



Recent Developments for Space-Time Modeling of Public-Use Federal Data

Jonathan R. Bradley

Department of Statistics

University of Missouri

(bradleyjr@missouri.edu)

Enhancing Spatial and Spatio-Temporal Analysis for Federal
Data via Academic—Research Partnerships: The NSF—Census
Research Network

October 16, 2015

Collaborators

- ▶ **Collaborators:**

- ▶ Scott H. Holan, U. Missouri
- ▶ Christopher K. Wikle, U. Missouri

- ▶ **Support:**

- ▶ NSF-Census Research Network

- ▶ **Spatio-Temporal Statistics NSF-Census Research Network (STSN):**

- ▶ Please visit our website,
<http://stsn.missouri.edu/index.shtml>

Motivation: Example Federal Data

- ▶ **The American Community Survey (ACS):**
 - ▶ An ongoing survey administered by the U.S. Census Bureau that provides timely information on several key demographic variables.
 - ▶ The ACS produces 1-year, 3-year, and 5-year “period-estimates,” and corresponding margins of errors, for the published demographic and socio-economic variables recorded over predefined geographies within the United States.
- ▶ **Quarterly Workforce Indicators (QWI):**
 - ▶ The Longitudinal Employer-Household Dynamics (LEHD) program, managed by the US Census Bureau, produces QWIs for key economic variables.

Motivation: Example Federal Data (Cont'd)

- ▶ Screenshot of <http://factfinder.census.gov/>.

The screenshot shows the American FactFinder search page in a Google Chrome browser. The address bar displays the URL `factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t`. The page header includes the U.S. Department of Commerce logo, the American FactFinder title, and a map of the United States with 'KANSAS' highlighted. Navigation tabs include MAIN, COMMUNITY FACTS, GUIDED SEARCH, **ADVANCED SEARCH**, and DOWNLOAD CENTER. A 'Feedback FAQs Glossary Help' link is also present.

The main content area is titled 'Search - Use the options on the left (topics, geographies, ...) to narrow your search results'. It features a 'Your Selections' box on the left, which is currently empty. Below this are five search filters: Topics (age, income, year, dataset, ...), Geographies (states, counties, places, ...), Race and Ethnic Groups (race, ancestry, tribe), Industry Codes (NAICS industry, ...), and EEO Occupation Codes (executives, analysts, ...).

The central search area contains the following instructions and form:

To search for tables and other files in American FactFinder:

- 1** Enter search terms and an optional geography and click GO

The search form includes two input fields: 'topic or table name' and 'state, county or place (optional)'. A 'GO' button and a help icon are to the right. Below the fields are radio buttons for 'topics', 'race/ancestry', 'industries', and 'occupations', with 'topics' selected. A separator line reads '-- or --'. Below this, it says 'Select from Topics, Race and Ethnic Groups, Industry Codes, EEO Occupation Codes.' followed by two bullet points: '• these are added to 'Your Selections'' and '• the Search Results are updated'.

- 2** Next, select **Geographies** (states, counties, cities, towns, etc.)

This step is followed by the same two bullet points: '• these are added to 'Your Selections'' and '• the Search Results are updated'.

Motivation: Example Federal Data (Cont'd)

The screenshot shows the American FactFinder search interface. On the left, there are navigation buttons for 'Your Selections', 'Search using the options below', 'Topics', 'Geographies', 'Race and Ethnic Groups', 'Industry Codes', and 'EEO Occupation Codes'. The main search area contains a search bar and a 'GO' button. A 'Select Topics' dialog box is open, listing various topics to be added to the search. The topics are grouped into categories: People, Housing, Business and Industry, and Governments. The 'People' category includes: Basic Count/Estimate, Age & Sex, Age Group, Disability, Education, Employment, Income & Earnings, Insurance Coverage, Language, Marital & Fertility Status, Origins, Population Change, Poverty, Relationship, and Veterans. The 'Housing' category includes: Housing. The 'Business and Industry' category includes: Business and Industry. The 'Governments' category includes: Governments. A note at the bottom of the dialog box states: 'Note: The Race & Ethnicity topic is available under the Race and Ethnic Groups button on the left.' There is a checkbox for 'Include archived products in your search'. On the right side of the screen, there is a large yellow rectangular area. Red text 'Continuous ACS Estimates' is overlaid on the right side. Two blue arrows point from this text to the 'Age & Sex' and 'Income & Earnings' topics in the dialog box. The text 'Industry Codes, EEO Occupation Codes.' is also visible on the right side.

Continuous ACS Estimates

Motivation: Example Federal Data (Cont'd)

The screenshot shows the American FactFinder search interface. A 'Select Topics' dialog box is open, listing various data categories. Blue arrows point from several topics in the dialog to a red text box on the right that reads 'Mostly Count-Valued Estimates!'. The topics being pointed to are:

- Basic Count/Estimate
- Age & Sex
- Age Group
- Disability
- Education
- Employment
- Income & Earnings
- Insurance Coverage
- Language
- Marital & Fertility Status
- Origins
- Population Change
- Poverty
- Relationship
- Veterans

The dialog box also includes a note: 'Note: The Race & Ethnicity topic is available under the Race and Ethnic Groups button on the left.' and a checkbox for 'Include archived products in your search'.

