Hierarchical Bayesian spatio-temporal models for mass imputation of CPI in small areas from sparse survey data

Vladislav Beresovsky [1], Terrance D. Savitsky [1] and Jeff Gonzalez [1]

[1] U.S. Bureau of Labor Statistics
Office of Survey Methods Research

2026 COPAFS Quaterly Meeting

March 13, 2026

***Disclaimer**: This presentation provides a summary of research results. The information is being released for statistical purposes, to inform interested parties, and to encourage discussion of work in progress. The presentation does not represent an existing, or a forthcoming new, official BLS statistical data product or production series.*

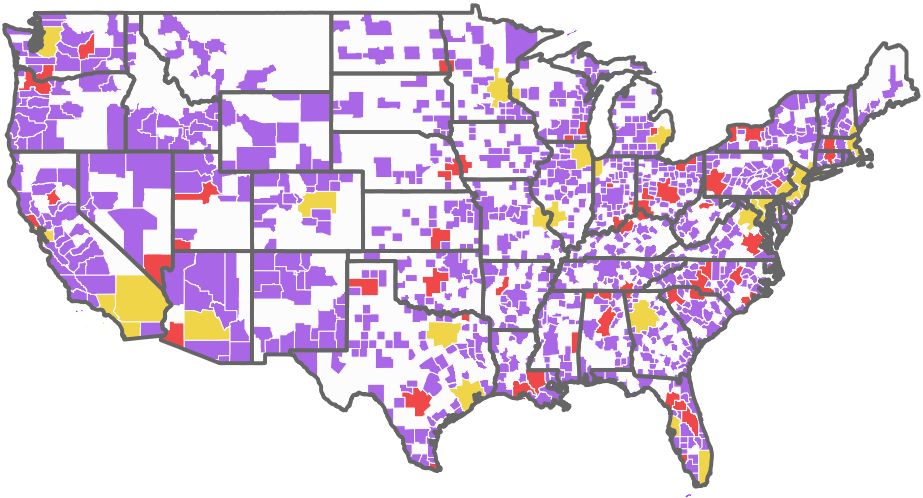
Outline

CPI program overview

- CPI-U population covers 93% of the U.S. population
- BLS collects prices for $\approx 100,000$ goods and services paid by urban households in core-based statistical areas (CBSA) (BLS Handbook of Methods)
- BLS calculates 7,776 basic indexes series $I^c_{a[t]}$ for
  - a - 32 geographic index areas in contiguous USA (9 Census Divisions + 23 large CBSA)
  - c - 243 commodity items
  - t - month of data collection
- Indexes are normalized to 100 starting from a base period (originated in 1983 for most items)
- Percent change (PC) over the last $p = 1$ or $p = 12$ months

$$\left(\frac{I^c_{a[t]}}{I^c_{a[t-p]}} - 1 \right) * 100\%$$

Sparse geographic coverage of the CPI survey



Outline

CBSA are building blocks for SAE

- $\hat{Y}_j, \hat{V}_j$ sampling estimates and sampling errors in $j \in 1, \ldots, n = 73$ sampled CBSA in contiguous USA
- Model estimates $\hat{\theta}_i$ in $i \in 1, \ldots, N = 894$ population CBSA
- $\hat{\theta}_i$ are building blocks for aggregation of any domain d

$$\hat{\theta}_d = \frac{\sum_{i \in d} N_i \hat{\theta}_i}{\sum_{i \in d} N_i}$$

N_i - population in CBSA
d - state, Census Division or any aggregation of CBSA

Borrowing strength: fixed effects, space and time

Model mean in all population CBSA: $\theta_i, i \in 1, \ldots, N = 894$

1. FH: Fay-Herriot area-level model $\left(\Sigma_p \in \mathbb{R}^{P \times P}, \sigma_u^2\right)$:

$$\theta_i = \mathbf{X}_i \boldsymbol{\beta} + u_i$$

$$(\beta_1, \ldots, \beta_P) \sim \mathcal{N}(0, \Sigma_p),\ (u_1, \ldots, u_N) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2 \mathbf{I}),\ \Sigma_p \in \mathbb{R}^{P \times P}$$

