Augmenting Analysis with Data from the US Census

Data Institute SF Annual Conference

James Livsey

Center for Statistical Research and Methodology
U.S. Census Bureau

March 10, 2019

Disclaimer

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the author and not necessarily those of the U.S. Census Bureau.

Census Bureau

- Headquartered in Suitland, MD since 1942
- Currently employs about 4,285 staff members

Workshop Outline

1. Census data and products
   - Geographies (Shape files)
   - Surveys
   - Data access (American Fact Finder, API)
   - Miscellaneous Products
     - PUMs
     - 1940 Census data
     - Seasonal Adjustment software
2. Spatial-temporal Change of Support
   - Bayesian hierarchical model
   - Custom Geographies

- I will be using R in this workshop

United States™ Census Bureau

The Census Bureau

- Largest federal government statistical agency
- Dedicated to providing current facts and figures about Americas people, places, and economy
- Federal law protects the confidentiality of all the information the Census Bureau collects

Mission: "To serve as the nation's leading provider of quality data about its people and economy. We honor privacy, protect confidentiality, share our expertise globally and conduct our work openly."

Census Geographies

- Central to the work of the Census
- Provides the framework for survey design, sample selection, data collection, tabulation, and dissemination
- Provides meaning and context to statistical data

Census Geographies

TIGER = Topologically Integrated Geographic Encoding and Referencing

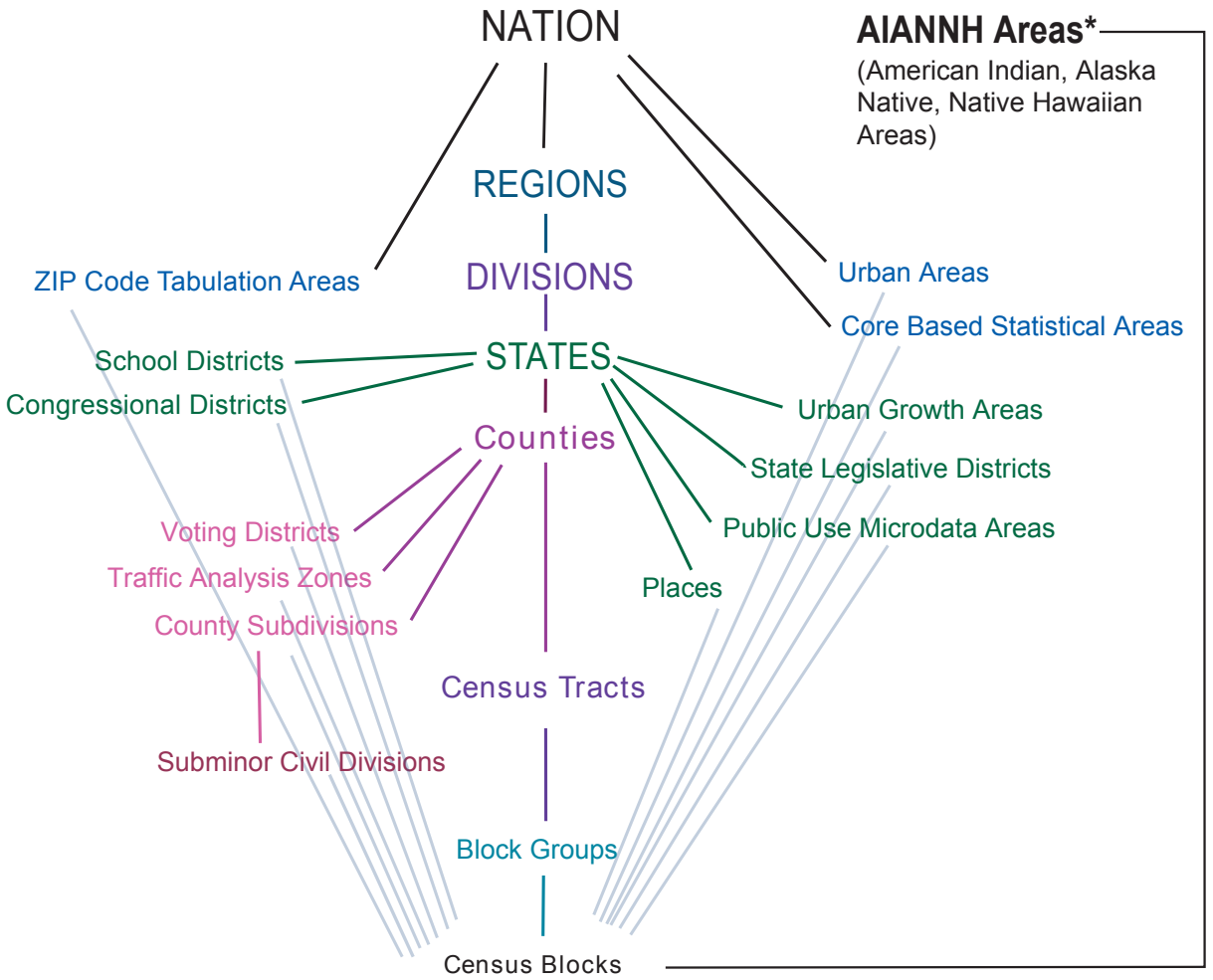
- TIGER products are spatial extracts from the Census Bureau's MAF/TIGER database, containing features such as roads, railroads, rivers, as well as legal and statistical geographic areas
- The Census Bureau offers several file types and online mapping applications
- Disseminated for use in statistical software as shape files

▸ https://www.census.gov/geo/maps-data/data/tiger.html

Census Geographies

Census Geographies

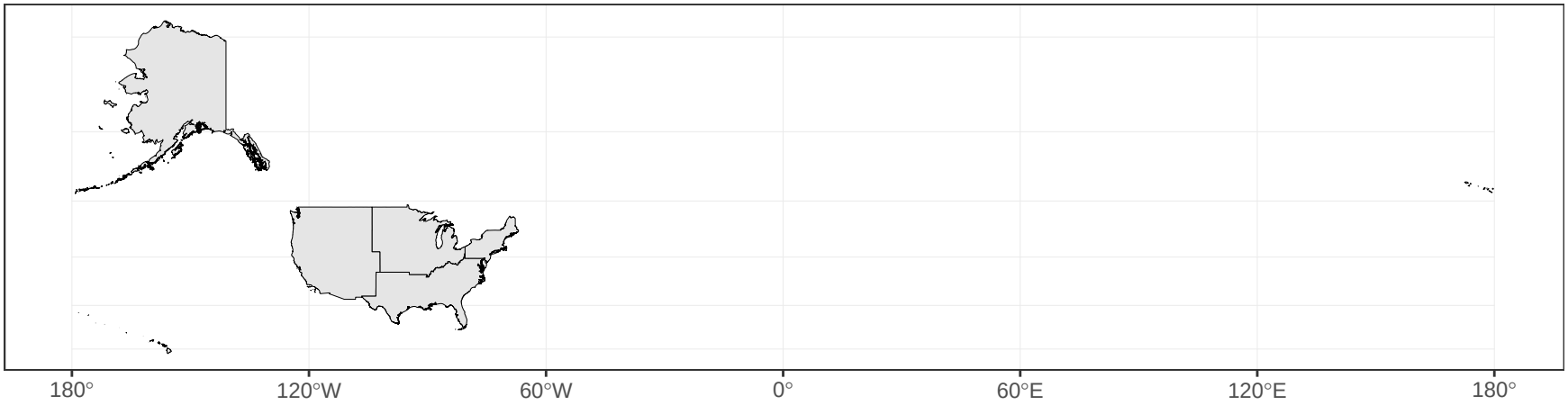
Standard Hierarchy of Census Geographic Entities



