Model-based hybrid small area population estimates: Combining disparate sources of population data

Lance A. Waller Emily Peterson

Department of Biostatistics and Bioinformatics

Rollins School of Public Health

Emory University

lwaller@emory.edu

Outline

“Design what you can, model the rest.”

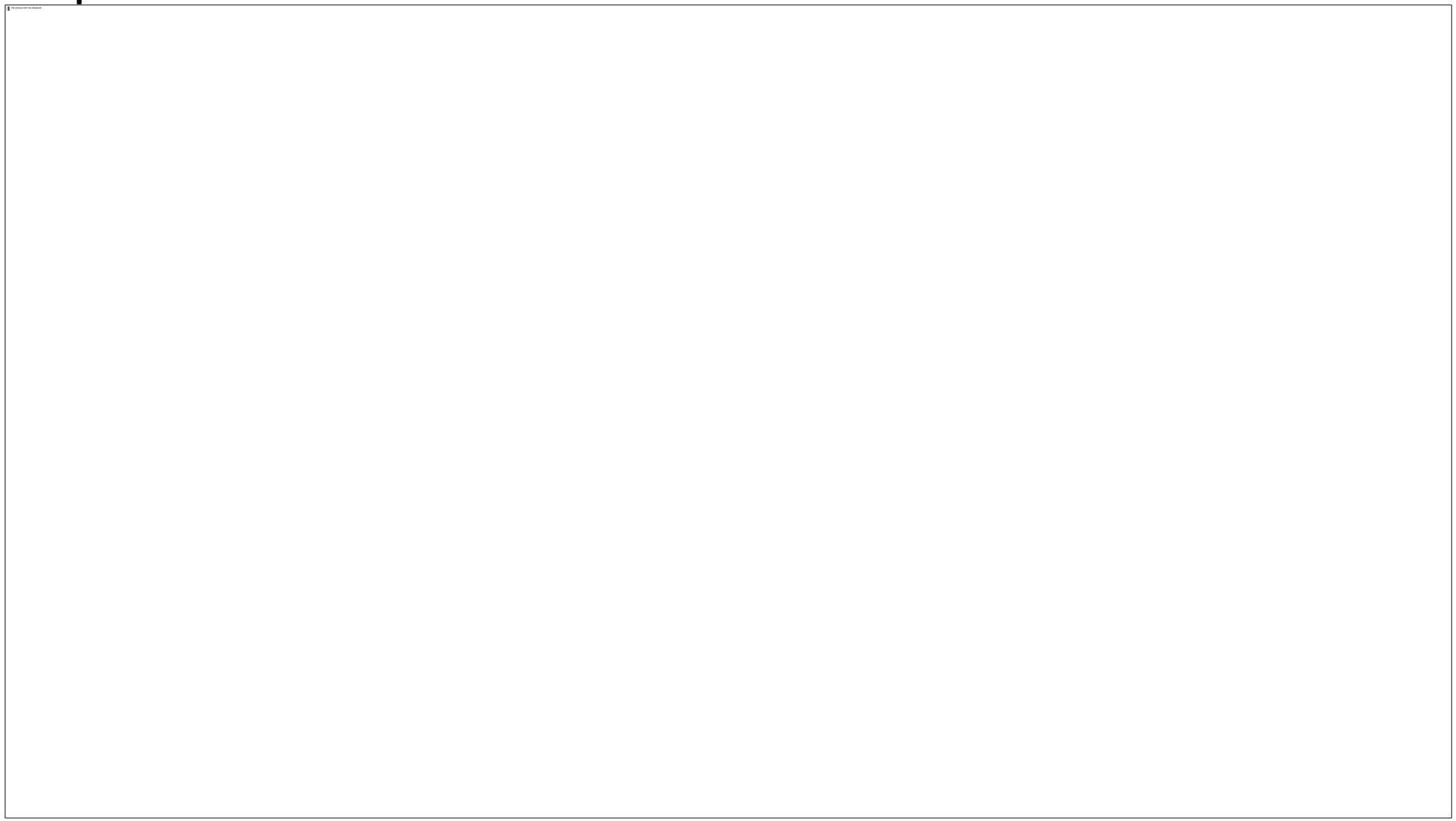
- Compromises: Differential privacy and small area estimation
- Differential Privacy and Small Area Estimates of Health Disparities
- Hybrid, model-based estimation of small area population sizes.

“Design what you can, model the rest.”

Acknowledgements

- Support: NICHHD R01: 5R01HD092580
- Emory University, Rollins School of Public Health
  - Lance Waller, Emily Peterson
- Harvard T.H. Chan School of Public Health
  - Brent Coull, Nancy Krieger, Jarvis Chen, Rachel Nethery, Yanran Li, Christian Testa, Nick Link
- Imperial College, London
  - Fred Piel, Marta Blangiardo, Paul Elliott, Mahboubeh Parsaeian
- Drexel University
  - Loni Tabb

Compromises: SAE and DP



Small Area Estimation

- Fay & Herriot (1979, *JASA*)
- Shrinkage estimation, James-Stein estimation
- Random intercepts borrow information
  - From overall mean (SAE)
  - From neighbors (Clayton and Kaldor 1987, *Bcs*)
  - From both (Besag, York, Mollié 1991, *Ann Inst Stat Math*)

- **Elegant mathematics to borrow the right amount of information.**

“Design what you can, model the rest.”