2. CS: Cross-sectional Spatial model $\left(\Sigma_p \in \mathbb{R}^{P \times P}, \Sigma_s \in \mathbb{R}^{N \times N}\right)$:

$$\theta_i = \mathbf{X}_i \boldsymbol{\beta} + u_i,$$

$$(\beta_1, \ldots, \beta_P) \sim \mathcal{N}(0, \Sigma_p),\ (u_1, \ldots, u_N) \sim \mathcal{N}(0, \Sigma_s),\ \Sigma_p \in \mathbb{R}^{P \times P}, \Sigma_s \in \mathbb{R}^{N \times N}$$

3. TS6 and TS12: Spatio-Temporal models, $t \in 1, \ldots, T = 6$ or 12 months:

$$\theta_{it} = \mathbf{X}_i \boldsymbol{\beta_t} + u_{it},$$

$$(u_{11}, \ldots, u_{1T}, u_{21}, \ldots, u_{NT}) \sim \mathcal{N}\left(0, \Sigma_{s \times t}\right),\ \Sigma_{s \times t} = \Sigma_s \otimes \Sigma_t \in \mathbb{R}^{NT \times NT}$$

$$(\beta_{11}, \ldots, \beta_{1T}, \beta_{21}, \ldots, \beta_{PT}) \sim \mathcal{N}\left(0, \Sigma_{p \times t}\right),\ \Sigma_{p \times t} = \Sigma_p \otimes \Sigma_t \in \mathbb{R}^{PT \times PT}$$

Outline

Spatial model: basic formulation

$\mathbf{W} = (W_1, \ldots, W_N)^T$ - field of spatially correlated random effects in CBSA
$i \in 1, \ldots, N$

Spatial correlations depend on distances s_{ij} between centroids of CBSA (i, j)
$\mathbf{A}_{ij} = s_{ij}^{-1}$ - spatial proximity matrix
$\mathbf{Q} = \text{diag}(\mathbf{A1}) - \mathbf{A} \in \mathbb{R}^{N \times N}$ - precision matrix

Model
$\theta_i = \mathbf{X}_i^T \boldsymbol{\beta}_x + \mathbf{W}_i$ - fixed effects + field of random effects
$p(\mathbf{W}|\tau) \propto \tau^{\text{rank}(\mathbf{Q})/2} \exp\left(-\frac{\tau}{2}\mathbf{W}^T\mathbf{Q}\mathbf{W}\right)$ - ICAR prior

Model dimension $D = N + P + 1$

Confounding between fixed and spatial effects

$C(\mathbf{X})$ - space of fixed effects with basis $\mathbf{K} \in \mathbb{R}^{N \times P}$
$C(\mathbf{X})^{\perp}$ - orthogonal space with basis $\mathbf{L} \in \mathbb{R}^{N \times (N-P)}$

$\mathbf{W}_i = \mathbf{K}_i \boldsymbol{\gamma} + \mathbf{L}_i \boldsymbol{\delta}$ - field of random effects decomposed

$\theta_i = \mathbf{X}_i^T \boldsymbol{\beta}_x + \mathbf{L}_i \boldsymbol{\delta}$ - model without confounding between $\mathbf{X}$ and $\mathbf{Q}$

$\mathbf{L} = \mathbf{P}^{\perp} = \mathbf{I} - \mathbf{P}$ - orthogonal basis, where $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is projection operator

$p(\boldsymbol{\delta}|\tau) \propto \tau^{(N-P)/2} \exp\left(-\frac{\tau}{2}\boldsymbol{\delta}^T\mathbf{Q}^{\perp}\boldsymbol{\delta}\right)$ - orthogonal ICAR prior
$\mathbf{Q}^{\perp} = \mathbf{L}^T\mathbf{Q}\mathbf{L}$ - orthogonal precision matrix

Model dimension $D = N + 1$

Sparse reparameterization of random effects

The reparameterized vector of random effects

1. have much smaller dimension $Q \ll N$
2. include only positive spatial dependence, i.e. attraction

$M_{\mathbf{X}} = \mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}$ is the Moran operator *with respect to* $\mathbf{X}$ and the corresponding Moran's I-statistic is

$$I_{\mathbf{X}}(\mathbf{W}, \mathbf{A}) = \frac{N}{\mathbf{1}^T\mathbf{A}\mathbf{1}} \frac{\mathbf{W}^T\left(\mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}\right)\mathbf{W}}{\mathbf{W}^T\mathbf{P}^{\perp}\mathbf{W}}.$$

