An R Package for Spatio-Temporal Change of Support

Andrew M. Raim

Center for Statistical Research and Methodology
U.S. Census Bureau
andrew.raim@census.gov

ICSA 2018, Kerala, India

Joint work with **Scott H. Holan** (U. of Missouri & U.S. Census Bureau), **Jonathan R. Bradley** (Florida State U.), and **Christopher K. Wikle** (U. of Missouri)

Disclaimer

This presentation is to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

The American Community Survey (ACS)

- The ACS is an ongoing survey administered by the U.S. Census Bureau to measure key socioeconomic and demographic variables for the U.S. population.
- ACS data is available to the public through the American FactFinder (`http://factfinder.census.gov`) for years 2005 through 2016.
- Estimates have been released for 1-year, 3-year, or 5-year periods. 3-year estimates were discontinued after 2013.
- Granularity is down to census block-groups. However, estimates for an area are suppressed unless the area meets certain criteria.
- For example, an area must have population $>$ 65,000 for 1-year estimates to be released, but there is no population requirement for 5-year estimates (U.S. Census Bureau, 2016).

Spatio-Temporal Change of Support in the ACS

- **Spatio-Temporal Change of Support (STCOS) Problem**: using all available ACS releases and their patterns over space and time, provide reasonable model-based estimates for user-specified geographies and periods.
- This work is based on models developed in Bradley et al. (2015, Stat). We develop the `stcos` R package to make the methodology widely accessible to data users.
- Statistical agencies have direct access to microdata, and can aggregate to any support and period without STCOS methodology.
- The methods and software are not limited to use with ACS data, but were developed with ACS in mind.
- See Bradley et al. (2015, Stat) and the references therein for a review of change-of-support literature.

The STCOS Problem

- There are three types of geographies involved.
  1. **Source supports** contain direct estimates, which will be used to train the STCOS model.
  2. **Target supports** are geographies on which we want to produce estimates and predictions.
  3. A **fine-level support**, which is used to “translate” estimates from source supports to target supports.

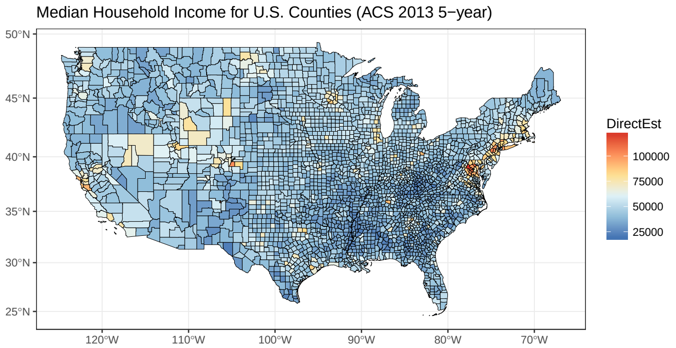
- In this work, a geography is represented by a shapefile.

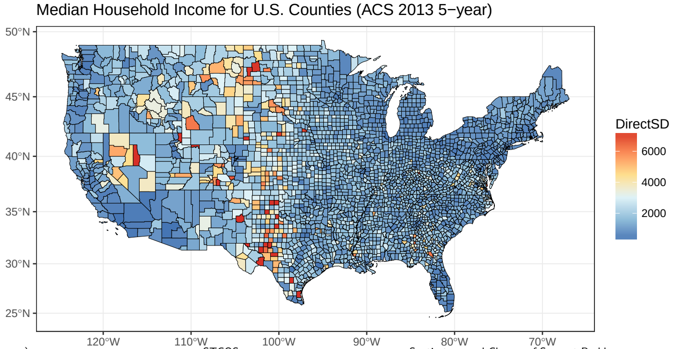
- The STCOS model is based on the amount of overlap between areas across supports.

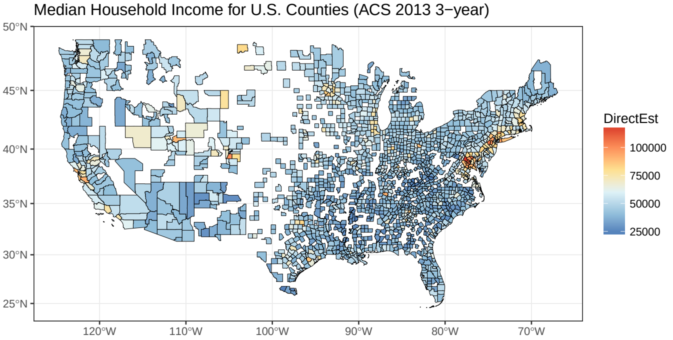
- Doing this within a statistical framework provides measures of variability along with estimates.

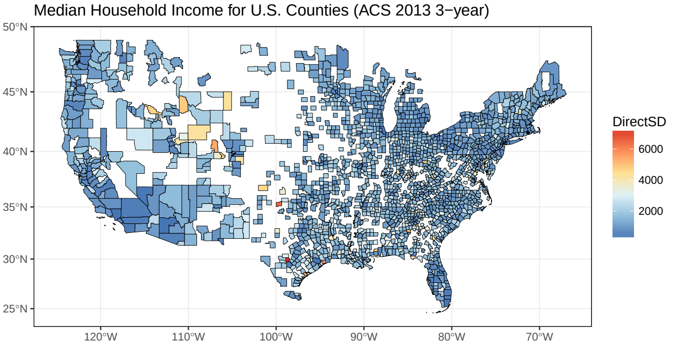
Congressional Districts

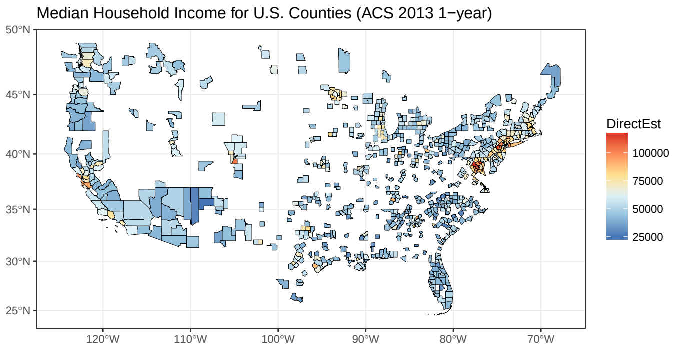
- Each state in the U.S. is apportioned into one or more congressional districts (CDs), which elect an official to the House of Representatives.
- The number of CDs for a state is determined by population counts from the decennial census.
- CDs do not necessarily align with other census geographies (tracts, block groups, counties, etc).
- ACS releases estimates for CDs, which provides a good benchmark to compare with STCOS model estimates.