Motivation: Spatial Statistics for Federal Data

- ▶ Federal data sources, such as ACS, share many common features.
 1. Very large datasets (**on the order of millions**).
 2. Multivariate data (**with a large number of variables**).
 3. Spatial referenced data over geographic regions (i.e., counties, states, etc.).
 4. Data recorded over discrete time.
 5. The data are often count-valued.
 6. There are multiple spatial-temporal scales.

Point of the talk:

- ▶ **Overview Talk:** Discuss some of the problems/applications that the **STSN node of the NCRN** has been working on.
- ▶ **Focus:** Multiscale Spatio-Temporal Analysis, and Multivariate Spatio-Temporal prediction.

Statistical Hierarchical Models

A Latent Gaussian Process Model

$$\text{Data Model : } \prod_{i=1}^n [Z_i | Y_i, \boldsymbol{\theta}_D];$$

$$\text{Process Model : Gaussian}(\mathbf{Y} | \boldsymbol{\mu}, \boldsymbol{\Sigma});$$

$$\text{Parameter Model : } [\boldsymbol{\mu}, \boldsymbol{\Sigma}],$$

where $\mathbf{Y} \equiv (Y_1, \dots, Y_n)'$.

- ▶ Z_i : i -th data point.
- ▶ Y_i : i -th latent quantity of interest.
- ▶ $\boldsymbol{\theta}_D$ is a set of real-valued data parameters.
- ▶ \mathbf{Y} is n -dimensional with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- ▶ See Cressie and Wikle (2011)

Statistical Hierarchical Models (Cont'd)

- ▶ When the dataset is continuous and normal it is typically assumed,

$$[Z_i|Y_i, \theta_D = \sigma^2] = \text{Gaussian}(Y_i, \sigma^2).$$

Main practical difficulty in fitting this HM:

- ▶ **Inverting a large $n \times n$ matrix Σ .**
- ▶ When the dataset is count-valued it is typically assumed,

$$[Z_i|Y_i, \theta_D = \emptyset] = \text{Poisson} \{ \exp(Y_i) \}.$$

Main practical difficulties in fitting this HM:

- ▶ Inverting a large $n \times n$ matrix Σ .
- ▶ **This distribution theory is computationally more difficult to use than the Gaussian data model/Gaussian process model specification.**

Important Problems for Federal Statistics

- ▶ **Multiscale Spatio-Temporal Analysis:**

- ▶ **Change of Support:** Can one produce estimates on user-defined geographies, and user-defined time-periods?
- ▶ **Regionalization:** Is there a “best” spatial support?

- ▶ **Multivariate Spatio-Temporal prediction:**

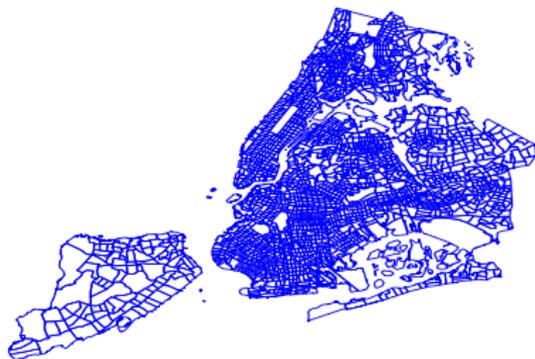
- ▶ **Leveraging Information:** Can one use dependencies between variables, times, and regions to predict “missing” values for federal Data?

Multiscale Spatio-Temporal Analysis: Spatial Change of Support

(a) Community District Boundaries in NYC



(b) Census Tract Boundaries in NYC



(c) NYC PUMA/Community District Overlap



Change of Support

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

1. **Bottom-up**: Define the stochastic process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, where $D \subset \mathbb{R}^d$ is a generic spatial domain. Then, COS is computed via the integral

$$Y(A) = \frac{1}{|A|} \int_A Y(\mathbf{s}) d\mathbf{s},$$

where $|A|$ is the cardinality of the set $A \in D$.

2. **Top-down**: Define the process by a partitioning of the source support and target support (Mugglin et al., 1998).
- ▶ For reviews see: Gelfand et al. (2001), Wikle and Berliner (2005), Gotway and Young (2002), and Trevisani and Gelfand (2013).

