Variances of Combined Consumer Price Index Survey and Alternative data

Onimissi M. Sheidu
Mathematical Statistician

FCSM 2023

Outline

- Introduction
- Methodology
- Simulation Results
- Conclusion

- *Disclaimer: Any opinions expressed in this presentation are those of the author and do not constitute policy of the Bureau of Labor Statistics.*

Definition

- Consumer Price Index data - a probability survey collected by random selection. Use as the study benchmark data

- Alternative data (Corp5) - a nonprobability sample units collected by nonrandom process

- Core of the study methodology: Propensity models (create pseudo weight)
  - Propensity score - a conditional probability of assignment to treated units given individual's covariate values: $e(x)_{psi}$ = pr(*S* = 1 | X =x).

- Motivation:
- Improve reliability and accuracy of computed CPI estimates
  - Provide credibility to alternative data estimates (often treated as suspects)
  - Enhance CPI data estimates at reduced cost - reduces nonresponse
- Task
  - Gauge accuracy of computed estimates for both blended and Corp5 data
  - Benchmarking estimates for both data groups against computed CPI estimates
  - Establish outperformance and provide answer on:
  - How best to obtain a more reliable estimates by using blended data

Data Selection

- *Data Sources:*

I. CPI Research database → CPI Survey data

II. Alternative (Corp5 data) → Non-probability data

- *Data Type :*

➢ Monthly average prices of gasoline

- *Study Period:*

➢ December 2017 – May 2021

Data Simulation Setups: Six data groups with varying sizes

- Unmixed groups:

I. CPI survey data - Benchmark estimates, $Total\ size_{42\ months}, N_a = 114284$

II. Corp5 data $Total\ size_{\ 42\ months}, N_b = 380930$

III. Corp5_Adjusted data treated with $1/\pi_i$, $Total\ size_{\ 42\ months}, n_b = 266466$

- Mixed data sets (CPI data plus Corp5 data):

I. Mix1 (50% each): CPI data $n_a = 114284$, Corp5 data $n_b = 114284, n = 228568$

II. Mix2 (30 -70%): CPI data $n_a = 114284$, Corp5 data $n_b = 266466, n = 380930$

III. Mix3 (70-30%): CPI data $n_a = 114284$, Corp5 data $n_b = 48992, \quad n = 163276$

- We use composite weights for Mixed data sets

Figure 1: Weights Adjustment Process $(w_{0i} = \frac{1}{1-\pi_i},\ w_{1i=} \frac{1}{\pi_i}, \widehat{w}_{si}=w_j * w_{ji})$