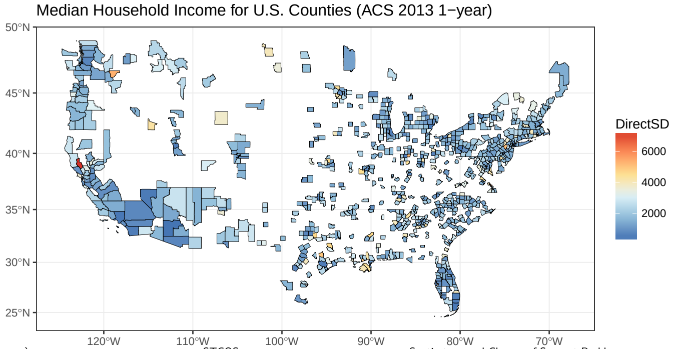

Median Household Income for U.S. Counties (ACS 2013 5−year)
50°N
45°N
40°N
35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
70°W
DirectEst
100000
75000
50000
25000


Median Household Income for U.S. Counties (ACS 2013 5−year)
50°N
45°N
40°N
35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
70°W
DirectSD
6000
4000
2000

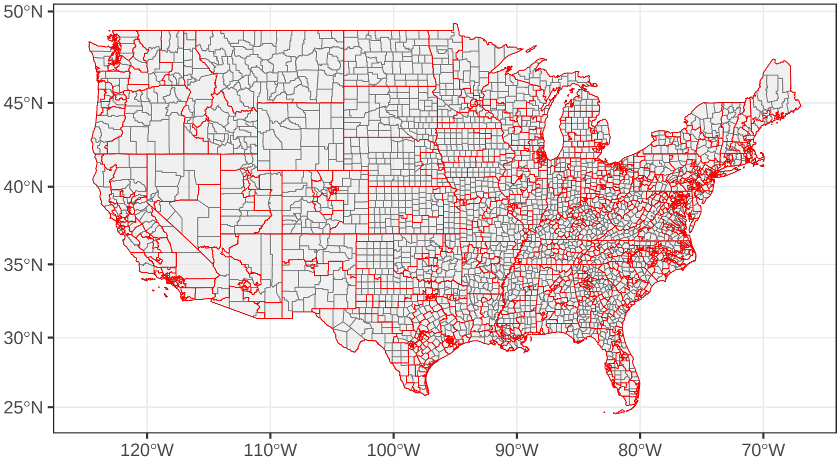

Median Household Income for U.S. Counties (ACS 2013 3−year)
50°N
45°N
40°N
35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
70°W
DirectEst
100000
75000
50000
25000


Median Household Income for U.S. Counties (ACS 2013 3−year)
50°N
45°N
40°N
35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
70°W
DirectSD
6000
4000
2000

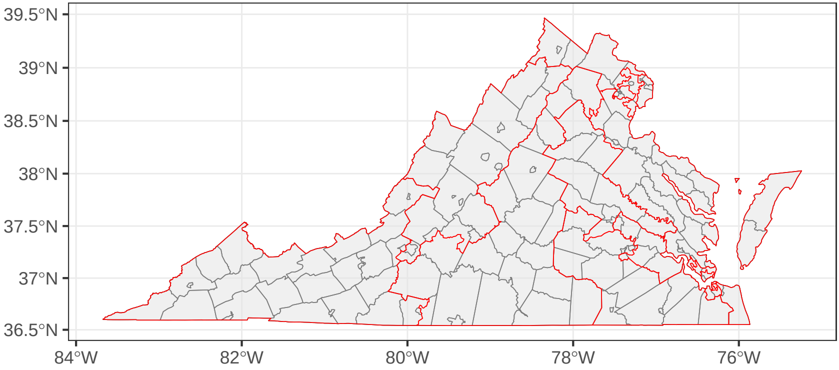

Median Household Income for U.S. Counties (ACS 2013 1−year)
50°N
45°N
40°N
35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
70°W
DirectEst
100000
75000
50000
25000


Median Household Income for U.S. Counties (ACS 2013 1−year)
50°N
45°N
40°N
35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
70°W
DirectSD
6000
4000
2000

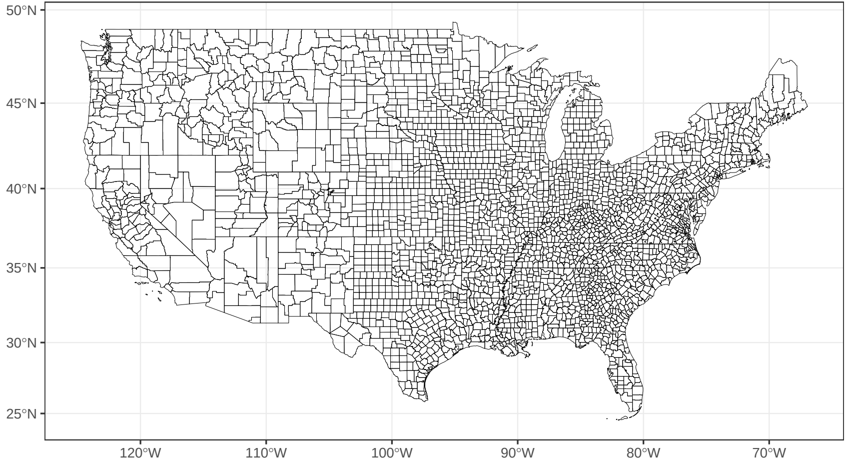
Congressional Districts in 2015



Congressional Districts in 2015



U.S. Counties in 2015



The STCOS Model

- $\mathcal{T} = \{T_L, \dots, T_U\}$: times for which direct estimates are available.
- $\mathcal{L}$: set of lookback periods. For ACS data, $\mathcal{L} = \{1, 3, 5\}$ are possible lookbacks.
- $D_{t\ell}$: source support — collection of areal units with direct estimates — for time $t \in \mathcal{T}$ and period $\ell \in \mathcal{L}$.
- $Z_t^{(\ell)}(A)$ and $\sigma^2_{t\ell}(A)$: direct survey estimate and associated variance for a survey variable of interest, $A \in D_{t\ell}$, $\ell \in \mathcal{L}$, $t \in \mathcal{T}$.
- $D_B = \{B_1, \dots, B_{n_B}\}$ is the fine level support.