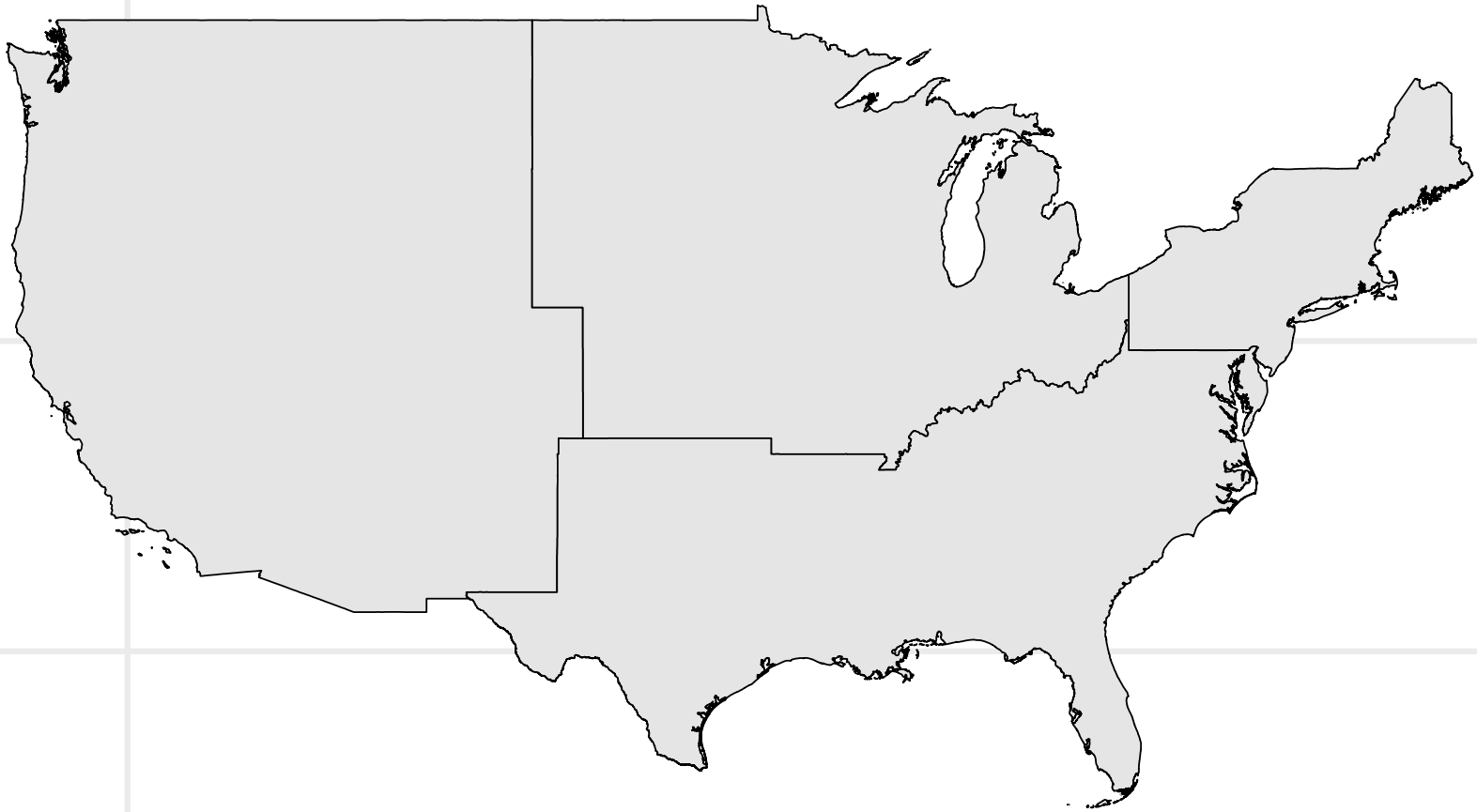
United States™ Census Bureau

Regions

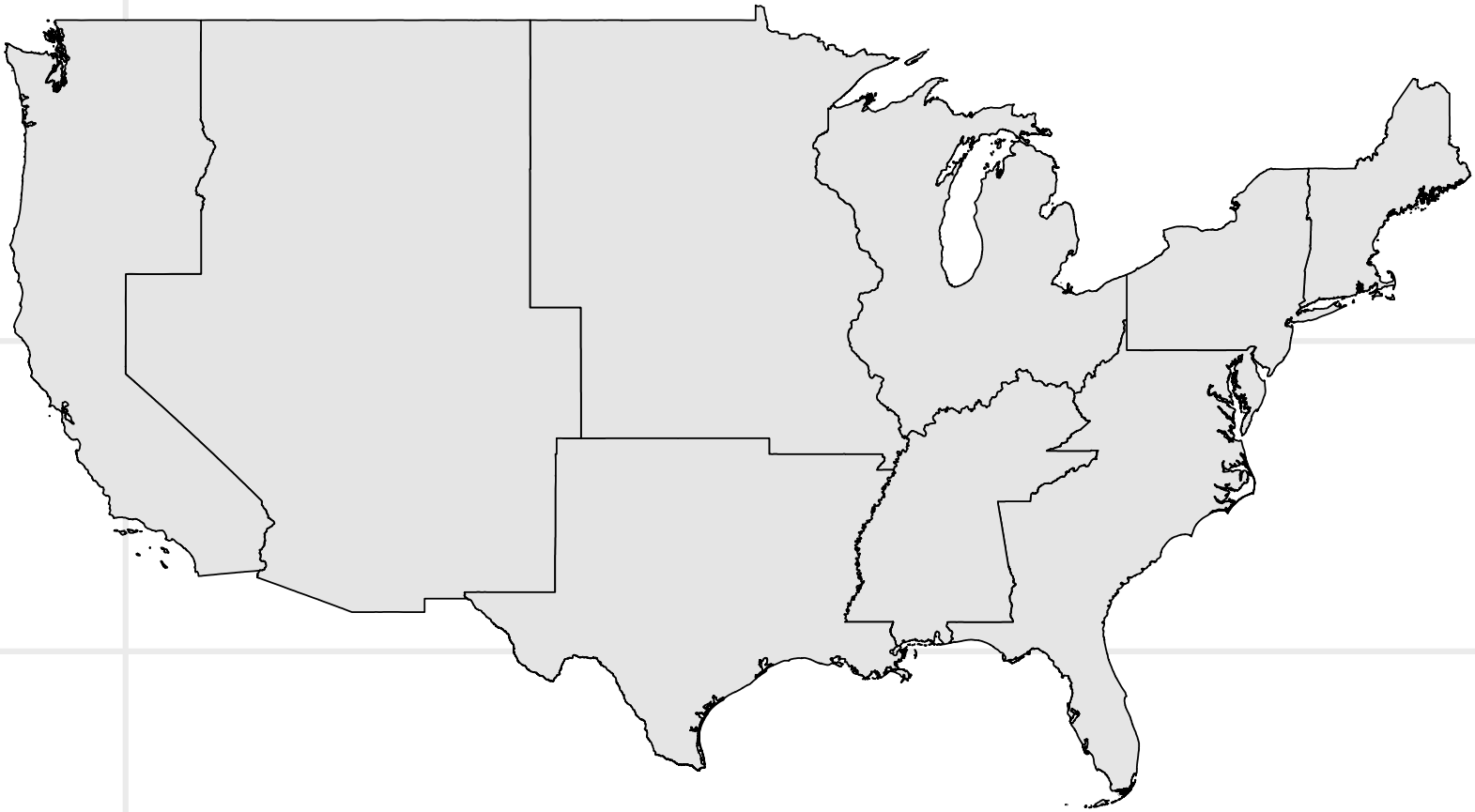
Census Regions



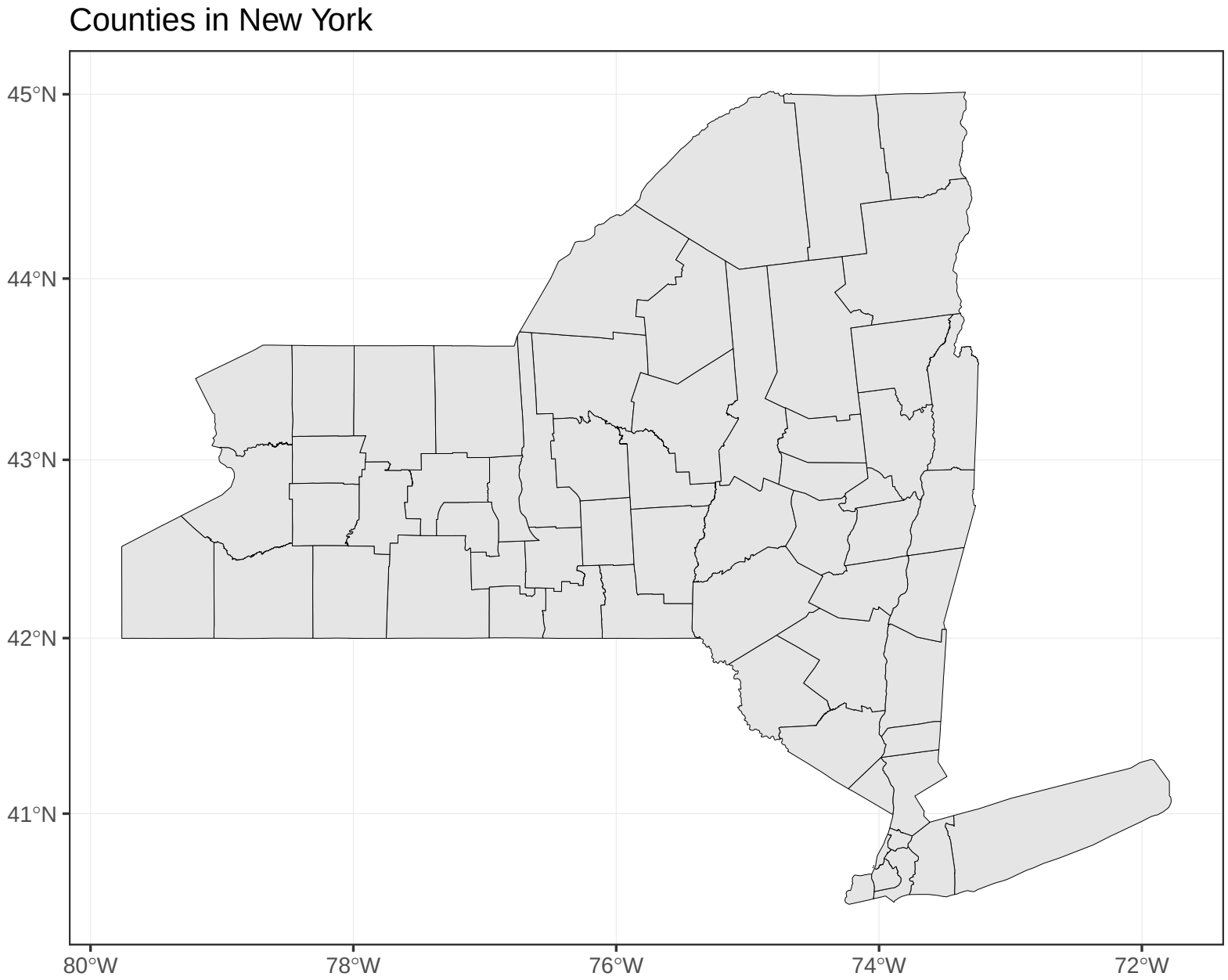
Regions



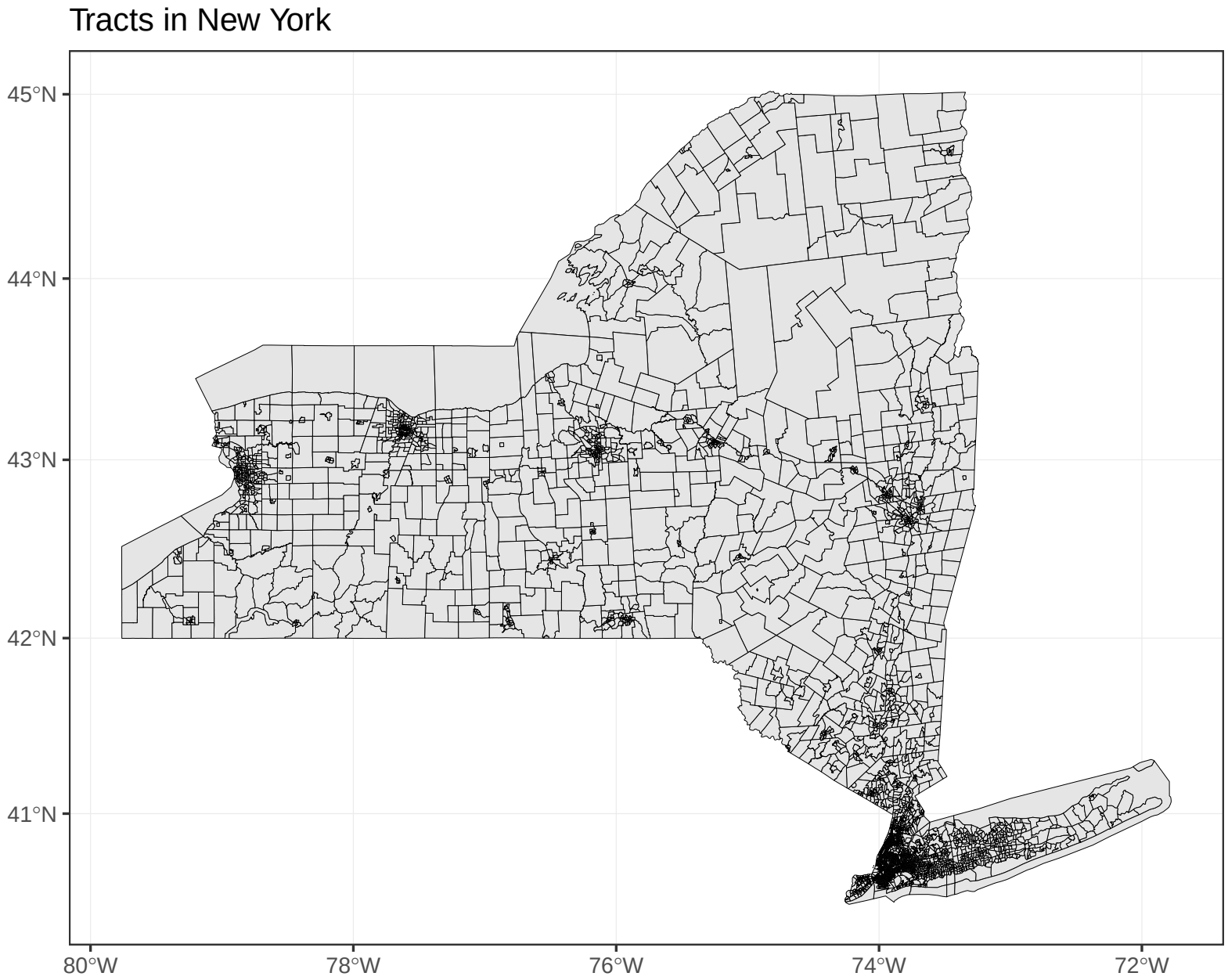
Divisions



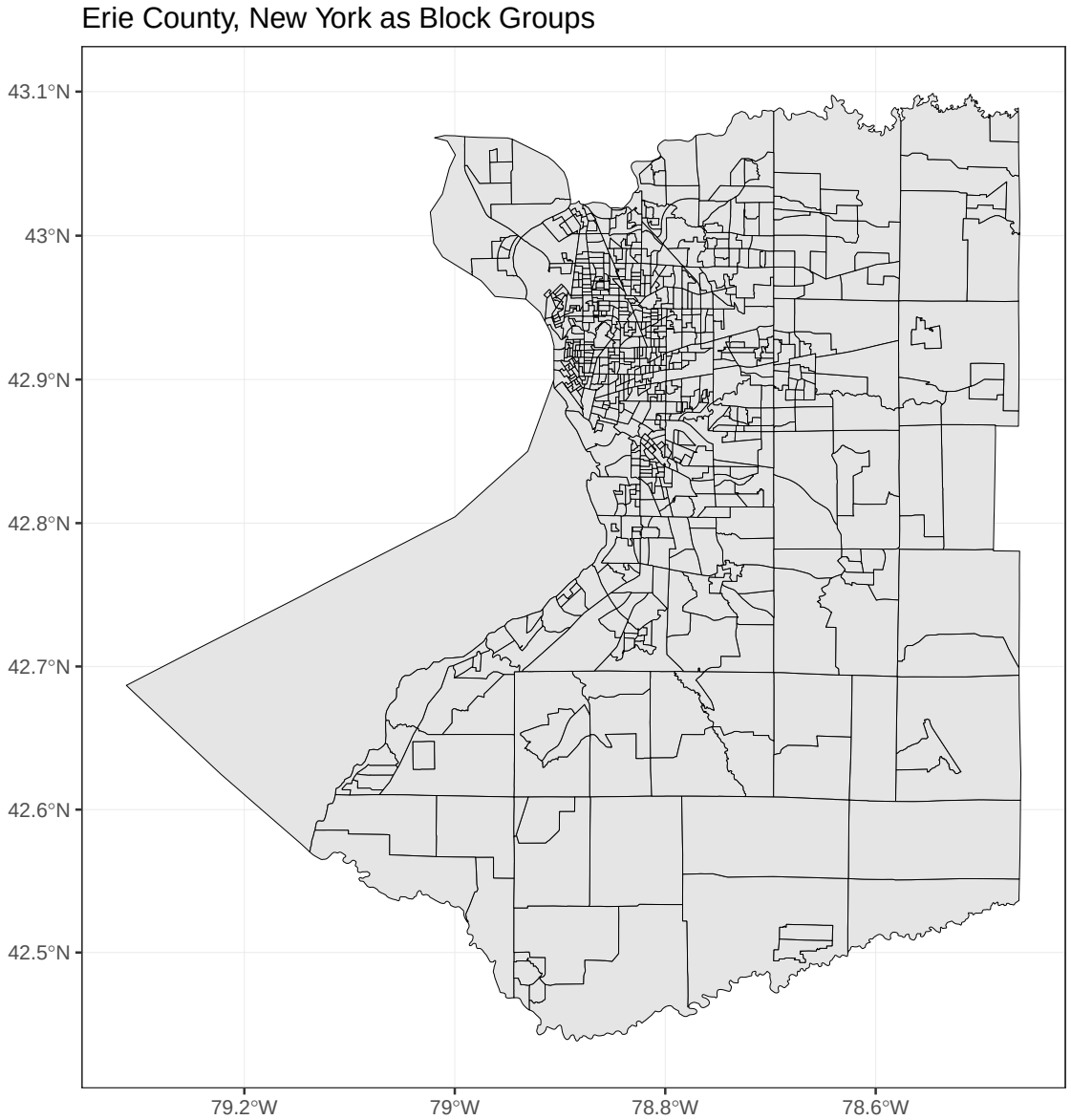
Counties



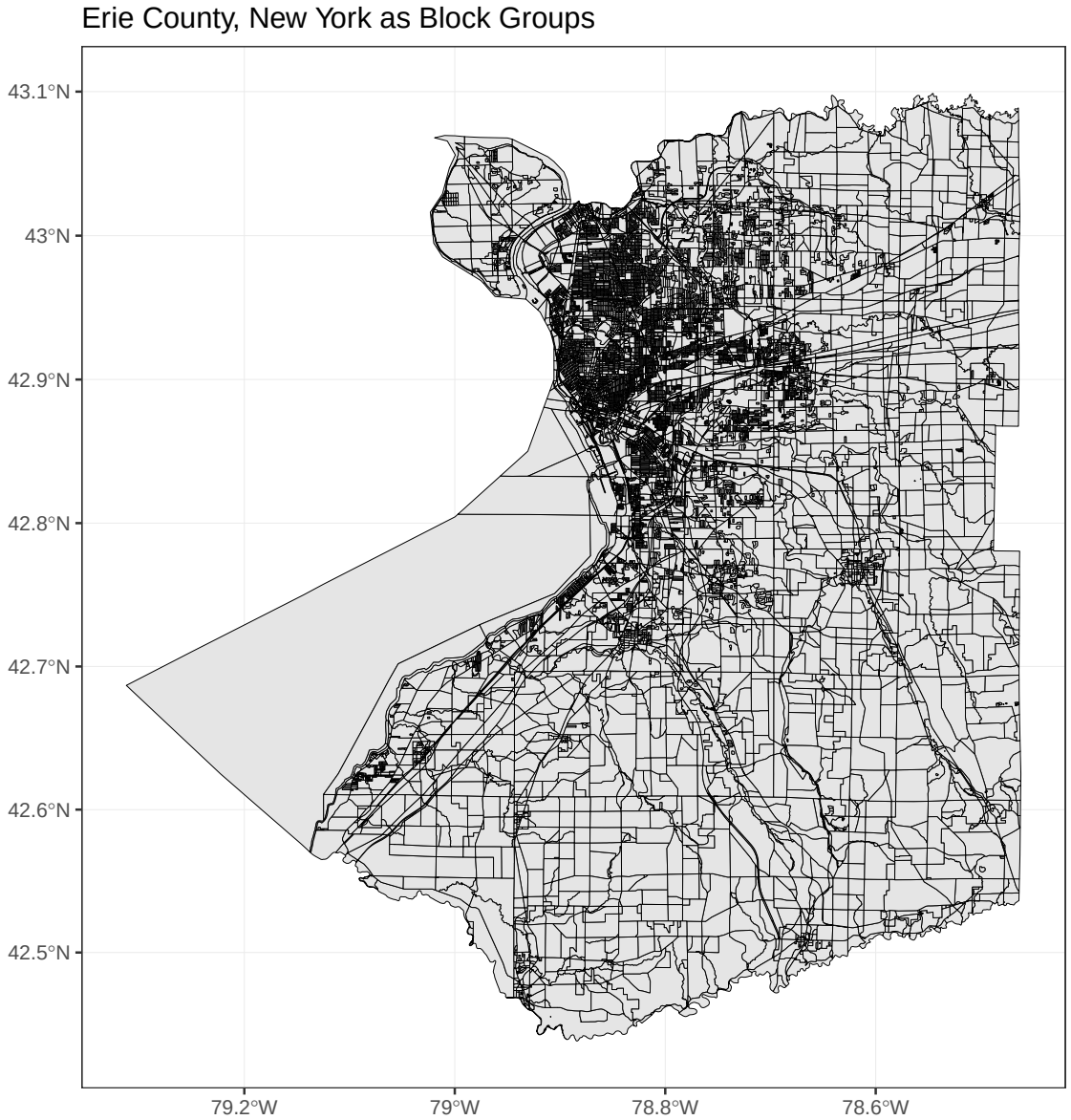
Tracts



Block Groups



Blocks


Erie County, New York as Block Groups
43.1°N
43°N
42.9°N
42.8°N
42.7°N
42.6°N
42.5°N
79.2°W
79°W
78.8°W
78.6°W

Census Geographies

Tracts:

- Around 1200-8000 people (optimal 4000)
- Nest within County boundaries
- Occasionally are split due to population growth or merged as a result of substantial population decline
- Each census tract contains at least one Block Group

Block Groups:

- Comprised of on average 39 blocks
- 600 - 3000 people

Block:

- In urban areas - small in area; for example, a block in a city bounded on all sides by streets.
- In suburban and rural areas - may be large, irregular, and bounded by a variety of features, such as roads, streams, and transmission lines.
- In remote areas - census blocks may encompass hundreds of square miles.