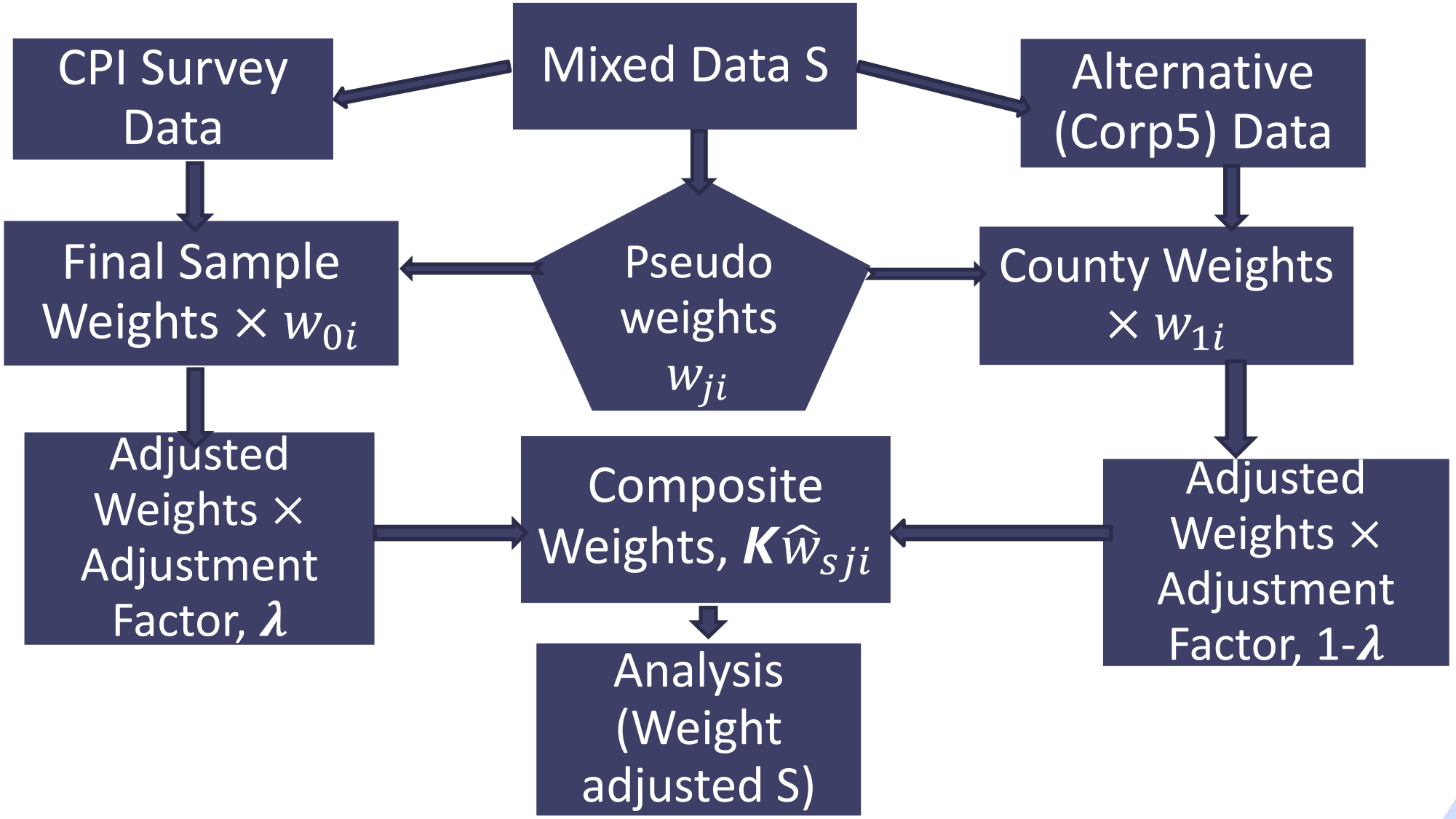


Figure 1a and 1b: Ex. of Common Support Validation for Model 1 (Mixed 1 Data)