STCOS Bayesian Hierarchical Model

- Data Model

$$
\begin{aligned}
&Z_t^{(\ell)}(A) = Y_t^{(\ell)}(A) + \varepsilon_t^{(\ell)}(A), \\
&\varepsilon_t^{(\ell)}(A) \overset{\text{ind}}{\sim} \text{N}(0, \sigma^2_{t\ell}(A)).
\end{aligned}
$$

- Process Model

$$
\begin{aligned}
&Y_t^{(\ell)}(A) = h(A)^\top \boldsymbol{\mu}_B + \psi_t^{(\ell)}(A)^\top \boldsymbol{\eta} + \xi_t^{(\ell)}(A), \\
&\boldsymbol{\eta} \sim \text{N}(0, \sigma^2_K \boldsymbol{K}), \\
&\xi_t^{(\ell)}(A) \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2_\xi).
\end{aligned}
$$

- Prior Model

$$
\begin{aligned}
&\boldsymbol{\mu}_B \sim \text{N}(0, \sigma^2_\mu \boldsymbol{I}), \\
&\sigma^2_\mu \sim \text{IG}(a_\mu, b_\mu), \quad \sigma^2_K \sim \text{IG}(a_K, b_K), \quad \sigma^2_\xi \sim \text{IG}(a_\xi, b_\xi).
\end{aligned}
$$

Latent Process Model

- Define a continuous-space discrete-time process on $\boldsymbol{u} \in \bigcup_{i=1}^{n_B} B_i$, $t \in \mathcal{T}$,

$$Y(\boldsymbol{u}; t) = \delta(\boldsymbol{u}) + \sum_{j=1}^{\infty} \psi_j(\boldsymbol{u}; t) \cdot \eta_j,$$

  where $\delta(\boldsymbol{u})$ is a large-scale spatial trend process and $\{\psi_j(\boldsymbol{u}, t)\}_{j=1}^{\infty}$ is a pre-specified set of spatio-temporal basis functions.

- Integrate $Y(\boldsymbol{u}; t)$ over $u \in A$ (wrt uniform density) and ℓ lookbacks,

$$Y_t^{(\ell)}(A) = \underbrace{\frac{1}{|A|} \int_A \delta(\boldsymbol{u}) d\boldsymbol{u}}_{\text{large-scale spatial trend}} + \underbrace{\frac{1}{\ell |A|} \sum_{k=t-\ell+1}^{t} \sum_{j=1}^{r} \int_A \psi_j(\boldsymbol{u}; k) \cdot \eta_j}_{\text{spatio-temporal random process}}$$

$$+ \underbrace{\frac{1}{\ell |A|} \sum_{k=t-\ell+1}^{t} \sum_{j=r+1}^{\infty} \int_A \psi_j(\boldsymbol{u}; k) \cdot \eta_j}_{\text{remainder}}$$

$$= \mu(A) + \psi_t^{(\ell)}(A)^\top \boldsymbol{\eta} + \xi_t^{(\ell)}(A).$$

- For the remainder, assume that $\xi_t^{(\ell)}(A) \stackrel{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\xi^2)$.

Basis Functions

- We make use of local bisquare basis functions,

$$\psi_j(\boldsymbol{u}, t) = \left[1 - \frac{\|\boldsymbol{u} - \boldsymbol{c}_j\|^2}{w_s^2} - \frac{|t - g_t|^2}{w_t^2}\right]^2 \times$$
$$I(\|\boldsymbol{u} - \boldsymbol{c}_j\| \leq w_s) \cdot I(|t - g_t| \leq w_t).$$

- Spatial knot points $\boldsymbol{c}_j$, $j = 1, \dots, r_{\text{space}}$, are selected via a space-filling design on D_B; see the `R` `fields` package (Nychka et al., 2015).
- Temporal knot points g_t, $t = 1, \dots, r_{\text{time}}$, are chosen to be equally spaced through $\mathcal{T}$.
- For area A and lookback period ℓ, we take a Monte Carlo approximation

$$\psi_{jt}^{(\ell)}(A) \approx \frac{1}{\ell Q} \sum_{k=t-\ell+1}^{t} \sum_{q=1}^{Q} \psi_j(\boldsymbol{u}_q, k),$$

using a uniform random sample $\boldsymbol{u}_1, \dots, \boldsymbol{u}_Q$ on A.

Change of Support Term

- Suppose for the large-scale spatial trend process that

$$\delta(u) = \sum_{i=1}^{n_B} \mu_i I(u \in A \cap B_i), \quad \text{for a given area } A.$$

- Then, integrating over $u \in A$,

$$\mu(A) = \frac{1}{|A|} \sum_{i=1}^{n_B} \int_{A \cap B_i} \delta(u) du = \frac{1}{|A|} \sum_{i=1}^{n_B} \mu_i \int_{A \cap B_i} du = \sum_{i=1}^{n_B} \mu_i \frac{|A \cap B_i|}{|A|}$$
$$= h(A)^\top \boldsymbol{\mu}_B.$$

- $h(A) = (|A \cap B_1|/|A|, \ldots, |A \cap B_{n_B}|/|A|)$ is computed from the source and fine-level supports.

- $\boldsymbol{\mu}_B = (\mu_1, \ldots, \mu_{n_B})$ is unknown, to be estimated from the data.

STCOS Model in Vector Form

We may write

$$\boldsymbol{Z} = \boldsymbol{H}\boldsymbol{\mu}_B + \boldsymbol{S}\boldsymbol{\eta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \text{N}(0, \boldsymbol{V}),$$

where

$$\begin{aligned}
\boldsymbol{Z} &= \text{vec}\left(Z_t^{(\ell)}(A) : \ell \in \mathcal{L},\ t \in \mathcal{T},\ A \in \mathcal{D}_{t\ell}\right), \\
\boldsymbol{H} &= \text{rbind}\left(h_t^{(\ell)}(A)^T : \ell \in \mathcal{L},\ t \in \mathcal{T},\ A \in \mathcal{D}_{t\ell}\right), \\
\boldsymbol{S} &= \text{rbind}\left(\psi_t^{(\ell)}(A)^T : \ell \in \mathcal{L},\ t \in \mathcal{T},\ A \in \mathcal{D}_{t\ell}\right), \\
\boldsymbol{\xi} &= \text{vec}\left(\xi_t^{(\ell)}(A) : \ell \in \mathcal{L},\ t \in \mathcal{T},\ A \in \mathcal{D}_{t\ell}\right), \\
\boldsymbol{\varepsilon} &= \text{vec}\left(\varepsilon_t^{(\ell)}(A) : \ell \in \mathcal{L},\ t \in \mathcal{T},\ A \in \mathcal{D}_{t\ell}\right), \\
\boldsymbol{V} &= \text{Diag}\left(\sigma_{t\ell}^2(A) : \ell \in \mathcal{L},\ t \in \mathcal{T},\ A \in \mathcal{D}_{t\ell}\right),
\end{aligned}$$

and $h_t^{(\ell)}(A) \equiv h(A)$.