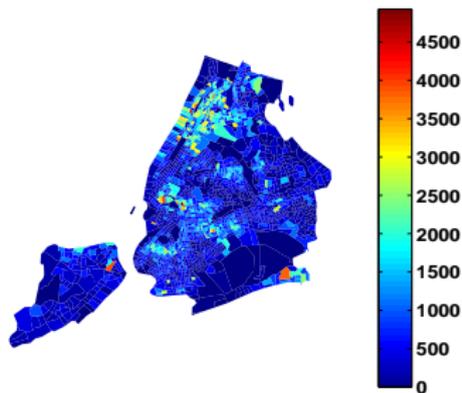
Spatial COS Count-Valued Survey Data (Bradley et al., 2014)

▶ **Summary of Methodology:**

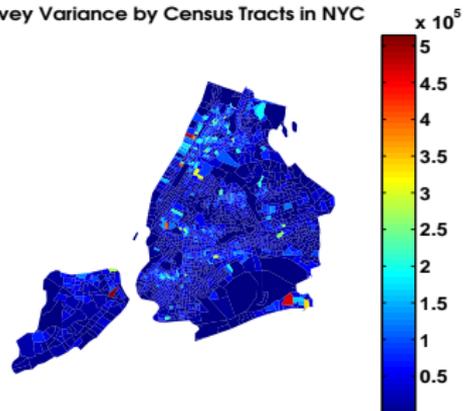
- ▶ Use a Poisson data model and a latent Gaussian process.
 - ▶ Include the sampling distribution of survey variances as a data model.
 - ▶ Introduce an extension of the Givens angle prior from Yang and Berger (1994).
 - ▶ Use the “bottom-up” approach for COS.
-
- ▶ **Paper:** Bradley, JR, Wikle, CK, and Holan, SH. (2015). Bayesian Spatial Change of Support for Count-Valued Survey Data. *arXiv preprint: 1405.7227. (Invited Revision – Journal of the American Statistical Association)*

Spatial COS for ACS

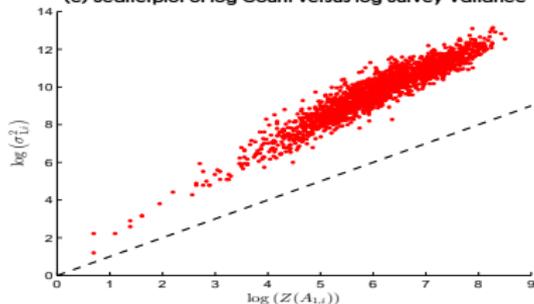
(a) Poverty by Census Tracts in NYC



(b) Survey Variance by Census Tracts in NYC



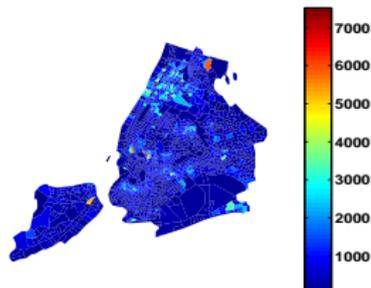
(c) Scatterplot of log Count versus log Survey Variance



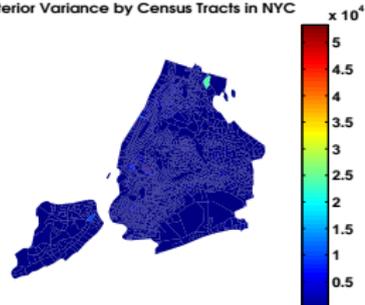
Spatial COS for ACS (Cont'd)

- ▶ The posterior predictive p-value (using the likelihood ratio as the discrepancy measure) is 0.60, which indicates no lack of fit; i.e., that we are obtaining a reasonable fit to the data.

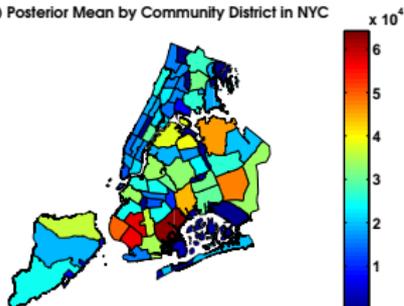
(a) Posterior Mean by Census Tracts in NYC



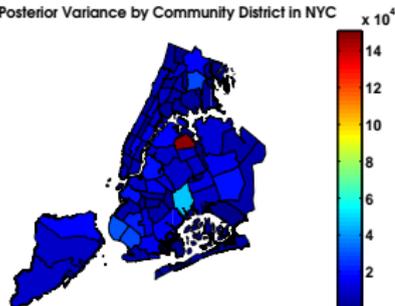
(b) Posterior Variance by Census Tracts in NYC



(c) Posterior Mean by Community District in NYC

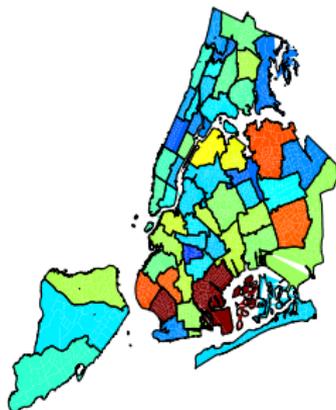


(d) Posterior Variance by Community District in NYC

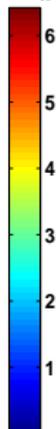


Spatial COS for ACS (Cont'd)

