Recent Advances in Spatial and Spatio-Temporal Change of Support for Official Statistics

Scott H. Holan
Department of Statistics
University of Missouri

Second FCSM/WSS Workshop on Quality of Integrated Data
January 25, 2018

Joint work with:

Jonathan R. Bradley (Florida State University)

Christopher K. Wikle (University of Missouri)

Motivating Data

- **The American Community Survey (ACS)**:
  - An ongoing survey administered by the US Census Bureau that provides timely information on many key demographic and socio-economic variables.
  - The ACS produces 1-year and 5-year "period-estimates," and corresponding margins of errors, for demographic and socio-economic variables recorded over predefined geographies within the United States.
- **Change of Support:** Producing estimates on multiple scales (i.e., user-defined geographies and/or time-periods).
  - **Example 1:** Provide estimates on user-defined geographies.
  - **Example 2:** Produce 3-year period estimates of ACS variables using 1-year and 5-year ACS estimates.

Change of Support

- There are two general approaches for spatial change of support.

  1. **Bottom-up**: Estimate the variable at a very fine resolution using the data defined on the source support **(i.e., regions associated with the data).** Then average the variables up to any target support **(i.e., regions that we would like to have estimates on).**

  2. **Top-down**: Define the process by a partitioning of the source support and target support (e.g., Mugglin et al., 1998).

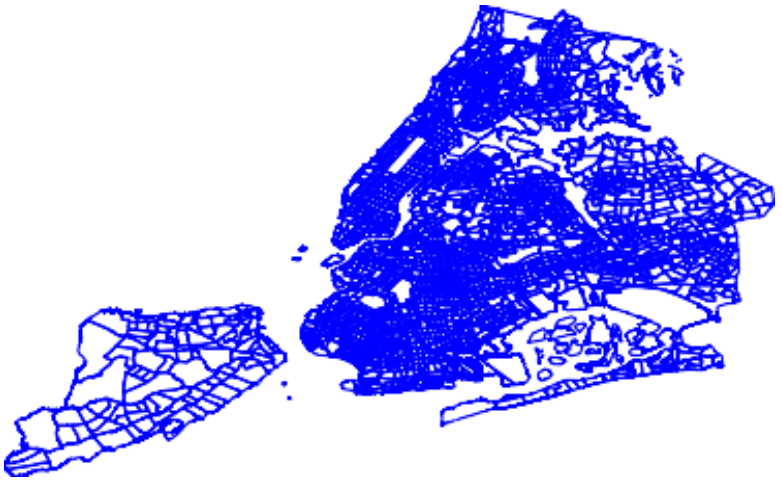
- For reviews see: Gelfand et al. (2001), Gotway and Young, (2002), Wikle and Berliner (2005), and Trevisani and Gelfand (2013).

Spatial Change of Support

(a) Community District Boundaries in NYC

"Target Support"

(b) Census Tract Boundaries in NYC



"Source Support"

(c) NYC PUMA/Community District Overlap

Community districts and census tracts are misaligned. The red lines are the boundaries of aggregate census tracts (PUMAs) and the black lines are the boundaries of community districts

Spatial COS for Count-Valued Survey Data

- **Summary of Methodology:**
  - Use a Bayesian statistical model that incorporates dependencies between different regions.
  - Use survey variances to improve the quality of estimates from our statistical model.
  - The **"bottom-up"** approach is used for COS.
- **Paper:** Bradley, J.R., Wikle C.K., and Holan, S.H. (2016) Bayesian Spatial Change of Support for Count-Valued Survey Data. *Journal of the American Statistical Association*