Differential Privacy

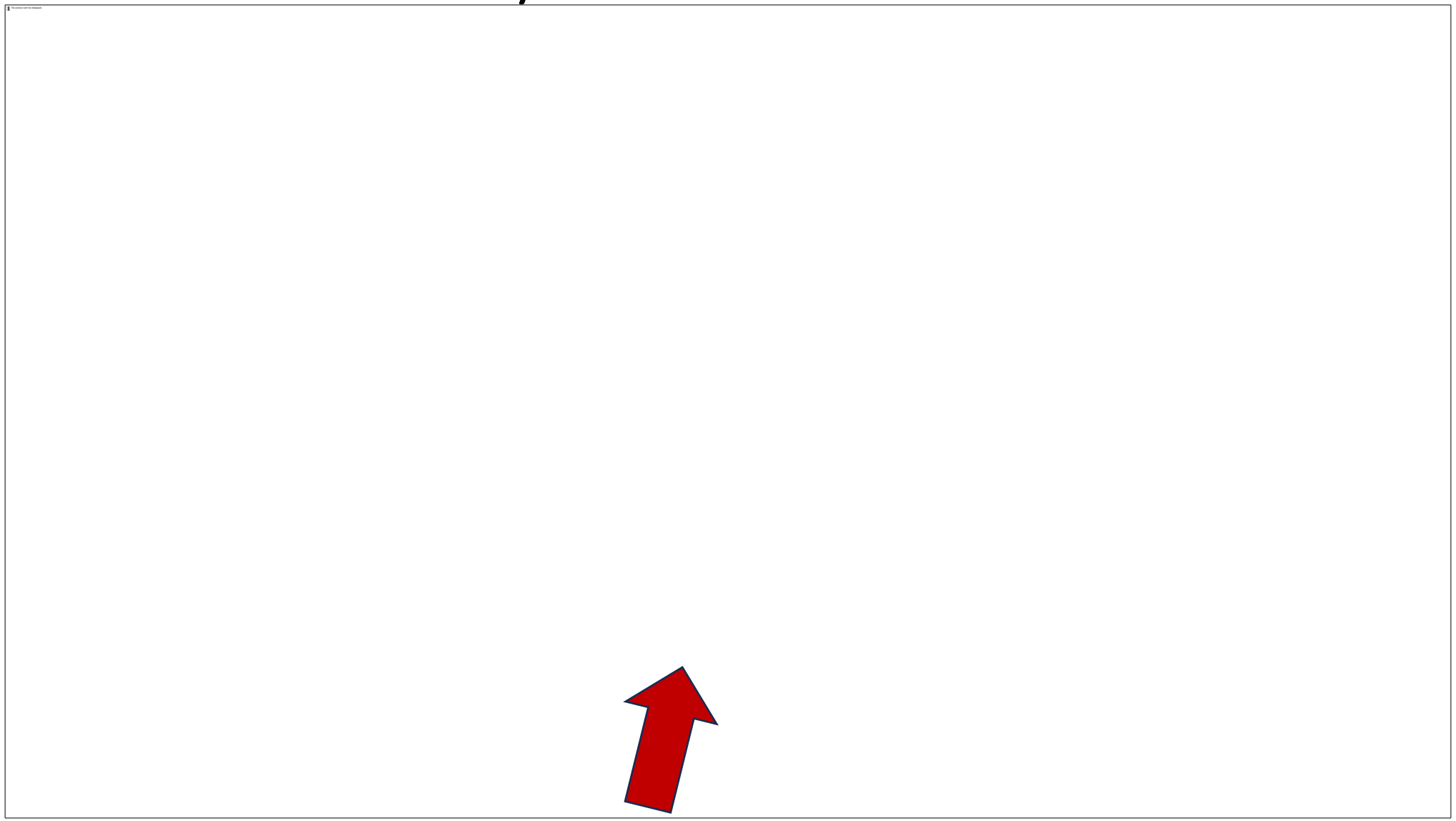
- Statistical summaries vs. needle-in-a-haystack
  - (Dwork IPAM-UCLA presentation on YouTube)
- Metaphor: 20 questions
  - Questions are statistical summaries
  - “20” = privacy budget (debit card vs. credit card)
- Differential privacy caps probability of discovery from statistical summaries
- **Elegant mathematics to add the right amount of noise.**

“Design what you can, model the rest.”

Putting data to use

- A lot of research on differential privacy
- US Census Bureau
  - Applied to 2020 US Census data products.
  - Provided in 2010 demonstration products.
- Impact on small area inference?

Differential Privacy and Small Area Estimation



Differential Privacy and Small Area Estimates

- 2010 (yes, 2010) US Census demonstration data product
  - Both old and new (DP) disclosure avoidance.
  - Releases in 2019, 2020, 2022 (spoiler alert: version is important!)
- Two recent assessments relating to small area health disparities:
  - Kurz et al (2022, *Health Services Research*), Waller (2022 commentary)
  - Li et al. (2023, *Science Advances*)

Kurz et al. (2022, *HSR*)

- Kurz et al (2022):
  - County-level 2010 population counts and local Medicaid participation.
  - Racialized subpopulations.
  - County-level: Differential privacy introduced errors up to 10% in counts and proportions of local Medicaid populations.
  - State-level: Negligible errors in race-specific Medicaid participation.

Waller (2022, *HSR*, Commentary)

- Decennial Census
  - Number of residents on particular day.
- American Community Survey
  - Rotating survey estimate averaged over a period of time.
  - Includes margins of error.
  - Spatial correlation in estimates (high values near high values).
  - Spatial correlation in margins of error (noisy values near noisy values).

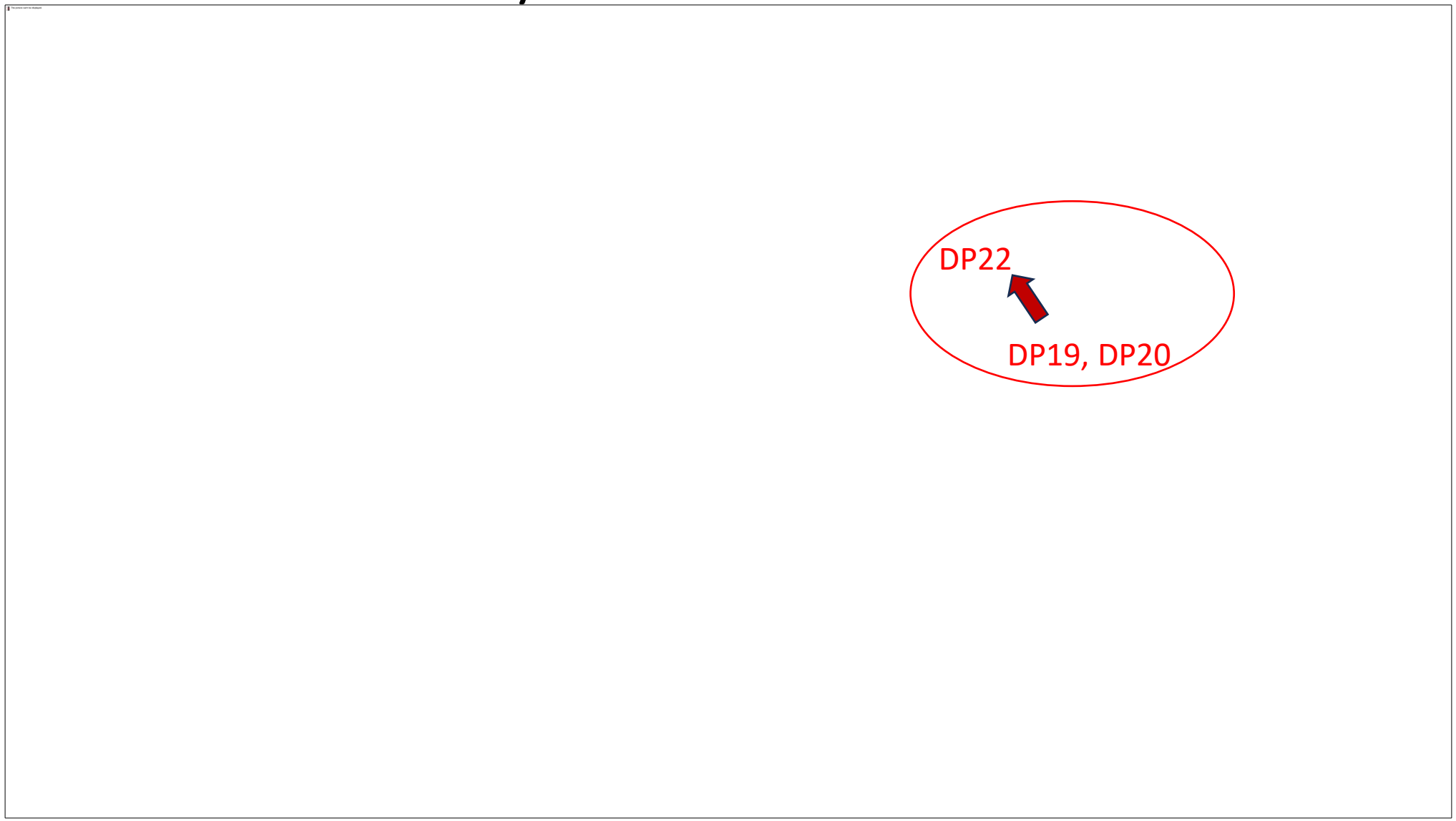
- Different data products require care in combining!