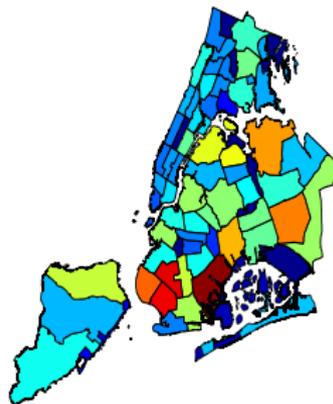
DCP estimated poverty (PUMA)



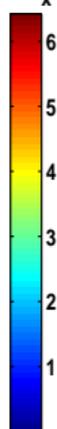
$\times 10^4$



(c) Posterior Mean by Community District in NYC



$\times 10^4$



Spatial COS for ACS (Cont'd)

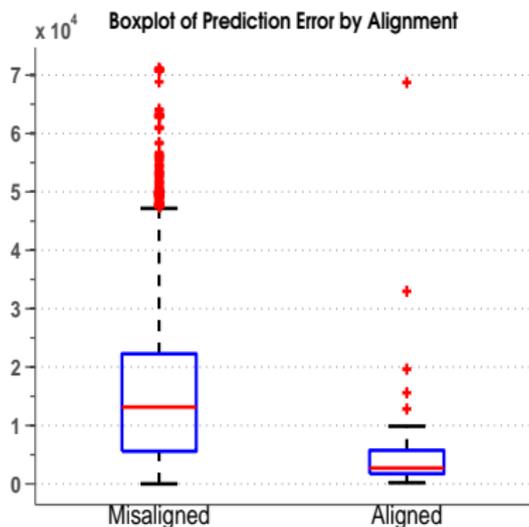


Figure: A generic community district B and a generic PUMA C are considered aligned if $0.7 \leq |B \cap C|/|B| \leq 1.3$, and are considered misaligned otherwise. Here we provide boxplots of $\text{abs} \{ E_{\text{CS}}(\mu(B)|\mathbf{Z}) - \hat{\mu}^{\text{DCP}}(C) \}$ by this categorization of alignment.

Spatio-Temporal COS for the American Community Survey (Bradley et al., 2015c)

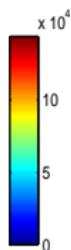
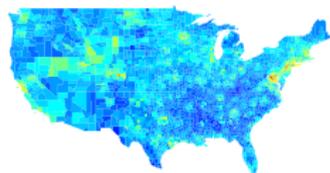
- ▶ **Need/Usefulness:**

- ▶ An ACS user might like to define their own geography, and define their own time-period.
- ▶ An ACS user might like to compare across different areal units.
- ▶ The latent Gaussian process framework leads to smaller measures of error.

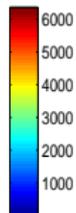
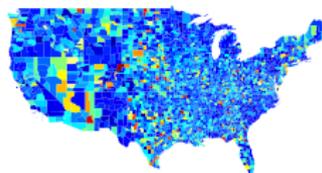
- ▶ **Paper:** Bradley, JR, Wikle, CK, and Holan SH. (2015 – To Appear). Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates. *Stat.*

Spatio-Temporal COS for ACS

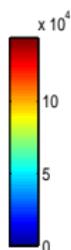
(a) 2013 5-year ACS Estimates



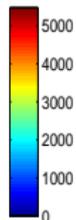
(b) 2013 5-year ACS Estimates of Std. Dev.



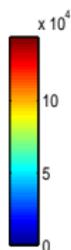
(c) 2013 3-year ACS Estimates



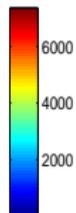
(d) 2013 3-year ACS Estimates of Std.Dev



(e) 2013 1-year ACS Estimates

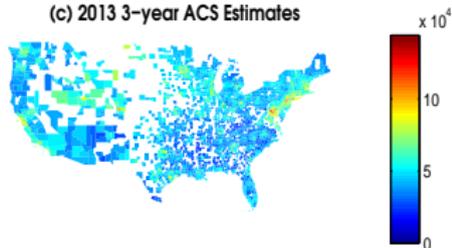


(f) 2013 1-year ACS Estimates of Std.Dev



Spatio-Temporal COS for ACS (Cont'd)

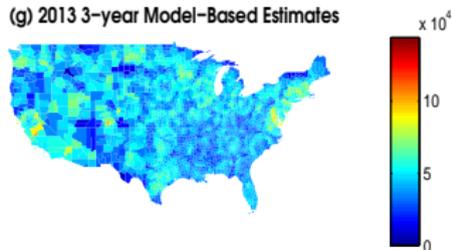
(c) 2013 3-year ACS Estimates



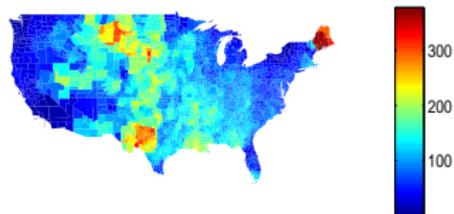
(d) 2013 3-year ACS Estimates of Std.Dev