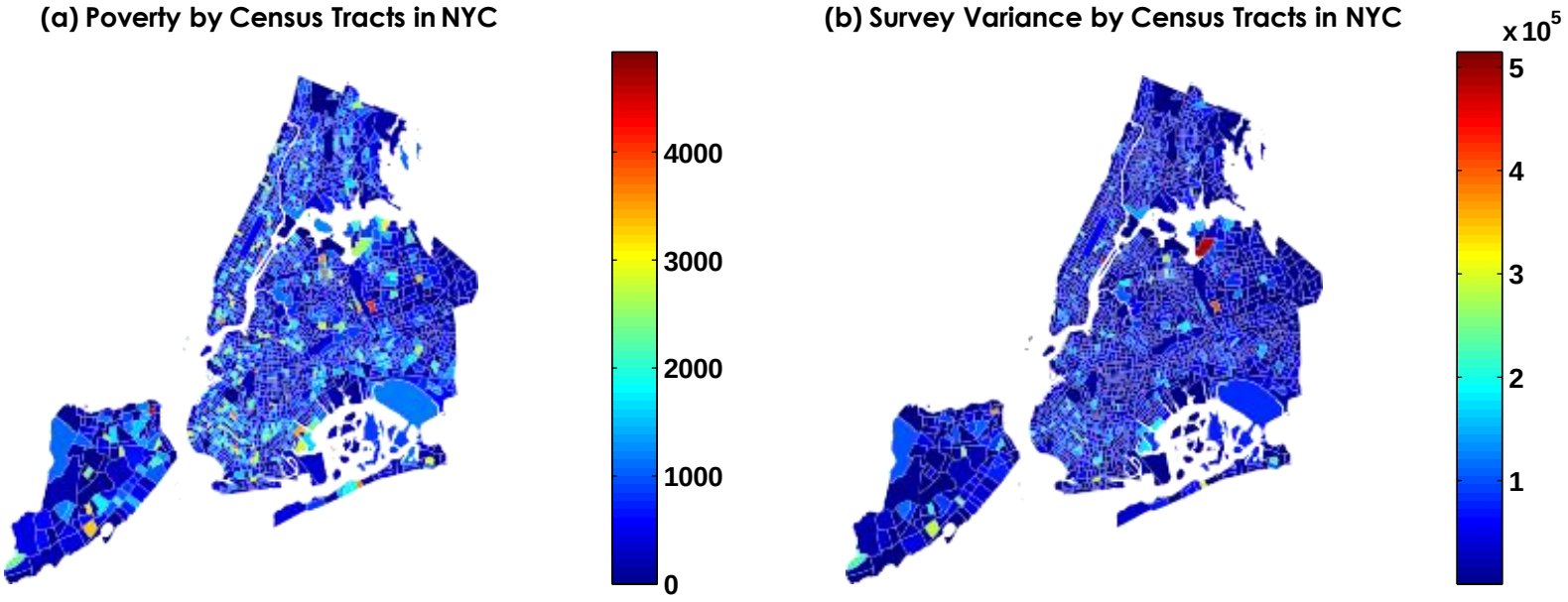
Application to ACS

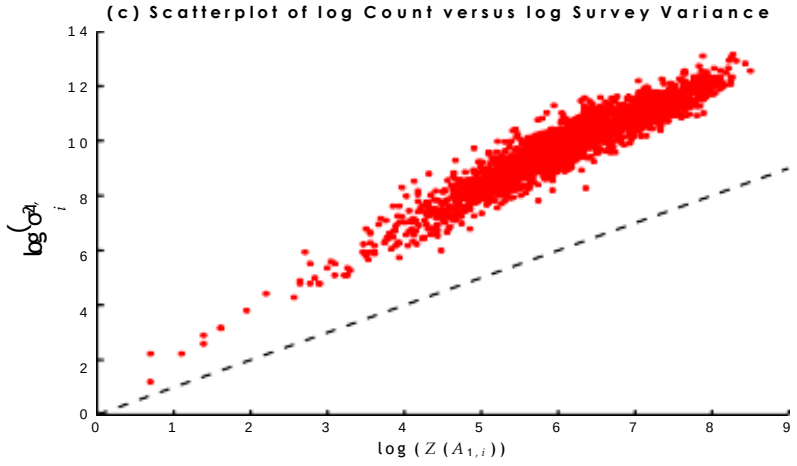
- The Department of City Planning in NYC use ACS period estimates of poverty, demographics, and social characteristics.

- They are interested in obtaining estimates of these variables defined on community districts (target support), but instead use aggregate census tracts (source support), since ACS data are not available on NYC's community districts.

- We use the proposed spatial COS methodology to change the spatial support of the 2012 5-year period estimates of poverty from census tracts (source support) to community districts (target support).

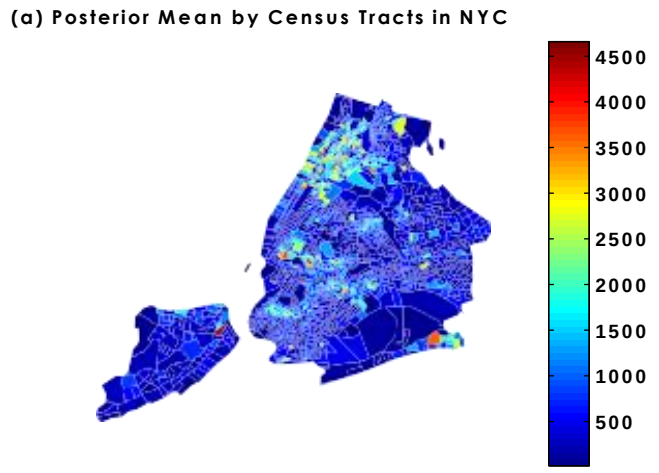
Application to ACS Continued

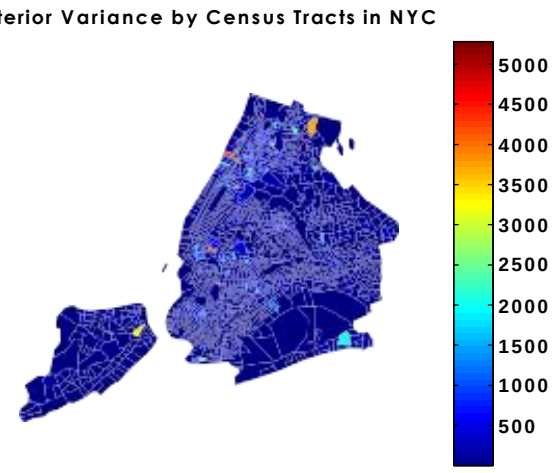


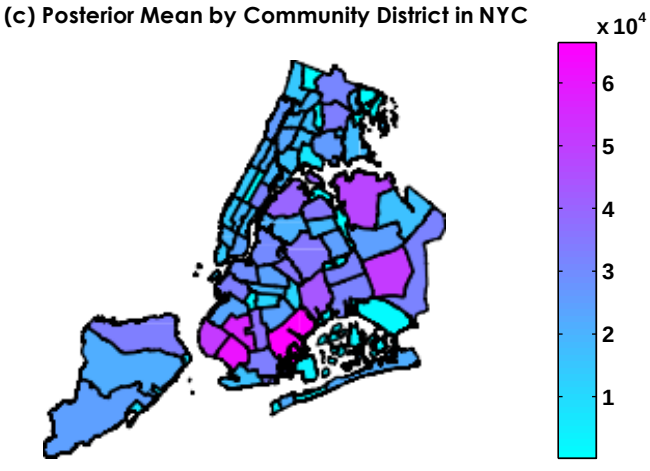


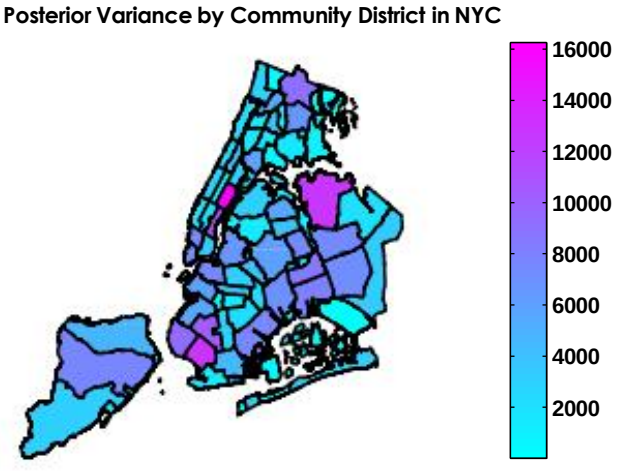
Application to ACS Continued

- Our hierarchical statistical analysis gives the posterior mean and posterior variances of the mean number of people in poverty defined on census tracts and community districts.
- Diagnostic measures were used to ensure that the quality of the estimates were reasonable (see, Bradley et al., 2016, JASA, for more details).