Specification of $\boldsymbol{K}$

- Suppose the fine-level support behaves according to the process

$$\boldsymbol{Y}_t^* = \boldsymbol{\mu}_B + \boldsymbol{\nu}_t, \quad \text{for } t \in \mathcal{T}$$

$$\boldsymbol{\nu}_t = \boldsymbol{M}\boldsymbol{\nu}_{t-1} + \boldsymbol{b}_t, \quad \boldsymbol{b}_t \stackrel{\text{iid}}{\sim} \text{N}(\boldsymbol{0}, \sigma_K^2(\boldsymbol{I} - \boldsymbol{A})^-).$$

  where $\boldsymbol{A}$ is the adjacency matrix of D_B.

- Let $\boldsymbol{\Sigma}_{y^*}$ denote the covariance matrix of $(\boldsymbol{Y}_t^* : t \in \mathcal{T})$.

- Obtain $\boldsymbol{C}^*$ by solving

$$\min \|\boldsymbol{\Sigma}_{y^*} - \boldsymbol{S}\boldsymbol{C}\boldsymbol{S}^\top\|_\text{F}, \quad \boldsymbol{C} \text{ is a } r \times r \text{ positive semidefinite matrix}$$

  which yields $\boldsymbol{C}^* = (\boldsymbol{S}^\top\boldsymbol{S})^{-1}\boldsymbol{S}^\top\boldsymbol{\Sigma}_{y^*}\boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{S})^{-1}$. The best positive approximant problem is discussed further in Bradley et al. (2015) and Higham (1988).

- Note that $\boldsymbol{\Sigma}_{y^*} = \sigma_K^2\tilde{\boldsymbol{\Sigma}}_{y^*}$ where $\tilde{\boldsymbol{\Sigma}}_{y^*}$ is free of unknown parameters and $\boldsymbol{M}$ does not need to be estimated. Then we have

$$\boldsymbol{C}^* = \sigma_K^2\boldsymbol{K}, \quad \boldsymbol{K} = (\boldsymbol{S}^\top\boldsymbol{S})^{-1}\boldsymbol{S}^\top\tilde{\boldsymbol{\Sigma}}_{y^*}\boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{S})^{-1}.$$

  and $\boldsymbol{K}$ can be precomputed outside of the MCMC.

Specification of $\boldsymbol{K}$

- **(Independence)** Taking $\boldsymbol{K} = \boldsymbol{I}$ assumes no spatio-temporal covariance in $\boldsymbol{\eta}$.

- **(Spatial-only)** Let $\boldsymbol{\Sigma}_{y^*} = \sigma_K^2(\boldsymbol{I} - \boldsymbol{A})^- \otimes \boldsymbol{I}_{|\mathcal{T}|}$ to ignore covariance in time.

- **(Random Walk)** If $\boldsymbol{M} = \boldsymbol{I}$, the process

$$\boldsymbol{Y}_t^* = \boldsymbol{\mu}_B + \boldsymbol{M}\boldsymbol{\nu}_{t-1} + \boldsymbol{b}_t, \quad \boldsymbol{b}_t \overset{\text{iid}}{\sim} \text{N}(\boldsymbol{0}, \sigma_K^2(\boldsymbol{I} - \boldsymbol{A})^-)$$

is a vector random walk with autocovariance

$$\boldsymbol{\Gamma}(t, h) = \begin{cases} t\sigma_K^2(\boldsymbol{I} - \boldsymbol{A})^- & \text{if } h \geq 0 \\ (t - |h|)\sigma_K^2(\boldsymbol{I} - \boldsymbol{A})^- & \text{if } -t < h < 0. \end{cases}$$

Take

$$\tilde{\boldsymbol{\Sigma}}_{y^*} = \begin{bmatrix} \boldsymbol{\Gamma}(1,1) & \boldsymbol{\Gamma}(1,2) & \cdots & \boldsymbol{\Gamma}(1,|\mathcal{T}|) \\ \boldsymbol{\Gamma}(2,1) & \boldsymbol{\Gamma}(2,2) & \cdots & \boldsymbol{\Gamma}(2,|\mathcal{T}|) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}(|\mathcal{T}|,1) & \boldsymbol{\Gamma}(|\mathcal{T}|,2) & \cdots & \boldsymbol{\Gamma}(|\mathcal{T}|,|\mathcal{T}|) \end{bmatrix}.$$

Basis Functions: Dimension Reduction

- The presence of multicollinearity can severely hinder convergence of the Markov-Chain Monte Carlo (MCMC) sampler.
- To protect against multicollinearity, we reduce the $n \times r$ matrix $\boldsymbol{S}$ using principal components analysis.
- Suppose $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ is the eigendecomposition of $\boldsymbol{S}^\top\boldsymbol{S}$, and $\tilde{\boldsymbol{U}}$ contains the r' columns of $\boldsymbol{U}$ corresponding the $r' \leq r$ largest magnitude eigenvalues in $\boldsymbol{D}$.
- The transformation $T(\boldsymbol{S}) = \boldsymbol{S}\tilde{\boldsymbol{U}}^\top$ is applied to all matrices computed from the basis functions.

Gibbs Sampler

- $[\boldsymbol{\mu}_B \mid \text{---}] \sim \text{N}(\boldsymbol{\vartheta}_\mu, \boldsymbol{\Omega}_\mu^{-1})$,

$$\begin{aligned}\boldsymbol{\vartheta}_\mu &= \boldsymbol{\Omega}_\mu^{-1} \boldsymbol{H}^\top \boldsymbol{V}^{-1}(\boldsymbol{Z} - \boldsymbol{S}\boldsymbol{\eta} - \boldsymbol{\xi}), \\ \boldsymbol{\Omega}_\mu &= \boldsymbol{H}^\top \boldsymbol{V}^{-1} \boldsymbol{H} + \sigma_\mu^{-2} \boldsymbol{I}.\end{aligned}$$

- $[\boldsymbol{\eta} \mid \text{---}] \sim \text{N}(\boldsymbol{\vartheta}_\eta, \boldsymbol{\Omega}_\eta^{-1})$,

$$\begin{aligned}\boldsymbol{\vartheta}_\eta &= \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{S}^\top \boldsymbol{V}^{-1}(\boldsymbol{Z} - \boldsymbol{H}\boldsymbol{\mu}_B - \boldsymbol{\xi}), \\ \boldsymbol{\Omega}_\eta &= \boldsymbol{S}^\top \boldsymbol{V}^{-1} \boldsymbol{S} + \sigma_K^{-2} \tilde{\boldsymbol{K}}^{-1}.\end{aligned}$$