(g) 2013 3-year Model-Based Estimates



(h) Posterior Standard Deviation



Important Problems for Federal Statistics

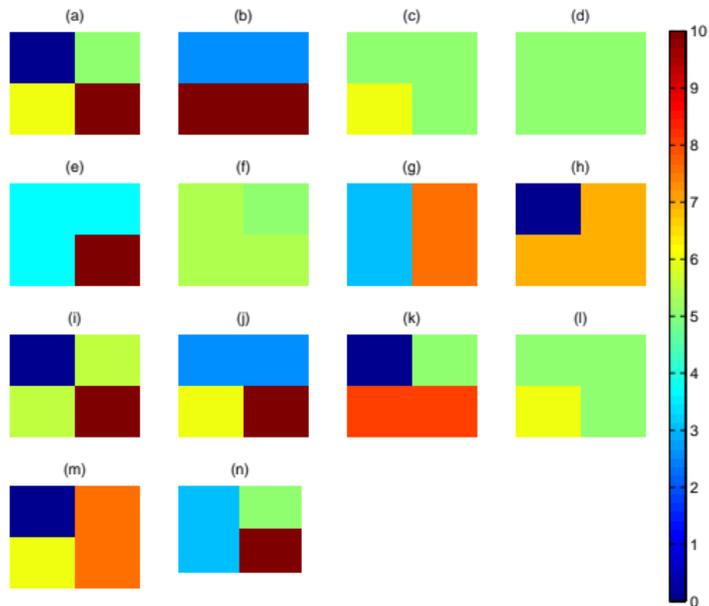
- ▶ **Multiscale Spatio-Temporal Analysis:**

- ▶ **Change of Support:** Can one produce estimates on user-defined geographies, and user-defined time-periods?
- ▶ **Regionalization:** Is there a “best” spatial support?

- ▶ **Multivariate Spatio-Temporal prediction:**

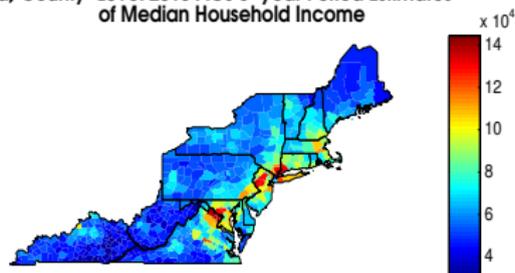
- ▶ **Leveraging Information:** Can one use dependencies between variables, times, and regions to predict “missing” values for federal Data?

Simple Example

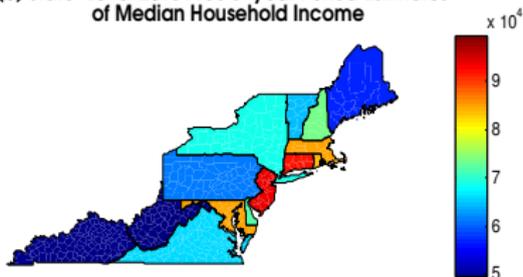


Another Example

(a) County-Level 2013 ACS 5-year Period Estimates of Median Household Income



(b) State-Level 2013 ACS 5-year Period Estimates of Median Household Income



(a), ACS estimates by counties (b) ACS estimates by state. Notice that the color-scales are different for each panel.

Regionalization (Bradley et al., 2015b)

The Karhunen-Loeve Expansion:

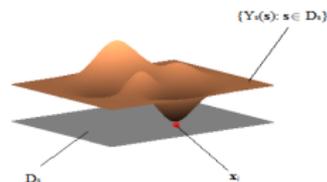
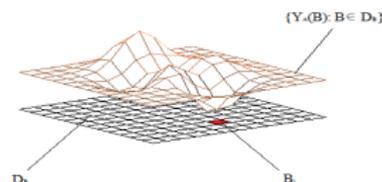
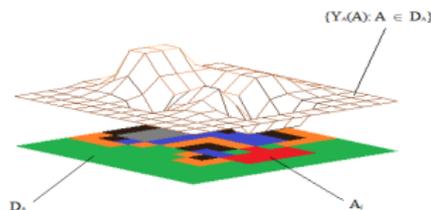
$$Y(\mathbf{s}) = \sum_j \phi_j(\mathbf{s}) \alpha_j; \quad \mathbf{s} \in D,$$

- ▶ where $\{\alpha_j : j = 1, 2, \dots\}$ are uncorrelated with variances $\{\lambda_j : j = 1, 2, \dots\}$ (eigenvalues);
- ▶ the orthonormal real-valued functions $\{\phi_j(\mathbf{s}) : j = 1, 2, \dots\}$ (eigenfunctions) have domain D , and satisfy a Fredholm integral equation.

