

# Challenges of estimating inflation in small areas in official statistics

Vladislav Beresovsky <sup>1</sup>, Terrance D. Savitsky <sup>1</sup>

<sup>1</sup> U.S. Bureau of Labor Statistics  
Office of Survey Methods Research

JSM 2024, Portland, OR

Aug 7, 2024

***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.*

— U.S. BUREAU OF LABOR STATISTICS • [bls.gov](https://bls.gov)

# Acknowledgement

I am grateful to colleagues from the Office of Prices and Living Conditions (OPLC) who introduced me to the CPI program and provided access to their data:

Bill Johnson

Jenny Fitzgerald

Alex Traczyk

Johanky Reyes

Steven Paben

Rob Cage

Thank you!

# Outline

Introduction to Consumer Price Index (CPI)

Modeling price changes (PC) in CBSA

Model fit evaluation

Discussion and next steps

References

# Consumer price Indexes and price changes (PC)

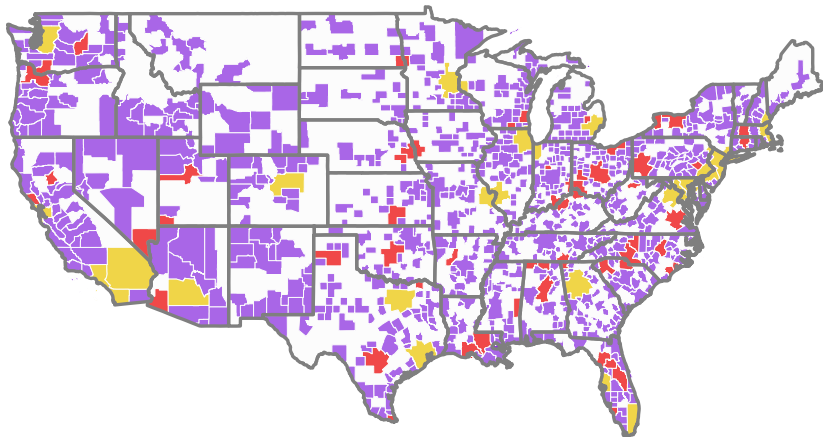
- ▶ The BLS collects prices for  $\approx 100,000$  goods and services paid by urban households of core-based statistical areas (CBSAs).
- ▶ CPI-U population covers 93% of the U.S. population.
- ▶ The BLS calculates 7,776 basic indexes series  $I_{a[t]}^c$  for
  - ▶ 32 geographic index areas  $a$
  - ▶ 243 commodities  $c$
  - ▶ month  $t$
- ▶ The published Indexes are normalized to 100 starting from a base period (originated in 1983 for many items).
- ▶ PC over the last  $p$ -months. Annual PC are the most referenced.

$$Y_{a[t]}^c = \frac{I_{a[t]}^c}{I_{a[t-p]}^c} - 1.$$

# Geography of CPI sampling and estimation

- ▶ **Self-representing (SR) CBSA**
  - ▶ 21 continental large metropolitan CBSA (pop > 2.5M) + AK & HI
  - ▶ SR CBSA represent 42% of CPI-U population
  - ▶ Estimates are published for all SR CBSA
- ▶ **Non self-representing (NSR) CBSA**
  - ▶ 52 NSR CBSA are sampled from the population of  $\approx$  900 metro- and micropolitan CBSA covering 58% of CPI-U population
  - ▶ Data collected in NSR CBSA is used to produce estimates in 9 Census Divisions
  - ▶ **NSR CBSA sample is sparse geographically**

# Sparse geographic coverage of the CBSA sample



Sampling Status ■ NSR ■ SR ■ Unsampled

# Outline

Introduction to Consumer Price Index (CPI)