Spatio-Temporal COS for the American Community Survey

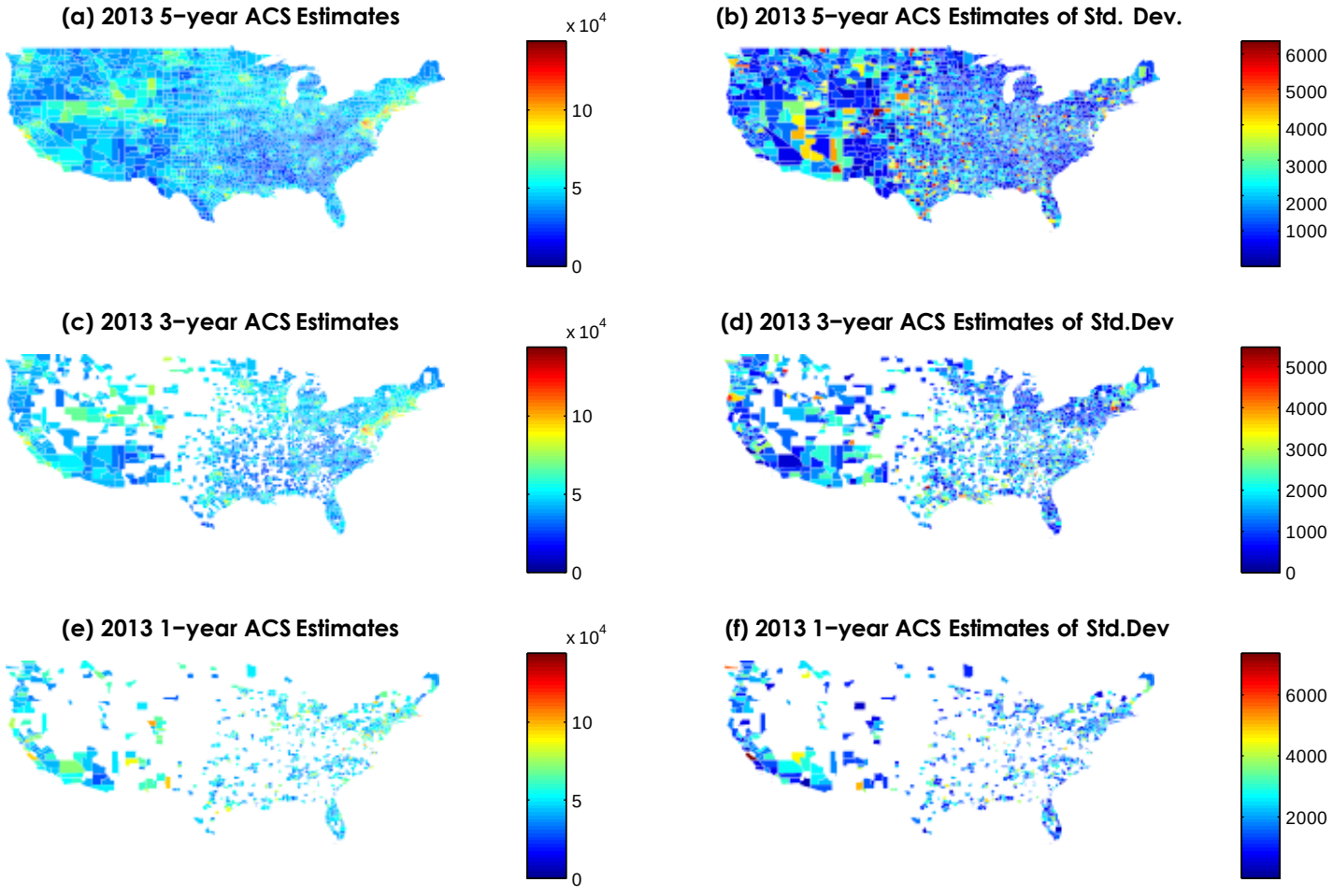
Spatio-Temporal COS is the focus of this talk.

Not only does our methodology allow an ACS user to define their own geography, but they can also **define their own time-period.**

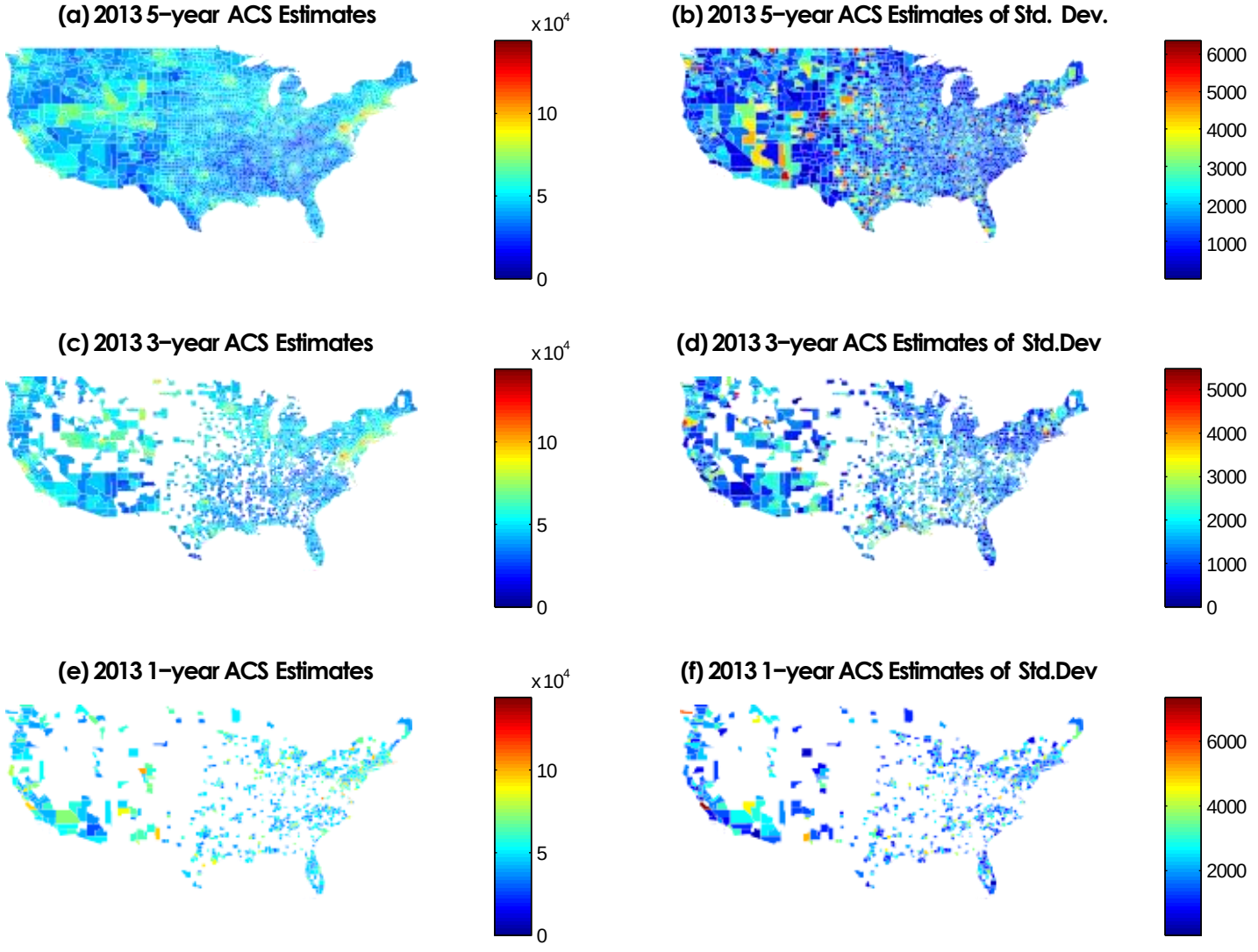
- **Need/Usefulness:**
  - Allows an ACS user to define geographies and time-periods that are meaningful to them.
  - Allows one to compare across different areal units by providing estimates on a common time period.
- **Paper:** Bradley, JR, Wikle, CK, and Holan SH. (2015; *Stat*) Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates.