Matrix $\mathbf{M}_S \in \mathbb{R}^{N\times Q}$ of eigenvectors (harmonics) of the Moran operator $M_{\mathbf{X}}$ corresponding to the first Q largest eigenvalues $\lambda_1 > \cdots > \lambda_Q > 0$.

$$\theta_i = \mathbf{X}_i^T\boldsymbol{\beta}_x + \mathbf{M}_{Si}\boldsymbol{\beta}_S$$

$$p\left(\boldsymbol{\beta}_S|\tau\right) \propto \tau^{Q/2}\exp\left(-\frac{\tau}{2}\boldsymbol{\beta}_S^T\mathbf{Q}_S\boldsymbol{\beta}_S\right)$$

$$\mathbf{Q}_S = \mathbf{M}_S^T\mathbf{Q}\mathbf{M}_S$$

Model dimension $D = P + Q + 1$

Temporal correlations with Gaussian Process (GP) priors

$\boldsymbol{\beta}_t$ - model parameters in collection periods $t \in 1, \ldots, T$.
GP priors allow modeling functional dependence $\boldsymbol{\beta}_t \sim \beta\,(t)$

$\boldsymbol{\beta}_t \sim \mathcal{N}\,(0, K\,(t|\alpha, \lambda))$

$K\,(t|\alpha, \lambda)$ - GP kernel function

α - marginal SD, controls the range of variability of β_t
λ - length-scale, controls the frequency of fluctuations

GP standard kernels

Exponential Quadratic

$$K^{\text{EQ}}(t_i, t_j|\alpha, \lambda) = \alpha^2 \exp\left(-\frac{(t_i - t_j)^2}{2\lambda^2}\right)$$

Matern 3/2

$$K^{\text{M32}}(t_i, t_j|\alpha, \lambda) = \alpha^2 \left(1 + \frac{\sqrt{3}|t_i - t_j|}{\lambda}\right) \exp\left(-\frac{\sqrt{3}|t_i - t_j|}{\lambda}\right)$$

Matern 5/2

$$K^{\text{M52}}(t_i, t_j|\alpha, \lambda) = \alpha^2 \left(1 + \frac{\sqrt{5}|t_i - t_j|}{\lambda} + \frac{5|t_i - t_j|^2}{3\lambda^2}\right) \exp\left(-\frac{\sqrt{5}|t_i - t_j|}{\lambda}\right).$$

Ideas for choosing the right kernel

1. Combine diverse kernels, use informative priors for hyper parameters. More flexible, runs very slow

$$\Sigma_{x|s} = K^{\text{M52}}\left(\alpha_{x|s}, \lambda_{x|s,1}\right) + K^{\text{EQ}}\left(\alpha_{x|s}, \lambda_{x|s,2}\right)$$
$$\lambda_{x|s,1} \sim IG(4,1), \bar{\lambda} = \frac{1}{3}$$
$$\lambda_{x|s.2} \sim IG(4,5), \bar{\lambda} = \frac{5}{3}$$
$$\alpha \sim \mathcal{N}^{+}(0,1)$$

2. Use standard kernels with *constant and reasonable* parameters. Less flexible, runs ≈ 80 times faster

$$\Sigma_{x|s} = K^{\text{EQ}}\left(\alpha_{x|s} = 1, \lambda_{x|s,2} = 1\right)$$

3. Depending on the data, Approach #2 may give acceptable results much faster. Need more experimenting.

Outline

Calibration to direct estimates

$(\hat{Y}_d, \hat{V}_d)$ - direct sampling estimates in Census Divisions $d \in 1, \ldots, 9$.

Calibration equation for CBSA estimates $\theta_i,\ i \in d$:

$$\hat{Y}_d \sim N\left(\theta_d, \hat{V}_d^2\right)$$

$$\theta_d = \frac{\sum_{i \in d} N_i \theta_i}{\sum_{i \in d} N_i}$$

Global-local priors for fixed effects

Fixed effects $\mathbf{X} \in \mathbb{R}^{N \times P}$, $\boldsymbol{\beta}_x^p$, $p \in 1, \ldots, P$ - model coefficients

$$\boldsymbol{\beta}_x^p \sim N(0, \sigma_p^2)$$

$$\boldsymbol{\sigma}_p | \sigma_G \sim C^+(0, \sigma_G), \ \sigma_G | \sigma \sim C^+(0, \sigma)$$