- $[\boldsymbol{\xi} \mid \text{---}] \sim \text{N}(\boldsymbol{\vartheta}_\xi, \boldsymbol{\Omega}_\xi^{-1})$,

$$\begin{aligned}\boldsymbol{\vartheta}_\xi &= \boldsymbol{\Omega}_\xi \boldsymbol{V}^{-1}(\boldsymbol{Z} - \boldsymbol{H}\boldsymbol{\mu}_B - \boldsymbol{S}\boldsymbol{\eta}), \\ \boldsymbol{\Omega}_\xi^{-1} &= \boldsymbol{V}^{-1} + \sigma_\xi^{-2} \boldsymbol{I}.\end{aligned}$$

- $[\sigma_\mu^2 \mid \text{---}] \sim \text{IG}(\alpha_\mu, \beta_\mu)$, $\alpha_\mu = a_\mu + n_B/2$ and $\beta_\mu = b_\mu + \boldsymbol{\mu}_B^\top \boldsymbol{\mu}_B/2$.
- $[\sigma_K^2 \mid \text{---}] \sim \text{IG}(\alpha_K, \beta_K)$, $\alpha_K = a_K + r/2$ and $\beta_K = b_K + \boldsymbol{\eta}^\top \tilde{\boldsymbol{K}}^{-1} \boldsymbol{\eta}/2$.
- $[\sigma_\xi^2 \mid \text{---}] \sim \text{IG}(\alpha_\xi, \beta_\xi)$, $\alpha_\xi = a_\xi + N/2$ and $\beta_\xi = b_\xi + \boldsymbol{\xi}^\top \boldsymbol{\xi}/2$.

Using MCMC Draws

- Suppose A is an area of interest, not necessarily one of the source supports.
- We are primarily interested in draws of the mean

$$\mathrm{E}(Y_t^{(\ell)}(A) \mid \boldsymbol{\mu}_B, \boldsymbol{\eta}) = \boldsymbol{h}(A)^\top \boldsymbol{\mu}_B + \boldsymbol{\psi}_t^{(\ell)}(A)^\top \boldsymbol{\eta},$$

or draws from the posterior predictive distribution

$$[Y_t^{(\ell)}(A) \mid \boldsymbol{\mu}_B, \boldsymbol{\eta}, \sigma_\xi^2] \sim \mathrm{N}\left(\boldsymbol{h}(A)^\top \boldsymbol{\mu}_B + \boldsymbol{\psi}_t^{(\ell)}(A)^\top \boldsymbol{\eta}, \sigma_\xi^2\right)$$

using draws from the posterior distribution.

- We take the sample mean of MCMC draws from either distribution as a point estimate, and the sample standard deviation (SD) to measure variability in the respective distribution.

STCOS R Package

- The stcos R package facilitates application of the model.
  1. Preprocess: Prepare $\boldsymbol{Z}$, $\boldsymbol{V}$, $\boldsymbol{H}$, $\boldsymbol{S}$, and $\boldsymbol{K}^{-1}$ needed to fit the model.
  2. Fit the model via Gibbs sampler.
  3. Postprocess: Compute estimates and predictions on target supports using MCMC draws.
- Preprocess once for a given set of source supports. Redo model fit for each survey variable of interest. Redo postprocess for each target support of interest.
- We depend on several other R packages.
  1. Manipulation of shapefiles via the sf package (Pebesma, 2017).
  2. Object-oriented programming using the R6 package (Chang, 2017).
  3. Ability to call C++ code for performance via Rcpp (Eddelbuettel, 2013) and RcppArmadillo (Eddelbuettel and Sanderson, 2014).
  4. Sparse matrix computations in R via Matrix package (Bates and Maechler, 2017).
- Development version of ggplot2 (Wickham, 2016) can plot sf objects.

Source Supports: Shapefiles with Estimates

User provides all supports as shapefiles. Direct estimates and variance estimates should be embedded into source supports.

```
R> library(sf)
R> acs5.2013 <- st_read("county_acs_5yr2013.shp")
R> head(acs5.2013)
Simple feature collection with 6 features and 9 fields
geometry type:  MULTIPOLYGON
dimension:      XY
bbox:           xmin: -9799374 ymin: 3532006 xmax: -9468076 ymax: 4063675
epsg (SRID):    3857
proj4string:    +proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0
                +y_0=0 +k=1.0 +units=m +nadgrids=@null +wktext +no_defs
          GEO_ID STATE COUNTY    NAME   LSAD SHAPE_AREA SHAPE_LEN DirectEst
1 0500000US01001    01    001 Autauga County 2202587903  235761.0     53682
2 0500000US01003    01    003 Baldwin County 5913339907  493065.0     50221
3 0500000US01005    01    005 Barbour County 3262897491  275539.8     32911
4 0500000US01007    01    007    Bibb County 2309817471  223844.6     36447
5 0500000US01009    01    009  Blount County 2454704099  258633.9     44145
6 0500000US01011    01    011 Bullock County 2258507149  233574.8     32033
   DirectVar                       geometry
1   831625.9 MULTIPOLYGON(((-9675622.416...
2   915703.6 MULTIPOLYGON(((-9799373.733...
3  4437638.9 MULTIPOLYGON(((-9544726.836...
4  4323124.1 MULTIPOLYGON(((-9731765.399...
5  2150305.1 MULTIPOLYGON(((-9680577.914...
6 21959826.2 MULTIPOLYGON(((-9573382.365...
```

Load Source Supports

```
library(sf)
library(stcos)

# Fine-level support comes from ACS 5-year estimates for 2015
dom.fine <- st_read("shp/county_acs_5yr2015.shp")

# ACS 1-year source supports
acs1.2005 <- st_read("shp/county_acs_1yr2005.shp")
acs1.2006 <- st_read("shp/county_acs_1yr2006.shp")
...
acs1.2015 <- st_read("shp/county_acs_1yr2015.shp")

# ACS 3-year source supports
acs3.2007 <- st_read("shp/county_acs_3yr2007.shp")
acs3.2008 <- st_read("shp/county_acs_3yr2008.shp")
...
acs3.2013 <- st_read("shp/county_acs_3yr2013.shp")

# ACS 5-year source supports
acs5.2009 <- st_read("shp/county_acs_5yr2009.shp")
acs5.2010 <- st_read("shp/county_acs_5yr2010.shp")
...
acs5.2015 <- st_read("shp/county_acs_5yr2015.shp")
```

Prepare Basis

```r
library(fields)

# Spatial knots are selected via space-filling design
u <- st_sample(dom.fine, size = 5000)
M <- matrix(unlist(u), length(u), 2, byrow = TRUE)
out <- cover.design(M, 500)
knots.sp <- out$design

# Temporal knots are selected to be evenly spaced
knots.t <- c(2005, 2005.5, 2006, 2006.5, 2007, 2007.5, 2008, 2008.5,
    2009, 2009.5, 2010, 2010.5, 2011, 2011.5, 2012, 2012.5, 2013, 2013.5,
    2014, 2014.5, 2015)

# Combined spatio-temporal knots
knots <- merge(knots.sp, knots.t)
names(knots) <- c("x", "y", "t")

# Create a Basis object
basis <- SpaceTimeBisquareBasis$new(knots[,1], knots[,2], knots[,3], w.s = 1, w.t = 1)
```

