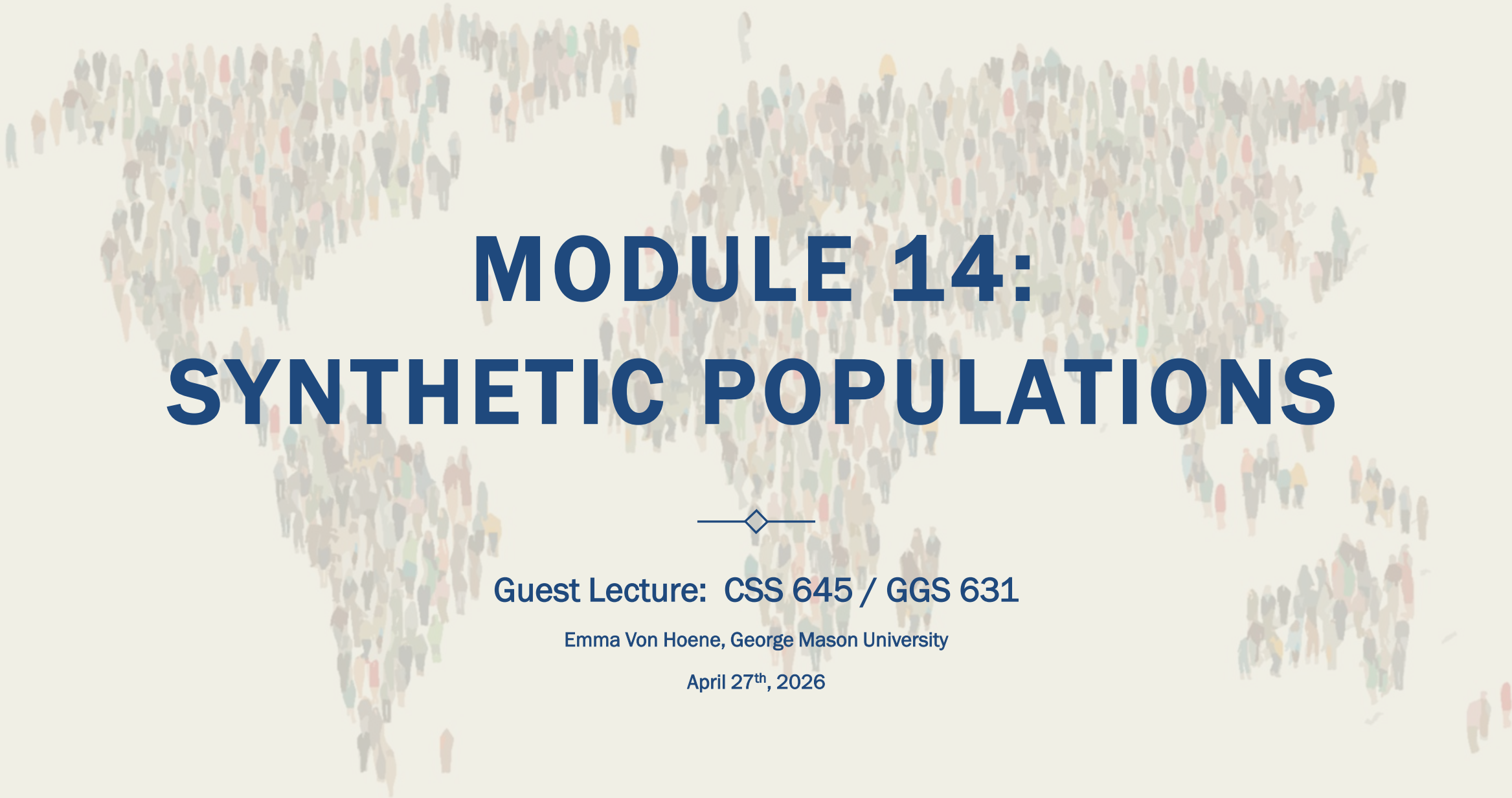


MODULE 14: SYNTHETIC POPULATIONS

Guest Lecture: CSS 645 / GGS 631

Emma Von Hoene, George Mason University

April 27th, 2026

Lecture Outline

- Introduction to Synthetic Populations
  - Motivation
  - Defining Synthetic Populations
  - Data and Methods
  - Applications
- Case Study: Synthetic Population Generation for Public Health
  - Motivation
  - Data and Methods
  - Results and Validation

Learning Objectives

- **Define** synthetic populations and explain their role in ABMs
- **Describe** the data sources used in population synthesis
- **Compare** major population synthesis approaches
- **Explain** the basic workflow of IPF for generating geographically explicit synthetic populations
- **Identify** common software tools and programming environments used in synthetic population generation
- **Evaluate** how synthetic populations can be applied to real-world applications
- **Interpret** a public health case study that integrates health behaviors and attitudes into a synthetic population

Motivation for Synthetic Populations

The Need



- Complex social systems require detailed, data-driven models at large-scales
- Models require real populations at the individual level for given purposes (e.g., forecasting)

The Challenge



- True individual-level data are:
  - Limited or incomplete
  - Difficult to access
  - Restricted due to privacy

The Solution

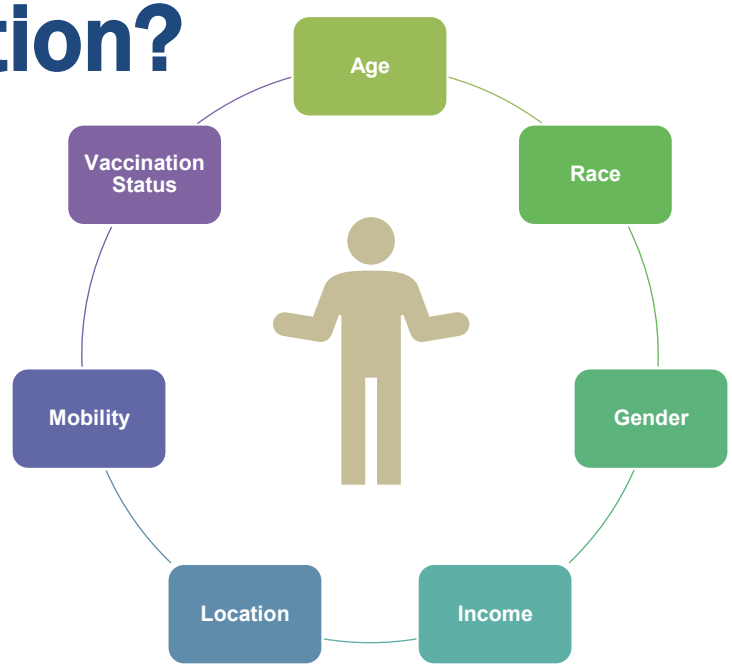


- **Synthetic populations** create realistic, individual-level representations of populations
  - Provides a foundation for agents, populations, and networks in agent-based models (ABMs)

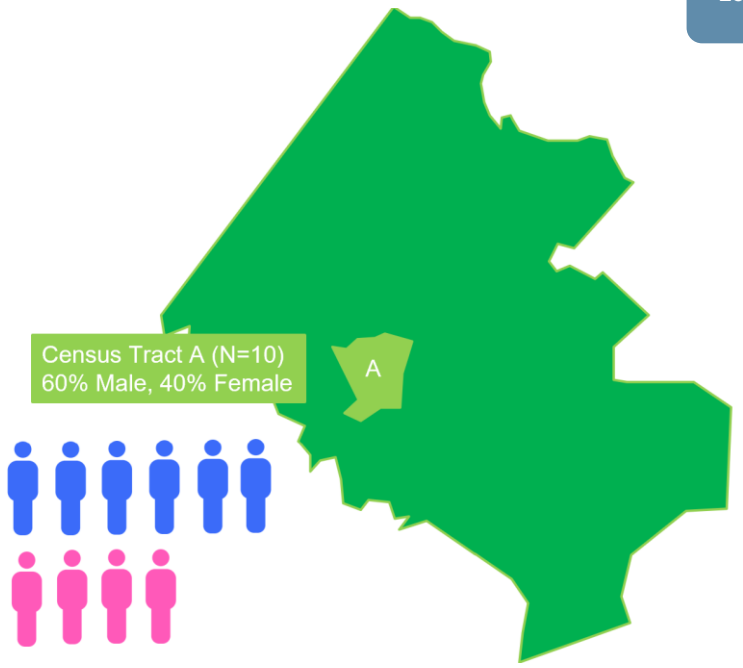
What is a Synthetic Population?

- **Synthetic agent:** represents an individual

- **Synthetic population:** A set of synthetic agents
  - Designed to match real-world statistical patterns
  - Each ‘agent’ has a realistic combination of attributes *(e.g., age, income, location)*
  - Agents may share common geographic, social, or demographic characteristics *(e.g., rural vs. urban populations, communities defined by region or demographic group)*



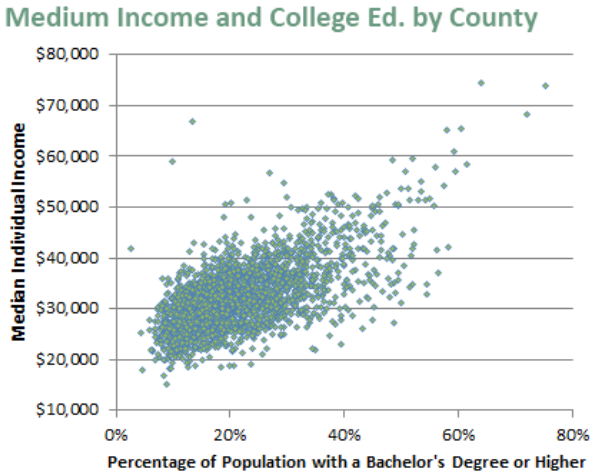
Conceptual representation of a synthetic agent.



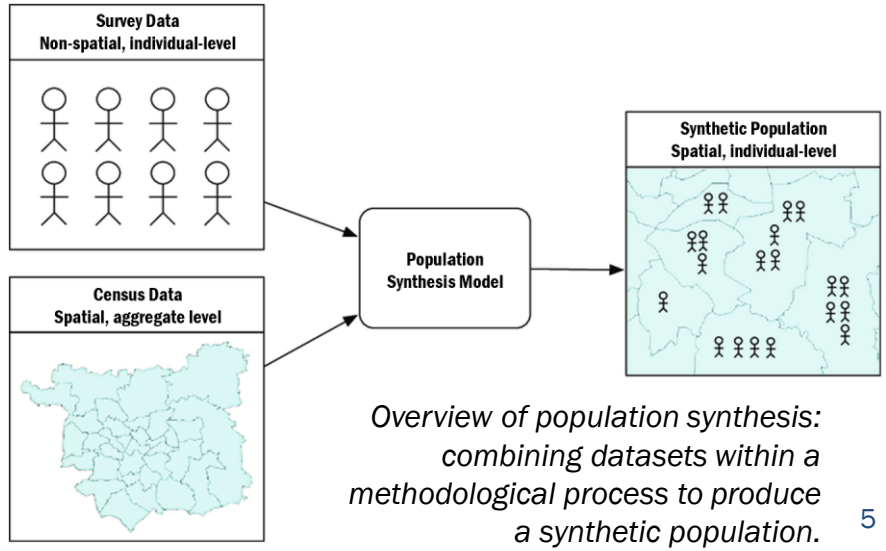
Subset of a synthetic population within a Fairfax County census tract, constructed to match the census gender distribution.

Properties of Synthetic Populations

- Attributes are based on real-world data sources
  - Preserve statistical relationships in the data
- Protect individual privacy (no real person is represented)
  - Synthetic agents are realistic but not real individuals
- Constructed by integrating data using various synthesis methods



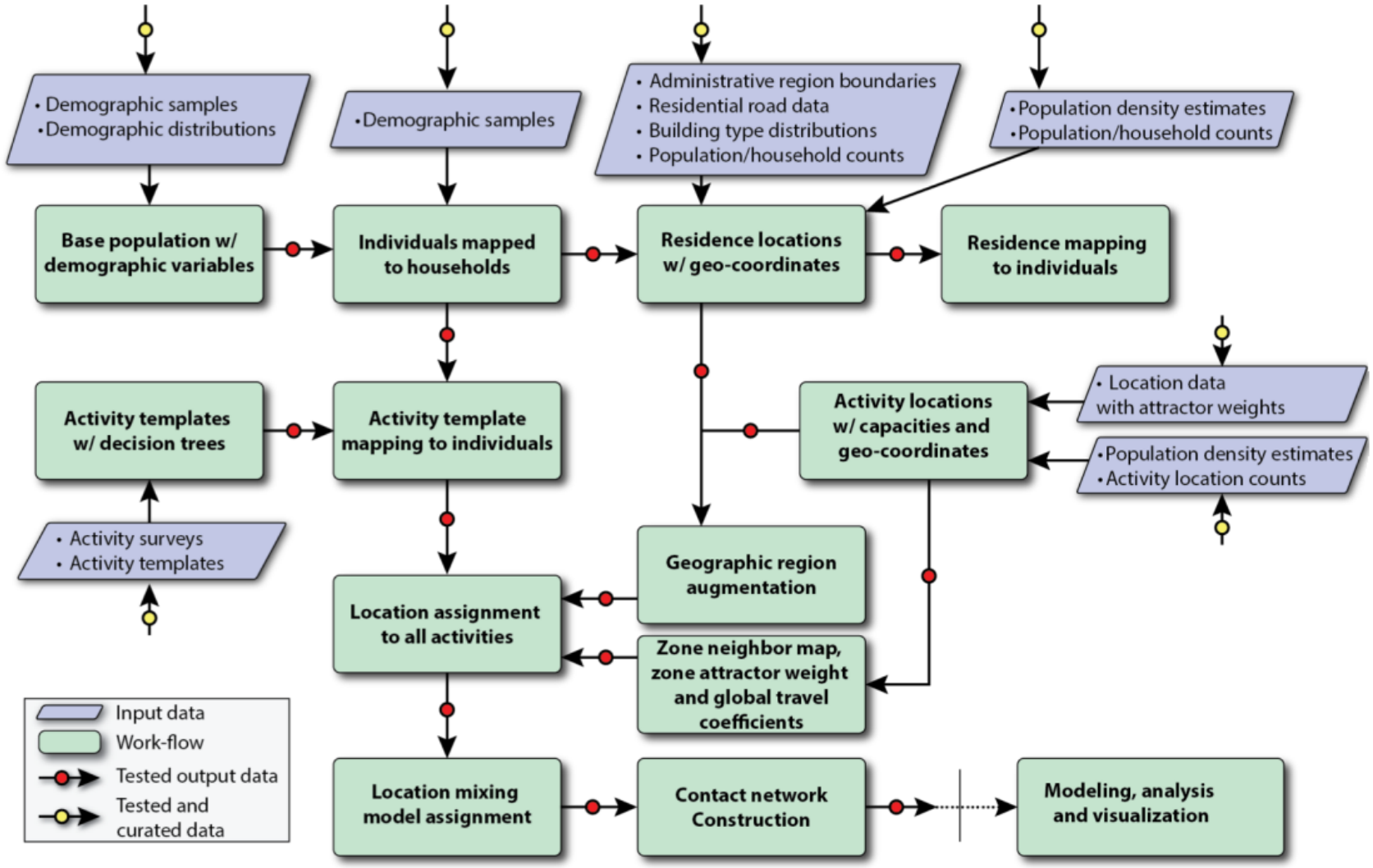
Correlation between median individual income and educational attainment → this relationship should hold true in synthetic populations.