$$C^+(x|0, a) = \frac{1}{\pi} \frac{a}{x^2 + a^2}, x > 0 \text{ - half-Cauchy distribution}$$

$$E(\boldsymbol{\beta}_x^p | \mathbf{Y}) = \{1 - E(\kappa_p | y)\} \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \left(\mathbf{X}^T \mathbf{Y}\right)_p$$

$$\kappa_p = 1 / \left(1 + \sigma_p^2\right) \sim \text{Beta}(0.5, 0.5)$$

κ_p - shrinkage coefficient with $\cup$-shaped "horseshoe" prior

Smoothed sampling variances via co-modeling (aka GVF)

Sampling equation of the Fay Herriot model

$$\hat{Y}_j|\theta_j, \hat{V}_j^2 \overset{\text{ind}}{\sim} N\left(\theta_j, \hat{V}_j^2\right) \text{ - no co-modeling}$$

Sampling equations for co-modeling of means and variances.
ν_j^2 - latent variance.

$$\begin{cases} \hat{Y}_j|\theta_j, \nu_j^2 \overset{\text{ind}}{\sim} N\left(\theta_j, \nu_j^2\right) \\ \hat{V}_j^2|a, b, \nu_j^2 \overset{\text{ind}}{\sim} G\left(\frac{an_j^*}{2}, \frac{an_j^*}{2b\nu_j^2}\right), \nu_j^2 \overset{\text{ind}}{\sim} IG\left(2, 1\right) \end{cases}$$

Outline

Fuel price change over 12-month (%)

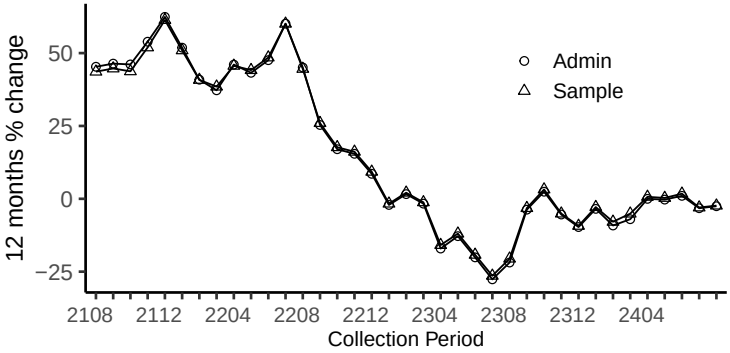


Figure: National average of 12-month change (%) in fuel prices in CBSA estimated from the data of CPI survey (△) and administrative dataset (○) by monthly collection periods from 202107 to 202407.

- Estimation from the CPI survey sample of fuel prices
- Validation against Admin data

Outline

Model fit averaged over collection periods

			Sampled CBSA				Non-sampled CBSA			
Model	Time (s)	WAIC	SD	95% CI	RMSE	R^2	SD	95% CI	RMSE	R^2
FH	60	199	2.6	.91	3.1	.68	3.5	.82	4.7	.08
CS	188	189	2.6	.93	2.4	.75	4.3	.86	4.4	.23
TS6	**921**	**166**	**2.6**	**.97**	**1.9**	**.83**	**5.4**	**.93**	**4.1**	**.31**
TS12	2661	170	2.7	.97	1.9	.83	6.6	.96	4.4	.28

Table: Summary of model fit averaged over collection periods from 202310 to 202407

WAIC - Widely Applicable Information Criteria

$$\bar{y}^{\text{adm}} = \frac{1}{N}\sum_i y_i^{\text{adm}}, \bar{\hat{\theta}}^{\text{mod}} = \frac{1}{N}\sum_i \hat{\theta}_i^{\text{mod}}$$

$$SS_{\text{tot}}^{\text{adm}} = \sum_i \left(y_i^{\text{adm}} - \bar{y}^{\text{adm}}\right)^2,\ SS_{\text{tot}}^{\text{mod}} = \sum_i \left(\hat{\theta}_i^{\text{mod}} - \bar{\hat{\theta}}^{\text{mod}}\right)^2,\ SS_{\text{res}} = \sum_i \left(y_i^{\text{adm}} - \hat{\theta}_i^{\text{mod}}\right)^2$$

$$\text{SD} = \sqrt{SS_{\text{tot}}^{\text{mod}}/N},\ \text{RMSE} = \sqrt{SS_{\text{res}}/N},\ R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}^{\text{adm}}}$$

