SPEW: Synthetic Populations and Ecosystems of the World

ICSP 2017 - Lucca, Italy

Shannon Gallagher

22 February 2017

Carnegie Mellon University
Department of Statistics
Lee Richardson
Samuel L. Ventura
William F. Eddy

Better Synthetic Ecosystems → Better Agent-Based Models

Agent-Based Models (ABMs) simulate phenomena

- Answer questions in ecology, epidemiology, sociology, and more
    - FRED, EpiSims
- Allow direct input from field experts
- Take advantage of low-cost and availability of modern computing
- Simulate of events that are not attainable from ordinary scientific methods
    - impractical, or even, unethical, scenarios

ABMs require agents and their environment as input

- Agents - a set of objects with possibly varying characteristics
  - e.g., mosquitoes, people, birds, cars
  - also known as synthetic {`individuals`, `people`, `households`, `etc.`}

- Environment - a constrained region containing loci of interaction
  - e.g., swamps, schools, forests, intersections

ABMs require agents and their environment as input

- Agents - a set of objects with possibly varying characteristics
  - e.g., mosquitoes, people, birds, cars
  - also known as synthetic {`individuals`, `people`, `households`, `etc.`}

- Environment - a constrained region containing loci of interaction
  - e.g., swamps, schools, forests, intersections

We call agents together with their environment an ecosystem

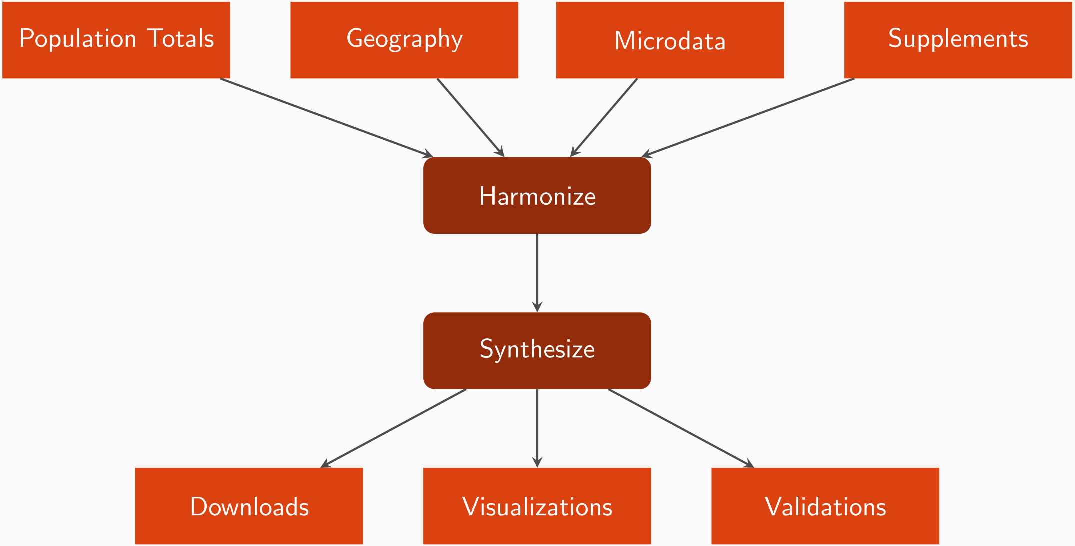
SPEW is an R package that produces synthetic ecosystems

stat.cmu.edu/~spew

- SPEW
  - Synthetic Populations and Ecosystems of the World
- “Synthetic” – created from data via statistics
- R
  - open-source, documented, easy data manipulation