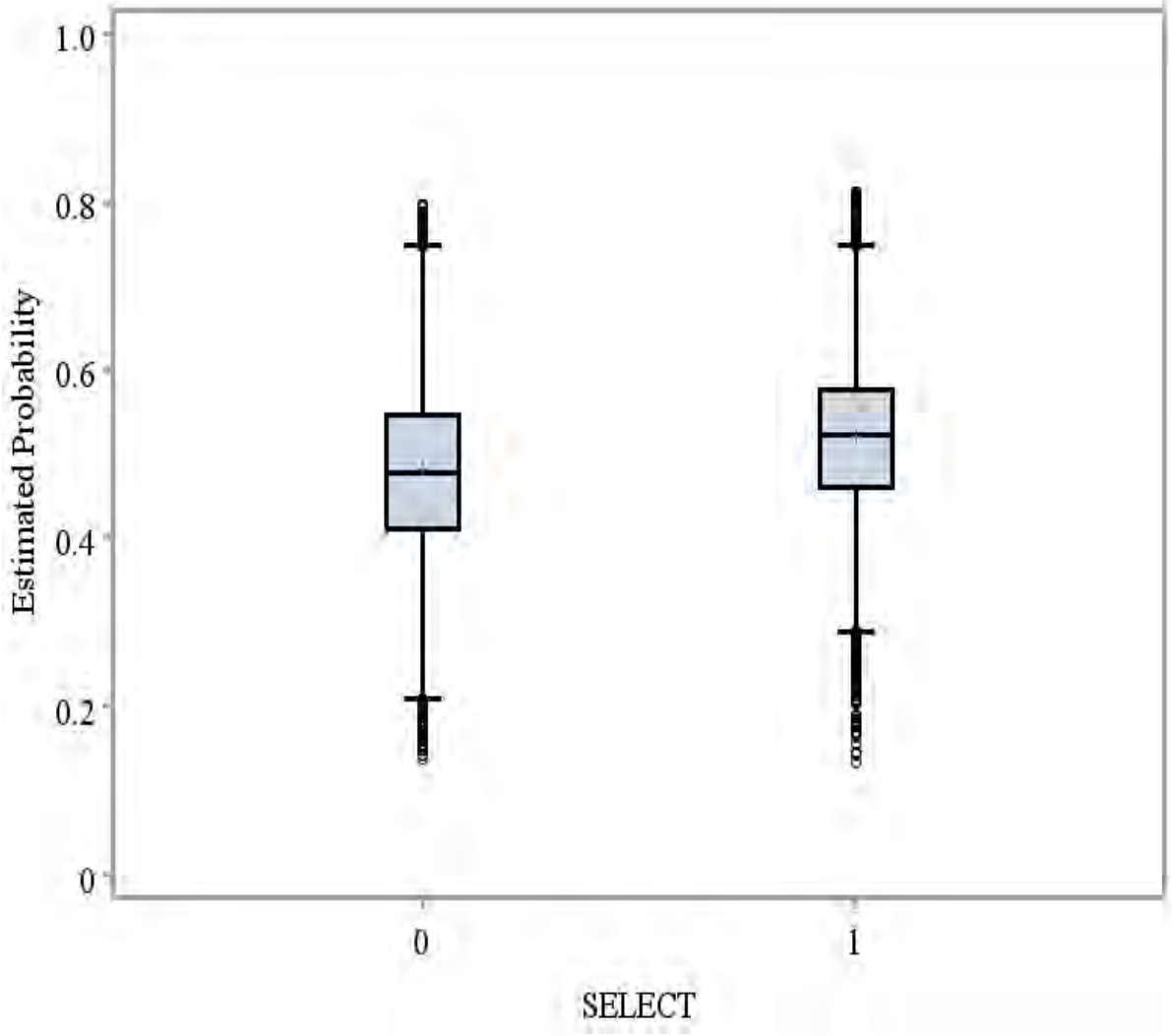


Figure 1a. Lacking Common