Closer look at 2010 Demonstration Products

- Li et al. (2022, *Sci Adv*) compare:
  - Original 2010 decennial census tract data (DC, data swapping)
  - Oct 2019 demonstration product (DP19, differential privacy)
  - May 2020 demonstration product (DP20, updated differential privacy)
  - August 2022 demonstration product (DP22, most recent differential privacy)

- Notes:
  - DP19, DP20: identical DP, distinct postprocessing
  - DP22: revised DP parameters in response to users (increased privacy-loss budget, giving *greater accuracy and weakening privacy guarantees*).

Differential Privacy and Small Area Estimation



% difference in age-standardized population

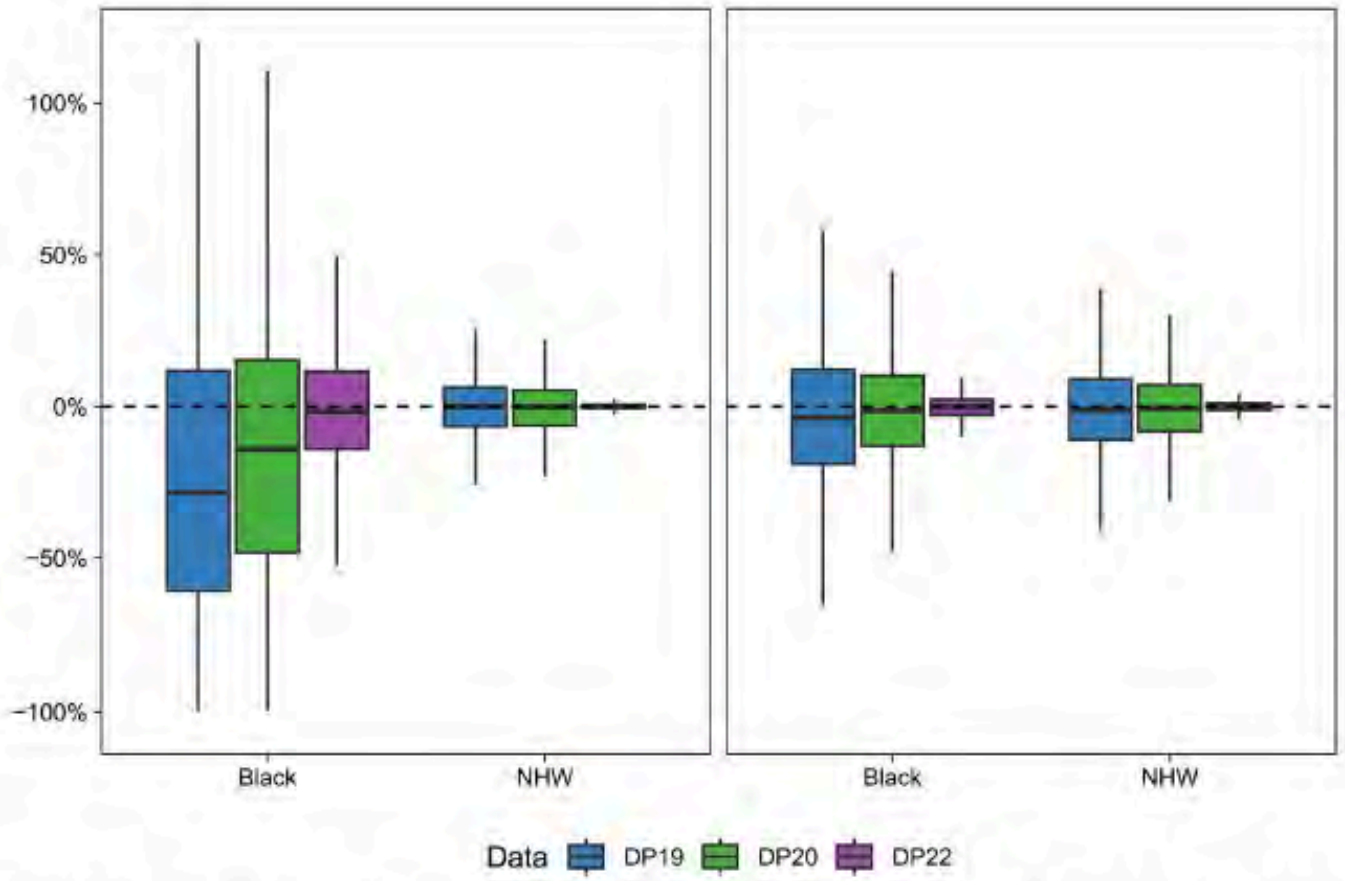


Fig. 2. Percent difference in age-standardized population denominators from each of the DAS demonstration products and the 2010 DC. Boxplots show the percent difference in census tract expected premature mortality counts for 2010 created from the U.S. Census Bureau DAS demonstration products released in 2019 (DP19), 2020 (DP20), and 2022 (DP22), relative to original 2010 DC-based expected counts, for Black and NHW populations in Massachusetts (**left**) and Georgia (**right**).