r-project.org

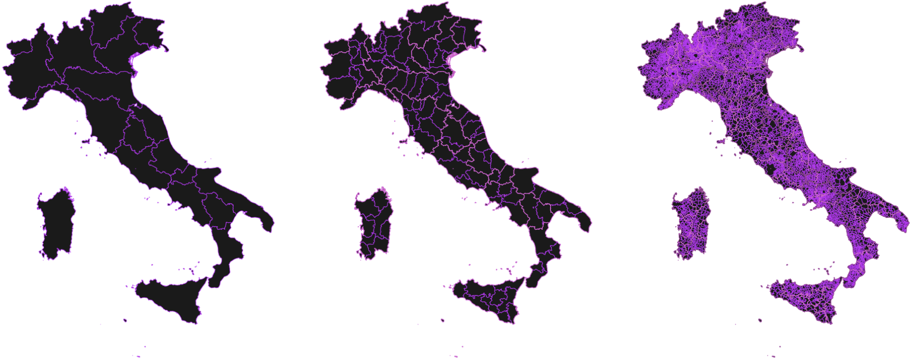
SPEW provides a framework for generating synthetic ecosystems



The framework partitions input data into 3 essential pieces and the rest

- Population Totals - how many synthetic individuals are in a region
- Geography - a digital representation of the region where the synthetic individuals are
- Microdata - samples of actual individuals with multiple features
- Supplements - schools, workplaces, churches, airports, etc.

SPEW works at the ‘sub-region‘ level



Italy partitioned into sub-regions at different administrative levels

Once SPEW has data, it synthesizes agents using sampling modules

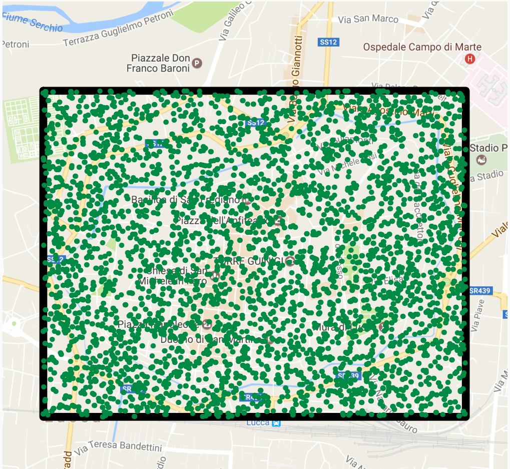
Algorithm 1 Process for synthesizing an ecosystem for a region

```
for every sub-region do
    Sample population characteristics of agents
    Sample locations of agents
    Assign environmental components to agents
end for
```

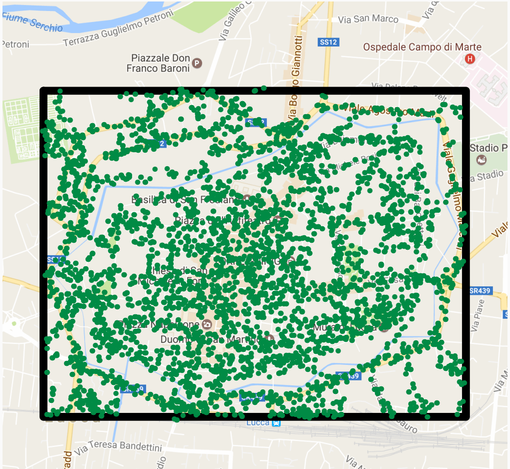
SPEW currently supports 3 methods of sampling individual characteristics

		Uniform	Moment Matching	Iterative Proportional Fitting
Data Required	microdata	✓	✓	✓
	moments of characteristics		✓	✓
	marginal/joint distribution			✓
Advantages	continuous characteristics	✓	✓	✓
	categorical characteristics	✓		✓
	accurate population totals		✓	✓
	ease of implementation	✓		
	use of non-rep. microdata		✓	✓
	emphasize characteristics		✓	✓
Disadvantages	flexibility	✓		
	curse of dimensionality		✓	✓
	reliance on small set of records			✓