Modeling price changes (PC) in CBSA

Model fit evaluation

Discussion and next steps

References

# Estimation goals

- ▶ Use cross-sectional data of 12-month fuel price changes from 73 sampled CBSA
- ▶ Smooth out estimates in 21 SR and 52 NSR CBSA
- ▶ Predict estimates in unsampled CBSA
- ▶ Aggregate estimates in CBSA to State and Division levels
- ▶ Validate model-based estimates by comparing with administrative data

# Main concepts implemented in our models

1. **Fay-Herriot** area model at CBSA level (Fay and Herriot, 1979)
2. **Bayesian calibration** of CBSA predictions to Census Divisions predictions (Savitsky, 2016)
3. **Co-modeling** of means and variances in small areas (Sugasawa, Tamae, and Kubokawa, 2017), (Savitsky and Gershunskaya, 2023).
4. Global-local (**Horseshoe**) shrinkage prior for regularized model selection (Carvalho, Polson, and Scott, 2010);
5. **Spatial** model incorporating dimension reduction and alleviation of confounding (Hughes and Haran, 2013)
6. **Multiplicative Gamma** shrinkage prior for selecting spatial factors (Bhattacharya and Dunson, 2011)

# FH HB area model with calibration to Census Division

Fay and Herriot (1979) hierarchical model for CBSAs

$$\hat{y}_{i(j)} | y_{i(j)} \stackrel{\text{ind}}{\sim} N \left( y_{i(j)}, \hat{V}_{i(j)}^2 \right), j \in 1, \dots, 73$$

$$y_i | \beta, \tau_\mu^2 \stackrel{\text{ind}}{\sim} N \left( \mathbf{x}_i^T \beta, \tau_\mu^2 \right), i \in 1, \dots, 894$$

$\hat{y}_{i(j)}, \hat{V}_{i(j)}^2, y_{i(j)}$  – direct and model-based estimates in sampled CBSA

$y_i$  – model-based predictions in all U.S. CBSA

Calibration of CBSA predictions  $y_i$  to Census Divisions

$$\hat{y}_d \stackrel{\text{ind}}{\sim} N \left( y_d, \hat{V}_d^2 \right), d \in 1, \dots, 9$$

$$y_d = (N_d)^{-1} \sum_{i \in d} y_i N_i, N_d = \sum_{i \in d} N_i$$

$\hat{y}_d, \hat{V}_d^2, y_d$  – direct and model-based estimates in Census Divisions

$N_i, N_d$  – CBSA and Division population counts

## Co-modeling of variances

Smoothing of direct variance estimates  $\hat{V}_{i(j)}^2$ , see Savitsky and Gershunskaya (2023)

$$\begin{aligned}\hat{y}_{i(j)}|y_{i(j)} &\stackrel{\text{ind}}{\sim} N\left(y_{i(j)}, \nu_{i(j)}^2\right), j \in 1, \dots, 73 \\ \hat{V}_{i(j)}^2|a, \nu_{i(j)}^2, b &\stackrel{\text{ind}}{\sim} G\left(\frac{an_{i(j)}^*}{2}, \frac{an_{i(j)}^*}{2b\nu_{i(j)}^2}\right) \\ \nu_{i(j)}^2 &\stackrel{\text{ind}}{\sim} IG\left(2, \exp\left(z_{i(j)}^T \gamma\right)\right)\end{aligned}$$

$\nu_{i(j)}^2$  - latent variances in sampled CBSA

$z_{i(j)}^T \gamma$  - linear modeling of latent variances

$n_{i(j)}^*$  - standardized CBSA sample size

$b \sim G(3, 3)$  - uniform bias factor of variance estimation

$a \sim \exp(N(0, 1))$  - uniform shape factor of variance estimation.

# Regularized model selection with horseshoe prior

$p \approx 30$  covariates are used to model 73 sampled CBSA. Model selection is regularized by the choice of prior for regression coefficients.