LOO-CV and WAIC by collection periods

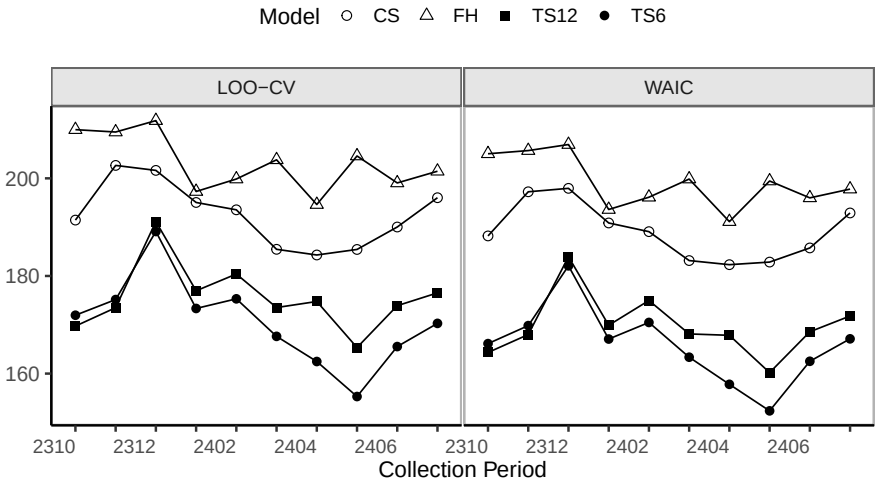


Figure: Leave One Out Cross Validation (LOO-CV) and Widely Applicable Information Criteria for *FH*, *CS*, *TS6* and *TS12* models. Data collection periods from 202310 to 202407.

SD and 95% CI coverage by collection periods

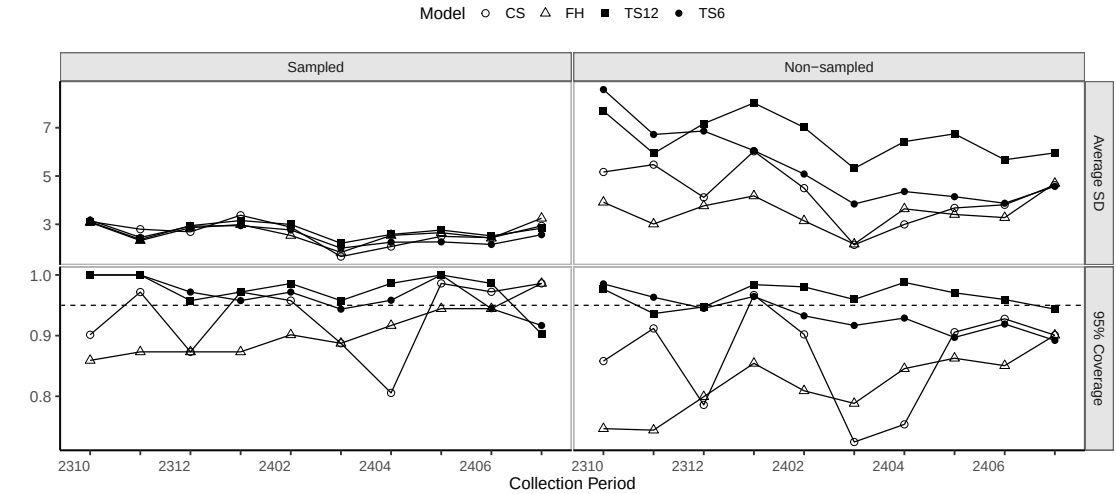


Figure: Average SD and 95% CI coverage rate of the estimates by *FH*, *CS*, *TS6* and *TS12* models for the sampled and non-sampled CBSA. Data collection periods from 202310 to 202407.

