

# Advances and Future Directions in Spatial and Spatio-Temporal Statistics for Federal Data: Academic Partnerships and the NCRN

**Scott H. Holan**  
*Department of Statistics*  
*University of Missouri*

**Research Supported by:** *NSF–Census Research Network*

*Geospatial Statistics, Tools, Data, Practices, Opportunities and Challenges in the Federal Agencies*  
*October 16, 2015*



# University of Missouri (MU) Node: Overview and Goals

- ▶ Develop **spatial and spatio-temporal (multiscale) statistical methodology** that improves the precision and interpretability of the **American Community Survey**.
- ▶ Broadly develop spatio-temporal statistical methodology for utilization in problems of official statistics (e.g., **LEHD, SAIPE, etc**).
- ▶ **Train undergraduate students, graduate students, and postdoctoral researchers** to become the next generation of Federal Statisticians and/or researchers working on problems in Federal Statistics.
- ▶ **Train Federal Statisticians (and other staff)** on statistical methodologies related to spatial statistics, spatio-temporal modeling, and hierarchical Bayesian modeling.

# STSN Team

## Current/Previous Researchers:

- ▶ PI: **Scott H. Holan** (Professor – MU)
- ▶ Co-PIs:
  - ▶ **Christopher K. Wikle** (Professor – MU)
  - ▶ **Noel Cressie** (Distinguished Professor – University of Wollongong; Adjunct Professor – MU)
- ▶ Current/Former Postdocs:
  - ▶ **Jonathan R. Bradley**
  - ▶ **Matthew Simpson**
  - ▶ **Aaron Porter** (Assistant Professor: Colorado School of Mines)
  - ▶ **Harrison Quick** (CDC)
- ▶ Current/Former GRAs:
  - ▶ **Christopher Hassett**
  - ▶ **Trevor Oswald**
  - ▶ **Guohui Wu** (SAS)
- ▶ URAs:
  - ▶ **Mary Ryan** (Journalism/Statistics)
  - ▶ **Ellynn Atkinson** (Statistics)

# Spatial and Spatio-Temporal Statistics for Federal Data

- ▶ Federal data sources can be broadly classified as **unit-level** (e.g., microdata) or **area-level** (e.g., ACS county estimates).
- ▶ Unit-level and area-level data often require different methodology.
- ▶ Aside from **Public Use Microdata Samples (PUMS)**, most publicly available data is area-level (e.g., spatially referenced over geographic regions).
- ▶ Both sources of federal data exhibit methodological challenges due to their expansive nature:
  1. Very large datasets (**on the order of millions**).
  2. Multivariate data (**large number of variables**).
  3. Spatially referenced.
  4. Recorded over discrete time.
  5. The data are often non-Gaussian.
- ▶ Data have multiple spatial-temporal scales.

# Some Important Problems for Federal Statistics

Note: This list is NOT meant to be exhaustive!

## 1. Multiscale Spatio-Temporal Analysis:

- ▶ **Change of Support:** Produce estimates on user-defined geographies, and user-defined time-periods – Jon Bradley's talk
- ▶ **Regionalization:** A “best” spatial support – Jon Bradley's talk

## 2. Multivariate Spatio-Temporal Prediction:

- ▶ **Leveraging Information:** Use dependencies between variables, times, and regions to predict “missing” values – Jon Bradley's talk and some discussion here

## 3. Spatial Small Area Estimation: Subset of Items 1 and 2

## 4. Synthetic Geographies: Imputation of locations for public-use microdata that preserves spatial attributes of original data

**Big Data:** Big Data responses and/or covariates (e.g. social media, satellite imagery, etc.) – Important in Items 1–4

## Small Area Estimation (SAE) – Fay-Herriot Model

- ▶ The **Fay-Herriot (FH) model** model is widely used in SAE and provides reduction in mean square error (MSE) by incorporating auxiliary information.
- ▶ The standard FH model has the form

$$Y_i = \theta_i + \epsilon_i,$$
$$\theta_i = \beta_0 + \mathbf{x}'_i \boldsymbol{\beta}_x + u_i.$$