SPEW can sample uniformly from a region or from known spatial information



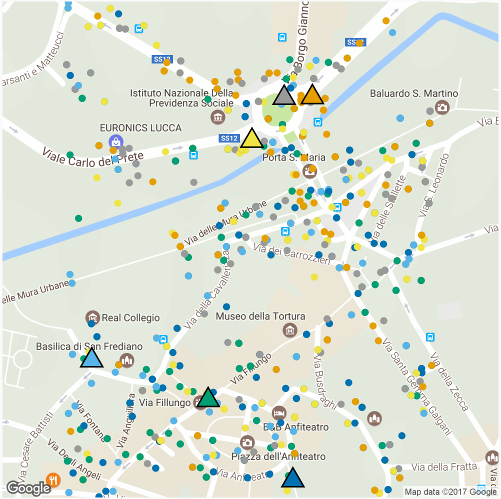
Uniform Sampling
Lucca, Italy



Road−based Sampling
Lucca, Italy

SPEW uses a ‘gravity model’ to assign agents to environmental components

Shop Assignments in Lucca, Italy



Outputs are disseminated as tables

Agent ID	Household Income ($)	Family Size	Sex	Age	School ID	Workplace ID
459799	0	2	2	32		
1065696	5400	4	1	10	100023000205	
1038094	34000	1	2	42		816264717
635925	48000	2	2	59		
1135185	49000	4	1	11	100020000229	
258679	26600	4	1	11	100002600259	
29921	104000	4	1	50		1765643
341548	129500	2	1	74		505047084

Example output from Tract 010101, DE

- Synthetic agent and household tables are both outputted (`.csv`)
- Output contains unique regional identifier
- Environments may be recovered from either input data or output

Outputs are disseminated as tables

Agent ID	Household Income ($)	Family Size	Sex	Age	School ID	Workplace ID
459799	0	2	2	32		
1065696	5400	4	1	10	100023000205	
1038094	34000	1	2	42		816264717
635925	48000	2	2	59		
1135185	49000	4	1	11	100020000229	
258679	26600	4	1	11	100002600259	
29921	104000	4	1	50		1765643
341548	129500	2	1	74		505047084

Example output from Tract 010101, DE

- Synthetic agent and household tables are both outputted (`.csv`)
- Output contains unique regional identifier
- Environments may be recovered from either input data or output

Outputs are disseminated as tables

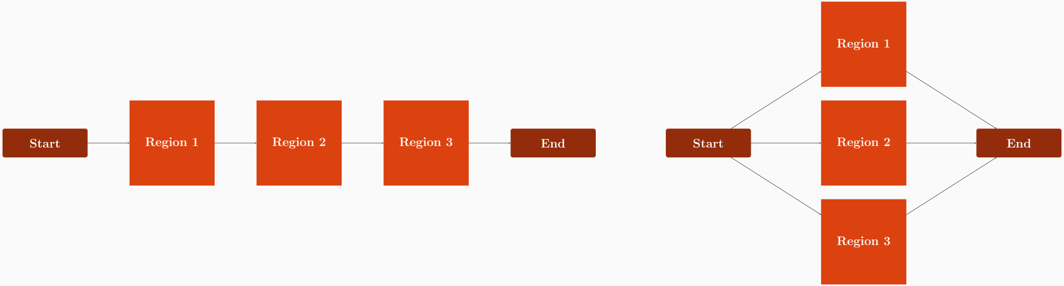
Agent ID	Household Income ($)	Family Size	Sex	Age	School ID	Workplace ID
459799	0	2	2	32		
1065696	5400	4	1	10	100023000205	
1038094	34000	1	2	42		816264717
635925	48000	2	2	59		
1135185	49000	4	1	11	100020000229	
258679	26600	4	1	11	100002600259	
29921	104000	4	1	50		1765643
341548	129500	2	1	74		505047084

Example output from Tract 010101, DE

- Synthetic agent and household tables are both outputted (`.csv`)
- Output contains unique regional identifier
- Environments may be recovered from either input data or output