Support

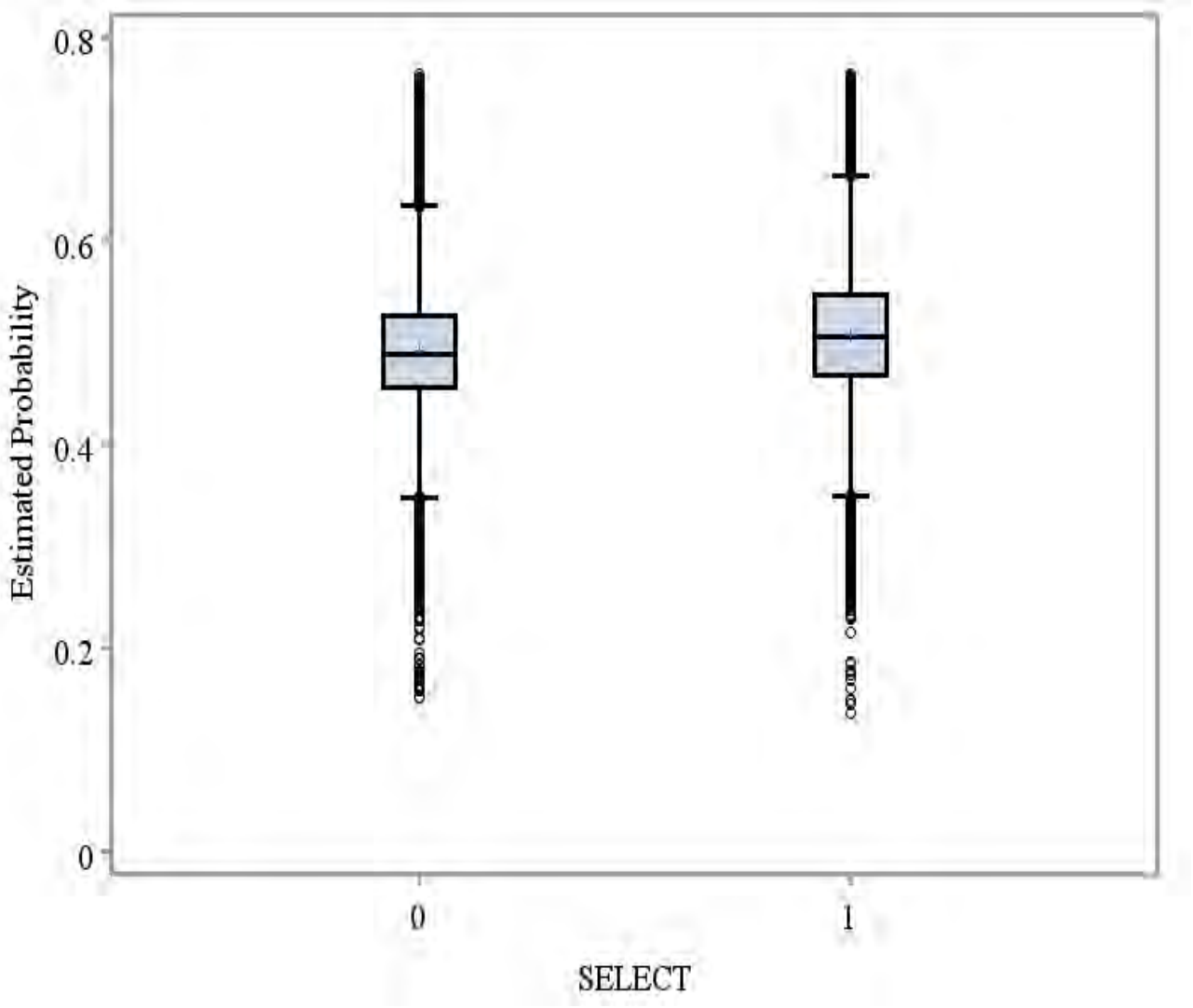