Simulation assessments of bias

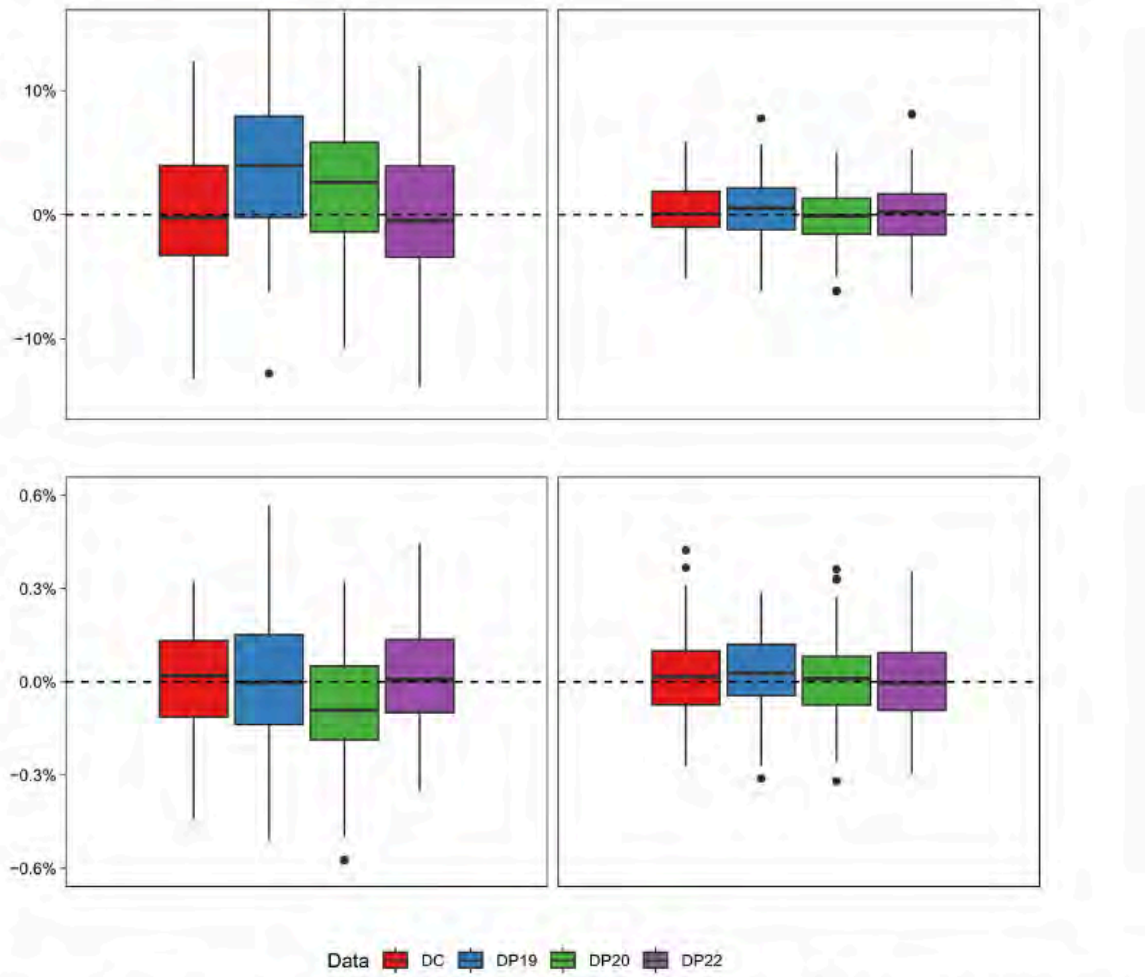


Fig. 3. Simulation results: Bias in estimated inequity model coefficients using different denominator data sources. Boxplots represent the distribution of bias in estimates of the Black versus NHW (**first row**) and percent poverty (**second row**) coefficients over 100 simulations, with simulated data mimicking patterns of premature mortality in Massachusetts (**left column**) and Georgia (**right column**) and models fit using each of the four denominator data sources: the 2010 DC and the U.S. Census Bureau DAS demonstration products released in 2019 (DP19), 2020 (DP20), and 2022 (DP22). Data were generated using DC as the true denominator data.

Next: combining disparate data products

- Small area population: How many people there? Two approaches:
- Nethery et al.: Comparing US Census data and non-Census data
  - American Community Survey
  - WorldPop
- Peterson et al.: Combining US Census Bureau products:
  - Decennial
  - PEP
  - American Community Survey

Comparing ACS and WorldPop

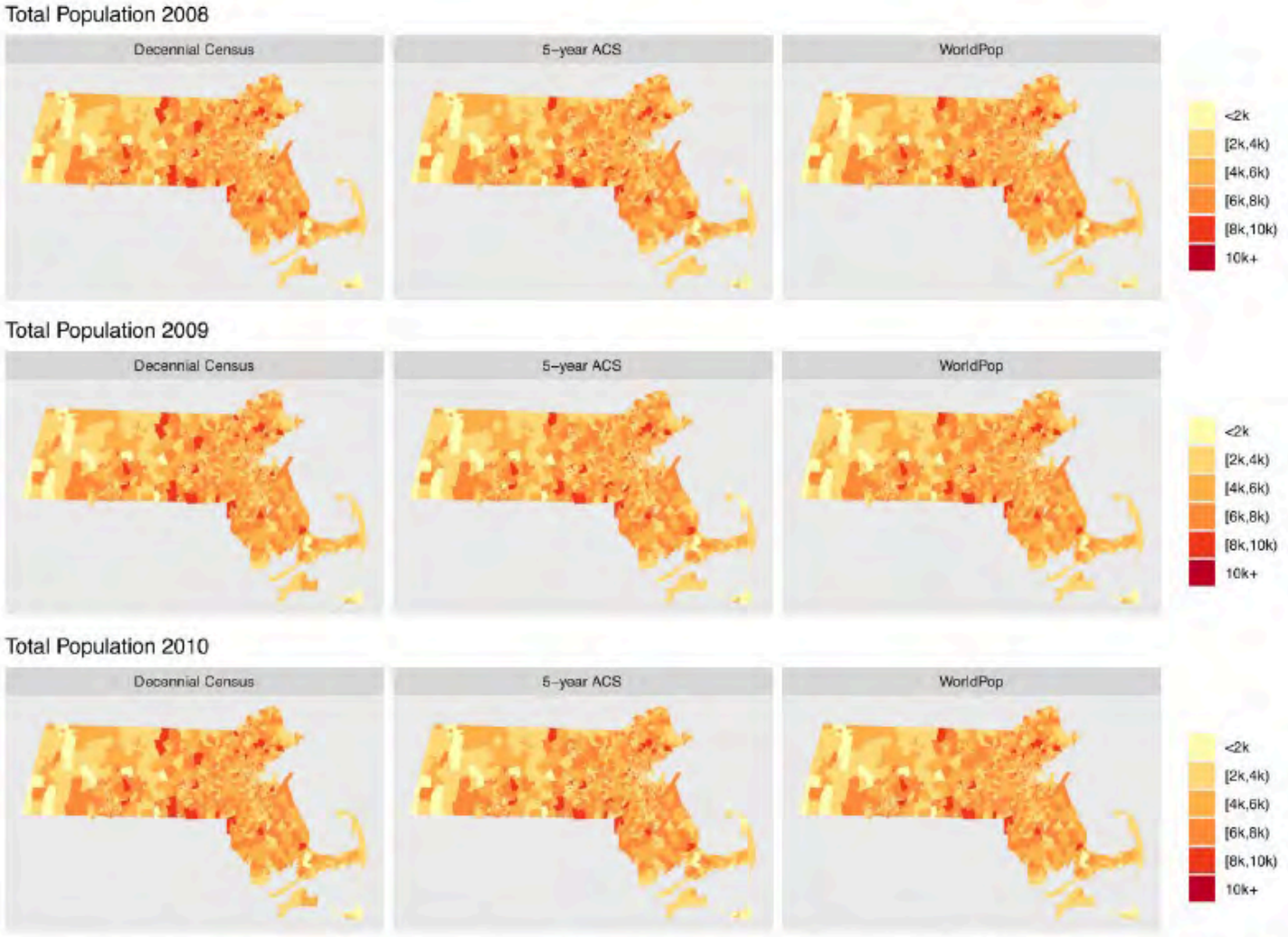


Fig. 1. Spatial distribution of census tract-level 2010 decennial census population counts compared to 5-year ACS and WorldPop population estimates for years 2008–2010, Massachusetts, USA.

Nethery et al. (2021, *SSM Population Health*)