Overview of population synthesis: combining datasets within a methodological process to produce a synthetic population.

Layers of Synthetic Population Modeling

- Population synthesis = who people are
- Activity assignment = what they do
- Location assignment = where they do it
- Network construction = who interacts with who



A full, systematic workflow example for generating a synthetic population and linking individuals to activities, locations, and contact networks for agent-based modeling.

Data Sources for Population Synthesis

MICRODATA

Individual-level data

Data about individuals and households from surveys or administrative records.

What it provides

- Joint distributions and relationships between characteristics
- Realistic combinations of attributes (e.g., age, income, race, vaccine attitude)
- Captures heterogeneity and diversity at the individual level

AGGREGATED DATA

Population-level summaries

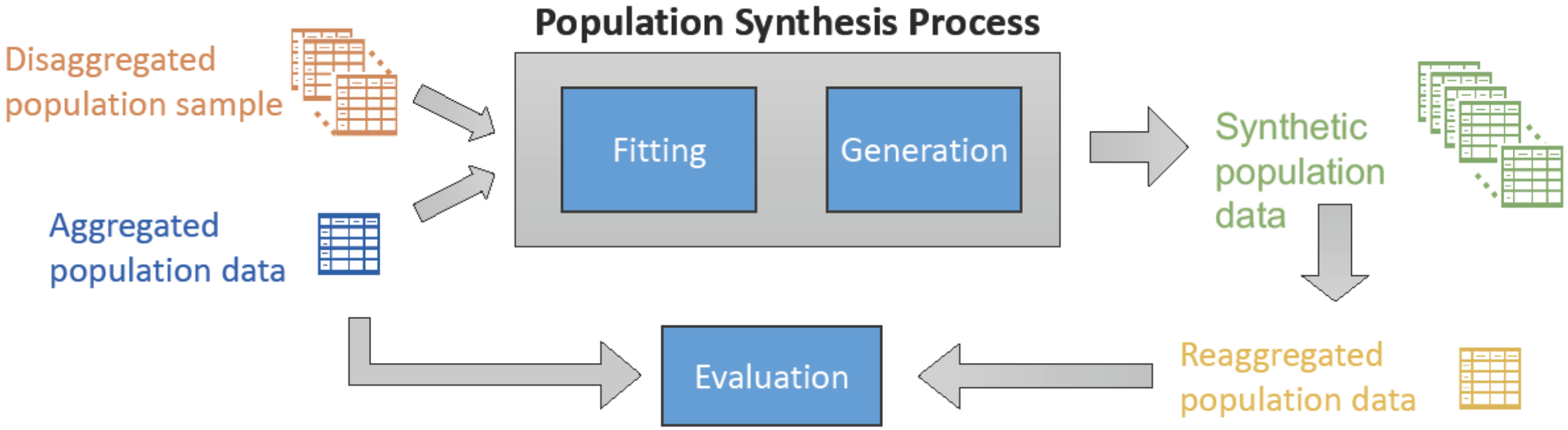
Summary statistics that describe the population in specific geographies or groups.

What it provides

- Population constraints and totals to be matched
- Distributions of key characteristics by geography or group
- Ensures the synthetic population reflects the known population structure

Data completeness, quality and compatibility are critical for reliable synthetic populations!

Synthetic Population Generation Process



High-level synthetic population generation process for any specific method. First, the two input datasets are fed into a fitting algorithm that captures the distribution of individuals' characteristics. Next, individuals are 'allocated' or replicated in the generation step. Data can be aggregated for evaluation.

An Abstract Example of Fitting in Population Synthesis

Example of 'marginal' data from census sources.

The number of Households with household size in given ranges							
Hsize	1	2	3	>3			
n	0	121	214	25			
The number of Households with age in given ranges							
Age	15-24	25-34	35-44	45-54	55-64	65-74	>74
n	44	134	94	46	46	36	0

For census tract 1, block group 2 of Los Alamos county, NM

Example of microdata from PUMS, but only a 5% sample!

	Householder' s age							
Hsize	15-24	25-34	35-44	45-54	55-64	65-74	>74	Total
1	2	11	9	3	26	64	42	157
2	11	108	122	48	80	61	18	448
3	28	135	274	156	85	22	6	706
>3	0	3	65	76	40	10	3	197
Total	41	257	470	283	231	157	69	1508

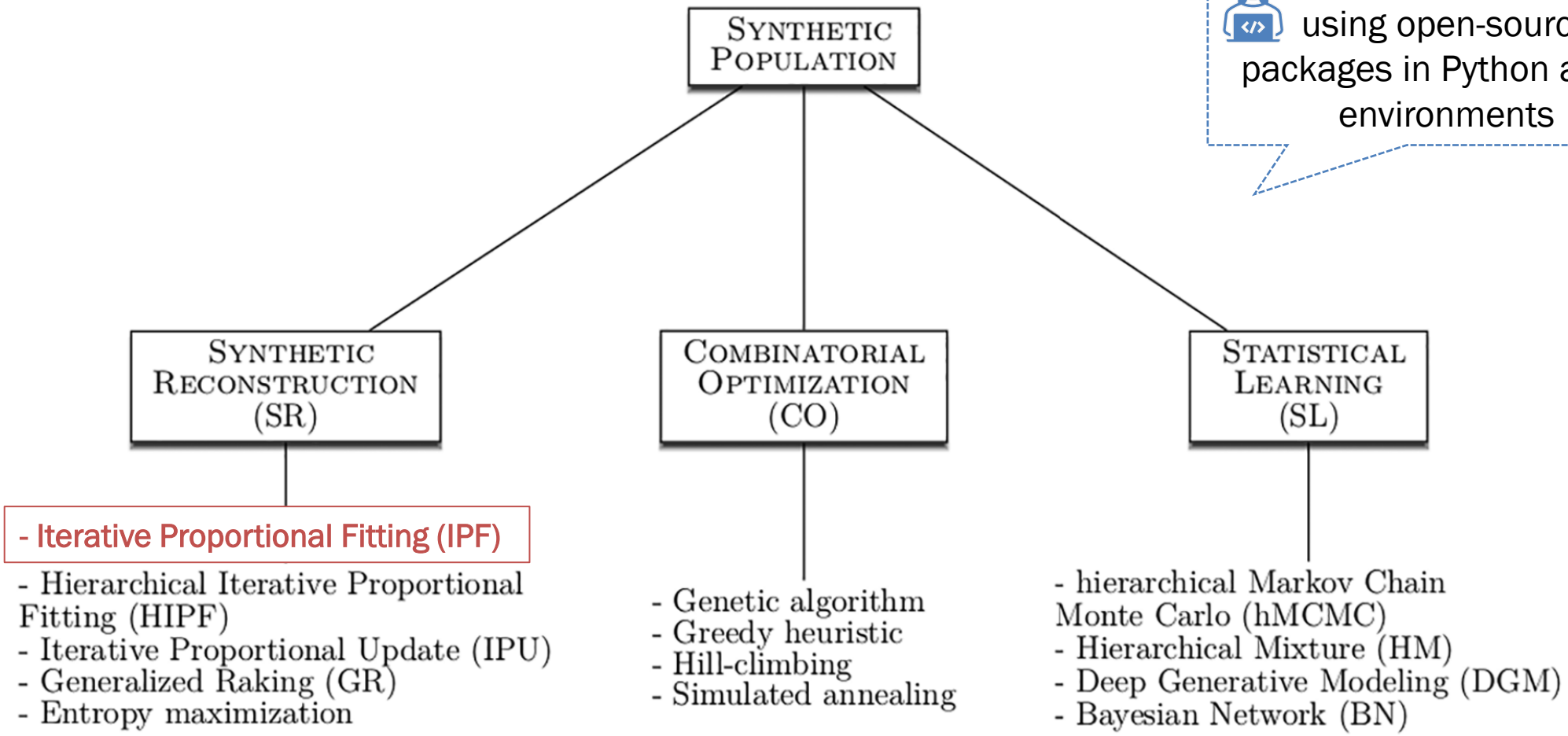
For PUMA containing census tract 1, block group 2 of Los Alamos county, NM

Ideally, what researchers want – similar to PUMS – but for each geographic zone,

	Householder' s age							
Hsize	15-24	25-34	35-44	45-54	55-64	65-74	>74	Total
1	?	?	?	?	?	?	?	0
2	?	?	?	?	?	?	?	121
3	?	?	?	?	?	?	?	214
>3	?	?	?	?	?	?	?	25
Total	44	134	94	46	46	36	0	

Population Synthesis Methods

Can be implemented using open-source packages in Python and R environments

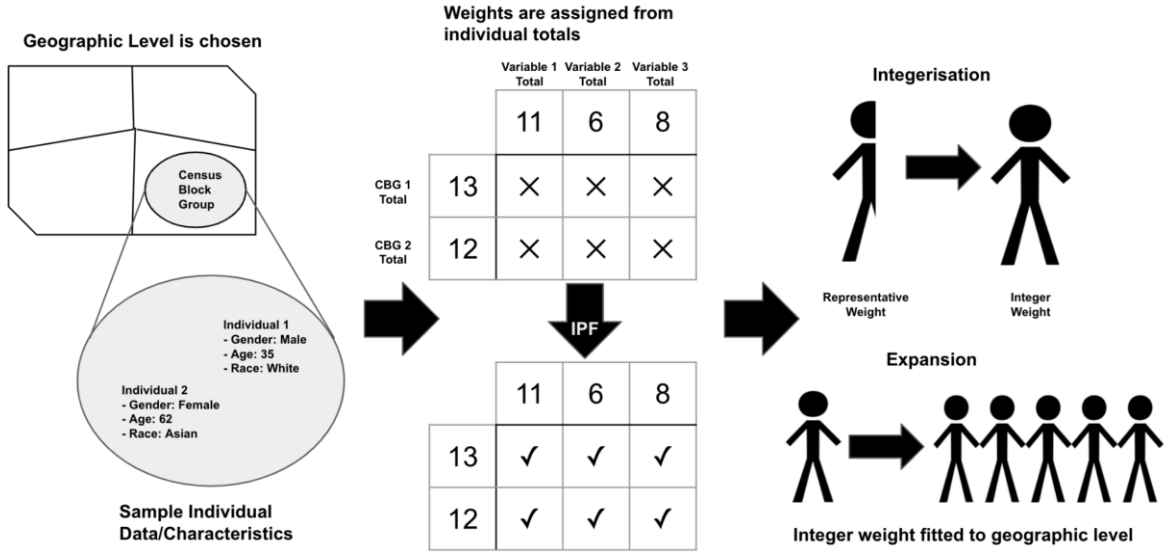


Breakdown of the main population synthesis categories, and specific methods that fall within each.

Population Synthesis Methods:

Synthetic Reconstruction (SR)

- **Main Idea:**
  Reconstruct individuals directly from survey microdata
- **Data Used:**
  Microdata and aggregated constraints
- **Mechanism:**
  Reweight / replicate survey records to match targets
- **Key Assumption:**
  Survey sample is representative enough to scale
- **Output:**
  Reweighted synthetic individuals
- **Strengths:**
  Simple, interpretable, lower computational cost
- **Limitations:**
  Limited to microdata representativeness (generalizability)

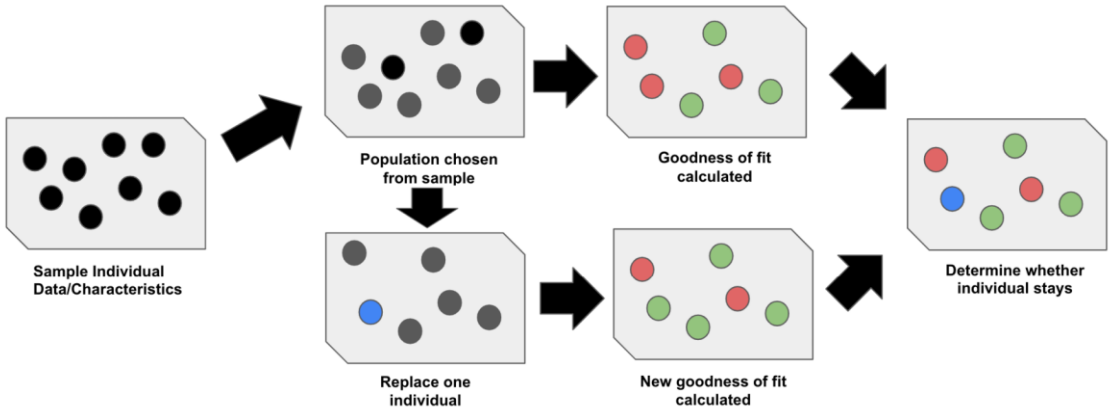


Representation of a SR method: Iterative Proportional Fitting (IPF).

From: D. Han, S. Islam, T. Anderson, A. T. Crooks and H. Kavak, "Quantitative Comparison of Population Synthesis Techniques," 2025 Winter Simulation Conference (WSC), Seattle, WA, USA, 2025, pp. 151-162, doi: 10.1109/WSC68292.2025.11338945.