SPEW is parallelizable

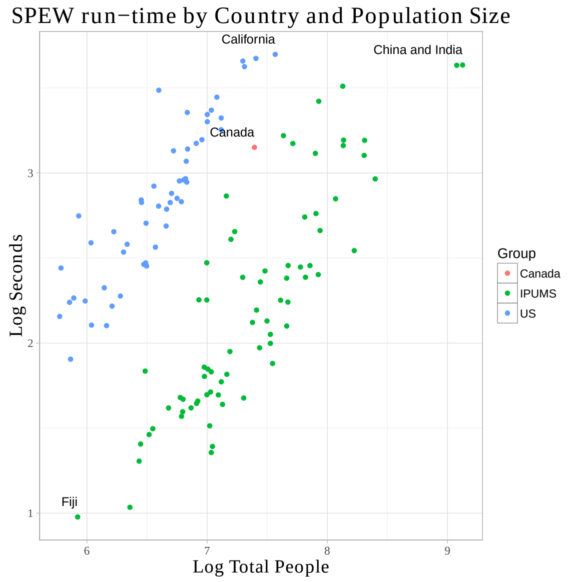
- SPEW can be parallelized on computers such as your laptop to supercomputers!
  - The process described earlier is embarrassingly parallel
  - i.e., generation of sub-regions are independent from one another



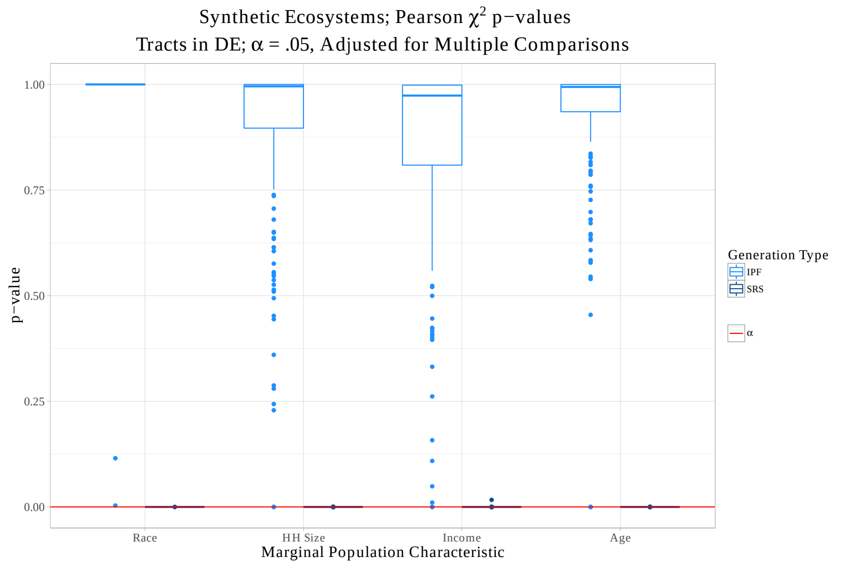
Left: Sequential generation. Right: Parallel generation.