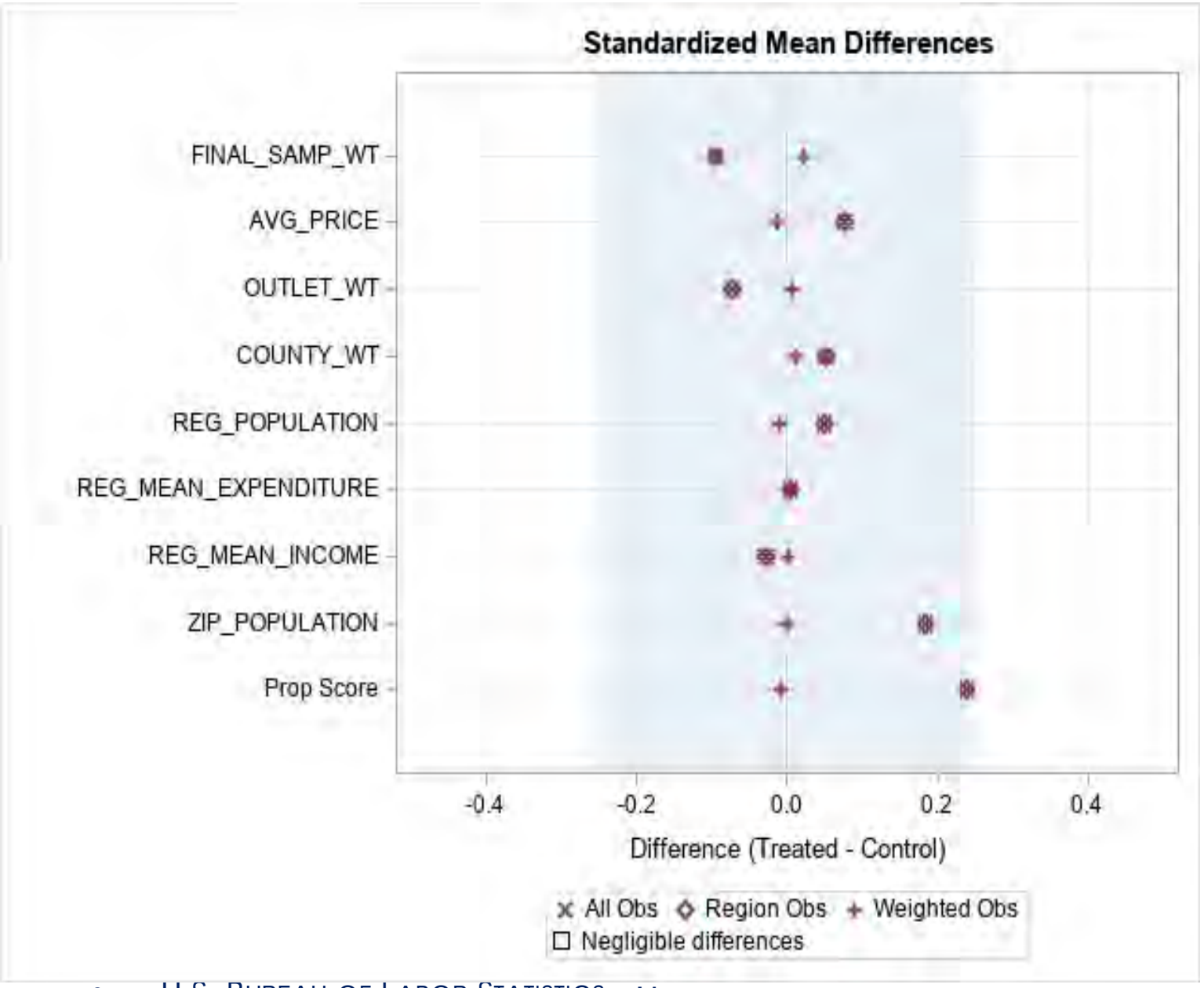
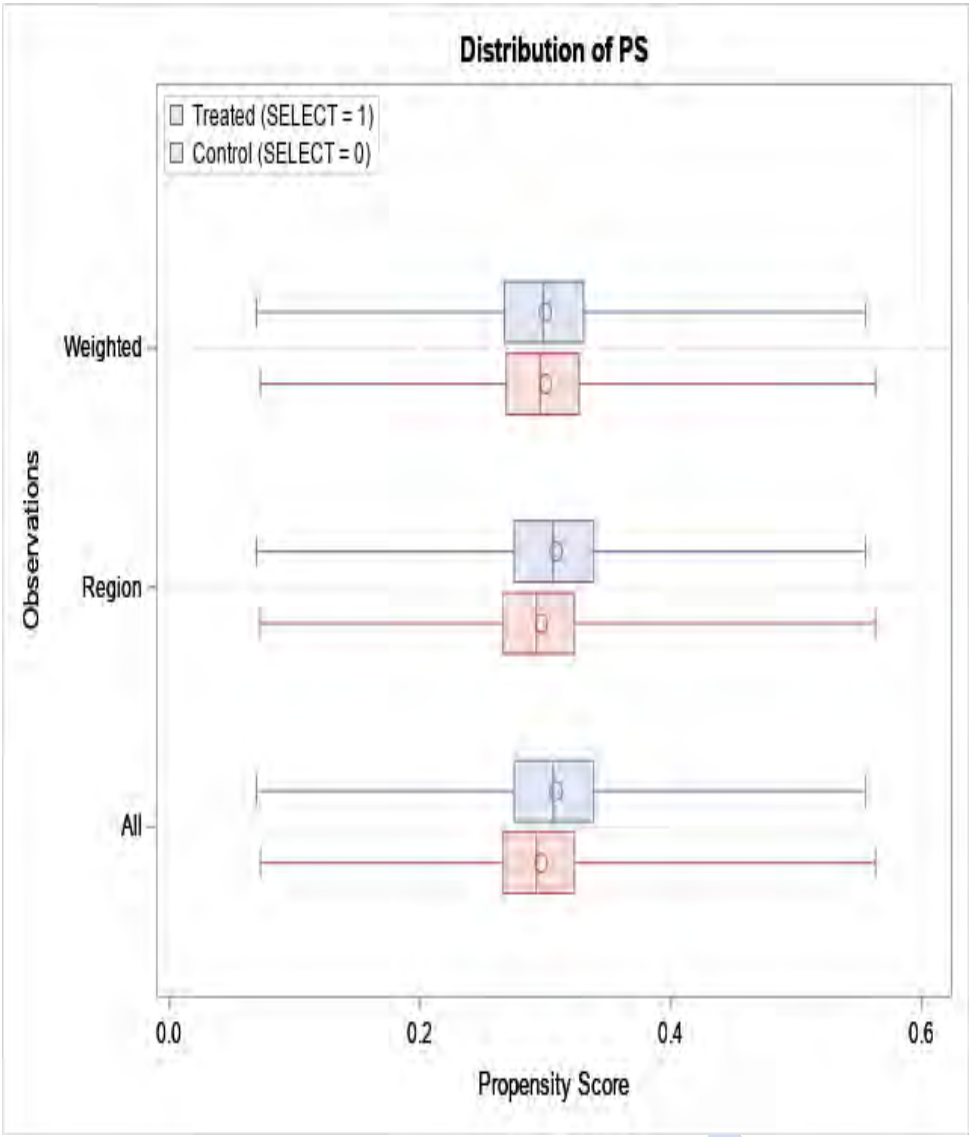