- ▶ **No regularization.** Multivariate normal prior with LKJ (Lewandowski, Kurowicka, and Joe (2010)) random correlation matrix

$$\beta \sim \text{MVN}(0, \Sigma(\eta)), \Sigma(\eta) \sim \det(E)^{(\eta-1)}$$

$\eta \in (1, \infty)$  for positive correlation. We used  $\eta = 4$  for moderate correlation.

- ▶ **Regularization.** Horseshoe global-local shrinkage prior (Carvalho, Polson, and Scott, 2010) shrinks parameter estimates to 0, leaving only statistically significant

$$\beta_p \sim N(0, (\sigma_p \sigma_G)^2),$$
$$\sigma_p \sim C^+(0, 1), \sigma_G \sim C^+(0, 1)$$

# Modeling spatial correlations between CBSA

Accounting for spatial correlations may improve SAE for sparsely sampled CBSA.

- ▶ **No spatial correlations.** Independently distributed random effects

$$E(y_i | \beta, u_i) = \mathbf{x}_i^T \beta + u_i, u_i \stackrel{\text{ind}}{\sim} N(0, \tau_u^2), i \in 1, \dots, n.$$

- ▶ **Spatial correlations,** see Hughes and Haran (2013)

$$E(y_i | \beta, \beta_s) = \mathbf{x}_i^T \beta + \mathbf{M}_i \beta_s$$

**A** -  $n \times n$  adjacency or proximity, e.g., inverse distance matrix

**Q** =  $\text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$   $n \times n$  precision matrix

**M** -  $n \times m$  spatial basis (harmonics) are the first  $m \approx 0.10n$  eigenvectors of Moran operator  $\mathbf{P}^\perp \mathbf{A} \mathbf{P}^\perp$ .  $\mathbf{P}^\perp$  is orthogonal projection complement for covariates **X**.

$\beta_s \sim \text{MVN}(0, \Sigma_s)$   $m$ -vector of random effects

$\Sigma_s = (\mathbf{M}^T \mathbf{Q} \mathbf{M})^{-1}$   $m \times m$  spatial correlation matrix

## Regularized spatial model

In case of factor analysis, Bhattacharya and Dunson (2011) used multiplicative shrinkage prior to regularize variable selection. It allows gradually restricting selection of spatial harmonics with higher indexes  $h$ .

We want to

$$\beta_{sh} \sim N(0, (\sigma_h \tau)^2)$$

$$\sigma_h = \prod_{l=1}^h \delta_l, \delta_l \sim G(2, 1) \quad \text{local shrinkage parameter}$$

$$\tau \sim C^+(0, 1) \quad \text{global shrinkage parameter}$$

# ACS and NAICS2 CBSA-level covariates

- ▶ American Community Survey (ACS) covariates  
low-income CBSA population \$15K-25K (%),  
high-income CBSA population \$75+K (%),  
median income value, property value,  
people with Bachelor degree or higher (%), computer users (%),  
young population of 25-31 years old(%),  
urban population (%) and population density.
- ▶ Employment distribution (%) by 21 **NAICS2 industry codes**
- ▶ Generated PC basis vectors and retained those explaining up to 95% variability between observations.

# Four Models

1. Model 1- FH area model, calibration to Census Divisions, co-modeling means and variances;
2. Model 2- Model 1 + Horseshoe shrinkage prior for regularized covariate selection;
3. Model 3- Model 2 + spatial harmonics;
4. Model 4- Model 3 + multiplicative Gamma shrinkage prior for spatial harmonics.