United States Census Bureau

Construct an STCOSPrep Object

```
sp <- STCOSPrep$new(fine_domain = dom.fine, fine_domain_geo_name = "GEO_ID",
    basis = basis, basis_mc_reps = 500)
sp$add_obs(acs1.2015, period = 2015, estimate_name = "DirectEst",
    variance_name = "DirectVar", geo_name = "GEO_ID")
sp$add_obs(acs3.2013, period = 2011:2013, estimate_name = "DirectEst",
    variance_name = "DirectVar", geo_name = "GEO_ID")
sp$add_obs(acs5.2013, period = 2009:2013, estimate_name = "DirectEst",
    variance_name = "DirectVar", geo_name = "GEO_ID")
...
Z <- sp$get_Z()
V <- sp$get_V()
H <- sp$get_H()
S <- sp$get_S()
```

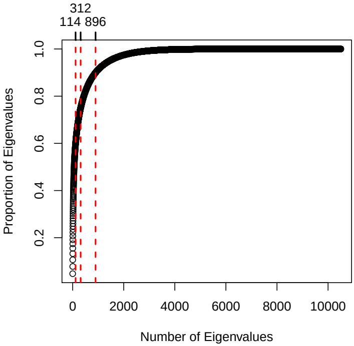
```
R> sp$add_obs(acs1.2015, period = 2015, estimate_name = "DirectEst",
    variance_name = "DirectVar", geo_name = "GEO_ID")
2017-07-13 15:03:18 - Begin adding observed space-time domain
2017-07-13 15:03:18 - Computing overlap matrix using field 'GEO_ID'
2017-07-13 15:03:22 - Computing basis functions
2017-07-13 15:03:30 - Computing basis for area 100 of 812
2017-07-13 15:03:37 - Computing basis for area 200 of 812
...
2017-07-13 15:04:12 - Computing basis for area 700 of 812
2017-07-13 15:04:19 - Computing basis for area 800 of 812
2017-07-13 15:04:20 - Extracting survey estimates from field 'DirectEst'
    and variance estimates from field 'DirectVar'
2017-07-13 15:04:20 - Finished adding observed space-time domain
```

Dimension Reduction for *S*

```
eig <- eigen(t(S) %*% S)
rho <- eig$values

idx.S <- which(cumsum(rho) / sum(rho) < 0.6)
Tx.S <- t(eig$vectors[idx.S,])
f <- function(S) { S %*% Tx.S }
sp$set_basis_reduction(f)

S.reduced <- sp$get_reduced_S()
```



Specification for K

- Independence

```
K.inv <- diag(x = 1, nrow = ncol(S.reduced))
```

- Spatial-only

```
K.inv <- sp$get_Kinv(2005:2015, autoreg = FALSE)
```

- Random Walk

```
K.inv <- sp$get_Kinv(2005:2015)
```

Gibbs Sampler

```r
# Std'ize before MCMC
D <- Diagonal(n = length(Z), x = 1/sd(Z))
Z.scaled <- (Z - mean(Z)) / sd(Z)
V.scaled <- V / var(Z)

# Use MLE as initial value for MCMC
mle.out <- mle.stcos(Z.scaled, S.reduced, V.scaled, H, init = list(sig2xi = 1))
init <- list(
    sig2xi = mle.out$sig2xi.hat,
    mu_B = mle.out$mu.hat,
    eta = mle.out$eta.hat
)

# Gibbs Sampler
gibbs.out <- gibbs.stcos.raw(Z.scaled, S.reduced, V.scaled, K.inv, H, R = 10000,
    report.period = 100, burn = 1000, thin = 10, init = init)
```

```
2017-07-06 13:21:58 - Begin Gibbs sampler
2017-07-06 13:24:09 - Begin iteration 100
2017-07-06 13:26:17 - Begin iteration 200
...
2017-07-06 16:46:22 - Begin iteration 10000
2017-07-06 16:46:23 - Finished Gibbs sampler
```

Estimation & Prediction on Target Support

```r
# Load a target support and transform to fine-level support's projection
cd1.2015 <- st_read("shp/cd_acs_1yr2015.shp")
dom <- st_transform(cd1.2015, crs = st_crs(dom.fine))

# Compute H and S matrices
target.out <- sp$domain2model(dom, period = 2015, geo_name = "GEO_ID")

# Posterior distribution for E(Y)
E.hat.scaled <- fitted(gibbs.out, target.out$H, target.out$S.reduced)
E.hat <- sd(Z) * E.hat.scaled + mean(Z)                    # Uncenter and unscale
dom$E.mean <- colMeans(E.hat)                              # Point estimates
dom$E.sd <- apply(E.hat, 2, sd)                            # SDs
dom$E.lo <- apply(E.hat, 2, quantile, prob = 0.025) # Credible interval lo
dom$E.hi <- apply(E.hat, 2, quantile, prob = 0.975) # Credible interval hi

# Posterior predictive distribution of Y
Y.pred.scaled <- predict(gibbs.out, target.out$H, target.out$S.reduced)
Y.pred <- sd(Z) * Y.pred.scaled + mean(Z)                    # Uncenter and unscale
dom$PP.mean <- colMeans(Y.pred)                              # Point estimates
dom$PP.sd <- apply(Y.pred, 2, sd)                            # SDs
dom$PP.lo <- apply(Y.pred, 2, quantile, prob = 0.025) # Prediction interval lo
dom$PP.hi <- apply(Y.pred, 2, quantile, prob = 0.975) # Prediction interval hi
```

```
> head(dom, 5)
         GEO_ID STATE CD NAME LSAD  SHAPE_AREA SHAPE_LEN   E.mean  PP.mean ...
1 5001400US0101    01 01    1   CD 22122432216 1493128.1 41729.51 41766.25 ...
2 5001400US0102    01 02    2   CD 36852408755 1334656.1 40387.58 40384.29 ...
3 5001400US0103    01 03    3   CD 28641422684 1176377.1 38615.31 38581.89 ...
4 5001400US0104    01 04    4   CD 34528166078 1454589.3 36302.97 36303.11 ...
5 5001400US0105    01 05    5   CD 14828071977  785480.4 43089.59 43065.39 ...
```

Model Selection Study

- Raim et al. (2017, JSM Proceedings) present a model selection study using Deviance Information Criterion (DIC). Tuning parameters are:
  1. the prior covariance structure for $\boldsymbol{K}$,
  2. the number of knot points used to define the basis functions,
  3. the radius parameters in the basis functions, and
  4. amount of dimension reduction on $\boldsymbol{S}$.