Accuracy vs. % poverty

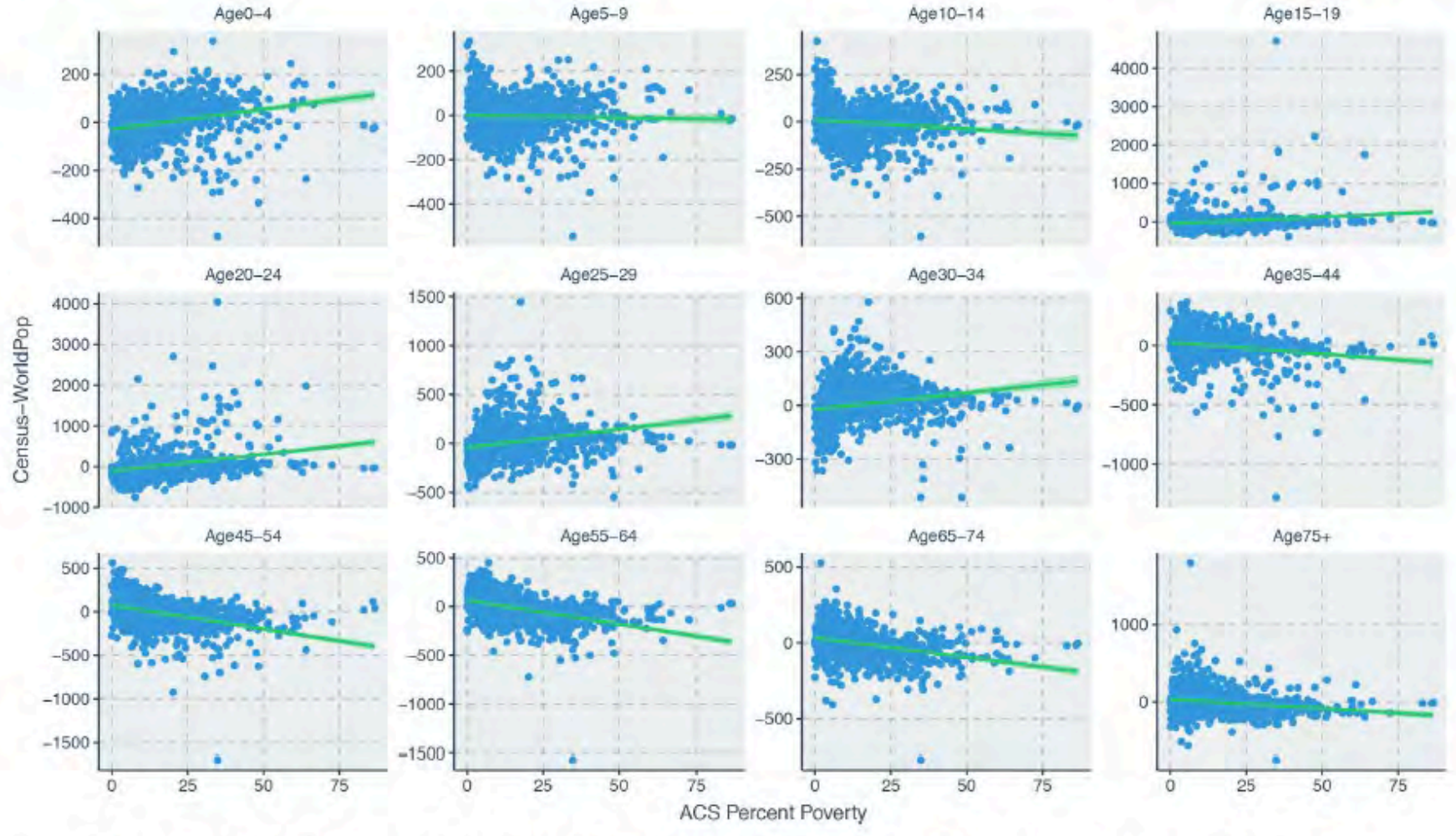


Fig. 2. Scatterplots of difference in WorldPop 2010 and decennial census age-stratified CT population estimates vs. percent of the CT in poverty.

Combining US Census Data Products

- Peterson et al.: Combining US Census Bureau products:
  - Decennial
  - PEP
  - American Community Survey
  - https://arxiv.org/abs/2112.09813

Decennial

PEP

ACS

Sources of Error

Undercount

Demographic errors

Sampling errors, Temporal aggregation

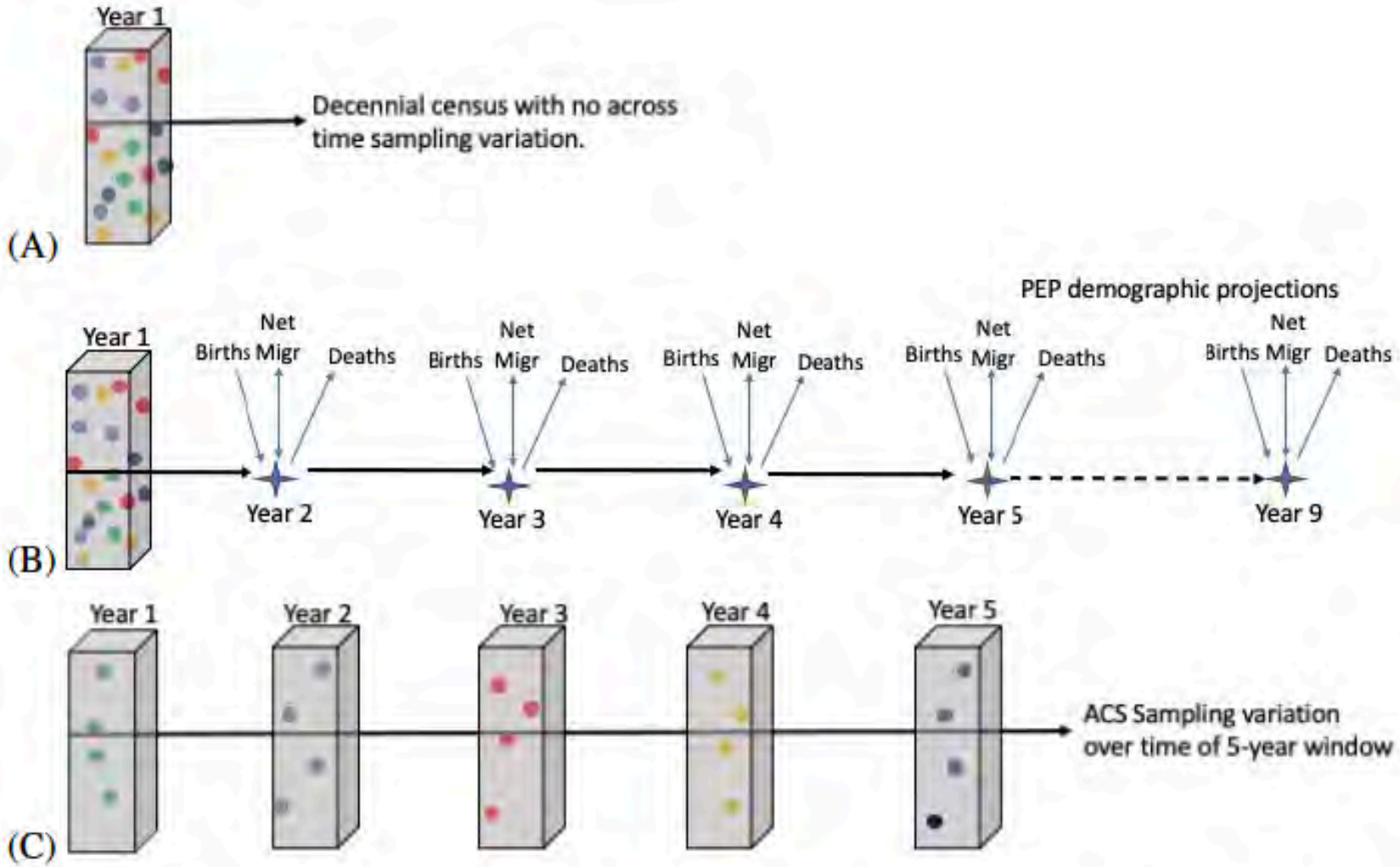


FIG 2. *Illustration of data generating processes for decennial census (A), PEP demographic projections (B), and ACS (C). Note: Net Migr refers to net migration, which is equal to in-migrations minus out-migrations. Boxes refer to data collection on individuals via surveys. Crosses refer to data generation through a demographic projection formula. Color dots refers to blocks of individuals.*

