individuals%20PUMF%20User%20Guide.pdf`

A.3 IPUMS: International Public Use Microdata Sample

The data dictionary for IPUMS is online at:

`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), Cooperative Agreement 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).