Higher quality input data → longer run-times

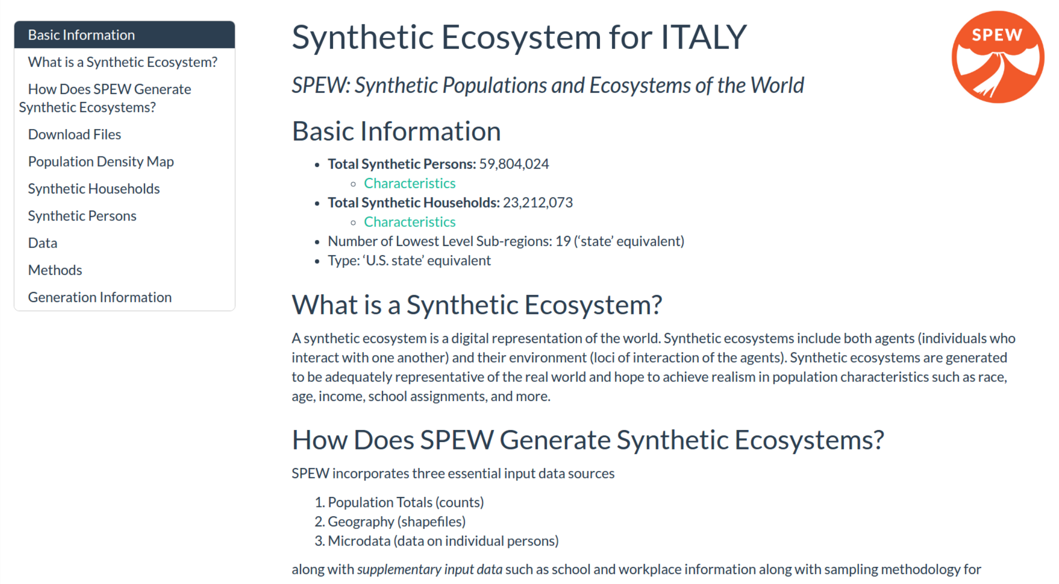
- U.S. synthetic ecosystems take longer
  - more available input data
- Run-time seems to be order O(*n*)
  - where *n* is # of agents
- Ran each country/state on one node
  - Run on Olympus
    - Pittsburgh Supercomputing Center
  - 24 nodes total
  - 64 cores/node



SPEW alerts us to synthetic ecosystems gone awry



We run reports to summarize SPEW synthetic ecosystems



Synthetic Ecosystem for ITALY

SPEW: Synthetic Populations and Ecosystems of the World

Basic Information

- **Total Synthetic Persons:** 59,804,024
  - Characteristics
- **Total Synthetic Households:** 23,212,073
  - Characteristics
- Number of Lowest Level Sub-regions: 19 ('state' equivalent)
- Type: 'U.S. state' equivalent

What is a Synthetic Ecosystem?

A synthetic ecosystem is a digital representation of the world. Synthetic ecosystems include both agents (individuals who interact with one another) and their environment (loci of interaction of the agents). Synthetic ecosystems are generated to be adequately representative of the real world and hope to achieve realism in population characteristics such as race, age, income, school assignments, and more.

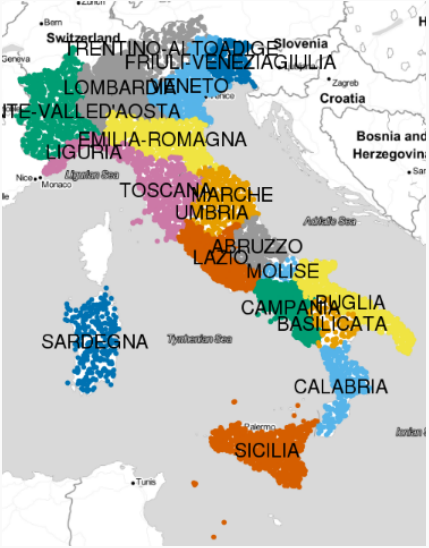
How Does SPEW Generate Synthetic Ecosystems?

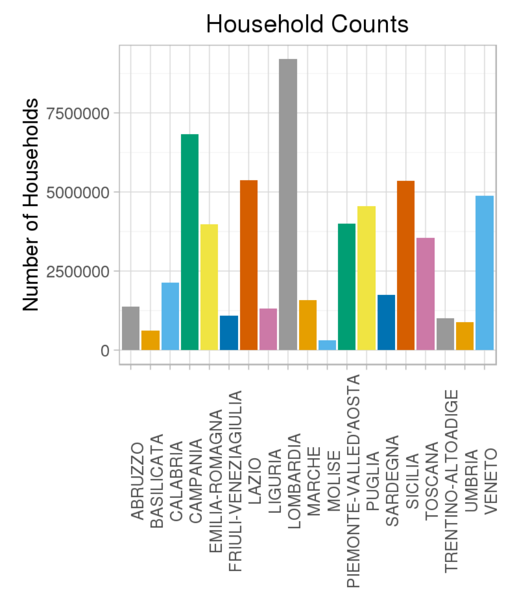
SPEW incorporates three essential input data sources