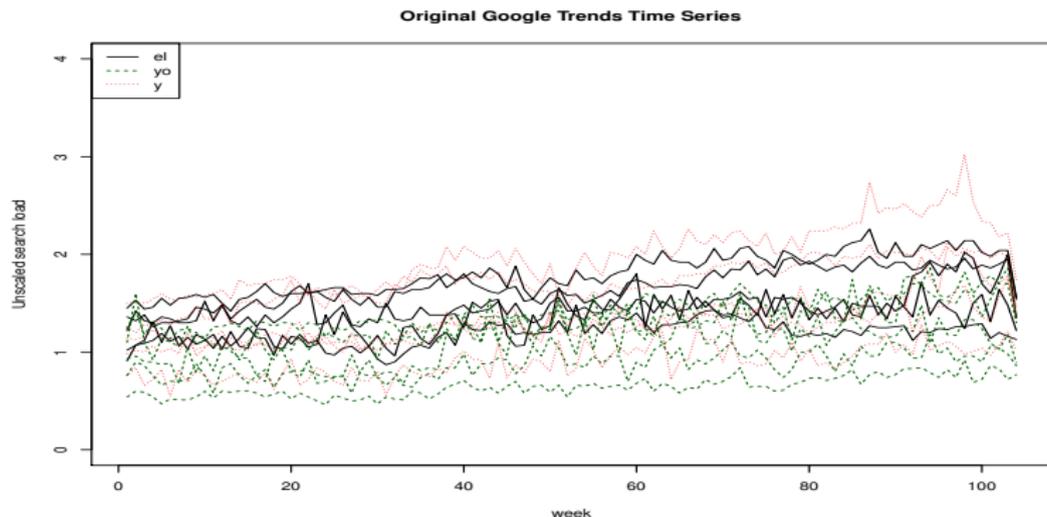
- ▶ In the observation equation,  $i = 1, \dots, n$  indexes location,  $\theta_i$  is the underlying parameter of interest for location  $i$  (e.g., the “superpopulation” mean), and  $\epsilon_i$  is random error (often assumed normally distributed) with variance  $\sigma_i^2$ .
- ▶  $\sigma_i^2$  is the sampling-error variance at location  $i$  and is often considered known.
- ▶ The model for  $\theta_i$  consists of auxiliary covariates  $\mathbf{x}'_i$  and a random spatially indexed effect  $u_i$ .

# Fay-Herriot Model Extensions: Functional Covariates and Spatial Dependence (SFFH)

- ▶ **FH model** is a natural choice for incorporating “Big Data” covariates, as it is a well established model for SAE that typically provides reduction in MSE.
- ▶ “**Big Data**” sources can often be viewed as **functional data** (curves and/or images).
- ▶ Extract information from entire curve or image, rather than using pre-defined summary measures of the function as covariates.
- ▶ Extend FH model to include **functional covariates** (e.g., Google Trends) and **spatial dependence** (Porter et al., 2014).
- ▶ We illustrate that **Google Trends** can be effectively utilized as auxiliary information to reduce the MSE in ACS estimates of **relative change in percent of household Spanish-speaking between 2008 and 2009**.

# Motivating Data – ACS and Google Trends

Functional covariates (temporal curves) for the Google Trends search loads of “el,” “yo,” and “y”. To avoid clutter, we show only the first five weekly time series, in alphabetical order (i.e., Alabama, Connecticut, District of Columbia, Florida, and Georgia), for each search term.



Comparing the variance of the estimates produced by the SFFH (i.e., using Google Trends data and accounting for spatial dependence) relative to the sampling error variances from the ACS we see a 21% reduction over the 21 states considered.

# FH Model Extensions: Semiparametric Hierarchical Empirical Likelihood (SHEL)

- ▶ In official statistics one often encounters datasets that **do not follow a common parametric form**.
- ▶ May be **due to a variety of issues** (e.g., outliers, heavy or light tails, abnormal skewness) mixtures of distributions may be present, but the underlying parametric forms may be unknown.
- ▶ **Transformations may be difficult to identify** for such data.
- ▶ **Porter et al. (2015a)** develops flexible methodology for such data in the presence of **spatial and other general dependence structures**.
- ▶ Methodology is general enough to handle **continuous or discrete data** on either a **continuous or discrete support**.
- ▶ Model illustrated using data concerning per capita income in Missouri counties from the ACS – collected on an **irregular areal support** and represents a **continuous outcome**.

# SHEL-FH: ACS Illustration

The difference of the squared deviations  $(Y_i - \hat{Y}_{(-i)})^2$  for each location of estimated per capita income for (a) the SHEL model versus the Chaudhuri and Ghosh (2011) independence model, (b) the SHEL model versus the Chaudhuri and Ghosh (2011) DP model, (c) the SHEL model versus the parametric model. The square represents Kansas City, MO and the triangle represents St. Louis, MO.



(a)



(b)



(c)



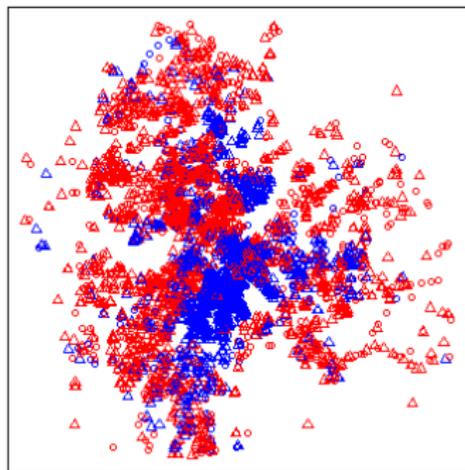
## FH Model Extensions: Multivariate Spatial

- ▶ SAE can be further improved by borrowing strength across spatial regions or by considering multiple outcomes simultaneously.
- ▶ We propose two different FH models that include both multivariate outcomes and latent spatial dependence (Porter et al., 2015b):
  - ▶ outcome-by-space dependence structure is separable,
  - ▶ cross-dependence through the use of a generalized multivariate conditional autoregressive (GMCAR) structure.
- ▶ State-level example: GMCAR model produces smaller MSPE, relative to equivalent census variables, than the separable model and the state-of-the-art multivariate model with unstructured dependence between outcomes and no spatial dependence.
- ▶ County-level example: GMCAR and separable models give smaller MSPE than the state-of-the-art model.