TIGER Shape files

- Core TIGER/Line Files and Shapefiles do not include demographic data
- They do contain geographic entity codes (GEOIDs) that can be linked to the Census Bureaus demographic data
- Available on American FactFinder.

My preferred way to access Census bureau shape files in R is through the `tigris` package.

Shape File Example

... Example in R ...

tigris-example.R

Creating Custom Shape file

Many times we want access to a shape file that is:

- Not directly accessible
- Made up of a collection of Census Geographies

For example,

1. the blocks that touch Alamo Square in San Francisco
2. the states with contact to the Canadian Boarder

Shape File Example

... Example in AFF then R ...

AFF-shapefile.R

Coordinate reference systems

Coordinate reference systems (CRS) provide a standardized way of describing locations. Many different CRS are used to describe geographic data. The CRS that is chosen depends on when the data was collected, the geographic extent of the data, the purpose of the data, etc. In R, when data with different CRS are combined it is important to transform them to a common CRS so they align with one another. This is similar to making sure that units are the same when measuring volume or distances.

EPSG codes A particular CRS can be referenced by its EPSG code (i.e., epsg:4121). The EPSG is a structured dataset of CRS and Coordinate Transformations. It was originally compiled by the, now defunct, European Petroleum Survey Group.

EPSG codes for commonly used CRS (in the U.S.)

Latitude/Longitude

- WGS84 (EPSG: 4326) - Commonly used by organizations that provide GIS data for the entire globe or many countries. CRS used by Google Earth
- NAD83 (EPSG:4269) - Most commonly used by U.S. federal agencies.
- NAD27 (EPSG: 4267) - Old version of NAD83

Projected (Easting/Northing)

- UTM, Zone 10 (EPSG: 32610) - Zone 10 is used in the Pacific Northwest
- Mercator (EPSG: 3857) - Tiles from Google Maps, Open Street Maps, Stamen Maps

United States™ Census Bureau

Coordinate Reference Systems

TIGER shape files are downloaded with NAD 1983 (EPSG:4269)

When data with different Coordinate Reference Systems (CRS) are combined it is important to transform them to a common CRS so they align with one another.

→ Same as making sure units are the same when comparing volumes or distances

In this workshop I will work primarily with CRS = 3857. This CRS is a *cylindrical map projection.*

- Many major online street mapping services use a variant of the Mercator projection for their map images

sp vs sf objects in R

sp objects

- Composed of lists inside lists (similar to SpatialPolygonsDataframe)
- Can be quite hard to decompose

sf objects

- Data-frames that are collections of spatial objects
- Each row is a spatial object (e.g. a polgyon)
- Each row may have data associated with it (e.g. its area) and a special geo variable that contains the coordinates.
- **easier to combine with our Census data**

- I will use sf objects in this workshop

United States™ Census Bureau

Example

... Code snip-it for conversion ...

sf-3857.R

Some of the Data We Collect

- Decennial Census of Population and Housing

  The U.S. census counts every resident in the United States. It is mandated by Article I, Section 2 of the Constitution and takes place every 10 years.
- Economic Census

  The Economic Census is the U.S. government's official five-year measure of American business and the economy.
- Census of Governments

  Identifies the scope and nature of the nation's state and local government sector including public finance and public employment and classifications.
- American Community Survey (ACS)

  The American Community Survey is the premier source for information about America's changing population, housing and workforce.
- Economic Indicators

  The Census Bureau releases 12 key economic indicators. Each indicator is released on a specific schedule.

Economic Indicators

- Data about economic activity
- Allows analysis of economic performance and/or predictions of future performance.
- Potential to move the financial markets

Principal Federal Economic Indicators

- Defined by the Office of Management and Budget
- Census produces 13 of the 38

US Economic Indicators

1. Advance Report on Durable Goods Manufacturers Shipments, Inventories, and Orders
2. Manufacturers Shipments, Inventories, and Orders
3. Monthly Wholesale Trade
4. Advance Monthly Sales for Retail and Food Services
5. Manufacturing and Trade Inventories and Sales
6. Quarterly Services Survey
7. Quarterly Financial Report Manufacturing, Mining, Wholesale Trade, and Selected Service Industries
8. Quarterly Financial Report Retail Trade
9. New Residential Construction
10. New Residential Sales
11. Construction Spending or Value of Construction Put in Place
12. Housing Vacancies and Homeownership
13. U.S. International Trade in Goods and Services

United States™ Census Bureau

List of all Surveys

- The amount of data is vast
- Unfortunately, there is not a single data-warehouse for each of our data products

→ In this workshop we will focus on a few that I believe are representative of the proceeder you might encounter while procuring our products.

▸ https://www.census.gov/programs-surveys/are-you-in-a-survey/survey-list.html

American Fact Finder

American FactFinder is the primary way to access data from:

- Decennial Census
- American Community Survey
- Puerto Rico Community Survey
- Economic Census
- Population Estimates Program
- Annual Economic Surveys

American Fact Finder In Action

Searching through AFF can be cumbersome unless you know the exact program and table you are looking for.

I will give some helpful tips to narrow down your selections

- Start with the most broad/important topic of interest
  - geography (county, block)
  - industry (NAICS code)
  - demographic (marital status, income)
- Narrow down what survey has your data
- Don't be afraid to start over

▸ https://factfinder.census.gov

Census AFF data merge with Shape file

... Direct download AFF data ...

shapefile-data-merge.R

Census API

Another way to access data directly through your favorite scripting language is thought the Census API.

- https://www.census.gov/developers/

In my experience the best way to use the API is as follows:

1. Get a key!
2. Look through available API's to see if your survey is included
3. 'Example and Supported Geographies' $\Longrightarrow$ 'variables' to find the variable code you want.
4. 'Example and Supported Geographies' $\Longrightarrow$ 'examples' for the calling string

... Access data in R through API ...

Census-API.R

- https://www.census.gov/developers/

Additional Data Products

What is the data you want is collected by a Census survey but not pretabulated on the American Fact Finder? For example,

- Customized table universe
  - → White females aged 40-50 vs all adults aged 18-65

Or alternatively,

- Sometimes it exists, but is not broken down by age or race or gender the way you want
- The table exist in exactly the format you want but there is not a comparable table for past years that you can use in a time series.

Public Use Microdata Sample (PUMS)

The Census Bureau releases PUMS

- A set of untabulated records about individual people or housing units
- Census Bureau produces the PUMS files so that data users can create custom tables that are not available through pretabulated (or summary) ACS data products

What you can expect if you work with PUMS files:

- Each row represents the responses to the Census questionnaire for a single person
- The first part of each row is information about the household (repeated for all members of the household)
- The second part contains responses to for individuals

How to access PUMs data

Obviously you can download directly from the Census Bureau...

I have found the most accessible way to access PUMS data is through IPUMS

▸ https://www.ipums.org/

- IPUMS is a part of the Institute for Social Research and Data Innovation at the University of Minnesota
- Provides census and survey data from around the world integrated across time and space
- Signature activity is harmonizing variable codes and documentation to be fully consistent across datasets
- Worlds largest accessible database of census microdata
- Includes almost a billion records from U.S. censuses from 1790 - present
- Over a billion records from the international censuses of over 100 countries

1940 Census Data

The '72-Year Rule' Governs Release of Census Records.