When does Spatial Aggregation Error Occur?:

$$Y(\mathbf{s}) - Y(A)$$

$$= \sum_j \left\{ \phi_j(\mathbf{s}) - \frac{1}{|A|} \int_A \phi_j(\mathbf{u}) d\mathbf{u} \right\} \alpha_j; \quad \mathbf{s} \in D.$$



Regionalization (Bradley et al., 2015b)

- ▶ We define the criterion for spatial aggregation error (CAGE) to measure the amount of spatial aggregation error as

$$\text{CAGE}(A) = \frac{1}{|A|} \sum_j \int_A \left\{ \phi_j(\mathbf{s}) - \frac{1}{|A|} \int_A \phi_j(\mathbf{u}) d\mathbf{u} \right\}^2 \lambda_j.$$

- ▶ **Regionalization Algorithm:**

1. Step 1: Obtain M MCMC replicates of $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ using a latent Gaussian process model.
2. Step 2: Apply a clustering algorithm to each of the M MCMC replicates of $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$. This yields a total of M candidate regionalizations.
3. Step 3: Choose the regionalization, from among the M candidates in Step 2, that minimizes CAGE.

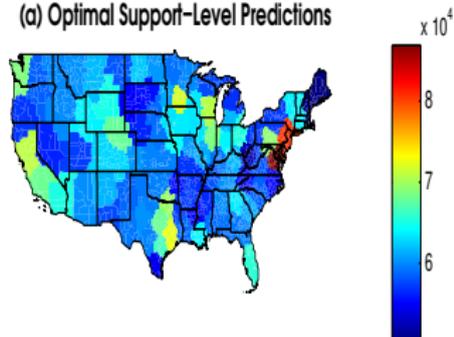
Regionalization (Bradley et al., 2015b)

- ▶ **Practical Conclusions:**

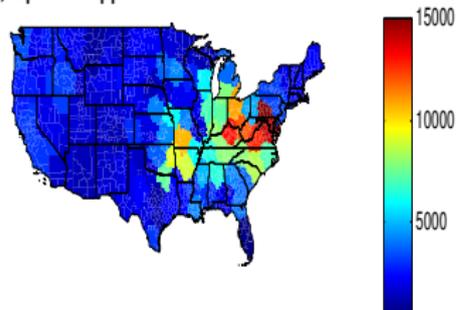
- ▶ CAGE allows us to find *optimal* regionalizations.
 - ▶ Evaluate the MAUP/ecological fallacy for a given spatial domain (i.e., uncertainty quantification).
 - ▶ Provides a way for dimension reduction.
-
- ▶ **Paper:** Bradley, JR, Wikle, CK, and Holan, SH. (2015). Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error. *arXiv preprint: 1502.01974*. (Invited Revision – *Journal of the Royal Statistics Society: Series B*)

ACS Example

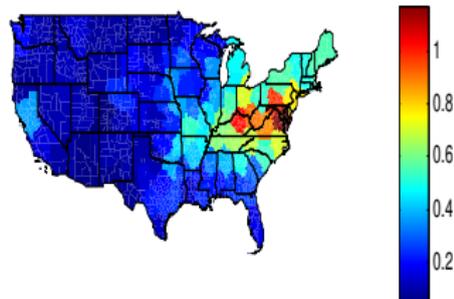
(a) Optimal Support-Level Predictions



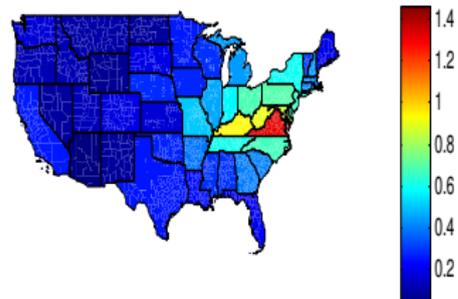
(b) Optimal Support-Level Root Prediction Error



(c) Square Root CAGE for Optimal Support



(d) Square Root CAGE for each State



Important Problems for Federal Statistics

- ▶ **Multiscale Spatio-Temporal Analysis:**

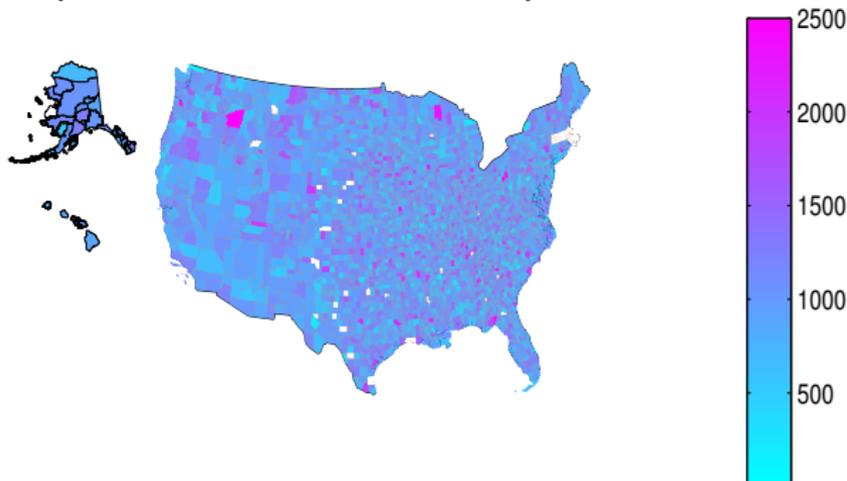
- ▶ **Change of Support:** Can one produce estimates on user-defined geographies, and user-defined time-periods?
- ▶ **Regionalization:** Is there a “best” spatial support?