# Bayesian Marked Point Process (MPP) Modeling: Generating Fully Synthetic Public Use Data with Point-Referenced Geography

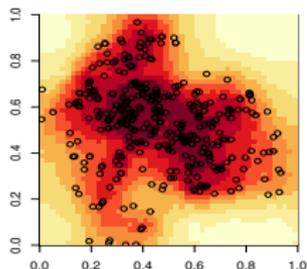
**Goal** – To publicly release data with the following information:

- ▶ Race
- ▶ Gender
- ▶ Income
- ▶ Spatial location
- ▶ Sensitive information such as disease status

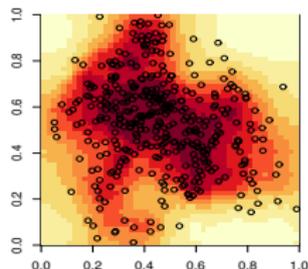


Releasing these data may present a disclosure risk. [Quick et al. \(2015\)](#) solves this problem.

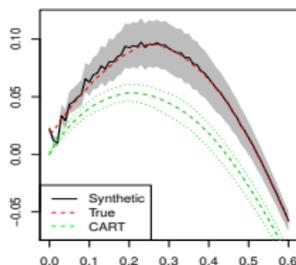
# MPP – Illustration



(a) True Locations



(b) Synthetic Locations



(c)  $L$ -Function

We generate fully synthetic, point-referenced data using a marked point process. We use a fully Bayesian hierarchical model that directly models the exact geographic locations and categorical and non-categorical marks.

# Discussion

- ▶ Briefly described NCRN and the importance of Academic/Government collaboration.
- ▶ 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
  - ▶ Spatial Small Area Estimation
  - ▶ Spatial and Spatio-Temporal Change of Support in the Presence of Sampling Error
  - ▶ Regionalization in the Presence of Sampling Error
  - ▶ Synthetic Geographies
  - ▶ Role of Big Data

**Overall Goal:** To facilitate meaningful discussion!

# Thank You!

holans@missouri.edu

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

<http://stsn.missouri.edu/index.shtml>

## Selected Relevant References:

- Bradley, J.R., Holan, S.H., and Wikle, C.K. (2015) (To Appear – *Annals of Applied Statistics*) Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics.
- Bradley, J.R., Wikle, C.K., and Holan, S.H. (2015) (To Appear – *STAT*) Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates.
- Bradley, J.R., Wikle, C.K., and Holan, S.H. (2015) Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error. (Under Invited Revision – *Journal of the Royal Statistical Society – Series B*).
- Bradley, J.R., Wikle, C.K., and Holan, S.H. (2015) Bayesian Spatial Change of Support for Count-Valued Survey Data. (Under Invited Revision – *Journal of the American Statistical Association*).
- Holan, S.H., McElroy, T.S., and Wu, G. (2015) The Cepstral Model for Multivariate Time Series: The Vector Exponential Model. (Under Invited Revision – *Statistica Sinica*).
- Porter, A.T., Holan, S.H., and Wikle, C.K. (2015a) Bayesian Semiparametric Hierarchical Empirical Likelihood Spatial Models. *Journal of Statistical Planning and Inference*, 165: 78–90.
- Porter, A.T., Holan, S.H., and Wikle, C.K. (2015b) Multivariate Spatial Hierarchical Bayesian Empirical Likelihood Methods for Small Area Estimation. *STAT*, 4: 108–116.
- Porter, A.T., Holan, S.H., Wikle, C.K., and Cressie, N. (2014) Spatial Fay-Herriot Models for Small Area Estimation With Functional Covariates. *Spatial Statistics*. 10: 27–42.
- Porter, A.T., Wikle, C.K., and Holan, S.H. (2015) Small Area Estimation via Multivariate Fay-Herriot Models With Latent Spatial Dependence. *Australian & New Zealand Journal of Statistics*. 57: 15–29.
- Quick, H., Holan, S.H., and Wikle, C.K. (2015) (To Appear – *STAT*) Zeros and Ones: A Case for Suppressing Zeros in Sensitive Count Data with an Application to Stroke Mortality.
- Quick, H., Holan, S.H., Wikle, C.K., and Reiter, J.P. (2015) (To Appear – *Spatial Statistics*) Bayesian Marked Point Process Modeling for Generating Fully Synthetic Public Use Data with Point-Referenced Geography.