- The National Archives released individual-level records from the 1930 Census in 2002
- The National Archives released individual-level records from the 1940 Census in 2012
- The National Archives will release individual-level records from the 190 Census in 2022

▸ https://1940census.archives.gov/

- Full access to the 1940 census
- 1940 census maps and descriptions
- Browse census images to locate a person in the 1940 census

Seasonal Adjustment Software

The Census Bureau actively develops and maintains the seasonal adjustment software X-13ARIMA-SEATS.

https://www.census.gov/srd/www/x13as/

This is the primary method of removing seasonality before public release of data

- Retail Sales
- Residential Construction, Construction Spending
- International Trade, Imports & Exports
- Manufacturing, Manufacturers' Shipments, Inventories, & Orders

United States™ Census Bureau

Seasonal Adjustment Software

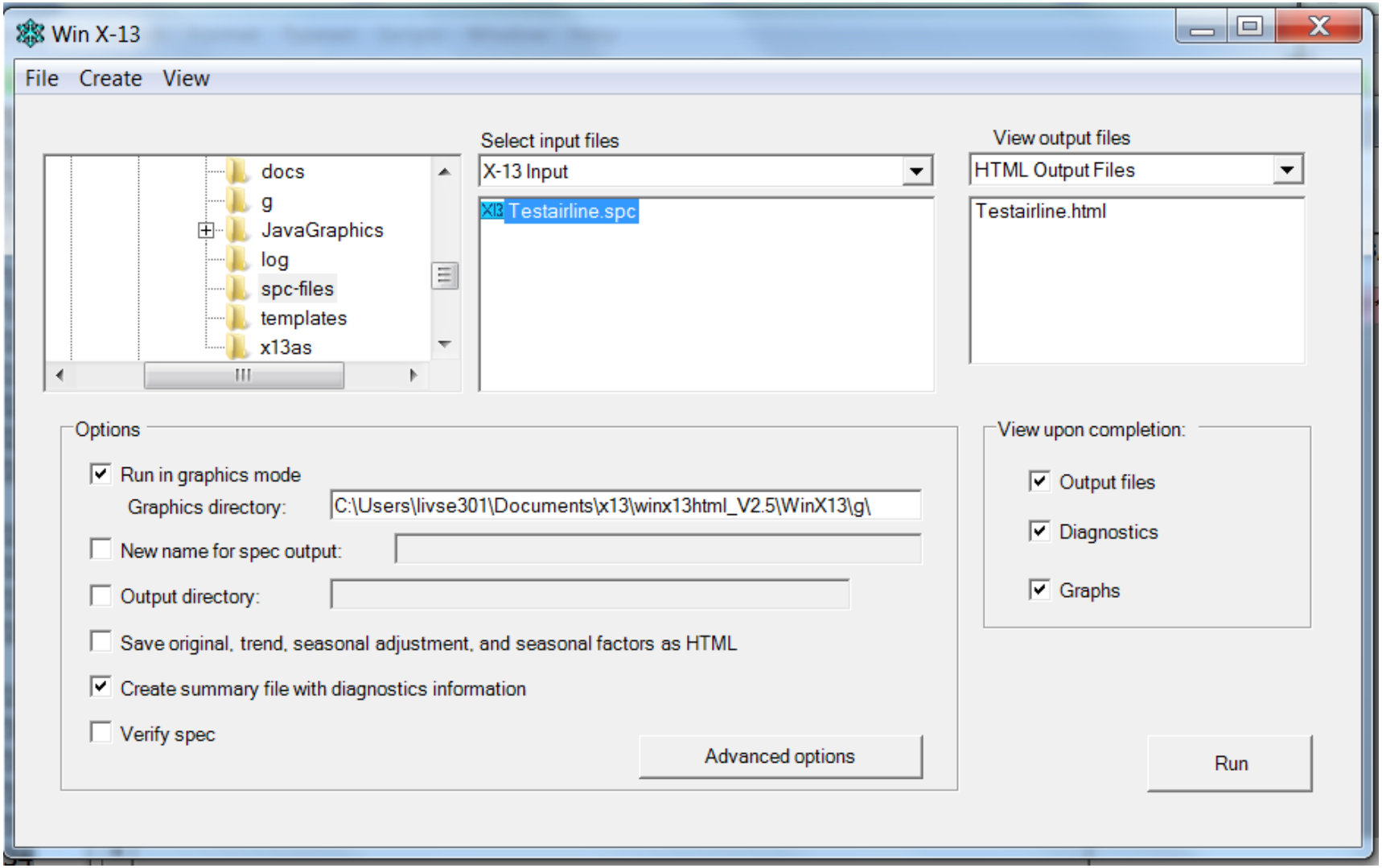
Many agencies release their X13AS input files (.spc files).

- https://www.bls.gov/web/empsit/cesseasadj.htm#sainputs

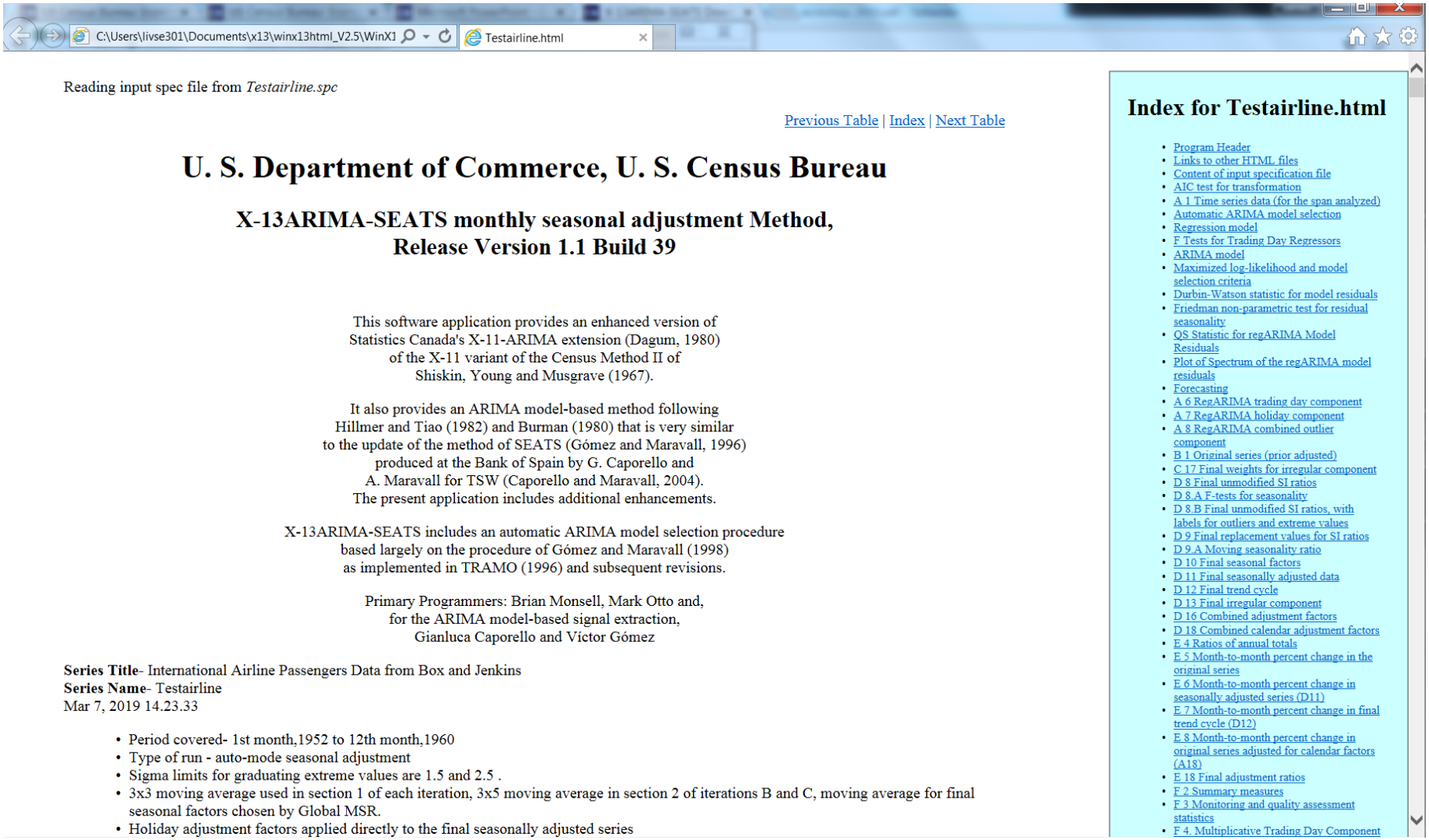
The X-13 program can be run as a stand alone program or in R with the `seasonal` package authored by Christoph Sax.