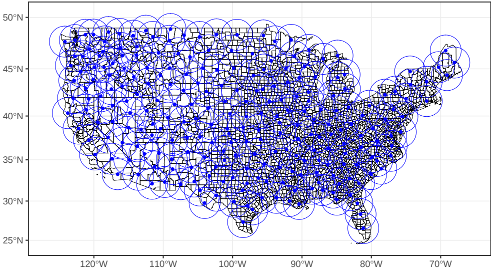
- Prior covariance structures: Independence (IND), Spatial-only (SP), and Spatial with Random-Walk (RW).

- Selection of basis functions:
  1. Knot points $\{\boldsymbol{c}_{jt}\}$, where $\boldsymbol{c}_{jt} = (\boldsymbol{c}_j, g_t)$ for $j = 1, \ldots, r_{\text{space}}$ and $t = 1, \ldots, r_{\text{time}}$.
  2. We fix $r_{\text{time}} = 21$ temporal cutpoints $g_1 = 2005$, $g_2 = 2005.5$, $\ldots$, $r_{20} = 2014.5$, $r_{21} = 2015$. For space, we consider $r_{\text{space}} \in \{250, 500\}$.
  3. To ensure that spatial radius w_s is compatible with the shapefile's projection, compute the distance matrix of $\{\boldsymbol{c}_{jt}\}$ and let $Q_{0.05}$ be the 0.05 quantile of the nonzero upper-triangular entries of the matrix. Take $w_s = \tau_s \cdot Q_{0.05}$, where $\tau_s \in \{0.5, 1.0\}$ is a selected multiplier.

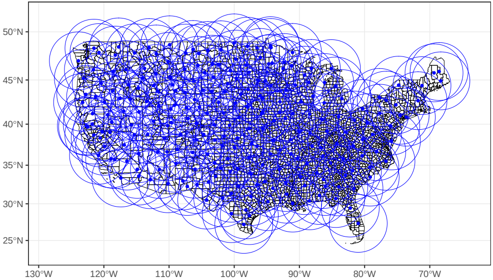
Model Selection Study

- We therefore consider four factors: prior covariance structures IND, SP and RW, $\tau_s \in \{0.5, 1.0\}$, $r_{\text{space}} \in \{250, 500\}$, and eigenvalue proportions 60%, 75%, and 90%.
- For each combination of factors, prepare the terms of the STCOS model and run Gibbs sampler for 2000 iterations. We discard the first 500 iterations as a burn-in period and thin by saving every 10th remaining iteration.
- The maximum likelihood estimator (MLE) is used as the initial value of the sampler in all cases.
- DIC is computed using the saved draws from MCMC sampling — smaller values of DIC indicate better fitting models.
- We also check convergence of the Gibbs sampler by visually examining trace plots of the sampled chains (not shown).

Space-filling designs for basis functions using 250 or 500 spatial knot points and radius $\tau_s \in \{0.5, 1.0\}$.



(a) 250 spatial knot points, $\tau_s = 0.5$



(b) 250 spatial knot points, $\tau_s = 1.0$

Data Sources

- County and CD level ACS estimates and shapefiles were obtained from the American FactFinder website (`http://factfinder.census.gov`).
- 2016 estimates were released between mid-Sept and mid-Dec, and were not used in this study.
- Fine-level support is from 2015 5-year ACS shapefile.
- Source supports are from the following files.

2015 ACS 5-year	2015 ACS 1-year	2014 ACS 5-year	2014 ACS 1-year
2013 ACS 5-year	2013 ACS 3-year	2013 ACS 1-year	2012 ACS 5-year
2012 ACS 3-year	2012 ACS 1-year	2011 ACS 5-year	2011 ACS 3-year
2011 ACS 1-year	2010 ACS 5-year	2010 ACS 3-year	2010 ACS 1-year
2009 ACS 5-year	2009 ACS 3-year	2009 ACS 1-year	2008 ACS 3-year
2008 ACS 1-year	2007 ACS 3-year	2007 ACS 1-year	2006 ACS 1-year
2005 ACS 1-year			

- Target supports are from 2015 CD 1-year and 2015 CD 5-year files.
- Need to ensure that the projection used in the source and target shapefiles matches the fine-level shapefile.

Model Selection Study

DIC for Fitted Models

			Prior Covariance		
r_{space}	τ_s	Eigval %	IND	SP	RW
250	0.5	60	-23213.99	-23213.42	-23208.09
500			-23747.65	-23745.41	-23738.00
250	1.0		-25167.15	-25167.26	-25165.68
500			-23535.65	-23536.62	-23533.74
250	0.5	75	-28011.33	-28001.35	-27994.73
500			-29870.56	-29853.79	-29845.44
250	1.0		-29415.06	-29425.25	-29407.43
500			**−30653.63**	-30652.96	-30652.96
250	0.5	90	-32298.14	-32235.30	-32306.21
500			-32989.99	-32799.24	-33104.12
250	1.0		-33951.35	-33918.53	-33903.14
500			-34533.95	-34453.51	-34465.37

Multicollinearity in the columns of $\boldsymbol{S}$ becomes pronounced when Eigval% $= 90$, and slows convergence of the MCMC (observed via trace plots).

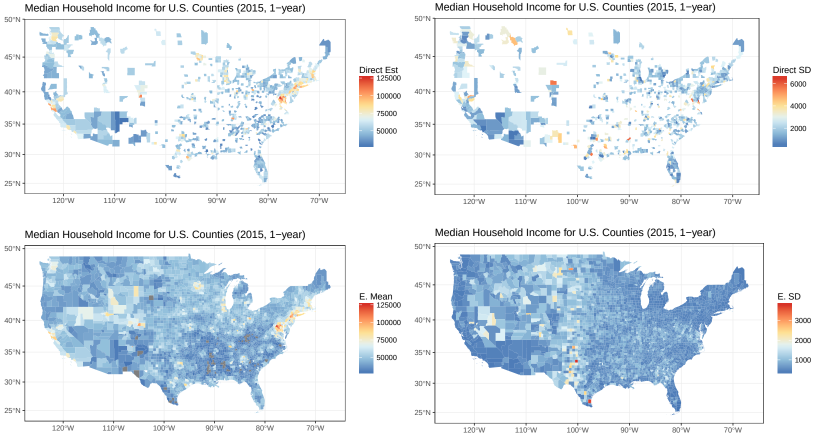
Model Selection Study

A Longer Chain with the Selected Model

- Using the selected model from Section 5, we ran a longer MCMC with 10,000 iterations, discarding the first 1,000 as a burn-in period, saving every 10th remaining iteration
- The MLE was used as the initial value.
- Visual inspection of trace plots was used to assess mixing of the sampled chains, with no lack of convergence detected.