Estimating 3-Year Period Estimates of Median Household Income

2013 ACS estimates of median household income, and their corresponding survey estimates of standard deviations.

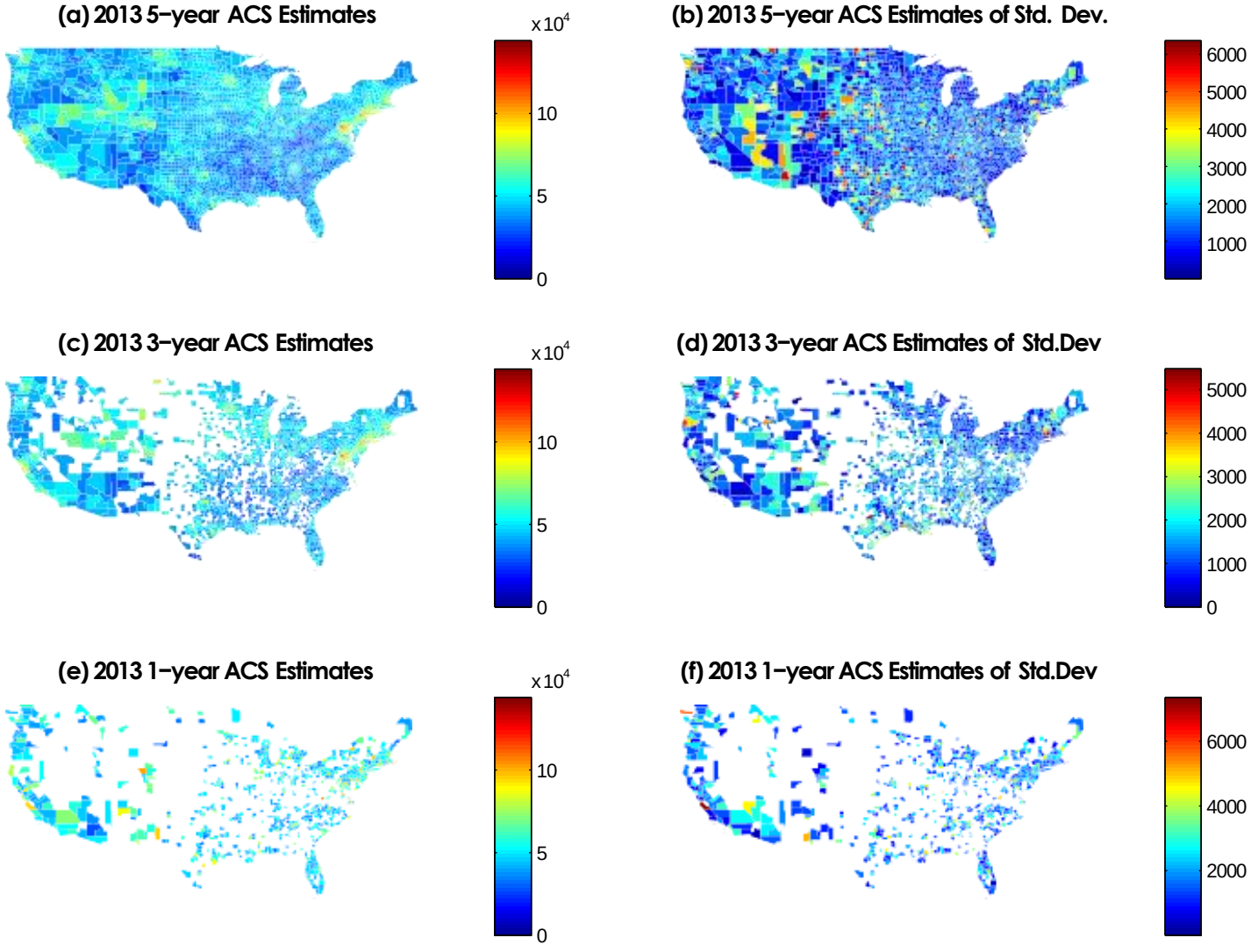


Estimating 3-Year Period Estimates of Median Household Income