Population Synthesis Methods:
Combinatorial Optimization (CO)

- **Main Idea:**
  Select best-fitting sample from microdata to match constraints
- **Data Used:**
  Microdata and aggregated constraints
- **Mechanism:**
  Iteratively optimize sample selection to match marginals
- **Key Assumption:**
  A good subset exists within microdata
- **Output:**
  Selected synthetic population from observed records
- **Strengths:**
  High constraint accuracy
- **Limitations:**
  Computationally intensive, limited by microdata

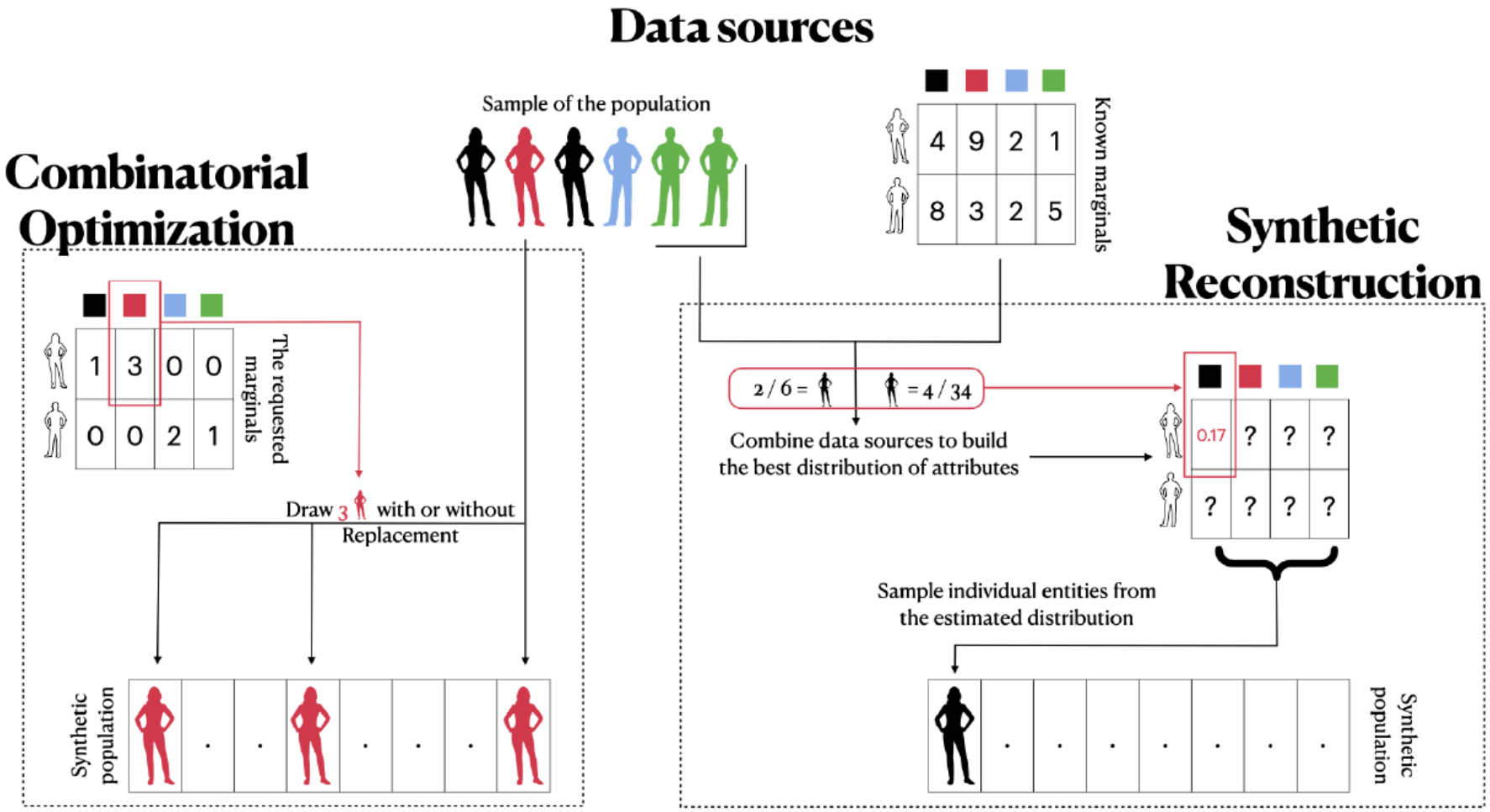


Representation of a CO algorithm: Hill Climbing (HC).

From: D. Han, S. Islam, T. Anderson, A. T. Crooks and H. Kavak, "Quantitative Comparison of Population Synthesis Techniques," 2025 Winter Simulation Conference (WSC), Seattle, WA, USA, 2025, pp. 151-162, doi: 10.1109/WSC68292.2025.11338945.

Population Synthesis Methods:

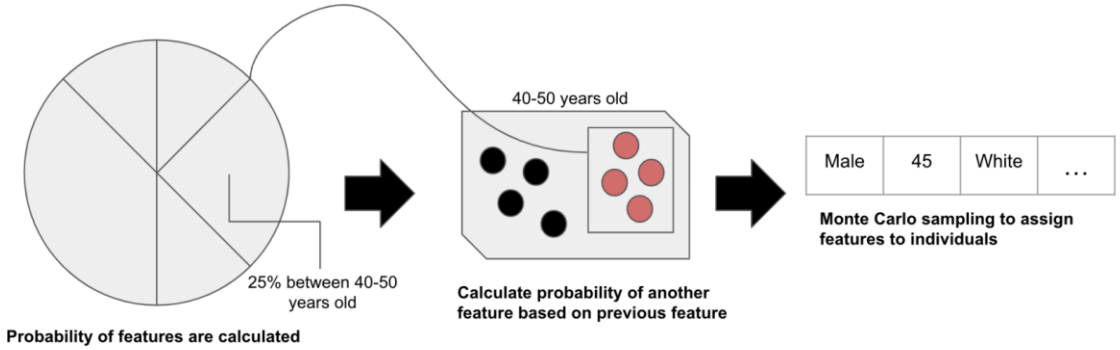
SR vs CO



Difference between CO and SR population synthesis methods.

Population Synthesis Methods:
Statistical Learning (SL)

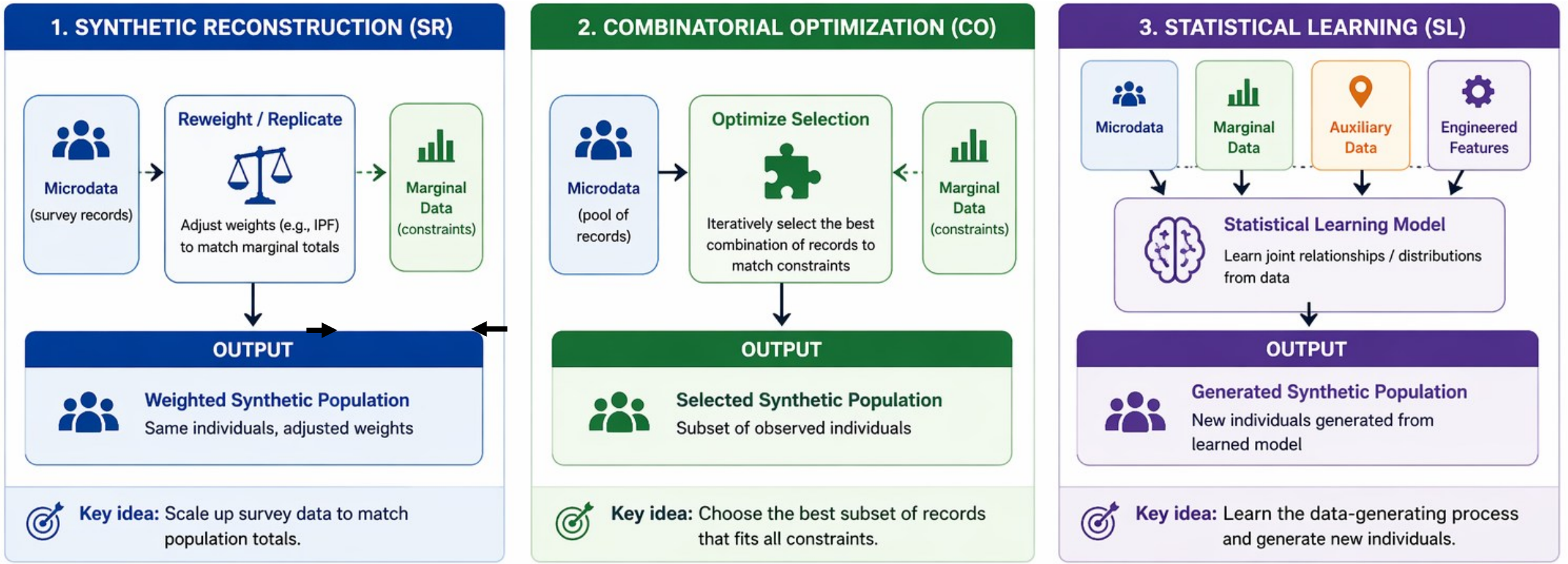
- **Main Idea:**
  Learn statistical relationships from data to generate agents
- **Data Used:**
  Microdata, aggregated constraints, external sources
- **Mechanism:**
  Uses probabilistic/ML models to learn joint distributions
- **Key Assumption:**
  Data distribution can be learned and generalized
- **Output:**
  Newly generated synthetic individuals
- **Strengths:**
  Can generate new combinations, captures complex relationships (flexible and generalizable)
- **Limitations:**
  Requires model training, less transparent



Representation of a SL algorithm: Conditional Probabilities (CP).

From: D. Han, S. Islam, T. Anderson, A. T. Crooks and H. Kavak, "Quantitative Comparison of Population Synthesis Techniques," 2025 Winter Simulation Conference (WSC), Seattle, WA, USA, 2025, pp. 151-162, doi: 10.1109/WSC68292.2025.11338945.

Population Synthesis Methods



AT A GLANCE

SR: Reweight what exists
Uses data directly to scale records.

CO: Select what fits
Finds the best combination from existing records.

SL: Learn and generate
Learns patterns and generates new records.

Schematic overview comparing the different groups of population synthesis methods.

Applications of Synthetic Population Generation

1. Transportation Modeling & Planning



- **Travel Demand:** Forecast traffic by simulating daily activity patterns (work, school, shopping) of individual agents.
- **Future Mobility:** Simulate the impact of autonomous vehicles or ride-hailing services on existing urban congestion.

2. Epidemiology & Public Health



- **Disease Transmission:** Trace how infectious diseases spread through a community by modeling interactions within realistic households and workplaces.
- **Intervention Testing:** Evaluate the potential effectiveness of policies like school closures or targeted vaccine distribution.

3. Urban Planning & Land Use



- **Demand Forecasting:** Estimate the need for schools, parks, and utilities based on projected localized growth.
- **Zoning Impacts:** Test how changes in housing policy or land-use might affect different types of households.

4. Disaster Management & Emergency Response



- **Evacuation Planning:** Estimate the number and specific needs of households in high-risk zones (e.g., flood/wildfire areas).
- **Recovery Prediction:** Forecast long-term resident displacement and socioeconomic recovery patterns following a disaster.

5. Market Research & Policy Analysis



- **Economic Impact:** Test how new tax laws or minimum wage changes might affect diverse household structures.
- **Hyper-Local Targeting:** Perform granular market segmentation and demand forecasting for specific products or services.

Case Study: Population Synthesis for Public Health

Study Overview:

- A novel application of population synthesis to initialize agents with health characteristics
- Extends Iterative Proportional Fitting (IPF)
- Completed using R: *https://github.com/evonhoene/Population-Generation-for-Public-Health-ABMs*



PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Synthetic population generation with public health characteristics for spatial agent-based models

Emma Von Hoene[1*], Amira Roess[2], Hamdi Kavak[3], Taylor Anderson[1]

1 Department of Geography and Geoinformation Science, George Mason University, Fairfax, Virginia, United States of America, **2** Department of Global and Community Health, George Mason University, Fairfax, Virginia, United States of America, **3** Department of Computational and Data Sciences, George Mason University, Fairfax, Virginia, United States of America

* evonhoen@gmu.edu

Check for updates

OPEN ACCESS

Citation: Von Hoene E, Roess A, Kavak H, Anderson T (2025) Synthetic population generation with public health characteristics for spatial agent-based models. PLoS Comput Biol 21(3): e1012439. https://doi.org/10.1371/journal.pcbi.1012439

Editor: Nicholas Geard, University of Melbourne, AUSTRALIA

Received: August 26, 2024

Accepted: February 10, 2025

et al. This is an under the terms ution License, , distribution, n, provided the redited.

he individual- urvey collected ataset, and the

Abstract

Agent-based models (ABMs) simulate the behaviors, interactions, and disease transmission between individual "agents" within their environment, enabling the investigation of the underlying processes driving disease dynamics and how these processes may be influenced by policy interventions. Despite the critical role that characteristics such as health attitudes and vaccination status play in disease outcomes, the initialization of agent populations with these variables is often oversimplified, overlooking statistical relationships between attitudes and other characteristics or lacking spatial heterogeneity. Leveraging population synthesis methods to create populations with realistic health attitudes and protective behaviors for spatial ABMs has yet to be fully explored. Therefore, this study introduces a novel application for generating synthetic populations with protective behaviors and associated attitudes using public health surveys instead of traditional individual-level survey datasets from the census. We test our approach using two different public health surveys to create two synthetic populations representing individuals aged 18 and over in Virginia, U.S., and their COVID-19 vaccine attitudes and uptake as of December 2021. Results show that integrating public health surveys into synthetic population generation processes preserves the statistical relationships between vaccine uptake and attitudes in different demographic groups while capturing spatial heterogeneity at fine scales. This approach can support disease simulations that aim to explore how real populations might respond to interventions and how these responses may lead to demographic or geographic health disparities. Our study also demonstrates the potential for initializing agents with variables relevant to public health domains that extend beyond infectious diseases, ultimately advancing data-driven ABMs for geographically targeted decision-making.

What characteristics should agents be assigned in health simulations?

Von Hoene et al. (2025), PLOS Computational Biology: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012439*

Previous Research: ABM of COVID-19 and Mask-Use

A Framework for Simulating Emergent Health Behaviors in Spatial Agent-Based Models of Disease Spread

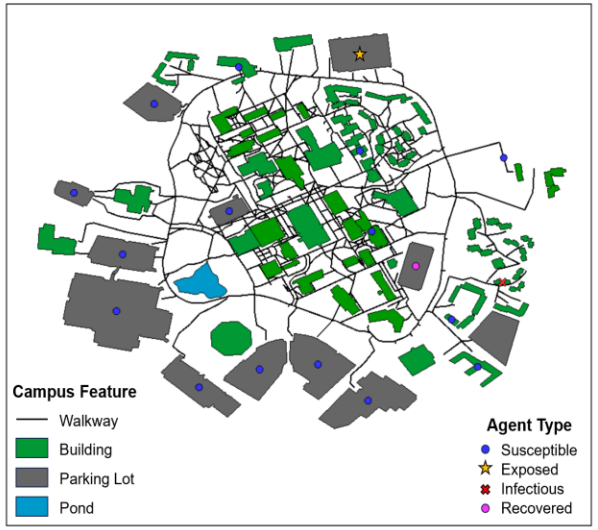
Emma Von Hoene
evonhoen@gmu.edu
Department of Geography and Geoinformation Science
George Mason University
Fairfax, Virginia, USA

Amira Roess
aroess@gmu.edu
Department of Global and Community Health
George Mason University
Fairfax, Virginia, USA

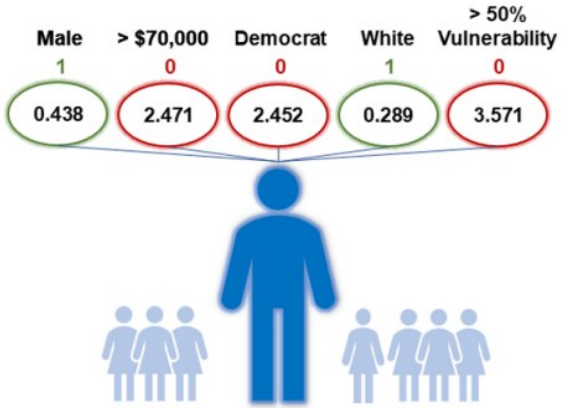
Shivani Achuthan
James Madison High School
Vienna, Virginia, USA
shivani@achuthans.com

Taylor Anderson
tander6@gmu.edu
Department of Geography and Geoinformation Science
George Mason University
Fairfax, Virginia, USA

Von Hoene et al. (2024), SIGSPATIAL International Workshop on GeoSim:
https://dl.acm.org/doi/abs/10.1145/3615891.3628010

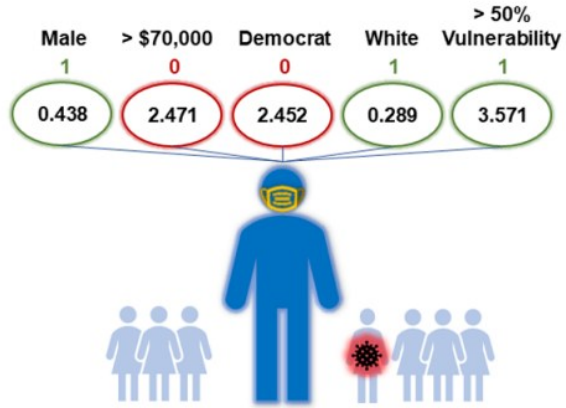


Environment of geospatial ABM simulating COVID-19 spread and mask behavior.



$$p_t(mask) = \frac{4.540 * (0.438^1 * 2.471^0 * 2.452^0 * 0.289^1 * 3.571^0)}{1 + 4.540 * (0.438^1 * 2.471^0 * 2.452^0 * 0.289^1 * 3.571^0)} = \mathbf{0.365}$$

There is a 36.5% chance I wear a mask!



$$p_{t+n}(mask) = \frac{4.540 * (0.438^1 * 2.471^0 * 2.452^0 * 0.289^1 * 3.571^1)}{1 + 4.540 * (0.438^1 * 2.471^0 * 2.452^0 * 0.289^1 * 3.571^1)} = \mathbf{0.672}$$

There is a 67.2% chance I wear a mask!