- ▶ **Multivariate Spatio-Temporal prediction:**

- ▶ **Leveraging Information:** Can one use dependencies between variables, times, and regions to predict “missing” values for federal Data?

Quarterly Workforce Indicators (QWIs)

Monthly Income For Women (Education Industry, 3rd Quarter of 2006)



The Longitudinal Employer-Household Dynamics (LEHD) program (US Census Bureau) produces QWIs for key economic variables for

- ▶ Each quarter in the years 1990–2013 (**92 discrete time-points**).
- ▶ Each of the **3,145 US counties**.
- ▶ Different genders and industries (**20 industries**).

Quarterly Workforce Indicators (QWIs)

► Need for Multivariate Spatio-Temporal Predictions:

1. **Memorandum of Understanding:** 35% of the QWIs are missing.
2. **Uncertainty Measurements:** Uncertainty measures are not made publicly available. Consequently, it is difficult for QWI data-users to assess the quality of the published estimates.
3. QWIs have had a significant impact on the economics literature: for example, see Davis et al. (2006), Thompson (2009), Dube et al. (2013), Allegretto et al. (2013), among others.

Quarterly Workforce Indicators (QWIs)

► **Big Data:**

1. There are a total of 7,530,037 observations;
2. There are a total of 3,680 spatial fields;
3. There are $2 \times 20 \times 3145 \times 92 = 11,573,600$ possible gender/industry/space/time combinations.

► **Complex Dependencies:**

1. Interactions between space, time, and variables.
2. Nonstationary in space and time.
3. Asymmetric Dependencies

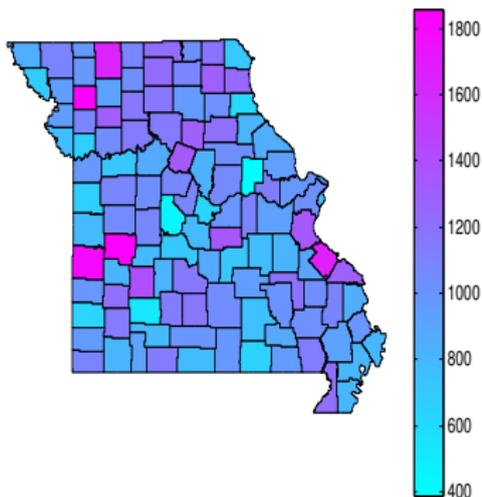
Multivariate Spatio Temporal Prediction for Gaussian Data (Bradley et al., 2015a)

▶ **Summary of Methodology:**

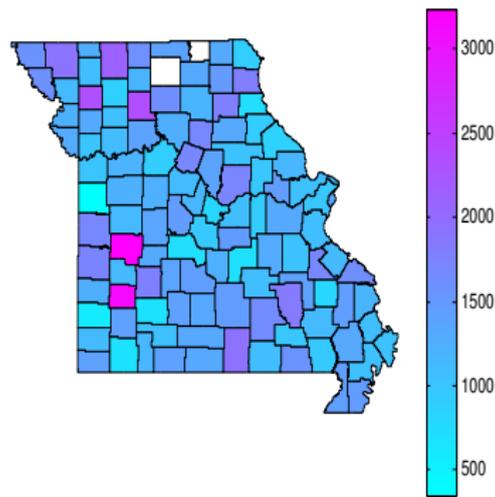
- ▶ We use a mixed effects modeling framework that we call the multivariate spatio-temporal mixed effects model (MSTM).
- ▶ Use covariate information to define temporal dynamics. This results in something we call the Moran's I propagator matrix.
- ▶ Introduce an extension of the Moran's I prior distribution from Hughes and Haran (2013).
- ▶ **Paper:** Bradley, JR, Holan, SH, and Wikle, CK. (2015 – To Appear). Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics. *The Annals of Applied Statistics*.

Predicting Quarterly Average Monthly Income

(a) Obs. Income For Women (Edc. Industry, 1st Quart. of 2013)



(b) Obs. Income For Men (Edc. Industry, 1st Quart. of 2013)



(a) and (b) present the QWI for quarterly average monthly income (US dollars) for the state of Missouri, for each gender, for the education industry, and for the first quarter of 2013. LEHD does not provide estimates at every county in the US at every quarter; these counties are shaded white.