# Outline

Introduction to Consumer Price Index (CPI)

Modeling price changes (PC) in CBSA

**Model fit evaluation**

Discussion and next steps

References

# Model fit for 12-month fuel price change

$\hat{Y}_i^{\text{pred}}$  - Models predictions in all U.S. CBSA

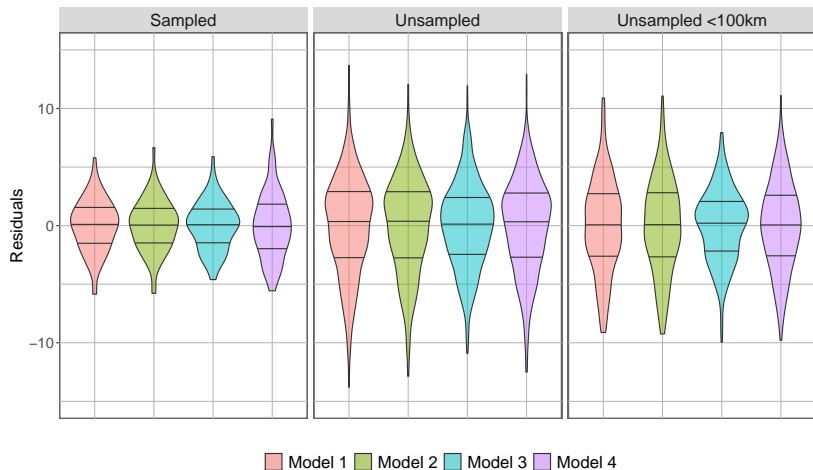
$Y_i^{\text{Admin}}$  - Administrative data in almost all CBSA

Regress Admin on predicted  $Y$ :  $Y_i^{\text{Admin}} \sim \beta * \hat{Y}_i^{\text{pred}} + \epsilon_i$

Model	RMSE	MAD	$\beta$	SD( $\epsilon_i$ )	$R^2$
73 sampled CBSA					
1	2.26	1.75	0.92	2.23	0.70
2	2.21	1.69	0.93	2.19	0.71
3	2.12	1.65	0.95	2.05	0.74
4	2.93	2.27	0.99	2.83	0.51
821 unsampled CBSA					
1	5.14	3.99	0.33	4.19	0.11
2	4.62	3.66	0.44	4.14	0.13
3	3.73	2.92	0.75	3.62	0.34
4	5.23	4.19	0.39	4.02	0.18
235 unsampled CBSA < 100km from the sampled					
1	4.97	3.96	0.36	3.99	0.11
2	4.51	3.64	0.46	3.97	0.12
3	3.09	2.52	0.93	3.05	0.48
4	5.06	4.16	0.46	3.72	0.23