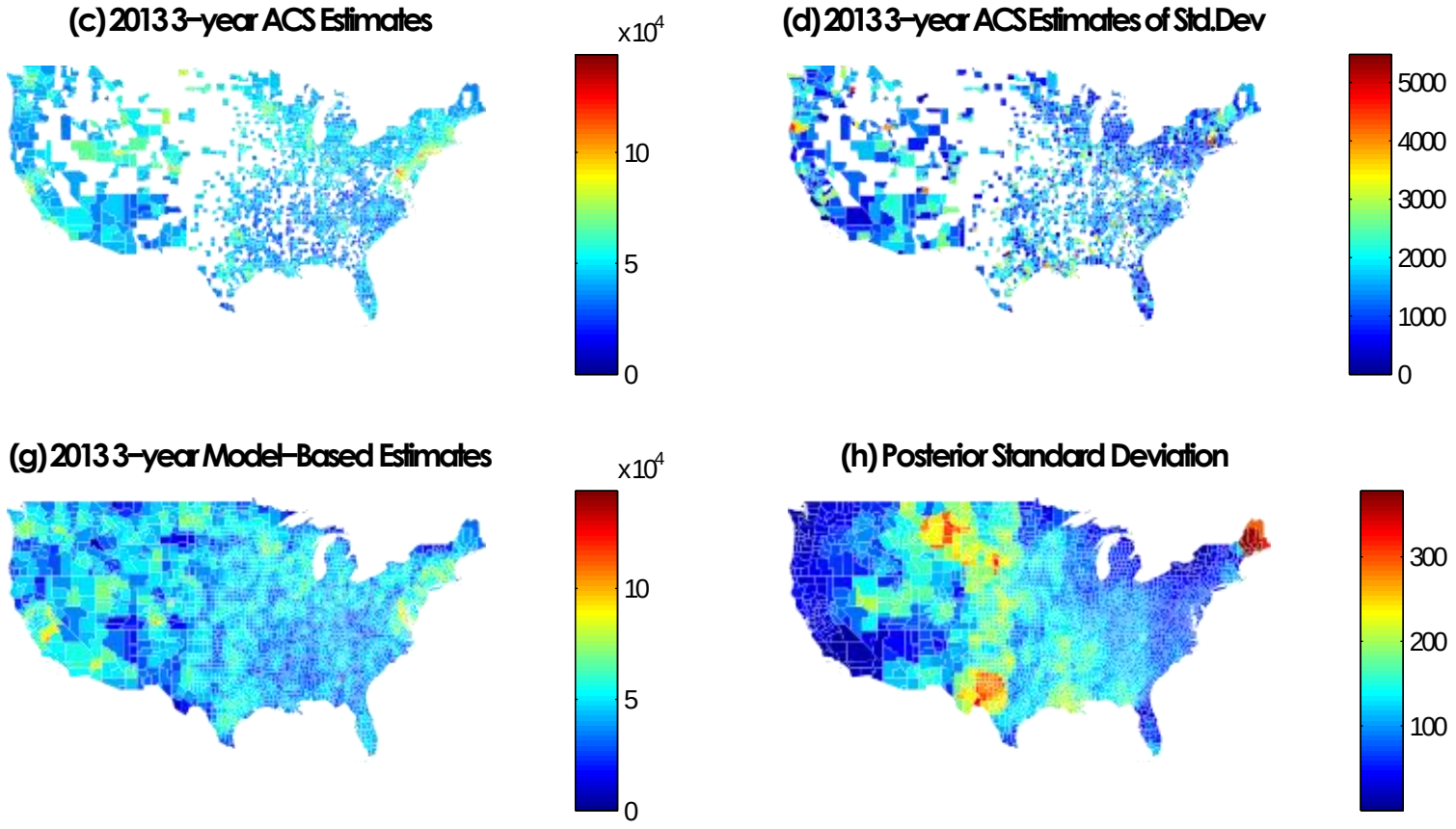
Longer time periods have fewer missing regions.

Estimating 3-Year Period Estimates of Median Household Income



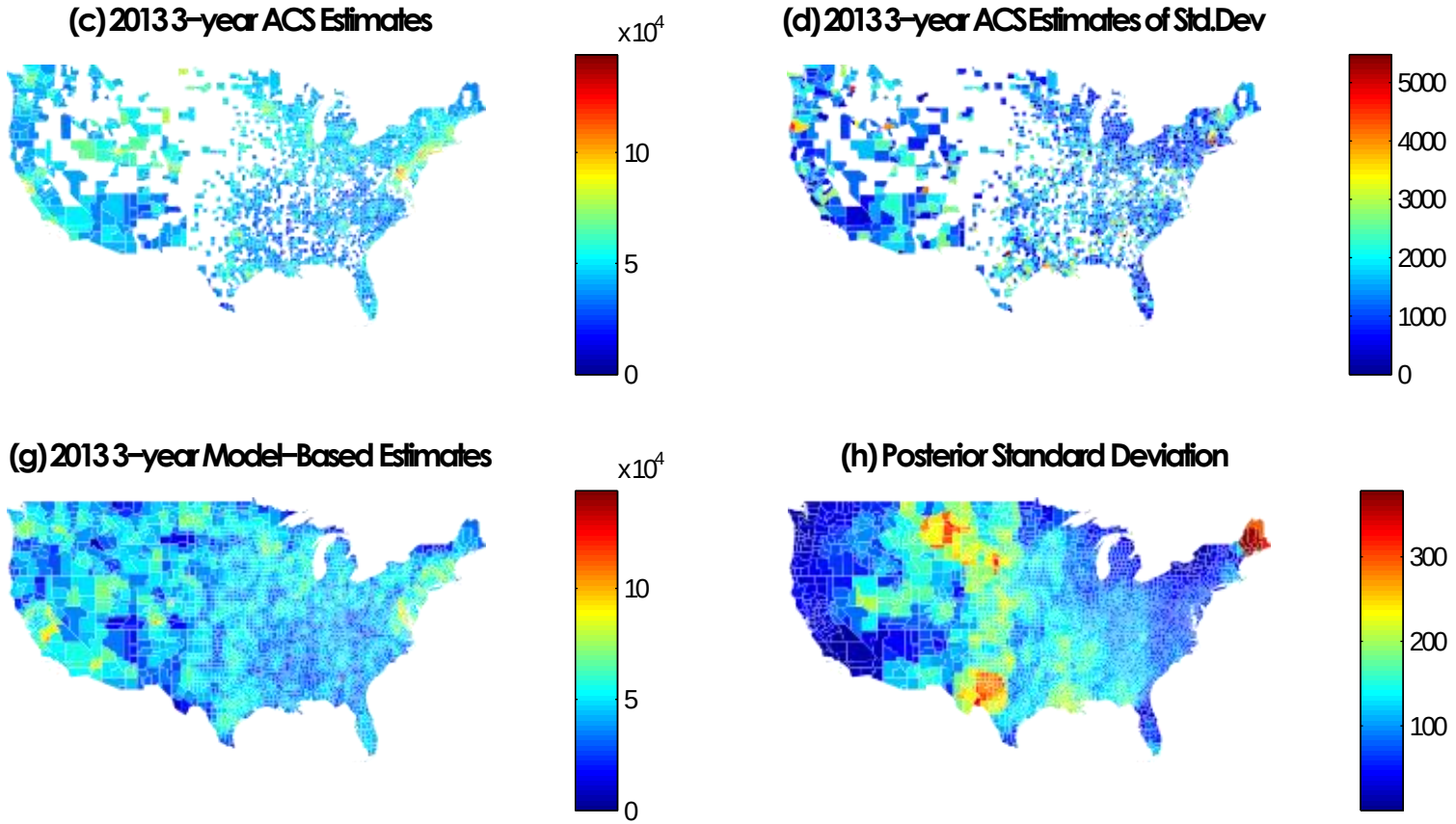
Each period estimate has a relatively large measure of uncertainty.