Modeling different types of error

- Hierarchical structure

Sampling error

$$p(\gamma, \sigma^2_{NS}, \Theta | \boldsymbol{n}^{(census)}, \boldsymbol{n}^{(PEP)}, \boldsymbol{n}^{(ACS)}, \boldsymbol{s}^2) = p(\boldsymbol{n}^{(ACS)} | \boldsymbol{s}^2, \boldsymbol{n}^{(PEP)}, \boldsymbol{n}^{(census)}, \gamma, \sigma^2_{NS}, \Theta)$$
$$\cdot \; p(\boldsymbol{n}^{(PEP)} | \boldsymbol{n}^{(census)}, \gamma, \sigma^2_{NS}, \Theta)$$
$$\cdot \; p(\boldsymbol{n}^{(census)} | \gamma, \sigma^2_{NS}, \Theta)$$
$$\cdot \; p(\gamma | \sigma^2_{NS}, \Theta) \cdot p(\sigma^2_{NS}, \Theta)$$

True population

Non-sampling error

1000 words

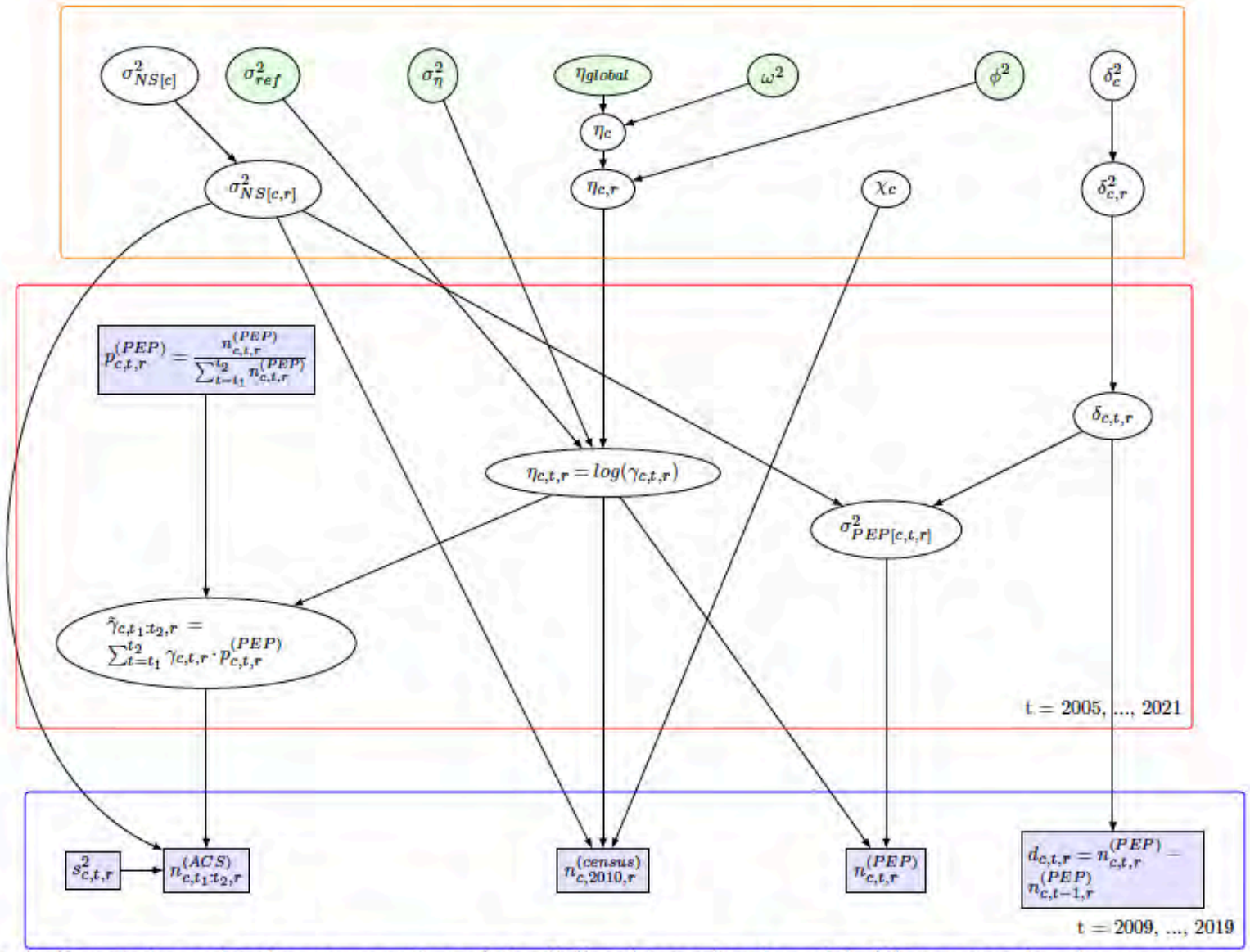
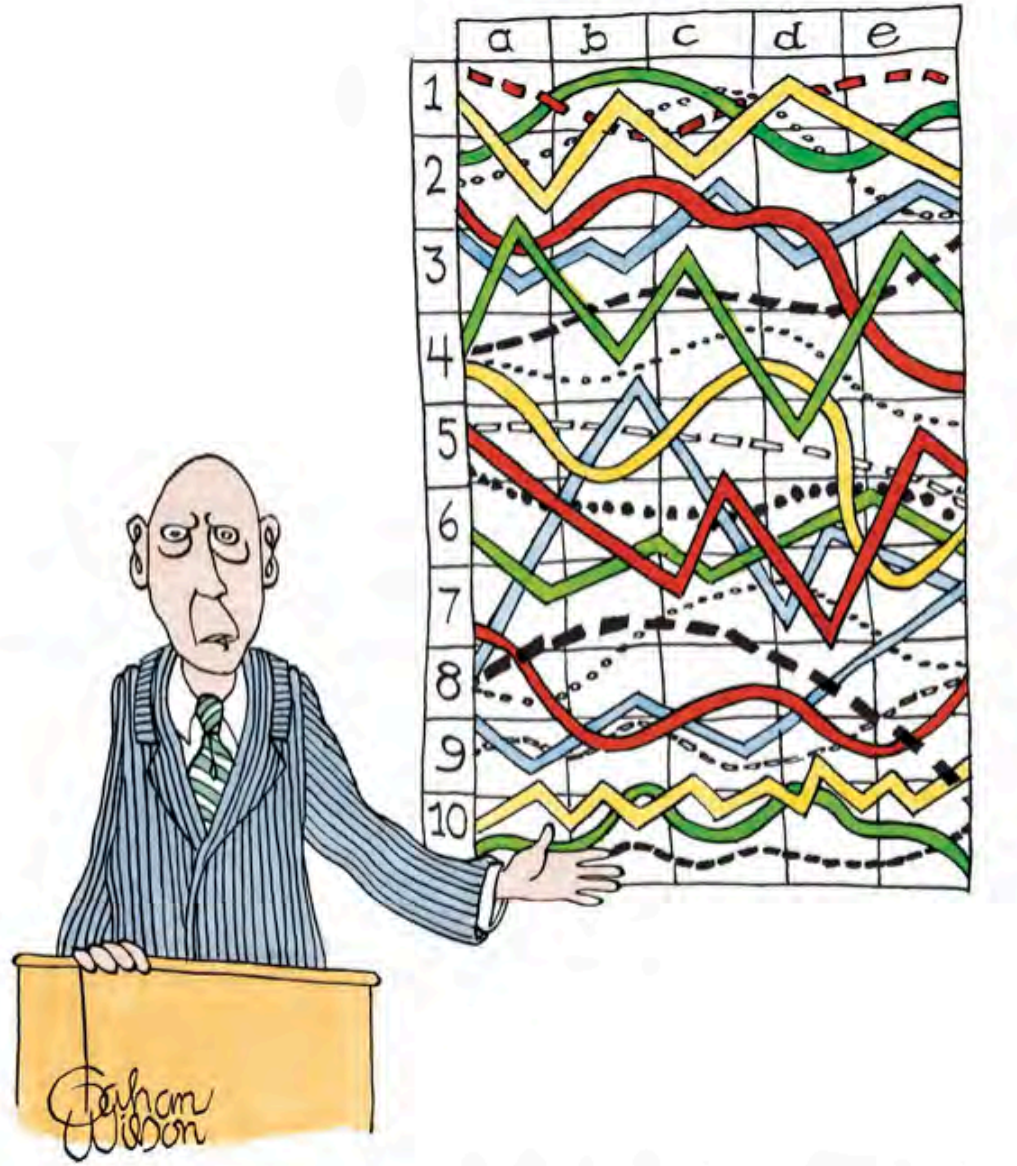