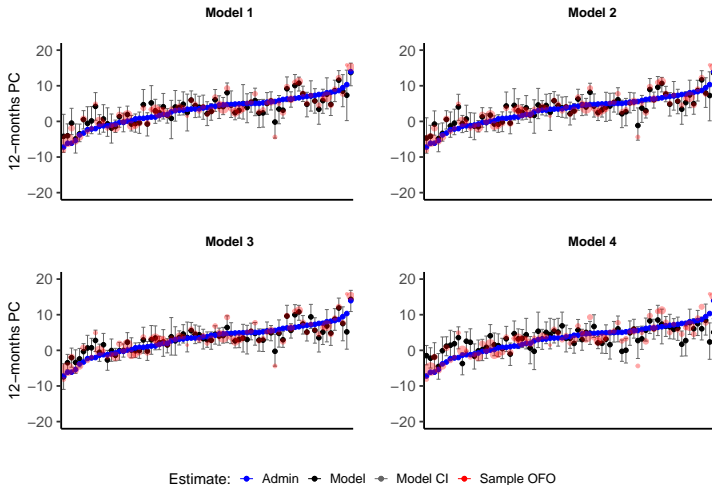
# Residuals distribution

Residuals by Model 3 (spatial correlations) are more concentrated around 0



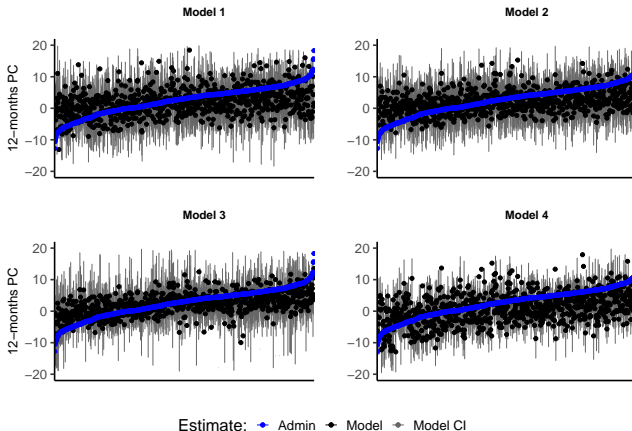
# Sampled CBSA ( $n = 73$ )

All models are validated equally well by Admin data



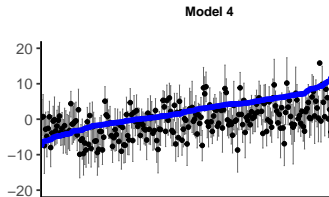
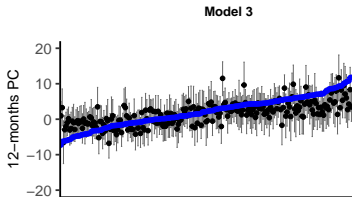
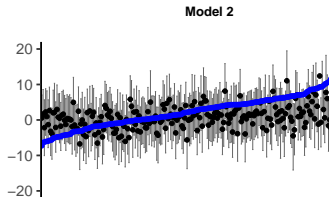
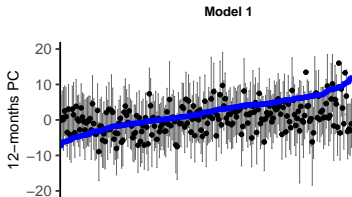
# Unsampled CBSA ( $n = 821$ )




Model 3 (spatial correlations) is **better** validated by Admin data



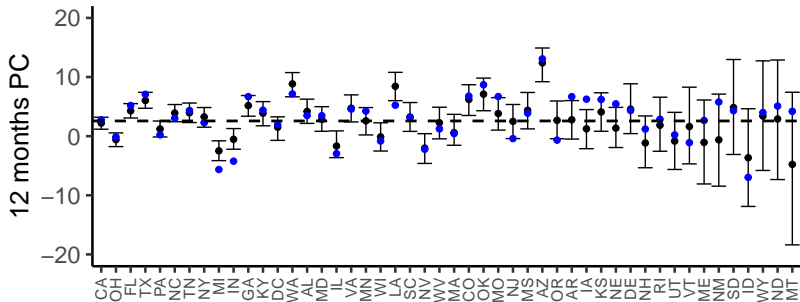
# Unsampled CBSA within 100km from sampled (235)

Model 3 (spatial correlations) is **even better** validated



Estimate:  Admin  Model  Model CI

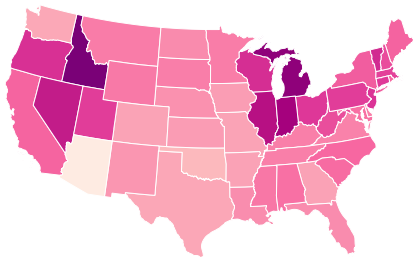
# State estimates by Model 3 (spatial correlations)