United States™ Census Bureau

Win X-13ARIMA-SEATS Program Interface



X-13ARIMA-SEATS Output



... Example: X-13ARIMA-SEATS in R ...

seasonal-example.R

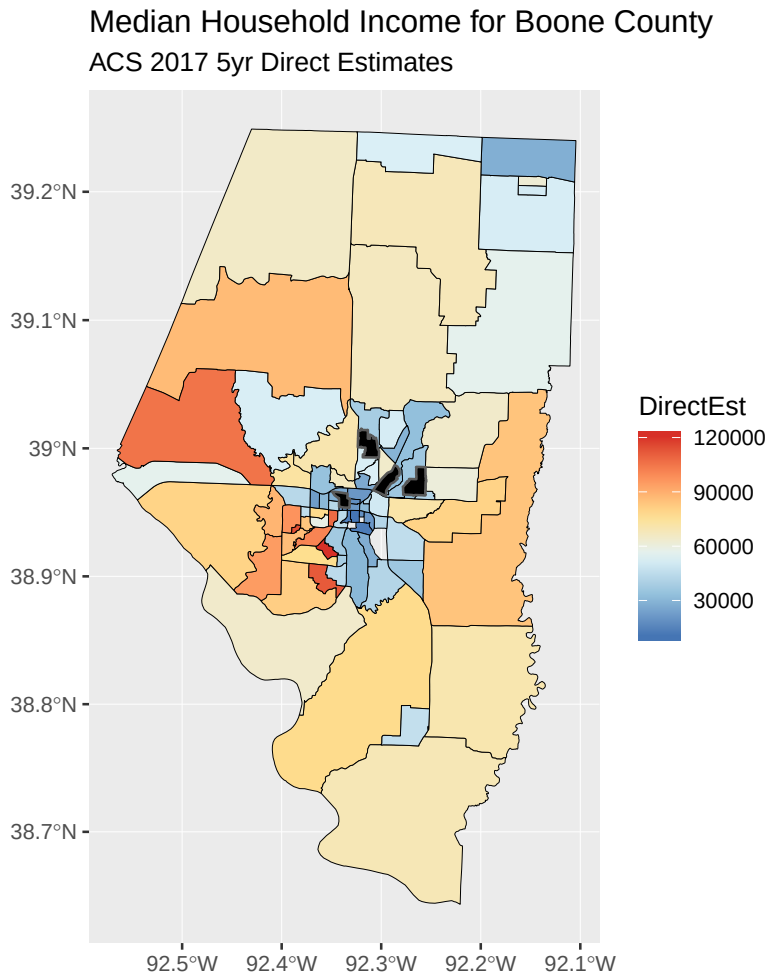
Switching gears...

Spatial Temporal Change of Support

Switching gears...

- Special Thank you to Andrew Raim and Scott Holan for there support...

Suppose we want estimates of these four neighborhoods in the City of Columbia in Boone County, Missouri.



The methods and software are not limited to use with ACS data, but I will focus on them in this workshop.

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 estimates for a user-specified support and period.
- Statistical agencies have direct access to microdata, and can aggregate to any support and period without STCOS methodology.
- See Bradley et al. (2015, Stat) and the references therein for a review of spatio(-temporal) change of support literature.

- Data users outside of the Census Bureau may be interested in custom geographies and/or nonstandard time periods which are not provided by the agency.
- Spatio-temporal change of support (STCOS) methodology provides model-based estimates for custom geographies and time periods using public-use ACS releases.
- This has recently been identified as an important problem by a National Academy of Sciences (NAS) panel (National Academy of Sciences 2015).

- The STCOS methodology makes use of spatio-temporal dependencies in the direct survey estimates and also incorporates associated direct survey variance estimates) through a Bayesian hierarchical model.
- Estimates, predictions, and appropriate measures of uncertainty can be extracted from the fitted model.

STCOS Geographies

- **Source supports**: Geographies on which direct estimates are available are used to fit the STCOS model
- **Target supports**: Geographies on which we want to produce estimates and predictions
- **fine-level support**: The STCOS methodology works by translating each of the source supports to the fine-level support during the model fitting process. Once the model has been fit, estimates and predictions on target supports of interest are obtained by translating from the fine-level support.

STCOS - Some notation

- $\mathcal{T} = \{T_L, \cdots, T_U\}$: times direct estimates available
- $\mathcal{L}$: set of lookback periods (For ACS can be 1,3,5)
- $D_{t\ell}$: source support (collection of areal units with direct estimates)
- $Z_t^{(\ell)}(A)$ and $\sigma_{t\ell}^2(A)$: direct survey estimate and associate variance, $A \in D_{t\ell}$
- $D_B = \{B_1, \cdots, B_{n_B}\}$: fine level support

The STCOS model is a Bayesian Hierarchical Model which can be described through three parts

1. Data model:

$$Z_t^{(\ell)}(A) = Y_t^{(\ell)}(A) + \epsilon_t^{(\ell)}(A)$$
$$\epsilon_t^{(\ell)}(A) \sim N(0, \sigma_{t\ell}^2(A))$$

2. Process model:

$$Y_t^{(\ell)}(A) = \underbrace{h(A)'\boldsymbol{\mu}_B}_{\text{course spatial trend}} + \underbrace{\psi_t^{(\ell)}(A)'\boldsymbol{\eta}}_{\text{fine spatio-temporal trend}} + \xi_t^{(\ell)}(A)$$

3. Prior model:

$$\boldsymbol{\mu}_B \sim N(0, \sigma_\mu^2 I), \boldsymbol{\eta} \sim N(0, \sigma_K^2 \mathbf{K}),$$
$$\sigma_\mu \sim IG(a_\mu, b_\mu), \sigma_K^2 \sim IG(a_K, b_K), \sigma_\xi^2 \sim IG(a_\xi, b_\xi)$$

Latent Process Model

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

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

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

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

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

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

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

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

Basis Functions

- We make use of local bisquare basis functions,

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

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

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

using a uniform random sample $\vec{u}_1, \ldots, \vec{u}_Q$ 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 $u \in A$,

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

- $h(A) = (|A \cap B_1|/|A|, \ldots, |A \cap B_{n_B}|/|A|)$ is computed from the source and fine-level supports.
- $\vec{\mu}_B = (\mu_1, \ldots, \mu_{n_B})$ is unknown, to be estimated from the data.

Specification of $\vec{K}$

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

$$\vec{Y}_t^* = \vec{\mu}_B + \vec{\nu}_t, \quad \text{for } t \in \mathcal{T}$$
$$\vec{\nu}_t = \vec{M}\vec{\nu}_{t-1} + \vec{b}_t, \quad \vec{b}_t \text{ iid } \mathrm{N}(\vec{0}, \sigma_K^2(\vec{I} - \vec{A})^-).$$

  where $\vec{A}$ is the adjacency matrix of D_B.
- Let $\vec{\Sigma}_{y^*}$ denote the covariance matrix of $(\vec{Y}_t^* : t \in \mathcal{T})$.
- Obtain $\vec{K}$ by solving

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

  which yields $\vec{K} = (\vec{S}^\top\vec{S})^{-1}\vec{S}^\top\vec{\Sigma}_{y^*}\vec{S}(\vec{S}^\top\vec{S})^{-1}$. The best positive approximation problem is discussed further in references.