Estimating 3-Year Period Estimates of Median Household Income



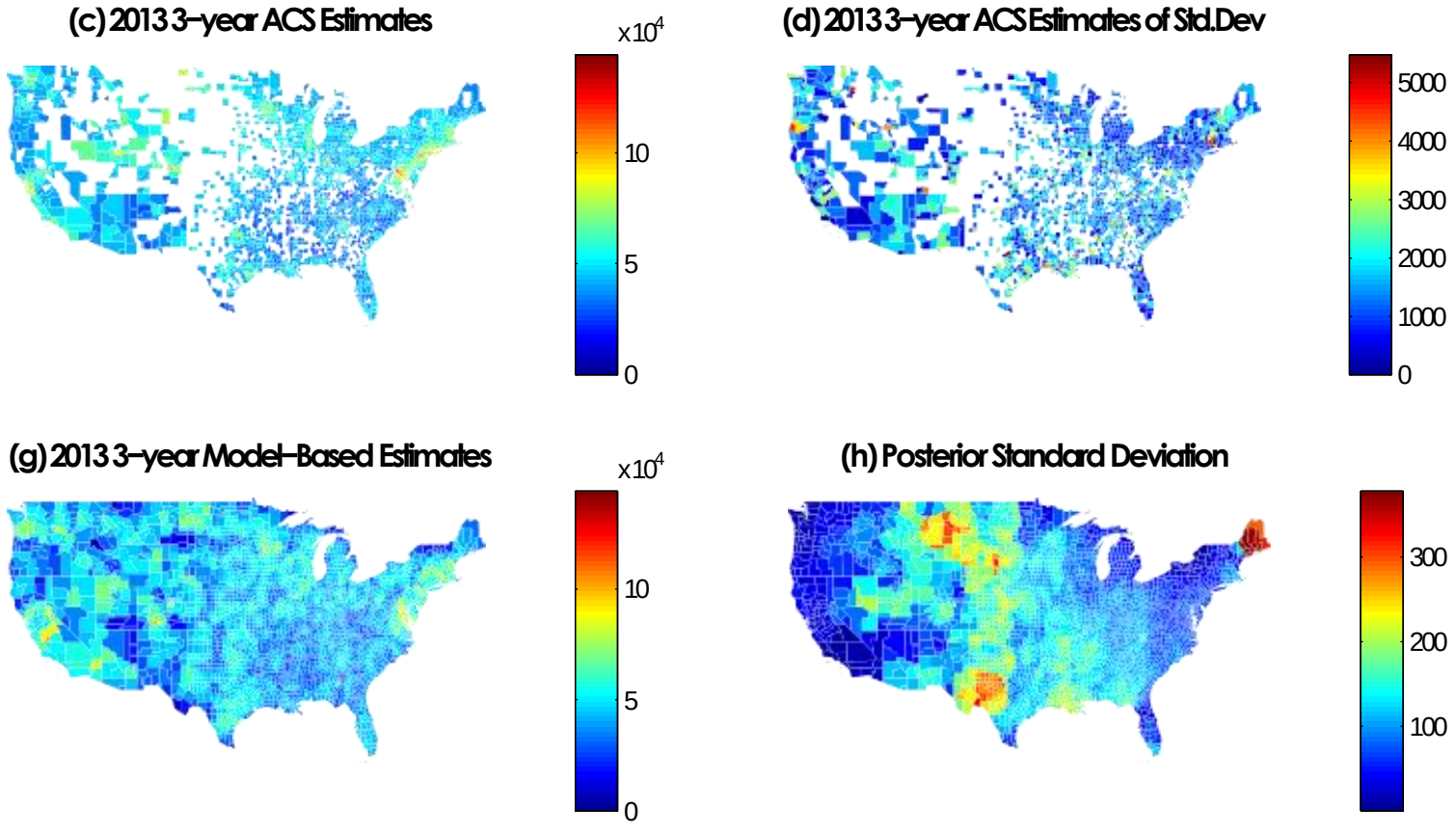
(Bayesian) Model-based based estimates (g) and (h) use the 1-year period estimates and the 5-year period estimates from the previous slide, but do not use the 3-year period estimates.

Estimating 3-Year Period Estimates of Median Household Income



We compute the ratios of the model-based estimates to the "hold-out" 3-year period ACS estimates. The median ratio is 1.04 indicating that the model-based estimates are very close to the "hold-out" 3-year period ACS estimates. See Bradley, Wikle and Holan (2015, *Stat*) for more diagnostic comparisons.

Estimating 3-Year Period Estimates of Median Household Income



The posterior standard deviations are considerably smaller than the standard deviations of the 2013 ACS estimates.