Figure 1b. Valid Common Support

Figure 2a & b: Ex. Balance Distribution for the Models covariates (Mixed Data)





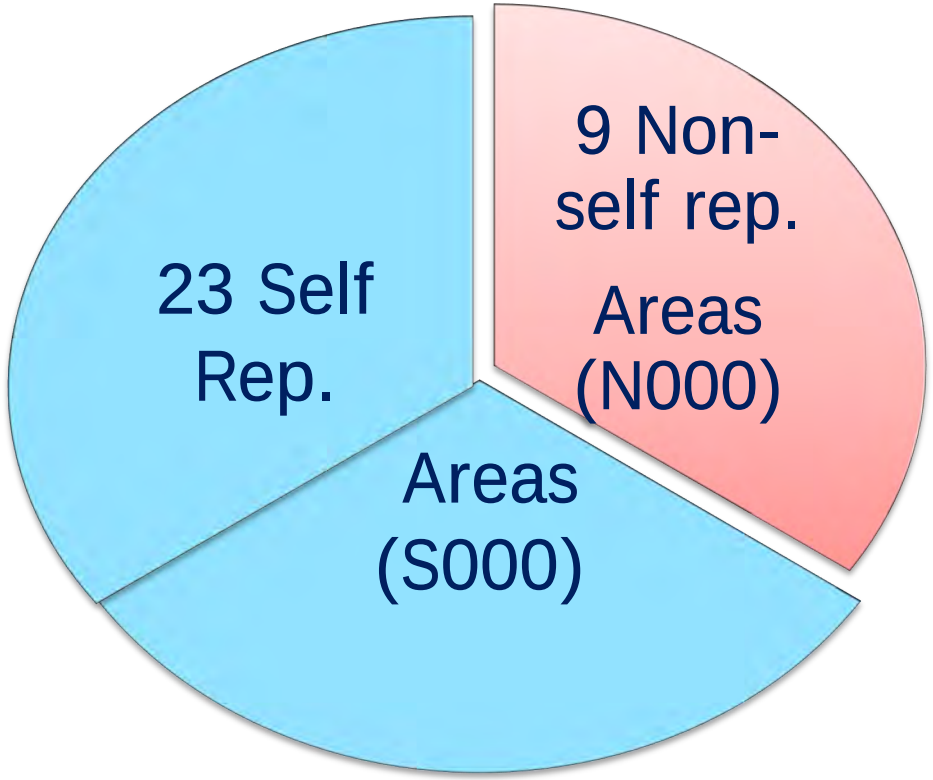
Calculated Composite Weights ($K\hat{w}_{sji}$)

- $\hat{t}_y = \sum_{i\epsilon \boldsymbol{s_b}} \lambda \hat{w}_{bi} y_{bi} + \sum_{i\epsilon \boldsymbol{s_a}} (1-\lambda) \hat{w}_{ai} y_{ai}$
  - $\lambda = \frac{\boldsymbol{n_b}}{\boldsymbol{n}}$, $n = n_a + n_b$, $\hat{t}_y$ - estimated population total
- here,
  - n_a = sample size for CPI survey data,
  - n_b = sample size for Corp5 data
- *Mixed 1 data*: $\hat{t}_y = \sum_{i\epsilon \mathbf{s_b}} 0.5 \hat{w}_{bi} y_{bi} + \sum_{i\epsilon \mathbf{s_a}} 0.5 \hat{w}_{ai} y_{ai}$
- *Mixed 2 data*: $\hat{t}_y = \sum_{i\epsilon \mathbf{s_b}} 0.7 \hat{w}_{bi} y_{bi} + \sum_{i\epsilon \mathbf{s_a}} 0.3 \hat{w}_{ai} y_{ai}$
- *Mixed 3 data*: $\hat{t}_y = \sum_{i\epsilon \mathbf{s_b}} 0.3 \hat{w}_{bi} y_{bi} + \sum_{i\epsilon \mathbf{s_a}} 0.7 \hat{w}_{ai} y_{ai}$
- *Corp5_Adjusted data*: $\hat{t}_y = \sum_{i\epsilon \mathbf{s_b}} \hat{w}_{bi} y_{bi}$,

➢ given $\hat{w}_{bi} = w_{bi} * w_{1i}$ and $w_{1i} = 1/\pi_i$; $\hat{w}_{ai} = w_{ai} * w_{0i}$ and $w_{0i} = 1/1-\pi_i$

Analysis Method

32 Areas (All-US cities,0000)



- Compute Index Area Percent Changes (PCs) for:
  - *1- month (PC01), 2- month (PC02),*
  - *6- month (PC06), 12- month (PC12).*
  - *All gasoline (item) & per gasoline grade (ELI)*
- Compute standard errors (SEs) of PCs
  - Using:

1) *Stratified Random Groups (SRG)*
2) *Bootstrap (BT)*
3) *Jackknife (JK)*

- Analyze groups by variance method:
  - All-US cities, All Self-representing and Non-self representing areas

Evaluation Method

1. $$Abs.Relative\ \beta ias(\Theta)_{t,s} = \left|\frac{\beta ias(\Theta)_{(t,t-k),s}}{\hat{\theta}cpi_{,t,t-k}}\right|$$