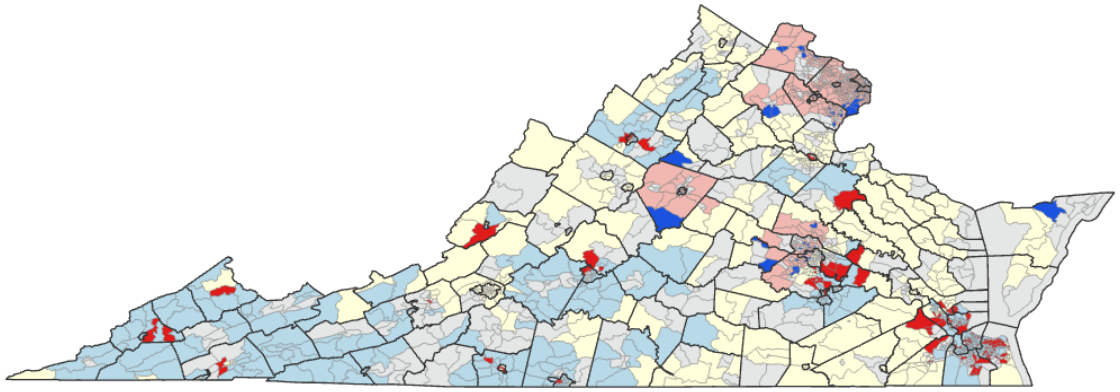
How can we realistically initialize agents with characteristics (other than demographics) that strongly influence their decision to adopt a health behavior (e.g., vaccines)?

Representation of the decision-making framework using odds ratios corresponding to agent attributes in a logistic regression model to predict dynamic mask-use.

Research Objectives

- **Research Question:** How can the **direct integration of survey data** into existing synthetic population generation approaches realistically reflect domain-specific characteristics, such as vaccine status and associated attitudes, and their spatial patterns in the agent population?

- **Research Goals:** To develop and validate a population synthesis approach for initializing agents with behavioral and related attitude or perceptions characteristics, specifically aiming to:
  - Capture the true **spatial heterogeneity** of such attributes
  - Preserve the **statistical relationship** of such attributes across demographic groups

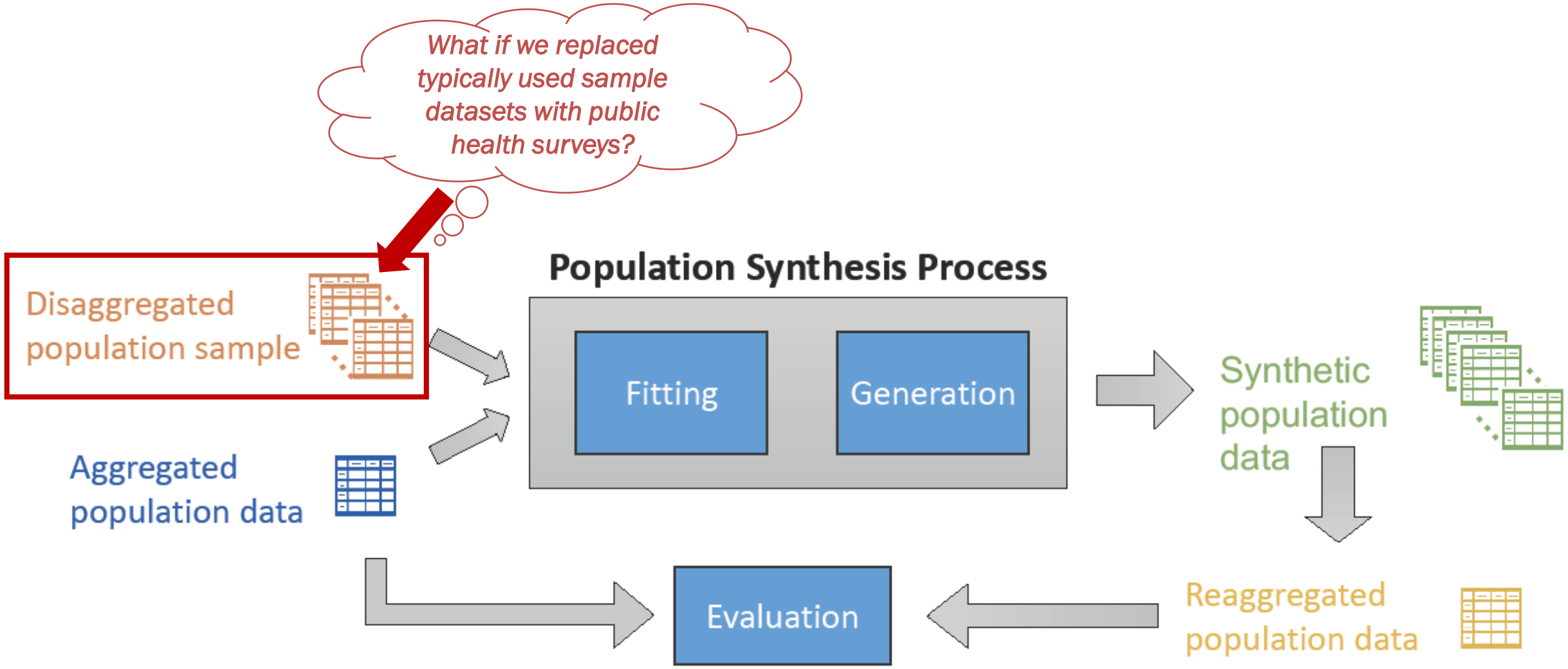


Clusters and outliers of observed vaccine uptake in December 2021.

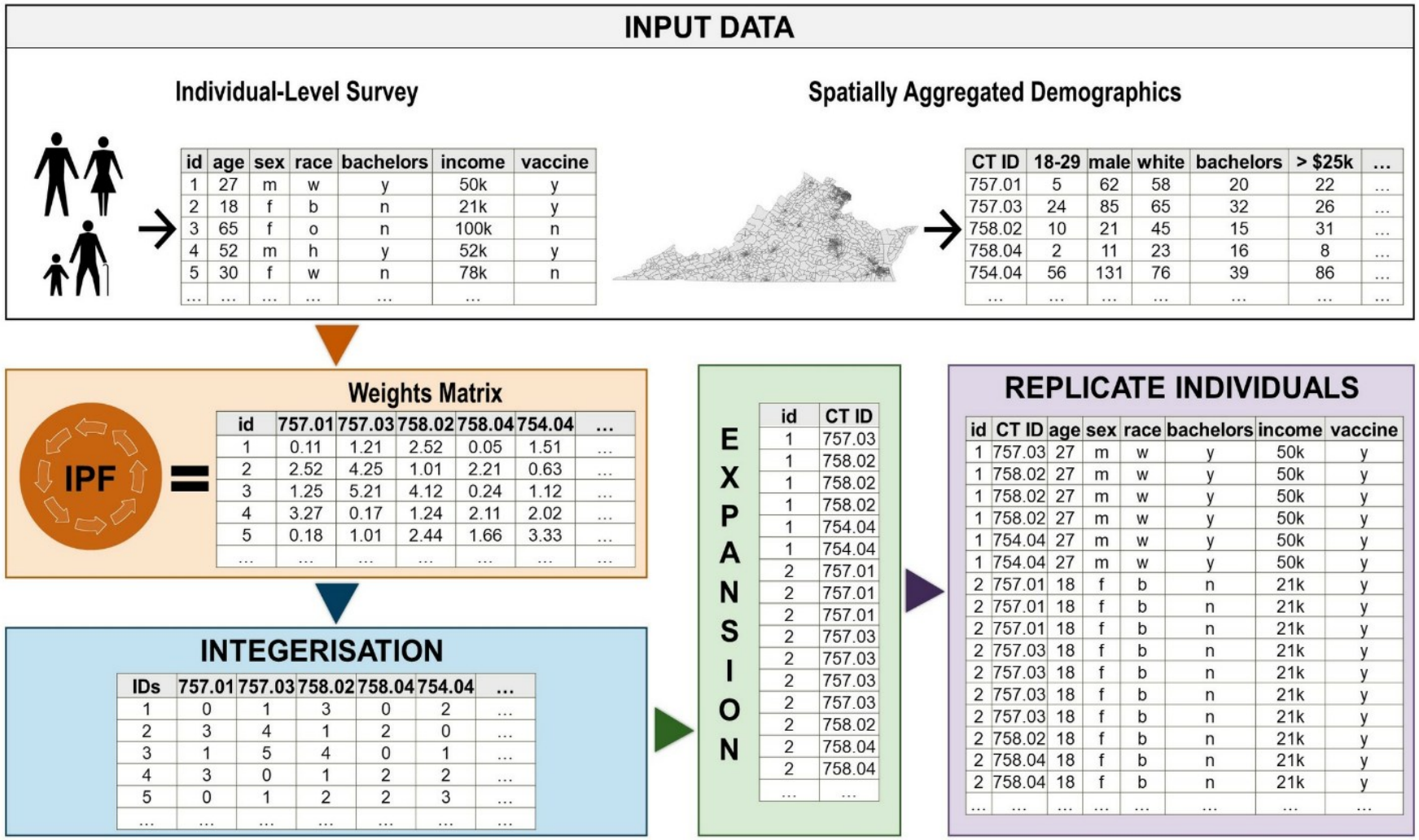
Relationships between demographic groups and attitudes towards COVID-19 vaccine intention [UAS 282, February – March 2021].

Independent Variable: Descriptor	Dependent Variable	Odds Ratio
Education: Bachelor's Degree	Positive Attitude	3.214
Gender: Male	Positive Attitude	1.855
Age: 65+	Positive Attitude	1.583
Income: Less than $25,000	Positive Attitude	0.443
Race: Black	Positive Attitude	0.410

Adaptation of the Population Synthesis Process



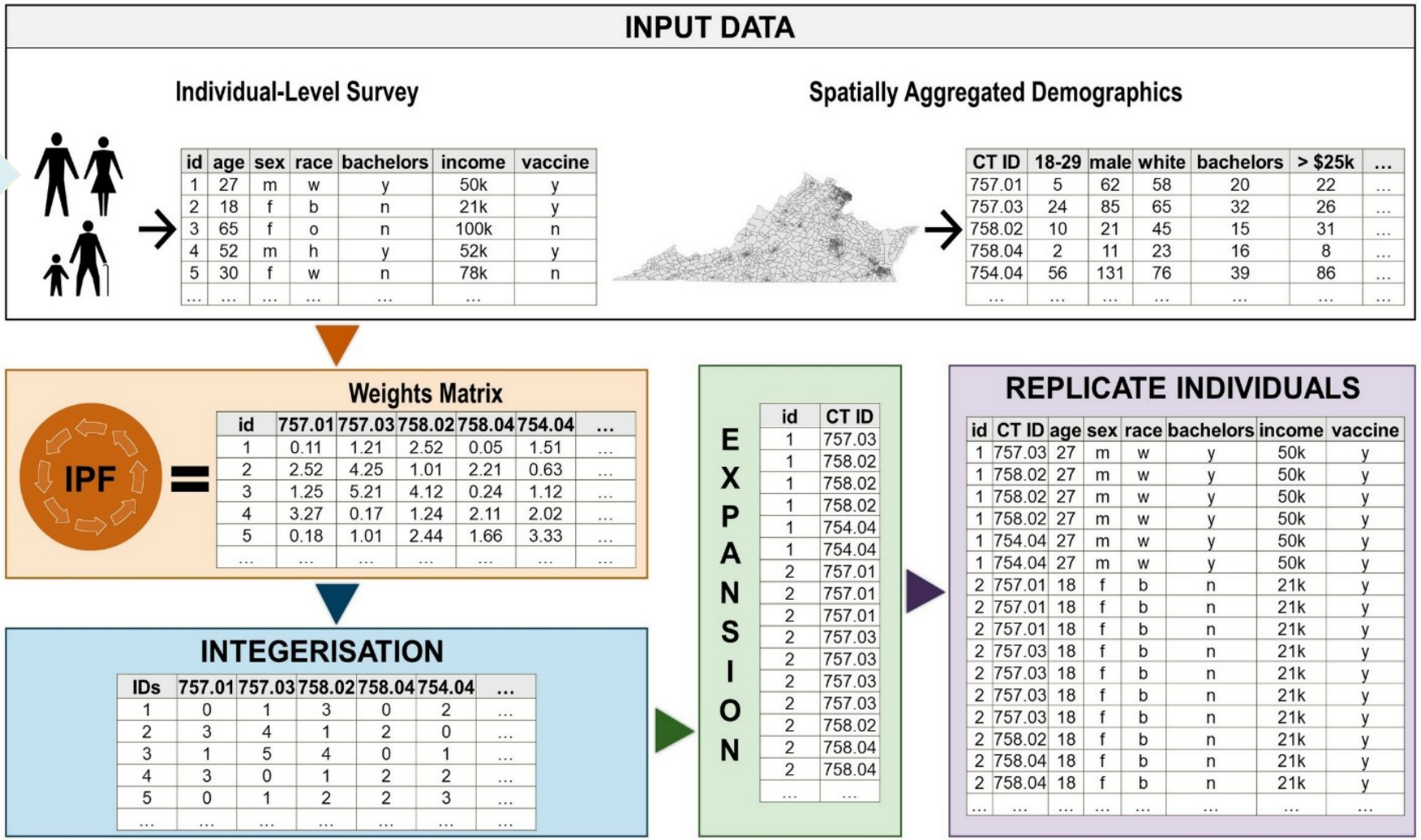
Proposed IPF-based Workflow



Overview of the direct survey-based synthetic population generation approach for initializing agents with health attributes.