RMSE and R^2 by collection periods

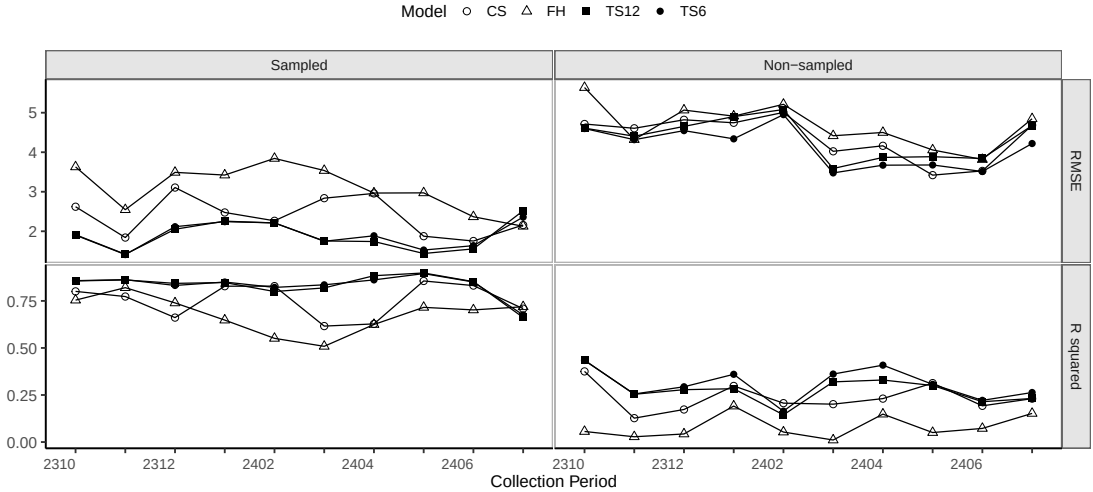


Figure: RMSE and coefficient of determination R^2 of the estimates by models *FH*, *CS*, *TS6* and *TS12* models relative the estimates from the administrative data (○). Data collection from 202310 to 202407.

States: Model vs Admin data

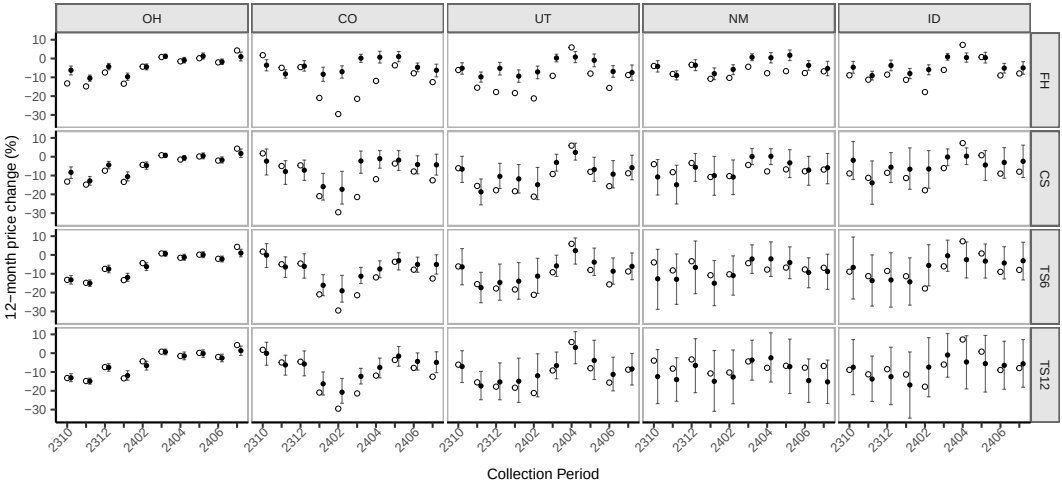


Figure: Model-based estimates (•) in Ohio (OH), Colorado (CO), Utah (UT), New Mexico(NM) and Idaho(ID) with 90% credible intervals by *FH*, *CS*, *TS6* and *TS12* models relative to the "gold-standard" estimates (○).