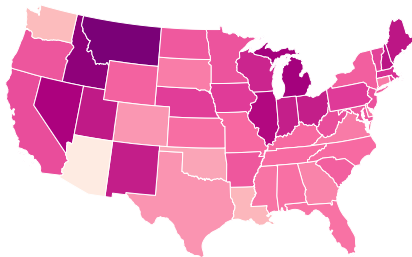
Estimate: • Admin • Model

# Model 3 (spatial correlations) vs Admin data

Administrative data



Model 3



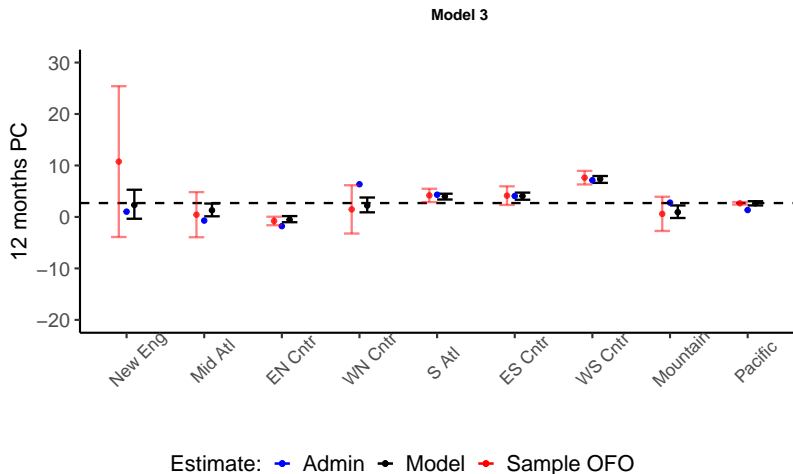
— U.S. BUREAU OF LABOR STATISTICS • [bls.gov](https://bls.gov)

PC (%)



-5 0 5 10

# Census Divisions estimates



# Outline

Introduction to Consumer Price Index (CPI)

Modeling price changes (PC) in CBSA

Model fit evaluation

Discussion and next steps

References

- ▶ Accounting for spatial correlations improved model fit to administrative data. Shrinkage of spatial harmonics with multiplicative Gamma shrinkage prior had negative effect on model fit.
- ▶ ACS and NAICS2 covariates don't explain fuel price changes across geographies. Horseshoe global-local shrinkage prior was ineffective for model selection.
- ▶ We plan fitting price change series using spatio-temporal models.

# Contact information

Vladislav Beresovsky  
Bureau of Labor Statistics (BLS)  
Office of Survey Methods Research (OSMR)

`beresovsky.vladislav@bls.gov`

# Outline

Introduction to Consumer Price Index (CPI)





Modeling price changes (PC) in CBSA

Model fit evaluation

Discussion and next steps

References

# References I

-  Bhattacharya, A. and D. B. Dunson (June 2011). “Sparse Bayesian infinite factor models”. In: *Biometrika* 98.2, pp. 291–306. eprint: <https://academic.oup.com/biomet/article-pdf/98/2/291/46695653/asr013.pdf>.
-  Carvalho, CARLOS M., NICHOLAS G. Polson, and JAMES G. Scott (2010). “The horseshoe estimator for sparse signals”. In: *Biometrika* 97.2, pp. 465–480.
-  Fay, R.E. and R.A. Herriot (1979). “Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data”. In: *Journal of the American Statistical Association* 74, pp. 269–277.
-  Hughes, John and Murali Haran (2013). “Dimension reduction and alleviation of confounding for spatial generalized linear mixed models”. In: *J. R. Statist. Soc. B* 75.1, pp. 139–159.

## References II

-  Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe (2010). “Generating Random Correlation Matrices Based on Vines and Extended Onion Method”. In: *Journal of Multivariate Analysis* 100, 1989–2001.
-  Savitsky, Terrance D. (2016). “Bayesian nonparametric multiresolution estimation for the American Community Survey”. In: *The Annals of Applied Statistics* 10.4, pp. 2157–2181.
-  Savitsky, Terrance D. and Julie Gershunskaya (2023). “Bayesian Nonparametric Joint Model for Domain Point Estimates AND Variances under Biased Observed Variances”. In: *Journal of Survey Statistics and Methodology* 11, 895–918.
-  Sugasawa, S., H. Tamae, and T. Kubokawa (2017). “Bayesian Estimators for Small Area Models Shrinking Both Means and Variances”. In: *Scandinavian Journal of Statistics* 44, pp. 150–167.