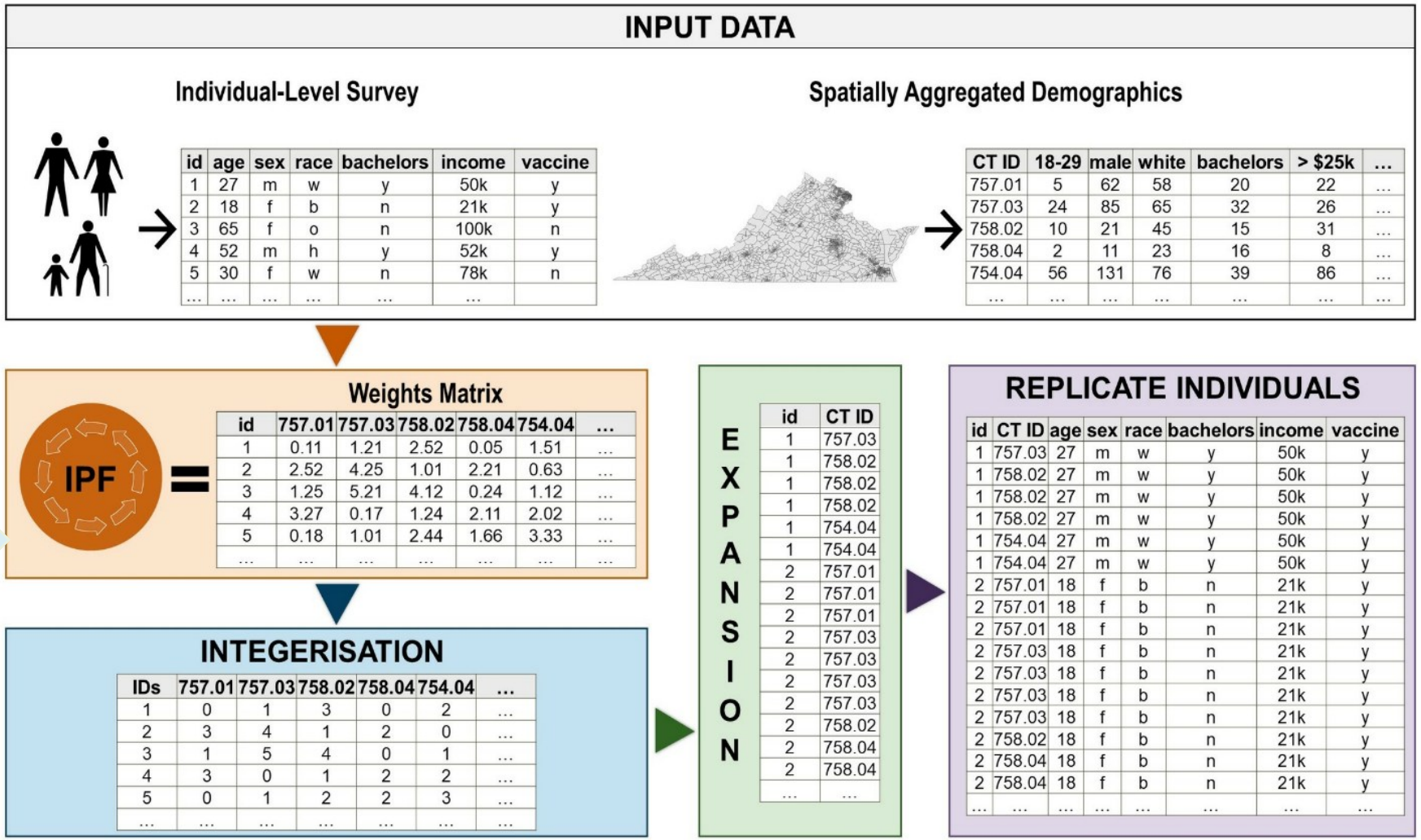
Stage 1: Input Data

- Focused on gender, race, age, education and income for individuals aged 18+
- **Spatially aggregate demographic data**
  - US Census Bureau, census tract level
- **Individual-level survey data**
  - HPS→ Household pulse survey (national, but not nationally representative)
  - NRS→ Nationally representative survey, collected by Dr. Amira Roess



Stage 2: IPF

- IPF assigns a weight to each individual from the survey in each geographic zone (in this case, census tracts) to match the overall population distribution.
- These weights represent how much each individual contributes to a census tract based on their demographic characteristics.
- The goal is to adjust the weights so that the final distribution of individuals across census tracts aligns with the observed or expected geographic data.



id	age	sex	race	bachelors	income	vaccine
1	27	m	w	y	50k	y
2	18	f	b	n	21k	y
3	65	f	o	n	100k	n
4	52	m	h	y	52k	y
5	30	f	w	n	78k	n
...	...	...	...	...	...	

CT ID	18-29	male	white	bachelors	> $25k	...
757.01	5	62	58	20	22	...
757.03	24	85	65	32	26	...
758.02	10	21	45	15	31	...
758.04	2	11	23	16	8	...
754.04	56	131	76	39	86	...
...	...	...	...	...	...	...

id	757.01	757.03	758.02	758.04	754.04	...
1	0.11	1.21	2.52	0.05	1.51	...
2	2.52	4.25	1.01	2.21	0.63	...
3	1.25	5.21	4.12	0.24	1.12	...
4	3.27	0.17	1.24	2.11	2.02	...
5	0.18	1.01	2.44	1.66	3.33	...
...	...	...	...	...	...	...

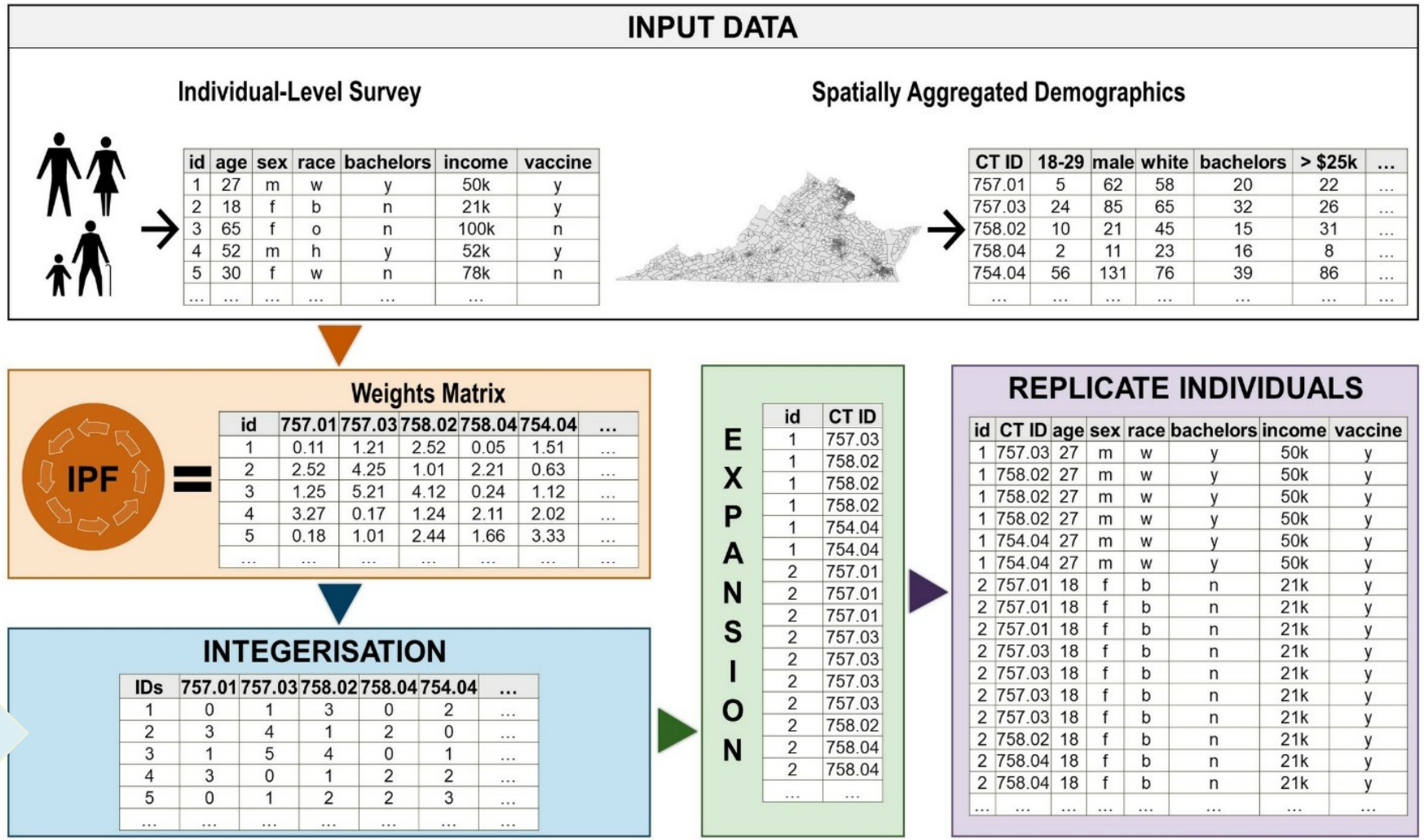
IDs	757.01	757.03	758.02	758.04	754.04	...
1	0	1	3	0	2	...
2	3	4	1	2	0	...
3	1	5	4	0	1	...
4	3	0	1	2	2	...
5	0	1	2	2	3	...
...	...	...	...	...	...	...

id	CT ID
1	757.03
1	758.02
1	758.02
1	758.02
1	754.04
1	754.04
2	757.01
2	757.01
2	757.01
2	757.03
2	757.03
2	757.03
2	757.03
2	758.02
2	758.04
2	758.04
...	...

id	CT ID	age	sex	race	bachelors	income	vaccine
1	757.03	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	754.04	27	m	w	y	50k	y
1	754.04	27	m	w	y	50k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	758.02	18	f	b	n	21k	y
2	758.04	18	f	b	n	21k	y
2	758.04	18	f	b	n	21k	y
...	...	...	...	...	...	...	...

Stage 3: Integerisation

- After IPF, a weight matrix is generated where individuals from the survey are represented as rows and geographic zones (i.e., census tracts) are columns
- Each cell in the matrix contains the weight assigned to the corresponding individual for the respective zone, which requires a *conversion to counts*



id	age	sex	race	bachelors	income	vaccine
1	27	m	w	y	50k	y
2	18	f	b	n	21k	y
3	65	f	o	n	100k	n
4	52	m	h	y	52k	y
5	30	f	w	n	78k	n
...	...	...	...	...	...	

CT ID	18-29	male	white	bachelors	> $25k	...
757.01	5	62	58	20	22	...
757.03	24	85	65	32	26	...
758.02	10	21	45	15	31	...
758.04	2	11	23	16	8	...
754.04	56	131	76	39	86	...
...	...	...	...	...	...	...

id	757.01	757.03	758.02	758.04	754.04	...
1	0.11	1.21	2.52	0.05	1.51	...
2	2.52	4.25	1.01	2.21	0.63	...
3	1.25	5.21	4.12	0.24	1.12	...
4	3.27	0.17	1.24	2.11	2.02	...
5	0.18	1.01	2.44	1.66	3.33	...
...	...	...	...	...	...	...

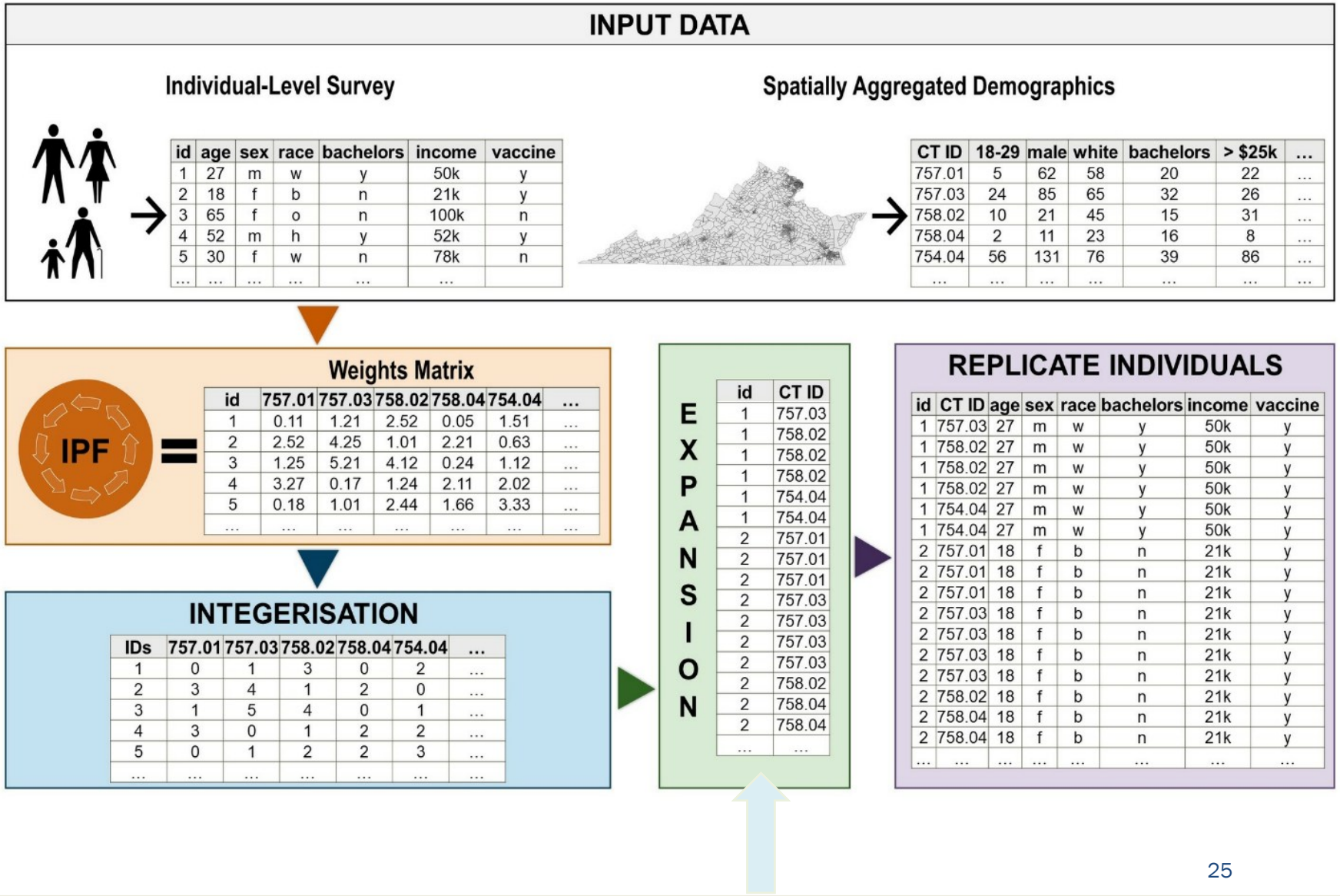
IDs	757.01	757.03	758.02	758.04	754.04	...
1	0	1	3	0	2	...
2	3	4	1	2	0	...
3	1	5	4	0	1	...
4	3	0	1	2	2	...
5	0	1	2	2	3	...
...	...	...	...	...	...	...

id	CT ID
1	757.03
1	758.02
1	758.02
1	758.02
1	754.04
1	754.04
2	757.01
2	757.01
2	757.01
2	757.03
2	757.03
2	757.03
2	757.03
2	758.02
2	758.04
2	758.04
...	...

id	CT ID	age	sex	race	bachelors	income	vaccine
1	757.03	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	754.04	27	m	w	y	50k	y
1	754.04	27	m	w	y	50k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	758.02	18	f	b	n	21k	y
2	758.04	18	f	b	n	21k	y
2	758.04	18	f	b	n	21k	y
...	...	...	...	...	...	...	...