Predicting Quarterly Average Monthly Income (Continued)

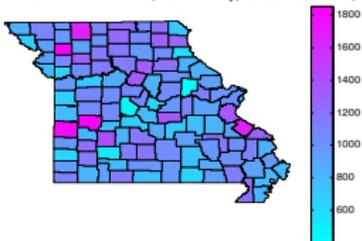
(a) Obs. Income For Women (Edc. Industry, 1st Quart. of 2013)



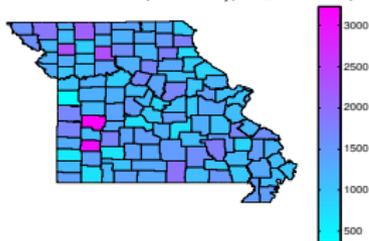
(b) Obs. Income For Men (Edc. Industry, 1st Quart. of 2013)



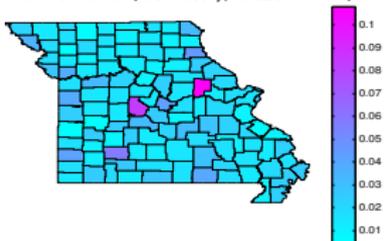
(c) Predicted Income For Women (Edc. Industry, 1st Quart. of 2013)



(d) Predicted Income For Men (Edc. Industry, 1st Quart. of 2013)



(e) Root MSPE For Women (Edc. Industry, 1st Quart. of 2013)



(f) Root MSPE For Men (Edc. Industry, 1st Quart. of 2013)



Discussion

- ▶ We described important features of public-use federal datasets from the point-of-view of spatio-temporal statistics.
- ▶ Described some recent work by the Spatio-Temporal Statistics node of the NCRN at the University of Missouri:
 - ▶ Multiscale Spatio-temporal Analysis
 - ▶ Multivariate Spatio-Temporal Prediction
- ▶ The recent research at STSN provides ways for data-users to:
 - ▶ Define their own geographies/time-periods.
 - ▶ Quantify the MAUP/ecological fallacy for a given geography.
 - ▶ Find an optimal regionalization.
 - ▶ Analyze high-dimensional multivariate spatio-temporal datasets.
- ▶ We have developed distribution theory to extend the MSTM to the **Poisson case** (Under Review) – **Stay Tuned!**

- Allegretto, S., Dube, A., Reich, M., and Zipperer, B. (2013). "Credible research designs for minimum wage studies." In *Working Paper Series*, 1–63. Institute for Research on Labor and Employment.
- Bradley, J., Holan, S., and Wikle, C. (2015a). "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics." *The Annals of Applied Statistics*, To Appear.
- Bradley, J., Wikle, C., and Holan, S. (2014). "Bayesian spatial change of support for count-valued survey data." *arXiv preprint: 1405.7227*.
- (2015b). "Regionalization of multiscale spatial processes using a criterion for spatial aggregation error." *arXiv preprint: 1502.01974*.
- (2015c). "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates." *Stat*, To Appear.
- Cressie, N. and Johannesson, G. (2008). "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society, Series B*, 70, 209–226.
- Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.
- Davis, E., Freedman, M., Lane, J., McCall, B., Nestoriak, N., and Park, T. (2006). "Supermarket human resource practices and competition from mass merchandisers." *American Journal of Agricultural Economics*, 88, 1289–1295.
- Dube, A., Lester, T., and Reich, M. (2013). "Minimum wage, labor market flows, job turnover, search frictions, monopsony, unemployment." In *Working Paper Series*, 1–63. Institute for Research on Labor and Employment.
- Gelfand, Zhu, L., and Carlin, B. (2001). "On the change of support problem for spatio-temporal data." *Biostatistics*, 2, 31–45.
- Gotway and Young (2002). "Combining incompatible spatial data." *Journal of the American Statistical Association*, 97, 632–648.
- Higham, N. (1988). "Computing a nearest symmetric positive semidefinite matrix." *Linear Algebra and its Applications*, 105, 103–118.
- Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed model." *Journal of the Royal Statistical Society, Series B*, 75, 139–159.
- Mugglin, A., Carlin, B., Zhu, L., and Conlon, E. (1998). "Bayesian areal interpolation, estimation, and smoothing: An inferential approach for Geographic Information Systems." *Environment and Planning A*, 31, 1337–1352.
- Oehlert, G. (1992). "A note on the delta method." *The American Statistician*, 46, 27–29.
- Porter, A., Holan, S. H., and Wikle, C. K. (2015). "Bayesian semiparametric hierarchical empirical likelihood spatial models." *Journal of Statistical Planning and Inference*, 165, 78–90.
- Thompson, J. (2009). "Using local labor market data to re-examine the employment effects of the minimum wage." *Industrial and Labor Relations Review*, 63, 343–366.
- Travisani, M. and Gelfand, A. (2013). "Sampling designs and prediction methods for Gaussian spatial processes." In *Advances in Theoretical and Applied Statistics*, eds. N. Torelli, F. Pesarin, and A. Bar-Hen, 269–279. Springer-Verlag Berlin Heidelberg.
- Wikle, C. and Berliner, M. (2005). "Combining information across spatial scales." *Technometrics*, 47, 80–91.
- Yang, R. and Berger, J. (1994). "Estimation of a covariance matrix using the reference prior." *Annals of Statistics*, 22, 1195–1211.

Thank you!

(bradleyjr@missouri.edu)