2. $$Abs.Diff.CV(\psi)_{(t),s} = \left|\frac{\hat{\delta}^2_{\theta^a_{(t,t-k),s}}}{\hat{\theta}^a_{(t,t-k),s}} - \frac{\hat{\sigma}^2_{cpi,t,t-k}}{\hat{\theta}cpi_{,t,t-k}}\right|$$

3. $$NRMSE\left(\hat{\theta}^a_{t,s}\right) = \sqrt{\frac{\left(\hat{\theta}^a_{(t,t-k),s} - \hat{\theta}cpi_{,t,t-k}\right)^2}{\hat{\sigma}^2_{cpi,t,t-k}}}$$

$\hat{\theta}^a_{(t,t-k),s}$ - t-month PC for data set *s*

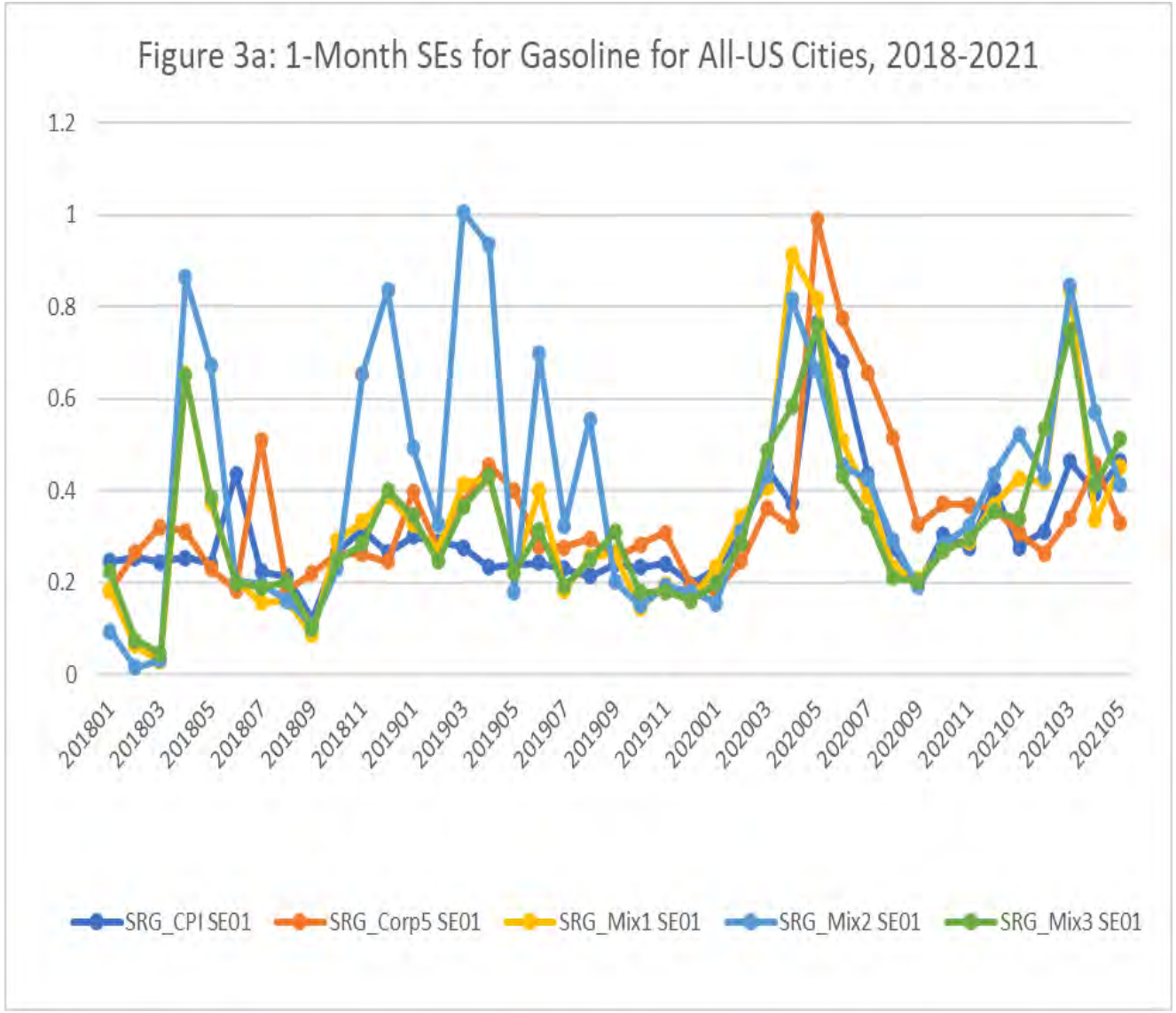
$\hat{\theta}cpi_{,t,t-k}$ - t-month PC for CPI data (Benchmark estimate)