United States™ Census Bureau

Specification of $\vec{K}$

- We propose several options where $\vec{\Sigma}_{y^*} = \sigma_K^2 \tilde{\vec{\Sigma}}_{y^*}$ such that $\tilde{\vec{\Sigma}}_{y^*}$ is free of unknown parameters and M does not need to be estimated. Here,

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

Specification of $\vec{K}$

- **(Independence)** Taking $\vec{K} = \vec{I}$ assumes no spatio-temporal covariance in $\vec{\eta}$.
- **(Spatial-only)** Let $\tilde{\vec{\Sigma}}_{y^*} = \sigma_K^2(\vec{I} - \vec{A})^- \otimes \vec{I}_{|\mathcal{T}|}$ to ignore covariance in time.
- **(Random Walk)** If $\vec{M} = \vec{I}$, the process

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

is a vector random walk with autocovariance

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

Take

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

Specification of $\vec{K}$

- **(Vector Autoregression)** Consider the VAR(1) process with $\vec{\mu_B} = \vec{X}\vec{\beta}$,

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

- One method suggests $\vec{M}$ as the eigenvectors of $(\vec{I} - \vec{P}_{\vec{X}})\vec{W}(\vec{I} - \vec{P}_{\vec{X}})$, where $\vec{P}_{\vec{X}} = \vec{X}(\vec{X}^\top\vec{X})^{-1}\vec{X}^\top$ and $\vec{W}$ is a pre-specified real-valued matrix.
- We take $\vec{X}$ to be a spatial-only bisquare basis expansion of the domain.

- Under VAR(1), the autocovariance becomes

$$\text{vec}(\vec{\Gamma}(0)) = [\vec{I} - \vec{M} \otimes \vec{M}]^{-1} \text{vec}(\sigma_K^2 (\vec{I} - \vec{A})^-)$$

$$\vec{\Gamma}(h) = \begin{cases} \vec{M}^h \vec{\Gamma}(0) & \text{if } h > 0, \\ \vec{\Gamma}(-h)^\top & \text{if } h < 0. \end{cases}$$

and therefore

$$\tilde{\vec{\Sigma}}_{y^*} = \begin{bmatrix} \vec{\Gamma}(0) & \vec{\Gamma}(-1) & \cdots & \vec{\Gamma}(-(|\mathcal{T}|-1)) \\ \vec{\Gamma}(1) & \vec{\Gamma}(0) & \cdots & \vec{\Gamma}(-(|\mathcal{T}|-2)) \\ \vdots & \vdots & \ddots & \vdots \\ \vec{\Gamma}(|\mathcal{T}|-1) & \vec{\Gamma}(|\mathcal{T}|-2) & \cdots & \vec{\Gamma}(0) \end{bmatrix}.$$

Specification of $\vec{K}$

- The usual formula

$$\text{vec}(\vec{\Gamma}(0)) = [\vec{I} - \vec{M} \otimes \vec{M}]^{-1}\text{vec}(\sigma_K^2(\vec{I} - \vec{A})^-)$$

  for the lag-0 autocovariance is computationally intractable for high dimensional series.

- A more tractable form is

$$\vec{\Gamma}(0) = \vec{V} \cdot \mathcal{M}\left(\text{Diag}(\vec{\Omega}) \circ \text{vec}(\vec{V}^{-1}\sigma_K^2(\vec{I} - \vec{A})^-\vec{V}^{-\top})\right) \cdot \vec{V}^\top,$$

  where $\vec{V}$ and $\vec{\lambda}$ are the eigenvectors/values of $\vec{M}$, $\circ$ is elementwise multiplication, $\mathcal{M} = \text{vec}^{-1}$, and $\vec{\Omega} = \text{Diag}(\vec{1} - \vec{\lambda} \otimes \vec{\lambda})^{-1}$.

We may write

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

where

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

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

United States™ Census Bureau

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 $\vec{S}$ using principal components analysis.
- Suppose $\vec{U}\vec{D}\vec{U}^\top$ is the eigen-decomposition of $\vec{S}^\top\vec{S}$, and $\tilde{\vec{U}}$ contains the r' columns of $\vec{U}$ corresponding the $r' \leq r$ largest magnitude eigenvalues in $\vec{D}$.
- The transformation $T(\vec{S}) = \vec{S}\tilde{\vec{U}}^\top$ is applied to all matrices computed from the basis functions.

United States™ Census Bureau

Gibbs Sampler

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

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

- $[\vec{\eta} \mid \cdot] \sim \mathrm{N}(\vec{\vartheta}_\eta, \vec{\Omega}_\eta^{-1})$,

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

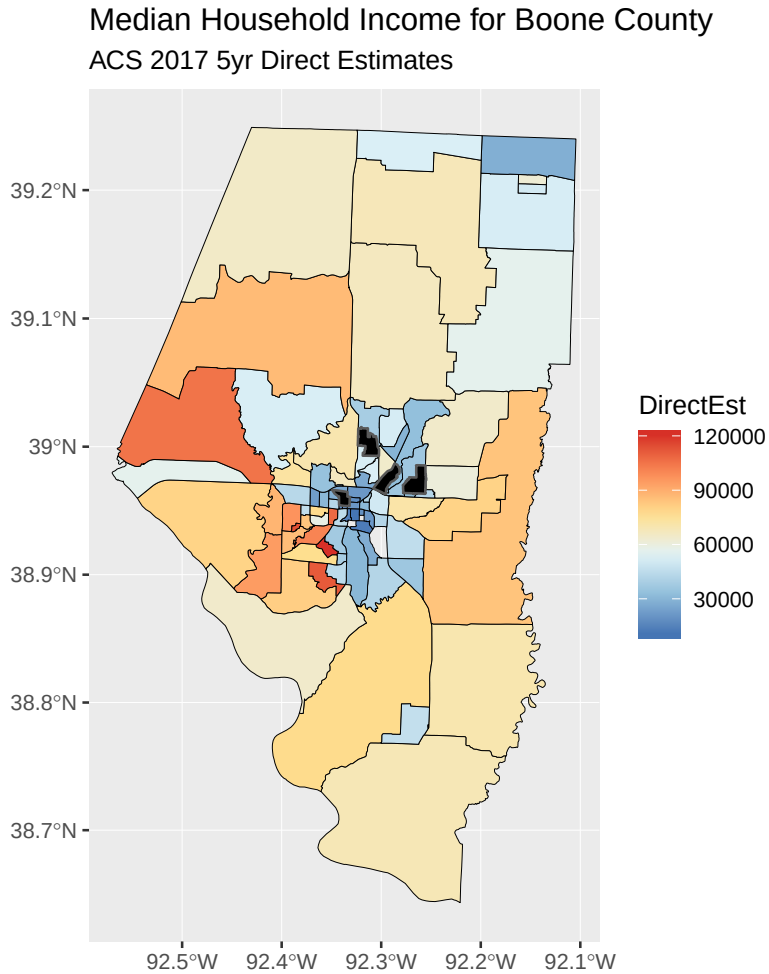
- $[\vec{\xi} \mid \cdot] \sim \mathrm{N}(\vec{\vartheta}_\xi, \vec{\Omega}_\xi^{-1})$,

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

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

Example

Want estimates of these four neighborhoods in the City of Columbia in Boone County, Missouri.



Example: Goals

- Want to produce model-based estimates of median household income
- Target Support: four specified neighborhoods
- Based on 5-year ACS estimates for block-groups in Boone County, Missouri from 2012, 2013, 2014, and 2015.
- Thus, 2012, 2013, 2014, 2015 year block-groups will be our source supports.

Example: Outline

1. Load Fine-level Support (block-groups Boone County, MO)
2. Use Census API to load median household income
3. Load target support (same CRS as fine-level)
4. Construct spatial-temporal knots for basis functions with space-filling swapping algorithm
5. Load "known" parts of model: Z, V, H, S
6. Dimension reduction through eigen-decomposition of $S'S$
7. Choose covariance structure of K
8. Run Gibbs sampler

United States™ Census Bureau

... STCOS Package ...

STCOS-example.R

Thank you.

Email: James.A.Livsey@census.gov

United States™ Census Bureau