Stage 4: Expansion

- Integerisation produces a matrix representing the number of times each individual needs to be replicated in a given geographic zone (i.e., census tract)
- Expansion is required to 'expand' the data based on the integer counts
- Transformation into a synthetic population output, which contains a row per individual within the whole population and a geographic zone for each person



id	age	sex	race	bachelors	income	vaccine
1	27	m	w	y	50k	y
2	18	f	b	n	21k	y
3	65	f	o	n	100k	n
4	52	m	h	y	52k	y
5	30	f	w	n	78k	n
...	...	...	...	...	...	

CT ID	18-29	male	white	bachelors	> $25k	...
757.01	5	62	58	20	22	...
757.03	24	85	65	32	26	...
758.02	10	21	45	15	31	...
758.04	2	11	23	16	8	...
754.04	56	131	76	39	86	...
...	...	...	...	...	...	...

id	757.01	757.03	758.02	758.04	754.04	...
1	0.11	1.21	2.52	0.05	1.51	...
2	2.52	4.25	1.01	2.21	0.63	...
3	1.25	5.21	4.12	0.24	1.12	...
4	3.27	0.17	1.24	2.11	2.02	...
5	0.18	1.01	2.44	1.66	3.33	...
...	...	...	...	...	...	...

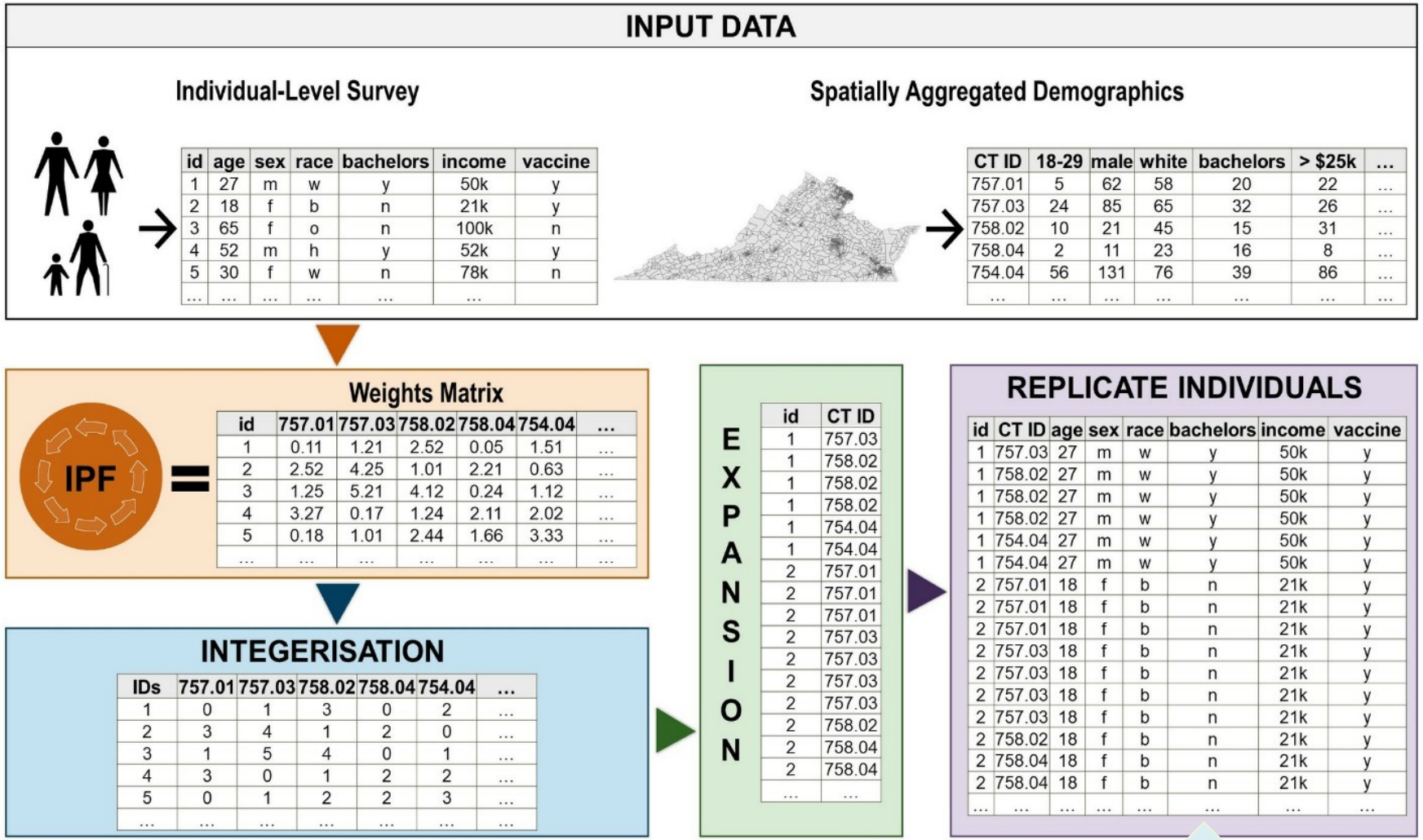
IDs	757.01	757.03	758.02	758.04	754.04	...
1	0	1	3	0	2	...
2	3	4	1	2	0	...
3	1	5	4	0	1	...
4	3	0	1	2	2	...
5	0	1	2	2	3	...
...	...	...	...	...	...	...

id	CT ID
1	757.03
1	758.02
1	758.02
1	758.02
1	754.04
1	754.04
2	757.01
2	757.01
2	757.01
2	757.03
2	757.03
2	757.03
2	757.03
2	758.02
2	758.04
2	758.04
...	...

id	CT ID	age	sex	race	bachelors	income	vaccine
1	757.03	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	754.04	27	m	w	y	50k	y
1	754.04	27	m	w	y	50k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	758.02	18	f	b	n	21k	y
2	758.04	18	f	b	n	21k	y
2	758.04	18	f	b	n	21k	y
...	...	...	...	...	...	...	...

Stage 5: Replication

- Any selected attributes from the survey data can be joined using the survey respondents' ID to the expanded dataset



id	age	sex	race	bachelors	income	vaccine
1	27	m	w	y	50k	y
2	18	f	b	n	21k	y
3	65	f	o	n	100k	n
4	52	m	h	y	52k	y
5	30	f	w	n	78k	n
...	...	...	...	...	...	

CT ID	18-29	male	white	bachelors	> $25k	...
757.01	5	62	58	20	22	...
757.03	24	85	65	32	26	...
758.02	10	21	45	15	31	...
758.04	2	11	23	16	8	...
754.04	56	131	76	39	86	...
...	...	...	...	...	...	...

id	757.01	757.03	758.02	758.04	754.04	...
1	0.11	1.21	2.52	0.05	1.51	...
2	2.52	4.25	1.01	2.21	0.63	...
3	1.25	5.21	4.12	0.24	1.12	...
4	3.27	0.17	1.24	2.11	2.02	...
5	0.18	1.01	2.44	1.66	3.33	...
...	...	...	...	...	...	...

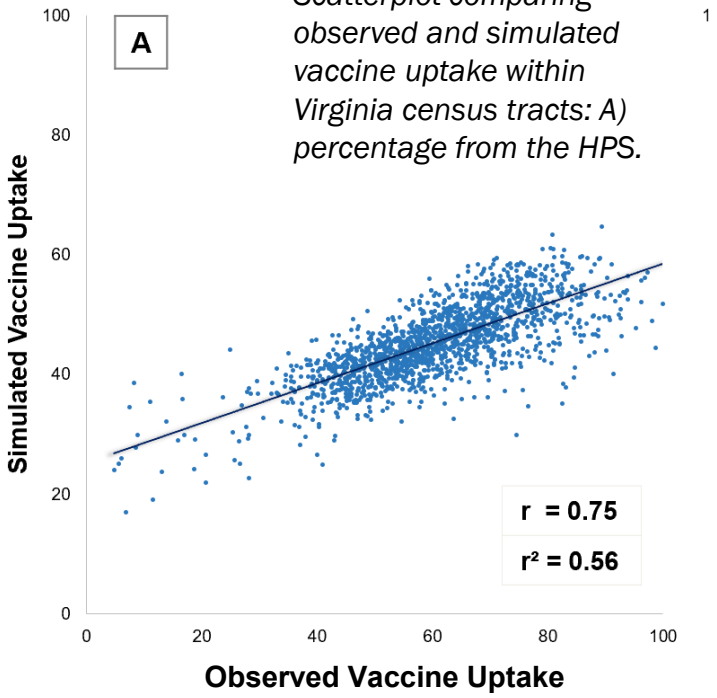
IDs	757.01	757.03	758.02	758.04	754.04	...
1	0	1	3	0	2	...
2	3	4	1	2	0	...
3	1	5	4	0	1	...
4	3	0	1	2	2	...
5	0	1	2	2	3	...
...	...	...	...	...	...	...

id	CT ID
1	757.03
1	758.02
1	758.02
1	758.02
1	754.04
1	754.04
2	757.01
2	757.01
2	757.01
2	757.03
2	757.03
2	757.03
2	757.03
2	758.02
2	758.04
2	758.04
...	...

id	CT ID	age	sex	race	bachelors	income	vaccine
1	757.03	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	758.02	27	m	w	y	50k	y
1	754.04	27	m	w	y	50k	y
1	754.04	27	m	w	y	50k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.01	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	757.03	18	f	b	n	21k	y
2	758.02	18	f	b	n	21k	y
2	758.04	18	f	b	n	21k	y
2	758.04	18	f	b	n	21k	y
...	...	...	...	...	...	...	...

Validation: Overall Accuracy

- Synthetic populations aggregated back to census tract level
  - Internally validated → against input US Census aggregated data
  - Externally validated→ against census tract vaccine uptake for Virginia
    - *(Virginia Department of Health – by request)*



Scatterplot comparing observed and simulated vaccine uptake within Virginia census tracts: A) percentage from the HPS.

Evaluation metrics comparing observed and simulated percentages of demographic and vaccine status variables across Virginia census tracts.

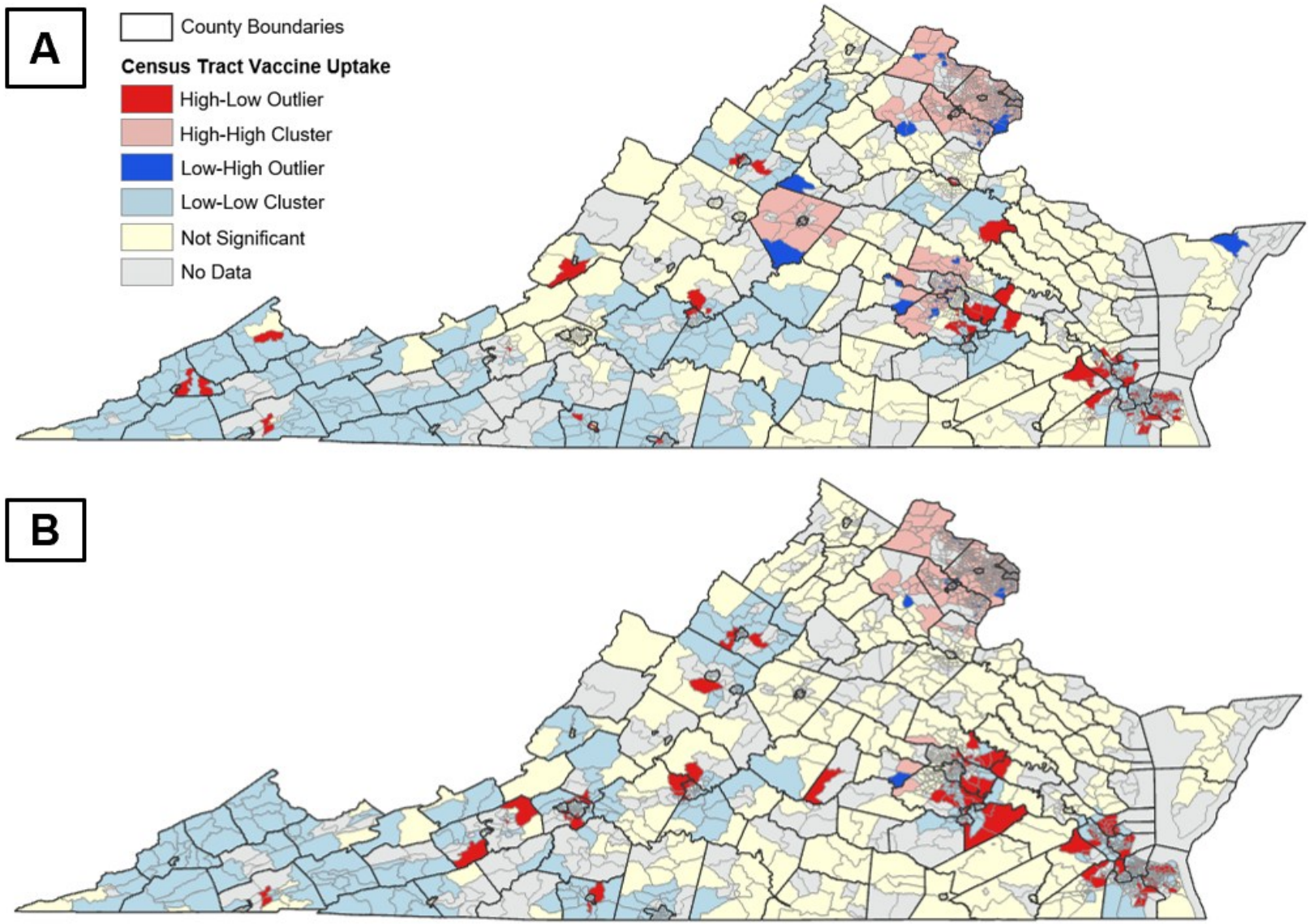
	HPS Dataset			
Variable: Descriptor	***Pearson's r***	***r^2***	***RMSE***	***MAE***
Gender: Male	0.9899	0.9799	0.7784	0.5867
Gender: Female	0.9899	0.9799	0.7784	0.5867
Race/Ethnicity: White	0.9991	0.9981	1.3229	1.0064
Race/Ethnicity: Black	0.9993	0.9987	1.0111	0.6690
Race/Ethnicity: Hispanic	0.999	0.9980	0.4696	0.3461
Race/Ethnicity: Other	0.9982	0.9965	0.9257	0.5872
Age: 18-29	0.9982	0.9963	0.758	0.5076
Age: 30-49	0.9948	0.9896	0.9428	0.7392
Age: 50-64	0.9945	0.9891	0.7982	0.6127
Age: 65 and over	0.9928	0.9857	1.2256	0.9309
Education: Bachelor's degree or higher	0.9969	0.9938	1.7922	1.5619
Education: No Bachelor's degree	0.9969	0.9938	1.7922	1.5619
Income: Less than $25,000	0.9992	0.9983	0.5555	0.4177
Income: $25,000 - $49,999	0.9984	0.9967	0.6097	0.4674
Income: $50,000 - $99,999	0.9974	0.9948	0.7428	0.5484
Income: Greater than $100,000	0.9994	0.9988	1.0354	0.7246
Vaccine Uptake: Received	0.7489	0.5608	18.2780	15.6970

Validation: Statistical Relationships

Coefficients from logistic regression models describing the relationship between demographic variables and vaccine uptake from the HPS, along with the synthetic population generated from the HPS. Significant coefficients are indicated with an asterisk (*) at the 90% confidence level (p-value < 0.10).