Direct vs. Model Estimates

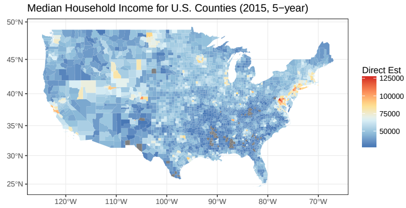
2015 Counties, 1-year

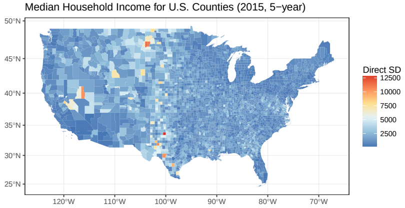


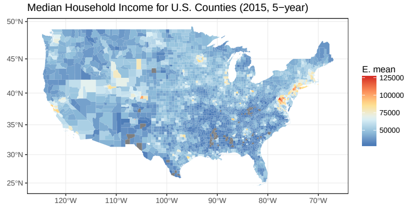
Model results shown are based on MCMC draws of $\text{E}(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \boldsymbol{H}\boldsymbol{\mu}_B + \boldsymbol{S}\boldsymbol{\eta}$.

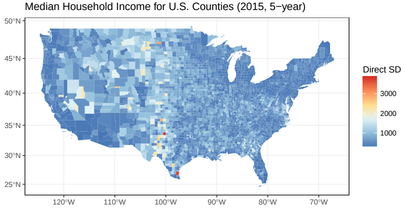
Direct vs. Model SDs

2015 Counties, 5-year





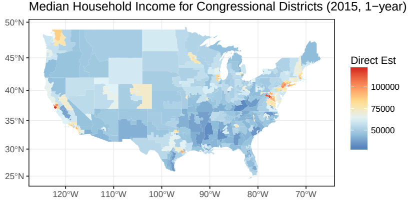


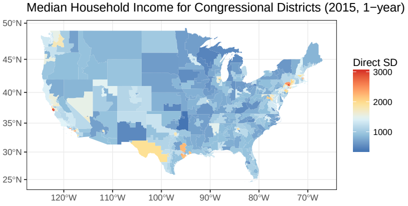


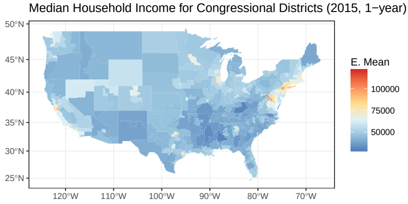
Model results shown are based on MCMC draws of $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \boldsymbol{H}\boldsymbol{\mu}_B + \boldsymbol{S}\boldsymbol{\eta}$.

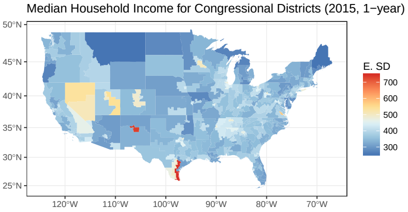
Direct vs. Model

2015 CDs, 1-year



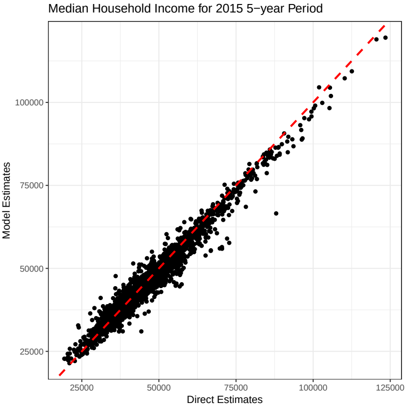




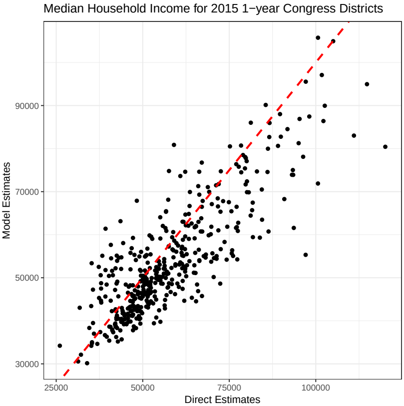


Model results shown are based on MCMC draws of $\text{E}(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \boldsymbol{H}\boldsymbol{\mu}_B + \boldsymbol{S}\boldsymbol{\eta}$.

Direct vs. Model Estimates



(a) 2015 Counties, 5-year.



(b) 2015 CDs, 1-year.

Scatter plots of 2015 direct ACS estimates versus estimates based on the posterior mean of $\mathrm{E}(Y_t^{(\ell)}(A))$. Sample correlation between the two sets of estimates in (a) is 0.9814, while in (b) the correlation 0.8295.

Conclusions

- We are developing the `stcos` `R` package based on methodology from Bradley et al. (2015).
- Improvements are underway to lower programming burden for users, and to increase performance (speed, memory usage) where possible.
- More extensive simulations are in progress.
- We are working toward a CRAN submission and a companion article.
- We hope that this software will facilitate exploration of official statistics on custom geographies and time periods.

See Raim et al. (2017) at `http://andrewraim.github.io`

References I

Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2017. URL `https://CRAN.R-project.org/package=Matrix`. R package version 1.2-10.

Jonathan R. Bradley, Christopher K. Wikle, and Scott H. Holan. Spatio-temporal change of support with application to American Community Survey multi-year period estimates. *Stat*, 4(1):255–270, 2015.

Winston Chang. *R6: Classes with Reference Semantics*, 2017. URL `https://CRAN.R-project.org/package=R6`. R package version 2.2.2.

Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer, 2013.

Dirk Eddelbuettel and Conrad Sanderson. Rcpparmadillo: Accelerating r with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, March 2014.

Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain. *fields: Tools for spatial data*. University Corporation for Atmospheric Research, Boulder, CO, USA, 2015. URL `www.image.ucar.edu/fields`. R package version 9.0.

References II

Edzer Pebesma. *sf: Simple Features for R*, 2017. URL https://CRAN.R-project.org/package=sf. R package version 0.5-1.

Andrew M. Raim, Scott H. Holan, Jonathan R. Bradley, and Christopher K. Wikle. A model selection study for spatio-temporal change of support. In *JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association*, pages 1524–1540, 2017.

U.S. Census Bureau. American Community Survey data suppression, September 2016. URL https://www.census.gov/programs-surveys/acs/technical-documentation/data-suppression.html.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.