Figure 4: Directed graphical representation of the B-Pop hierarchical model. Blue rectangles denote observed data quantities, and circles denote latent variables (green shaded circles for global hyper-parameters). Solid arrows denote stochastic dependency. Boxes group quantities by indices, i.e., (1) blue box contains observed population data, stratified by county-year-race for years 2009 to 2019, (2) red box contains both estimated parameters and observed data quantities stratified by county-year-race, for years 2005 to 2021, (3) orange box contains county-race, county-specific, and global parameters. Subscripts refer to county c, year t, race r.



"I'll pause for a moment so you can let this information sink in."

https://www.newyorker.com/culture/postscript/the-beautifully-macabre-cartoons-of-gahan-wilson

A new type of map (for me)….

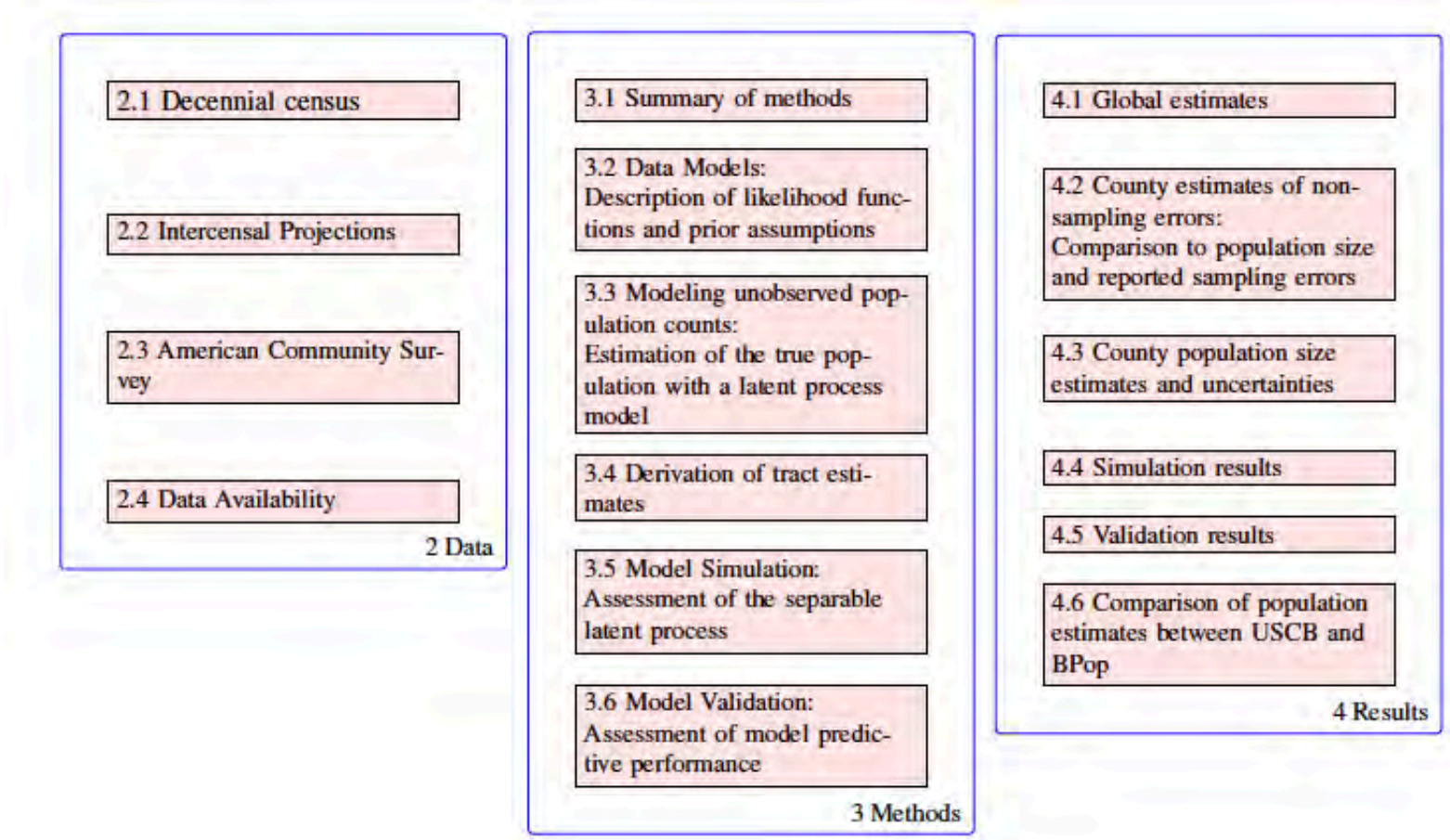


FIG 1. *Outline of the structure of sections and descriptions.*

Predictions: Baker County

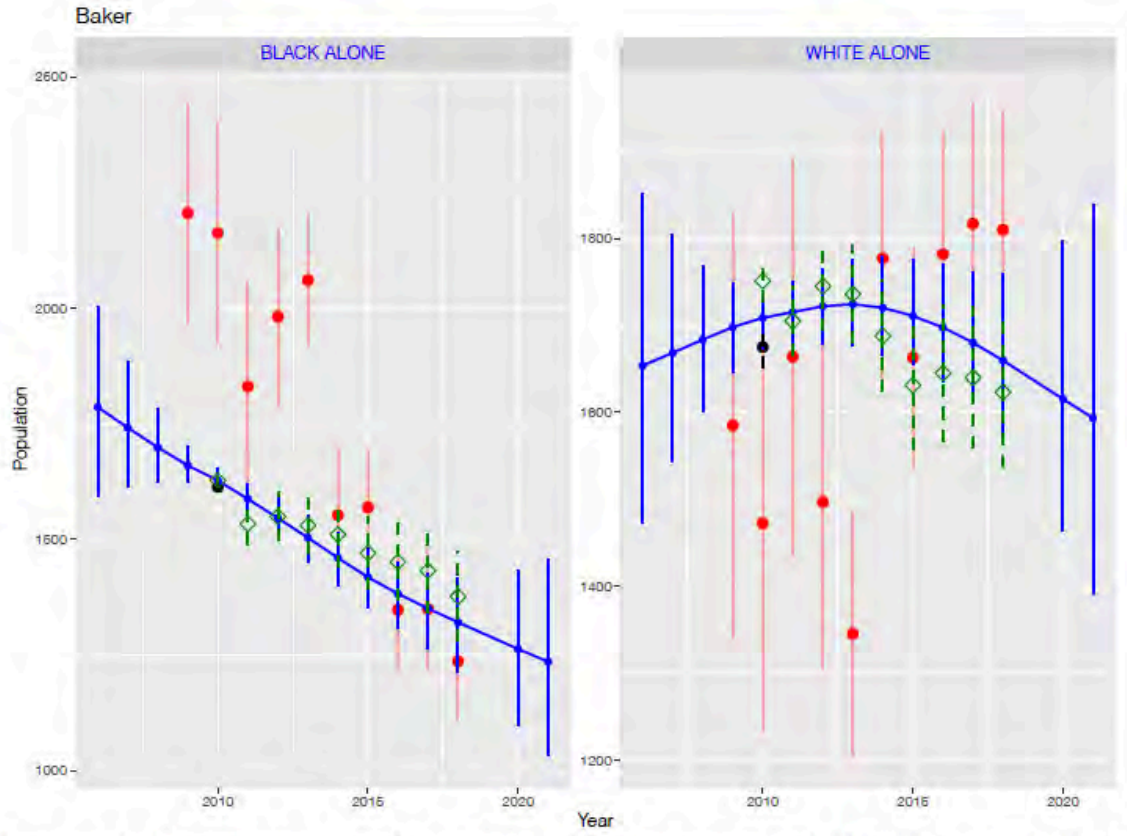
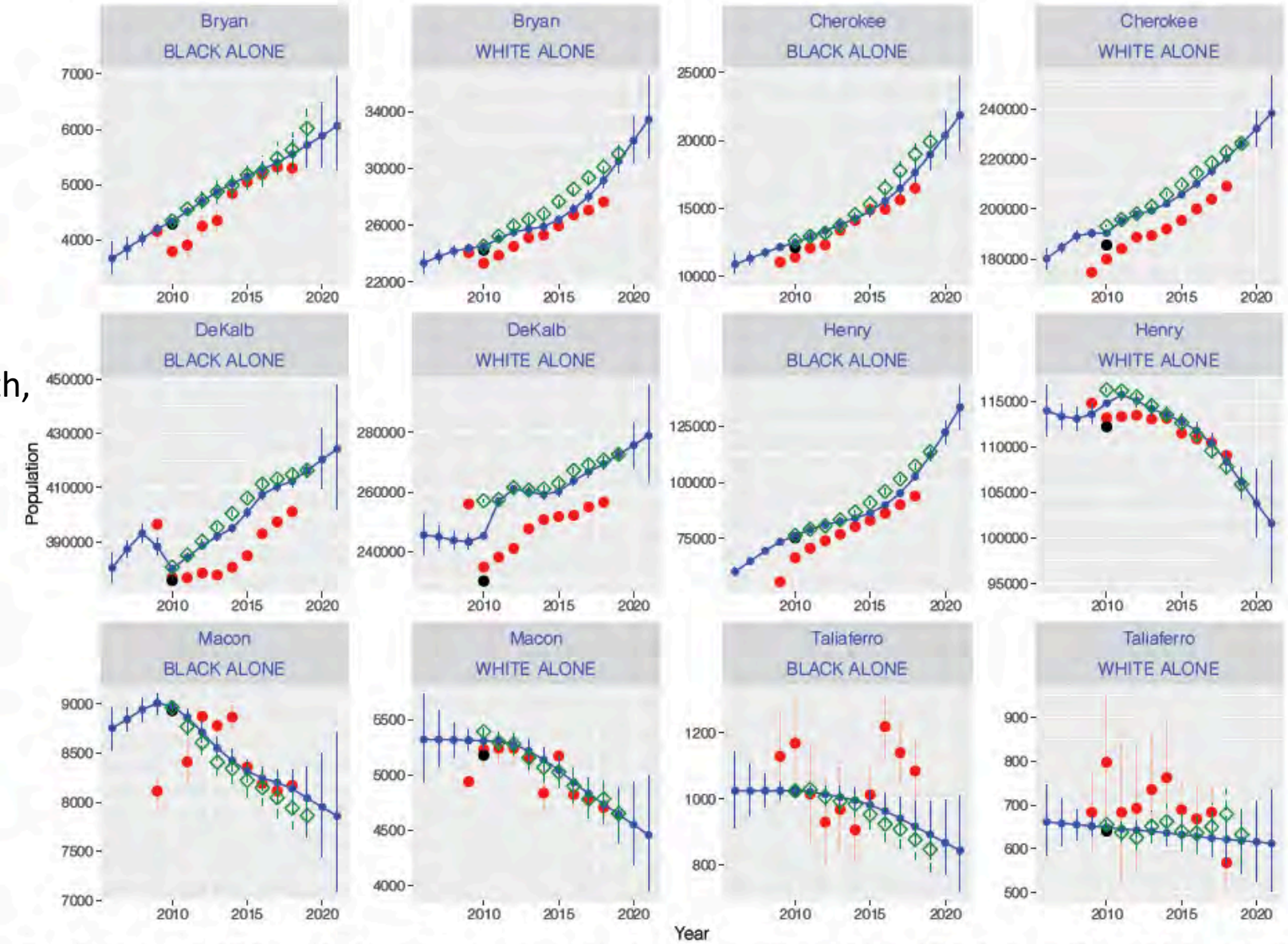


Figure 8: Time trends of county-race specific posterior median estimates, and 95% CIs, against observed data, for Baker County. B-Pop median estimates and 95% credible intervals are shown in blue. Estimates are obtained for years 2005-2021. Red refers to ACS reported counts, black refers to decennial census counts, and green refers to PEP projection estimates. Data are shown for years 2009-2019 refer to the end years of ACS 5-year period estimates.

High population:
PEP catches growth,
ACS lags



Low population:
PEP catches growth,
ACS noisier