Census Divisions: Model vs Admin data

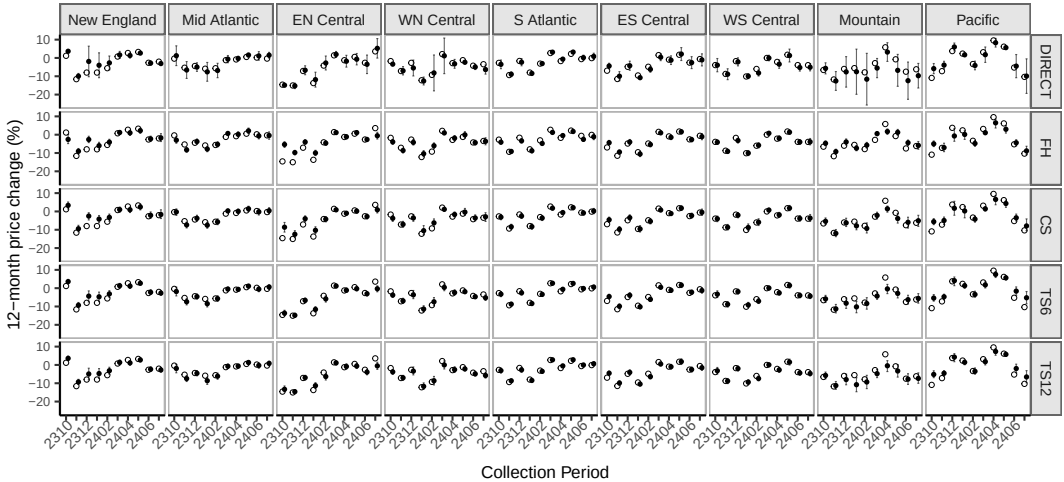


Figure: Model-based estimates (•) in Census Divisions with 90% credible intervals by *FH*, *CS*, *TS6* and *TS12* models relative to the "gold-standard" estimates from administrative data (○).

CBSA: socio-demographic groups

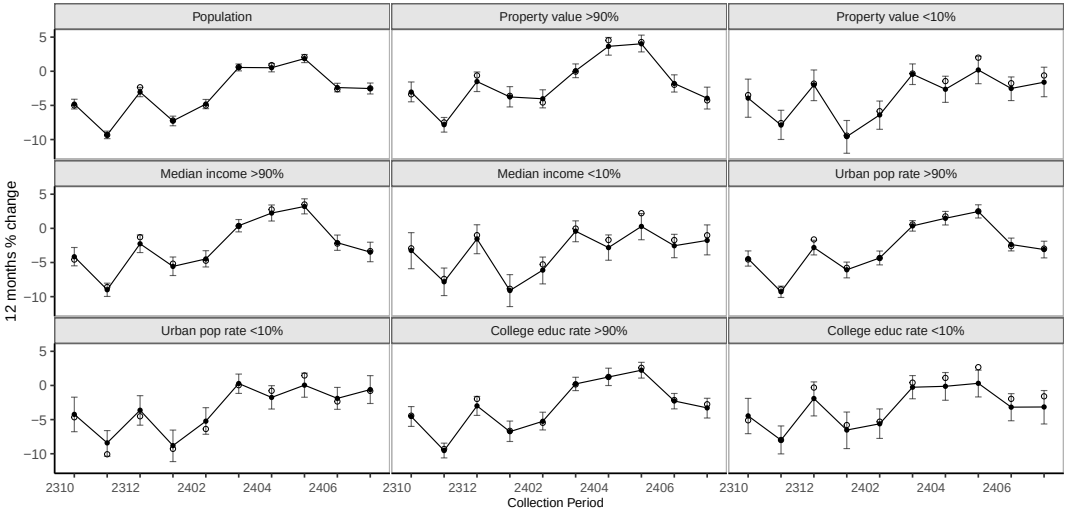


Figure: *TS6* model estimates (●) in socio-demographic groups with 90% credible intervals relative to the "gold-standard" estimates from administrative data (○).

Outline

Conclusions

- Why do spatio-temporal models improve over the Fay-Herriot model?
  - Sparse geography of CBSA sample
  - Low explanatory power of socio-demographic fixed effects
  - Strong spatio-temporal correlations of CPI survey data
- Why *TS6* model is better than *TS12*?
  - Data aggregation over collection periods reduces Bias and increases Variance. Same old Bias vs Variance dilemma
- New analyzes with CBSA building blocks
  - Estimates with proper uncertainty measures in any arbitrary geography: States, Census Divisions, North California, South Florida, North and South Carolinas, etc
  - Define any aggregation of CBSA based on socio-demographic characteristics: low income, high education, high % of manufacturing employment, etc. Make reliable inferences for time series

Vladislav Beresovsky
Office of Survey Research and Methodology
beresovsky.vladislav@bls.gov

Thanks to the colleagues from the Office of Prices and Living Conditions:

Bill Johnson
Jenny Fitzgerald
Steven Paben
Rob Cage
Alex Traczyk
Johanky Reyes