Estimating 3-Year Period Estimates of Median Household Income

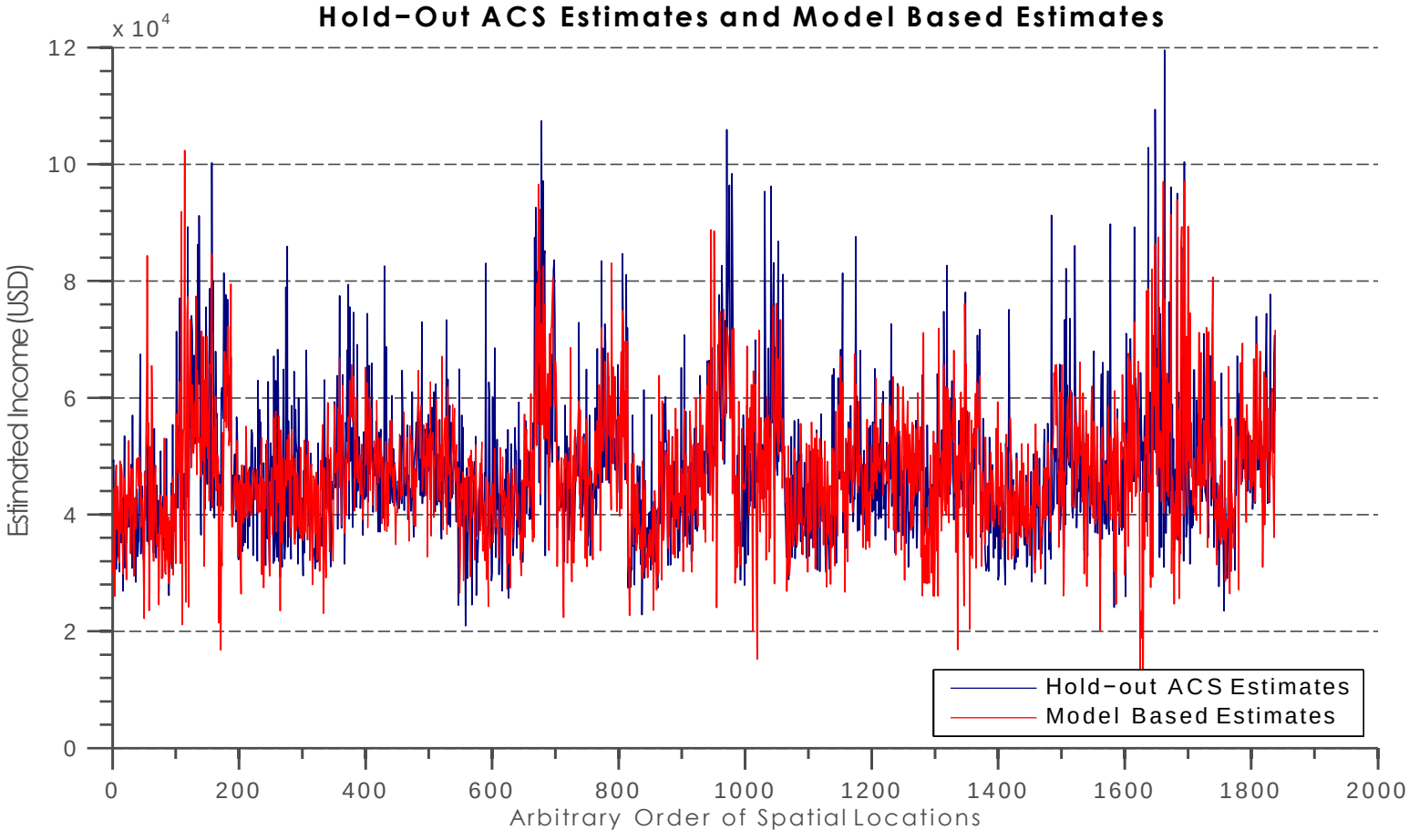
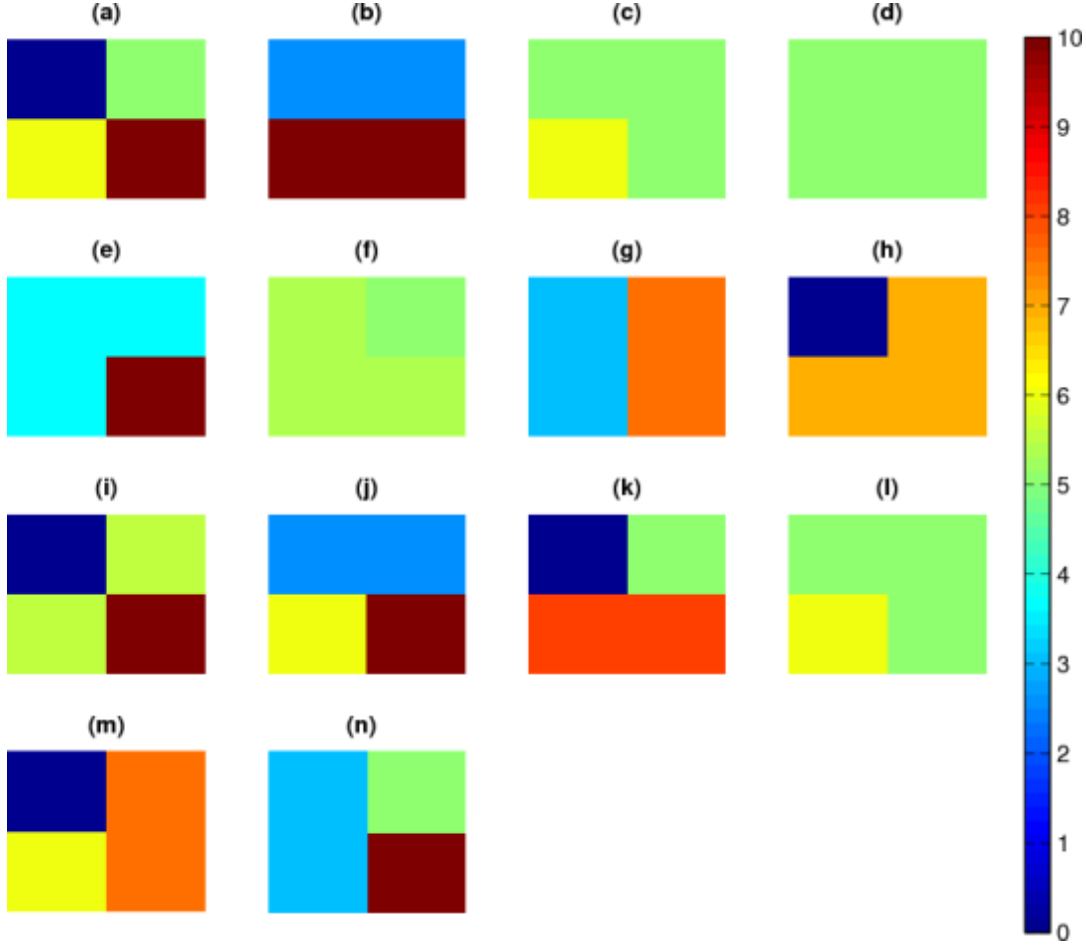


Illustration of Aggregation Error

Aggregation Error: Ecological Fallacy/Modifiable Areal Unit Problem (MAUP): when inference on the aggregate scale of spatial support differs from inference on another distinct spatial support.