Figure 9: Illustration of B-Pop model data and county estimates for selected counties (Bryan, Cherokee, DeKalb, Henry, Macon, Taliaferro) stratified by race (White only, Black only). Parameters plotted are estimated population counts. The plots include: 1. observed ACS data with associated observation-based 95% error intervals (red), 2. posterior estimates with 95% credible intervals (blue), 3. Decennial census population counts (black), and 4. PEP population projections (green).

Summary

- Compromises: Differential privacy and small area estimation
  - Tension in goals
- Differential Privacy and Small Area Estimates of Health Disparities
  - Reducing privacy budget improved accuracy
- Hybrid, model-based estimation of small area population sizes.
  - Tracking sources of uncertainty (challenging, but important)

- “Design what you can, model the rest.”

References

- Peterson EN et al. (2023) A Bayesian hierarchical small area population model accounting for data source specific methodologies from American Community Survey, Population Estimates Program, and Decennial Census Data. https://arxiv.org/abs/2112.09813
- Nethery et a. (2021) Comparing denominator sources for real-time disease incidence modeling: American Community Survey and WorldPop. *SSM-Population Health* 14, 100786. https://doi.org/10.1016/j.ssmph.2021.100786
- Li Y et al. (2023) Impacts of census differential privacy for small-area disease mapping to monitor health inequities. *Science Advances* 9, eade888.
- Kurz CF, König AN, Emmert-Fees KMF, Allen LD. The effect of differential privacy on Medicaid participation among racial and ethnic minority groups. *Health Serv Res*. 2022;57(Suppl. 2):207-213. doi:10.1111/1475-6773.14000
- Waller LA. Global and local impacts of differential privacy on estimates of health care inequity. *Health Serv Res*. 2022;57(Suppl. 2):204-206. doi:10.1111/1475-6773.14080