1. Population Totals (counts)
2. Geography (shapefiles)
3. Microdata (data on individual persons)

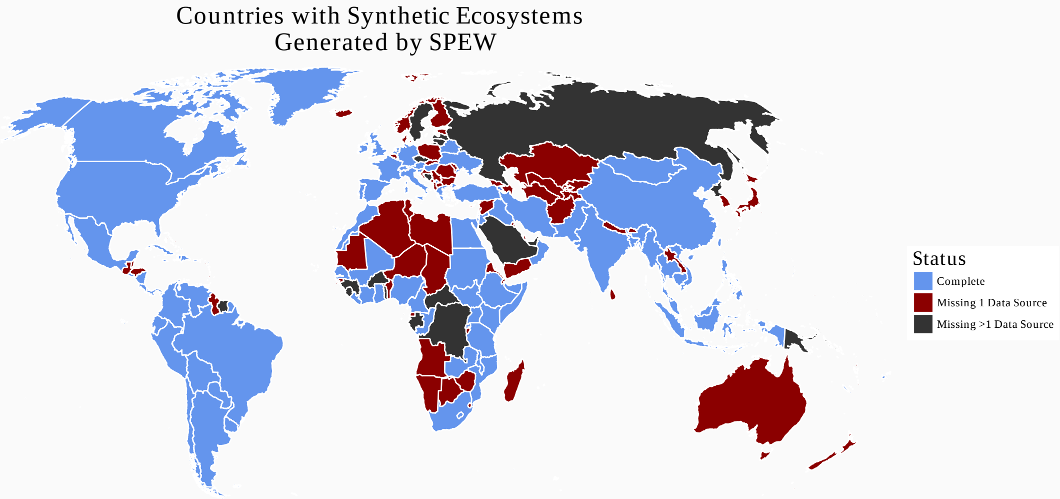
along with *supplementary input data* such as school and workplace information along with sampling methodology for

Maps and summary graphics are displayed





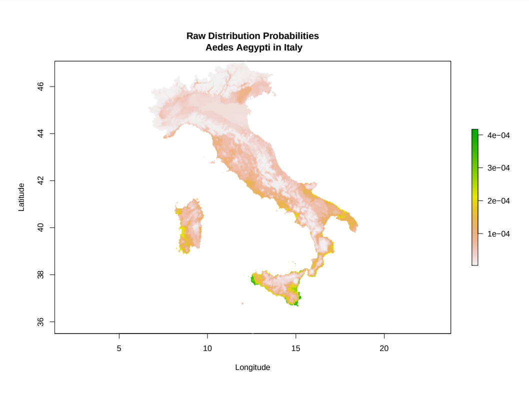
You can download SPEW synthetic ecosystems online



Download links available at shiny.stat.cmu.edu:3838/sgallagh/spew_dl

Current Work

- Incorporation of disease vectors
- Direct synchronization with ABMs
- Inclusion of data-driven interactions
- More and better data!



Acknowledgments

- MIDAS Informatics Systems Group
  - Grant 1 U24 GM110707-01 NIH/NIGMS v
- Pittsburgh Supercomputing Center
- International Conference on Synthetic Populations

Thank you.

Questions?

Resources

- stat.cmu.edu/~spew – main site
- data.olympus.psc.edu/syneco/spew_1.2.0/ – repo of completed synthetic ecosystems
- stat.cmu.edu/~spew/assets/spew_documentation.pdf – full documentation
- github.com/leerichardson/spew – github repo; coming soon to `CRAN`
- epimodels.org – Informatics Systems Group - MIDAS