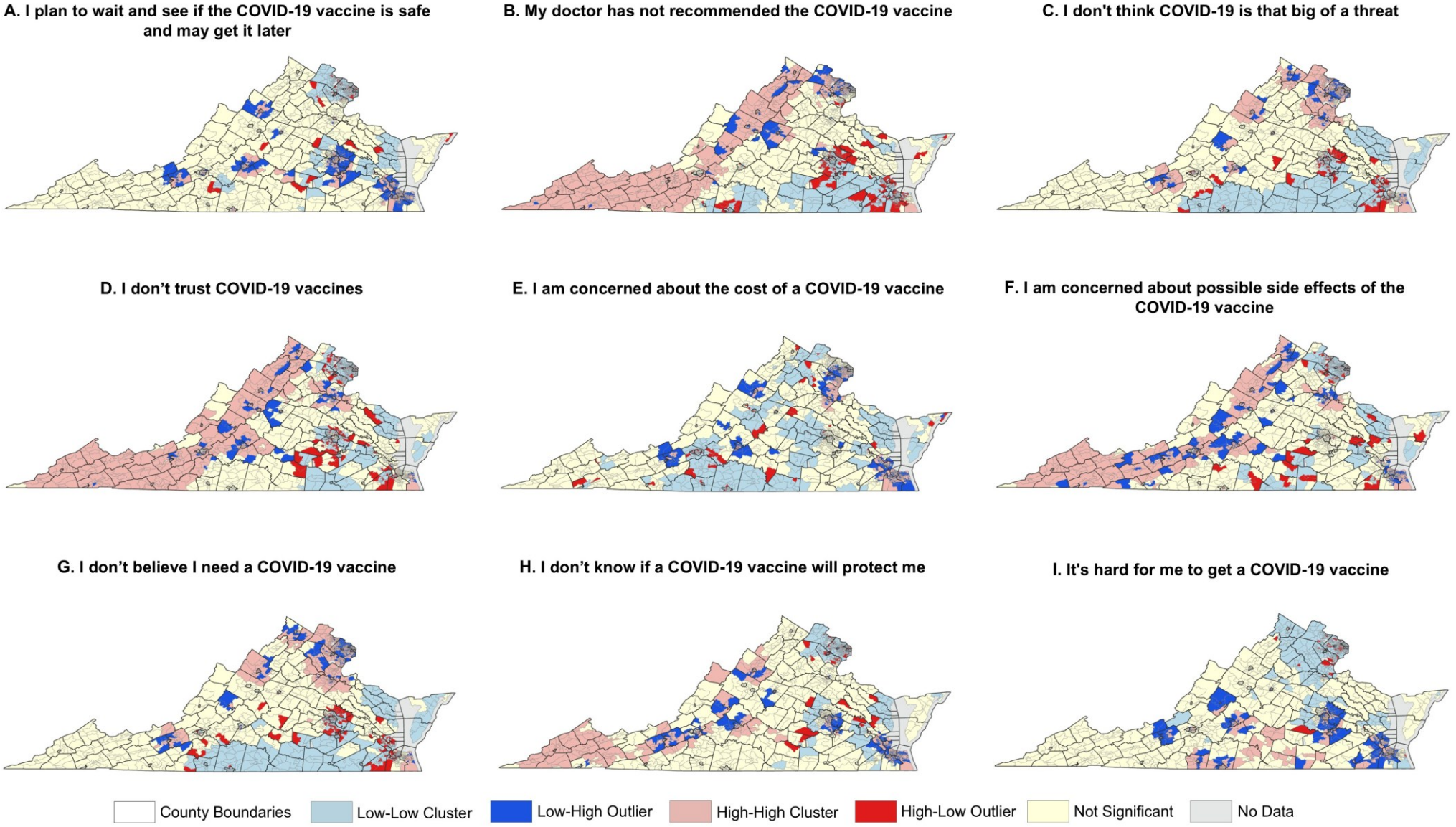
Variable: Descriptor	HPS Data: $R^2 = 0.08$ Coefficient β	Synthetic Population: $R^2 = 0.11$ Coefficient β
Intercept	1.8615*	- 0.3245*
Gender: Male	- 0.0595*	- 0.1872*
Race/Ethnicity: White	- 0.2277*	- 0.2758*
Age: 65 and over	1.1670*	1.4072*
Education: Bachelors or Higher	1.2429*	1.3435*
Income: $25,000 or less	- 0.5776*	- 0.8446*

Validation: Spatial Heterogeneity



Clusters and outliers of vaccination uptake in Virginia Census Tracts: A) observed, B) synthesized population with HPS data.

Spatial Patterns of Health Attitudes, Beliefs & Perceptions



Clusters and outliers of various health attitudes, beliefs or perceptions in the synthetized population from the HPS data.

Extending the Approach to Other Health Applications

Clusters and outliers of **depression** across census tracts for different locations:

A) Virginia CDC PLACES

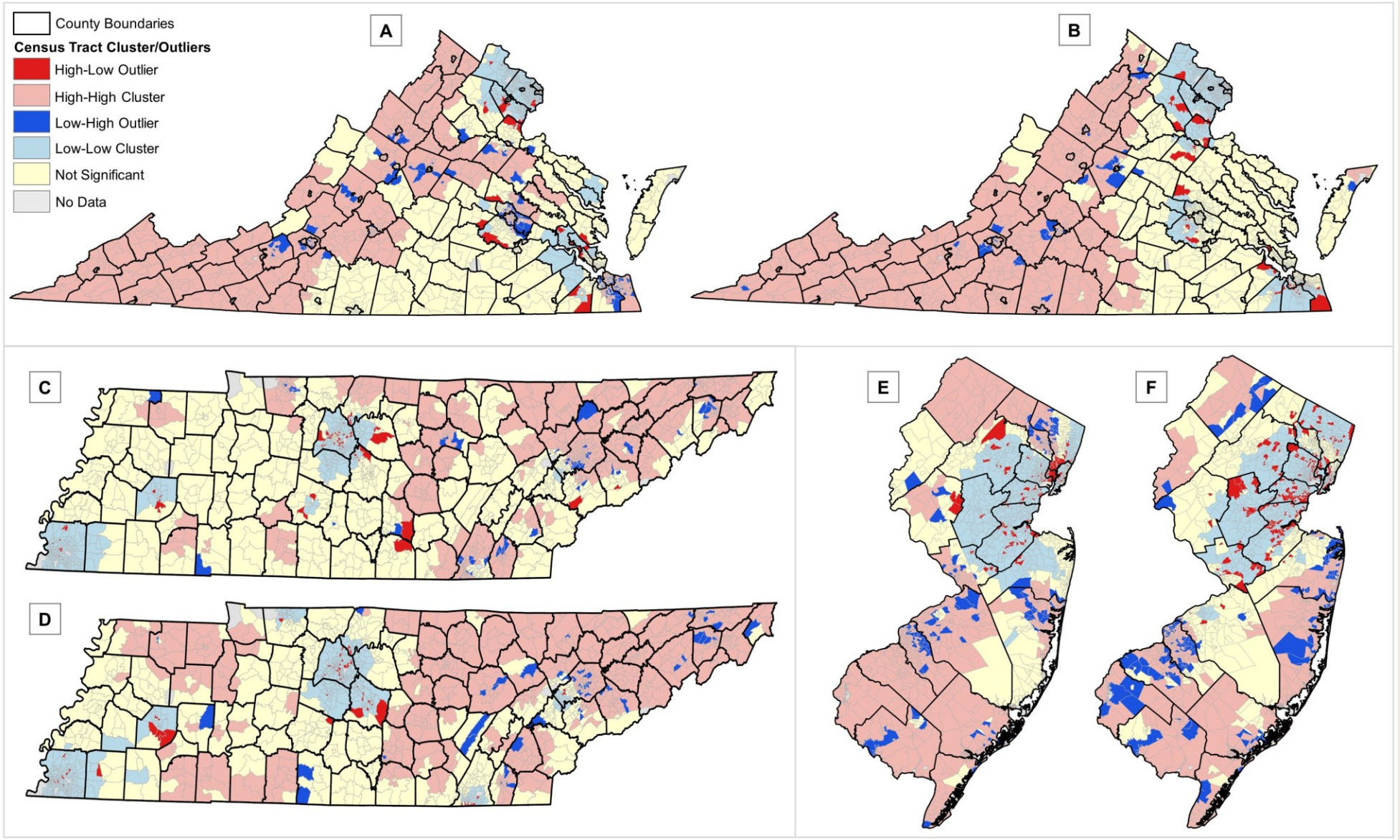
B) Virginia synthesized population

C) Tennessee CDC PLACES

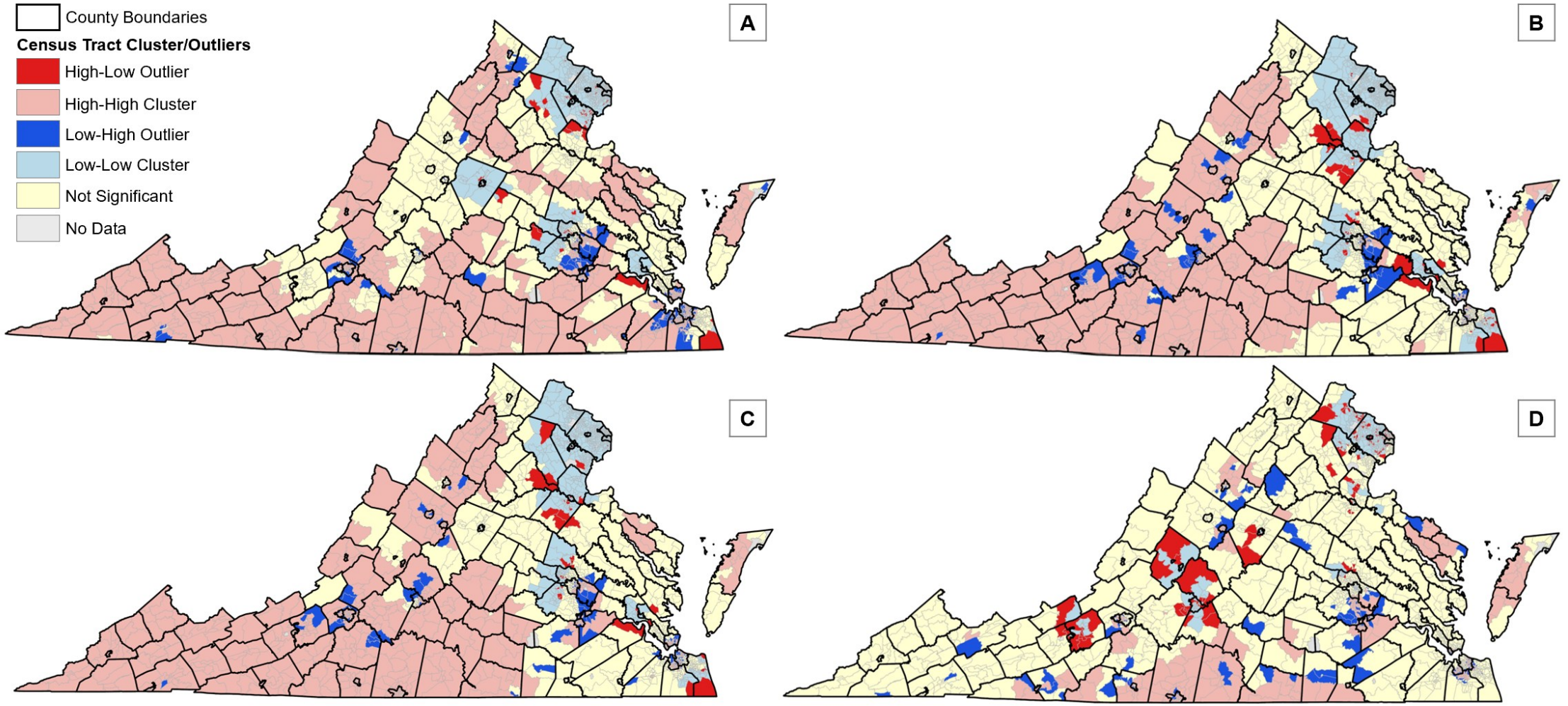
D) Tennessee synthesized population

E) New Jersey CDC PLACES

F) New Jersey synthesized population

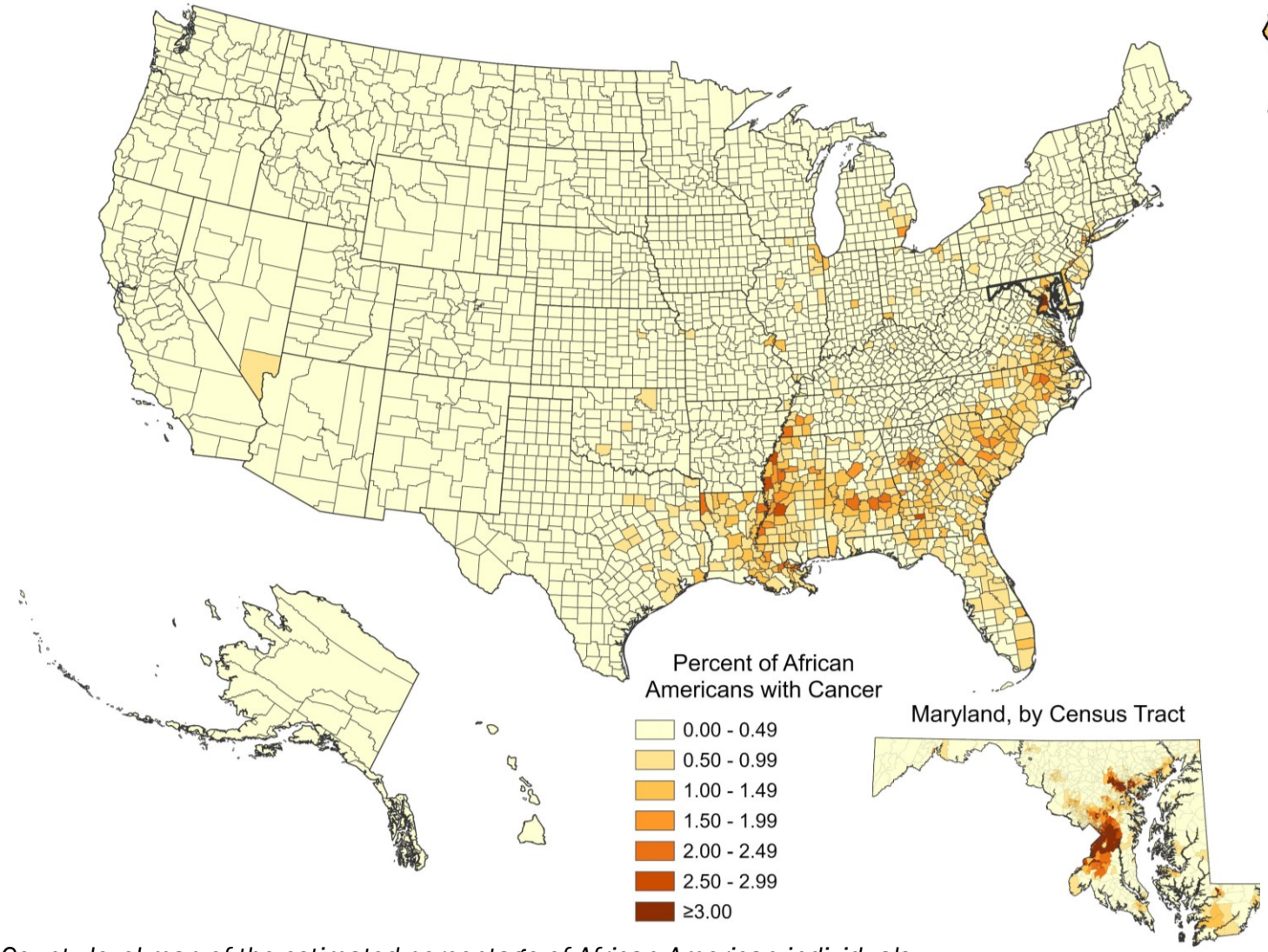


Extending the Approach to Other Health Applications

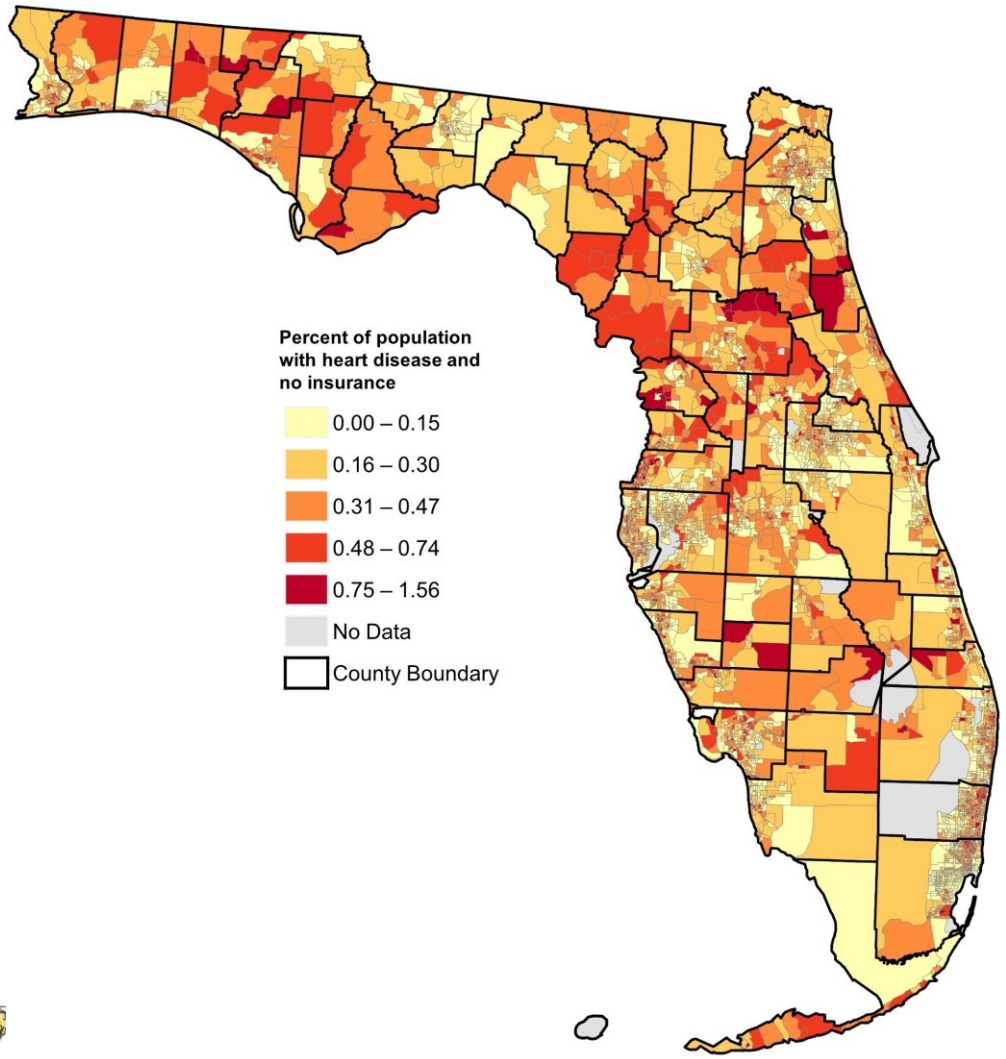


Clusters and outliers of the following: A) CDC PLACES estimates of ***current cigarette smokers****, B) 2021 synthesized population with smoking behavior C) 2022 synthesized population with smoking behavior D) Change in smoking from 2021 to 2022*

Extending the Approach to Other Health Applications



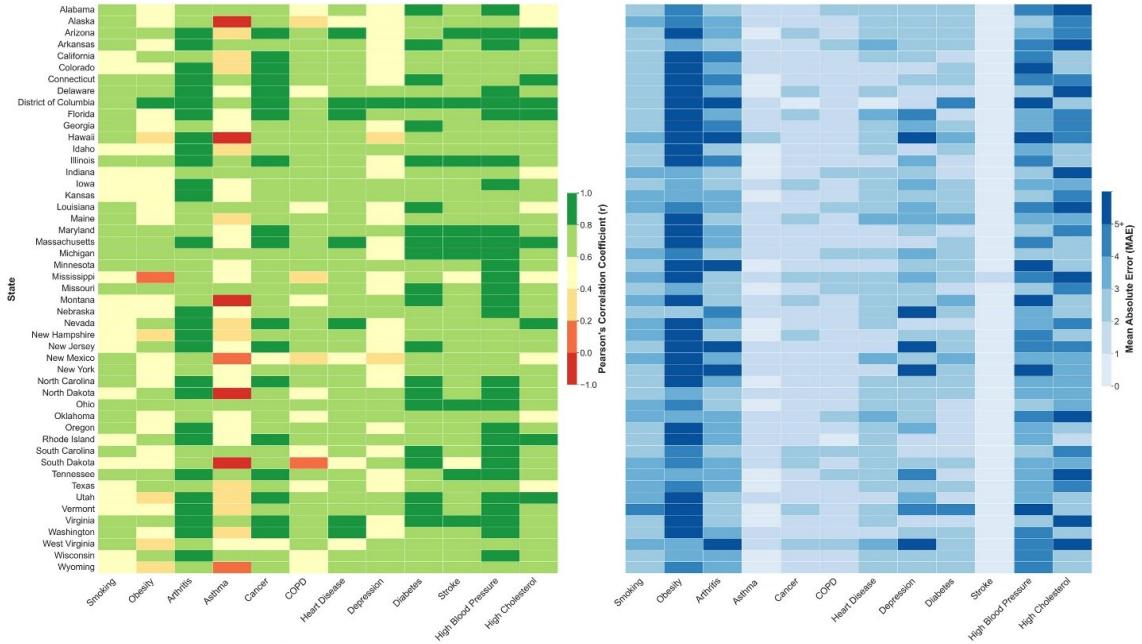
County-level map of the estimated percentage of African American individuals diagnosed with cancer, with an inset at the census-tract level for Maryland.



Percentage of population with heart disease and no health insurance across Florida census tracts (2019).

Extending the Approach to Other Health Applications

- *To our knowledge, no open-source, national individual-level synthetic population and SAE dataset exists for the U.S. that integrates sociodemographic attributes with multiple health behaviors and outcomes using transparent, reproducible methods and publicly available data sources*
  - **We generate a 2023 and 2024 spatially-explicit synthetic population (aged 18+) and corresponding SAEs at county and census tract scales for the entire U.S. (≈258 million agents)**
    - Seven sociodemographic attributes, two health behaviors, and eleven health outcomes
  - Evaluation with the latest CDC PLACES estimates (2023) demonstrates strong agreement, with variation by outcome and location



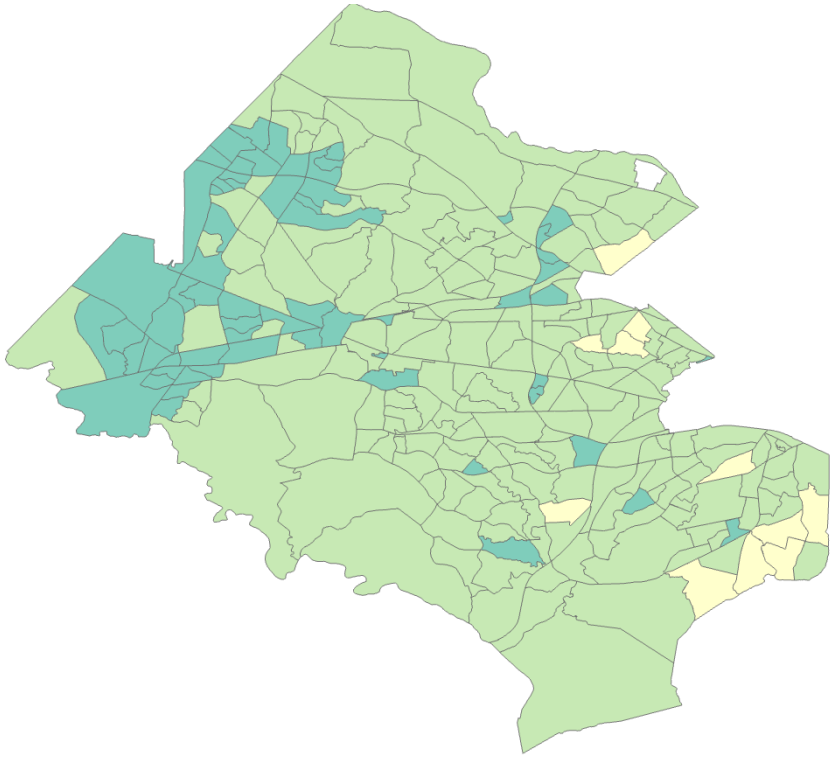
Heat map with the agreement between 2023 SHAPE and CDC PLACES estimates by state.

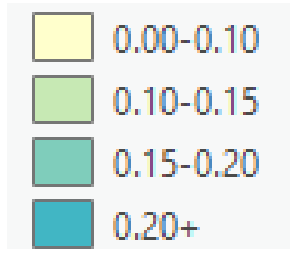
Health Behavior/Outcome	Pearson's r	MAE
Arthritis	0.862	1.732
Heart Disease	0.838	3.159
High Blood Pressure	0.819	2.527
Diabetes	0.815	1.885
Cancer	0.790	2.359
Smoking	0.771	1.703
COPD	0.635	1.781
Stroke	0.603	0.680
Depression	0.510	2.263
High Cholesterol	0.492	4.943
Obesity	0.420	3.503
Asthma	0.335	0.752

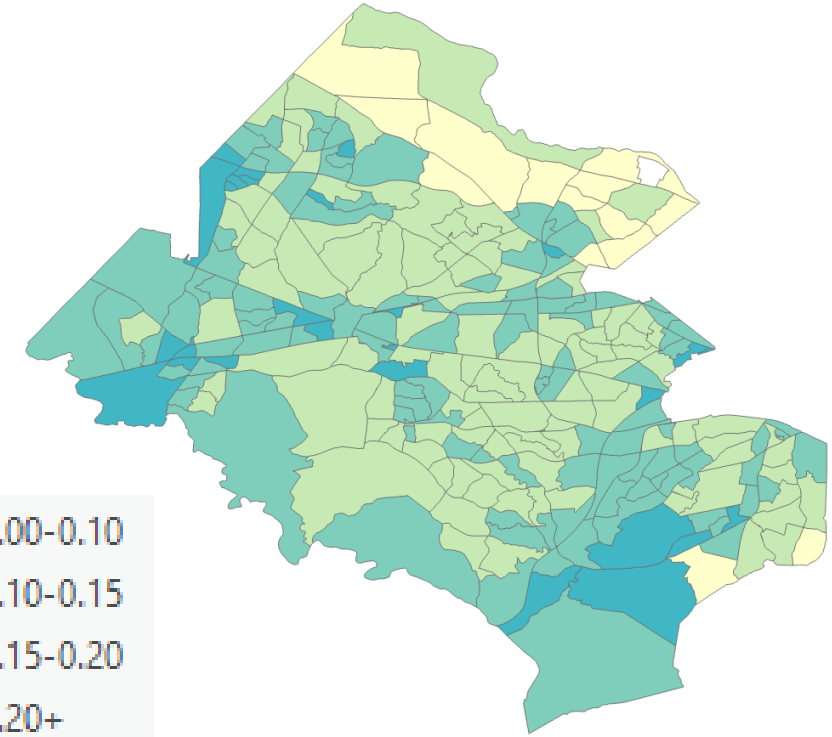
Evaluation metrics at the county level, comparing with corresponding CDC PLACES estimates.

How Do These Health-Enriched Synthetic Populations Influence ABM Outcomes?

Synthesized







*Percentage of population ever infected with COVID-19 across Fairfax County, VA census tracts when vaccination was **imposed** (70% of the population was randomly assigned as vaccinated).*

*Percentage of population ever infected with COVID-19 across Fairfax County, VA census tracts when vaccination was **synthesized using the IPF-based approach.***

How Do These Health-Enriched Synthetic Populations Influence ABM Outcomes?

Synthesized

group	Imposed	Synthesized
age1	16.472910	13.38906
age2	18.508743	13.65708
age3	13.627594	14.03755
age4	11.661374	13.85603
bach	8.506399	12.90204
income1	21.186660	14.16698
income2	16.881624	14.21399
income3	18.662131	14.12028
income4	13.669779	13.46592
male	14.273595	14.04939
race1	15.165753	13.43514
race2	25.057189	14.72453
race3	16.750593	14.33974
race4	11.478847	13.60521

Population making > $25k

Black Population

Percentage of population ever infected with COVID-19 across Fairfax County, VA census tracts when aggregated by demographic group. Differences are apparent when the population was imposed with vaccination (70% of the population was randomly assigned as vaccinated), compared to when the population was synthesized with vaccination using the IPF-based method.

Suggested Readings and Materials

- More on population synthesis and a tutorial in R: *https://spatial-microsim-book.robinlovelace.net/index.html*
- Literature review on synthetic population generation methods used in social simulation: *https://jasss.soc.surrey.ac.uk/25/2/6.html#remarks-on-the-data-processing-step-in-synthetic-population-generation*
- Comparison of different population synthesis methods: *https://ieeexplore.ieee.org/document/11338945*
- Population synthesis method for public health applications: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012439#sec002*