Example: (a) truth; (b)-(n) various 2-3 group realizations

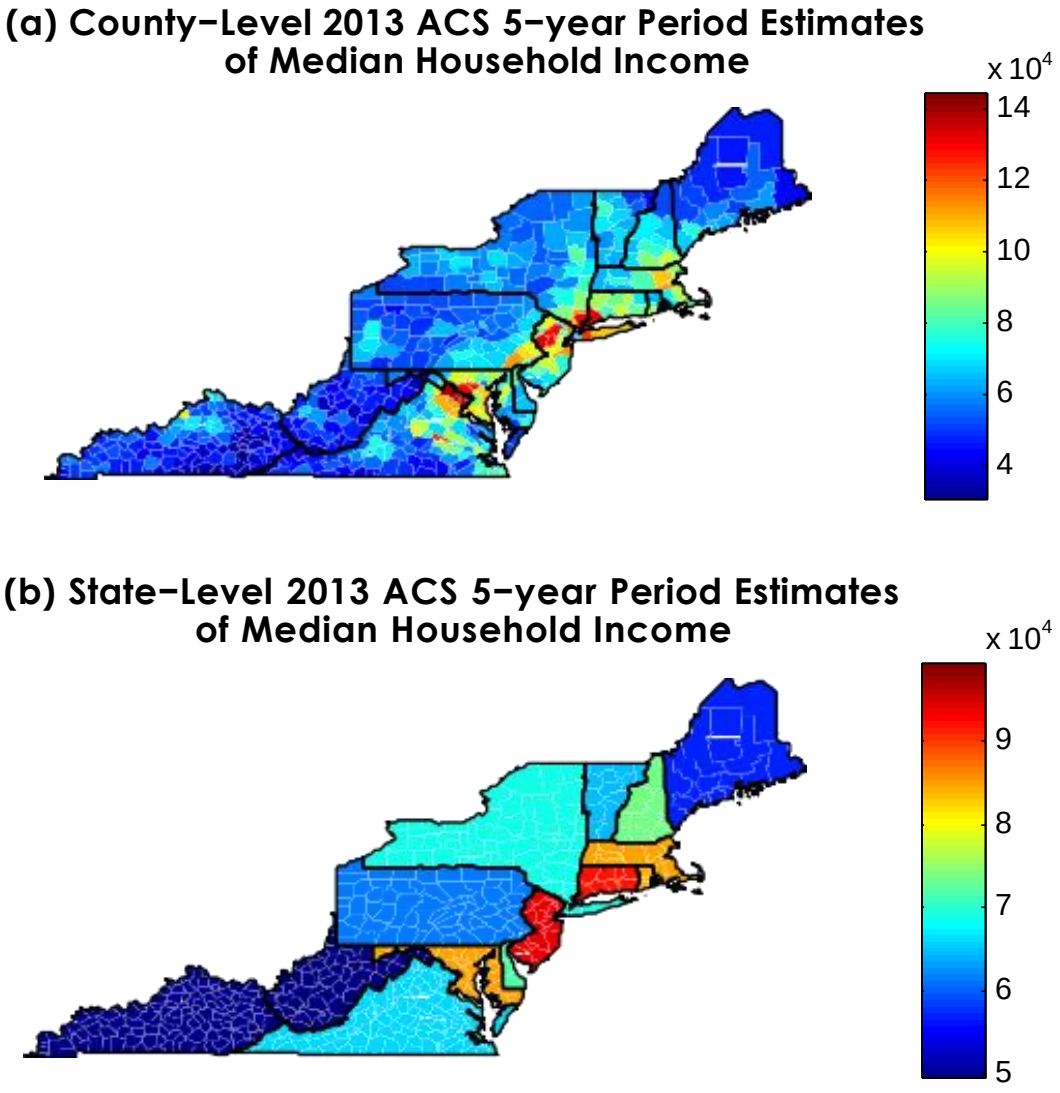
We seek to:
(i)quantify regionalization error,
(ii)select optimal regionalizations that minimize this error!



Another Example

ACS 5-year period estimates of median household income for 2013 over selected states in the NE US.

Panel (a), displays ACS estimates by counties, and panel (b) displays ACS estimates by state. The state boundaries are overlaid in each panel as a reference.



Southern Virginian counties have low median income, while northern Virginian counties have high median income. At the state level this cannot be seen.

Regionalization

- **In Bradley, Wikle, and Holan (2017, JRSS-B) we consider regionalizations motivated by mitigating the Modifiable Areal Unit Problem (MAUP).**

- We develop statistical theory behind the MAUP. The results are quite technical, and are based on a type of functional principal component decomposition called the Karhunen-Loeve expansion.

- These results motivate our criterion, which we call **the criterion for spatial aggregation error (CAGE):**

$$\text{CAGE} = \text{var}\{\text{Fine Scale Process}\} - \text{var}\{\text{Aggregate Scale Process}\}.$$

For example, the "Fine-Scale Process" could be county-level median income, and the "Aggregated-Level Process" could be state-level median income.

Regionalization Cont.

- **Practical Conclusions**
  - CAGE allows us to find optimal (minimizes MAUP) regionalizations.
  - Evaluate the severity of the MAUP for a given spatial domain (i.e., uncertainty quantification).
  - Provides a way for dimension reduction.
- See Bradley, Wikle, and Holan (2017, JRSS-B) Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error.

Discussion

- We have recently developed methodology that provides ways for data-users to:
  - Define their own geographies/time-periods.
  - Quantify the MAUP for a given geography.
  - Find an optimal regionalization.
  - Analyze high-dimensional multivariate spatio-temporal datasets.
- Other topics of interest include”
  - Combining data from multiple sources and different temporal sampling frequencies.
    - Combining data from multiple sources see Bradley, Holan, and Wikle, (2016, Stat), Wang et al. (2011, JABES), among others.
    - Combining data from different temporal sampling frequencies see Holan, Yang, Matteson, and Wikle (2012, ASMBI), Porter, Holan, Wikle, and Cressie (2014, Spatial Statistics), among others.

Thank You!

holans@missouri.edu