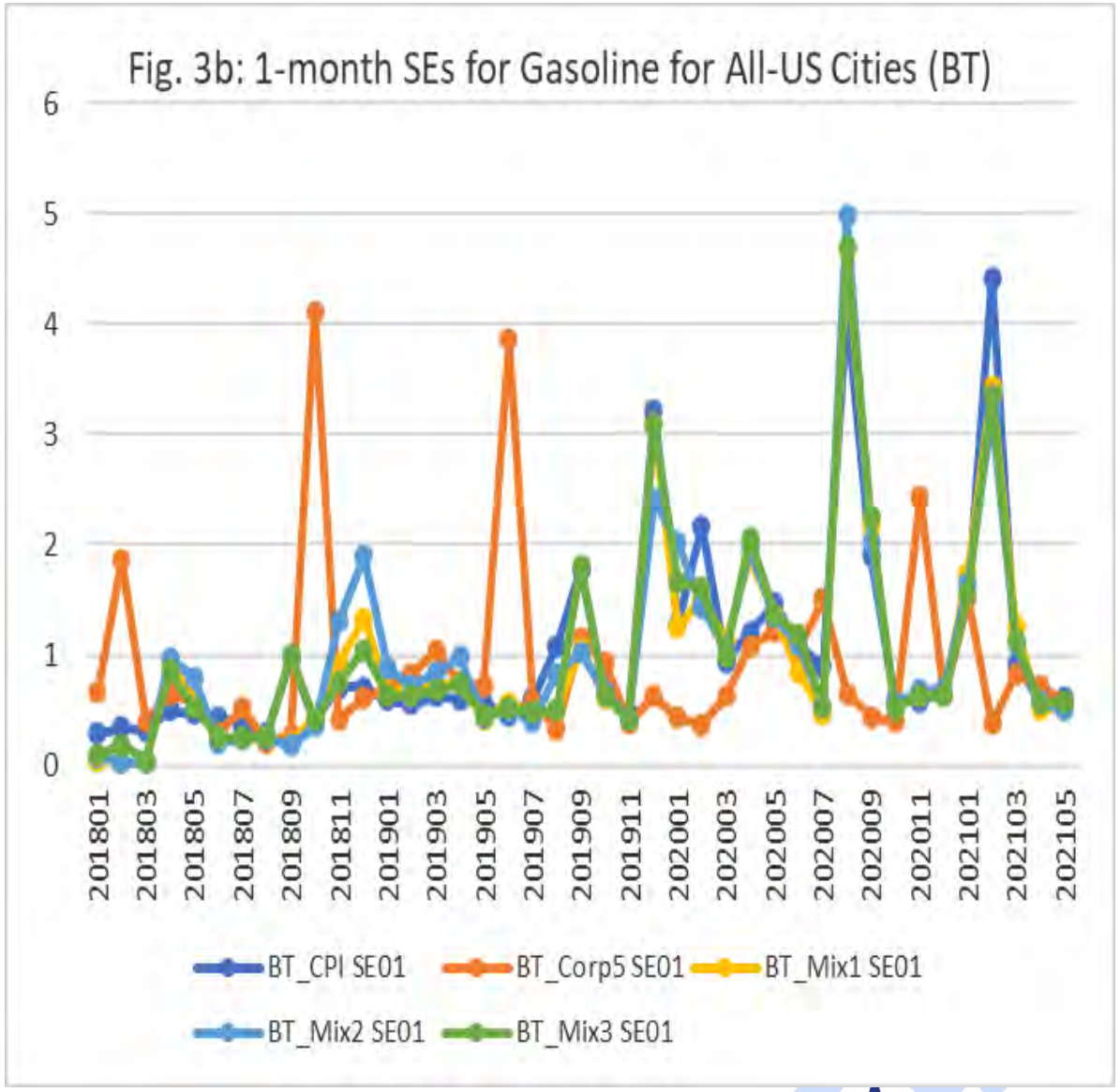
$\hat{\sigma}^2_{cpi,t,t-k}$ - t-month CPI variance estimate

$\hat{\delta}^2_{\theta^a_{(t,t-k),s}}$ - t-month variance estimate for data s

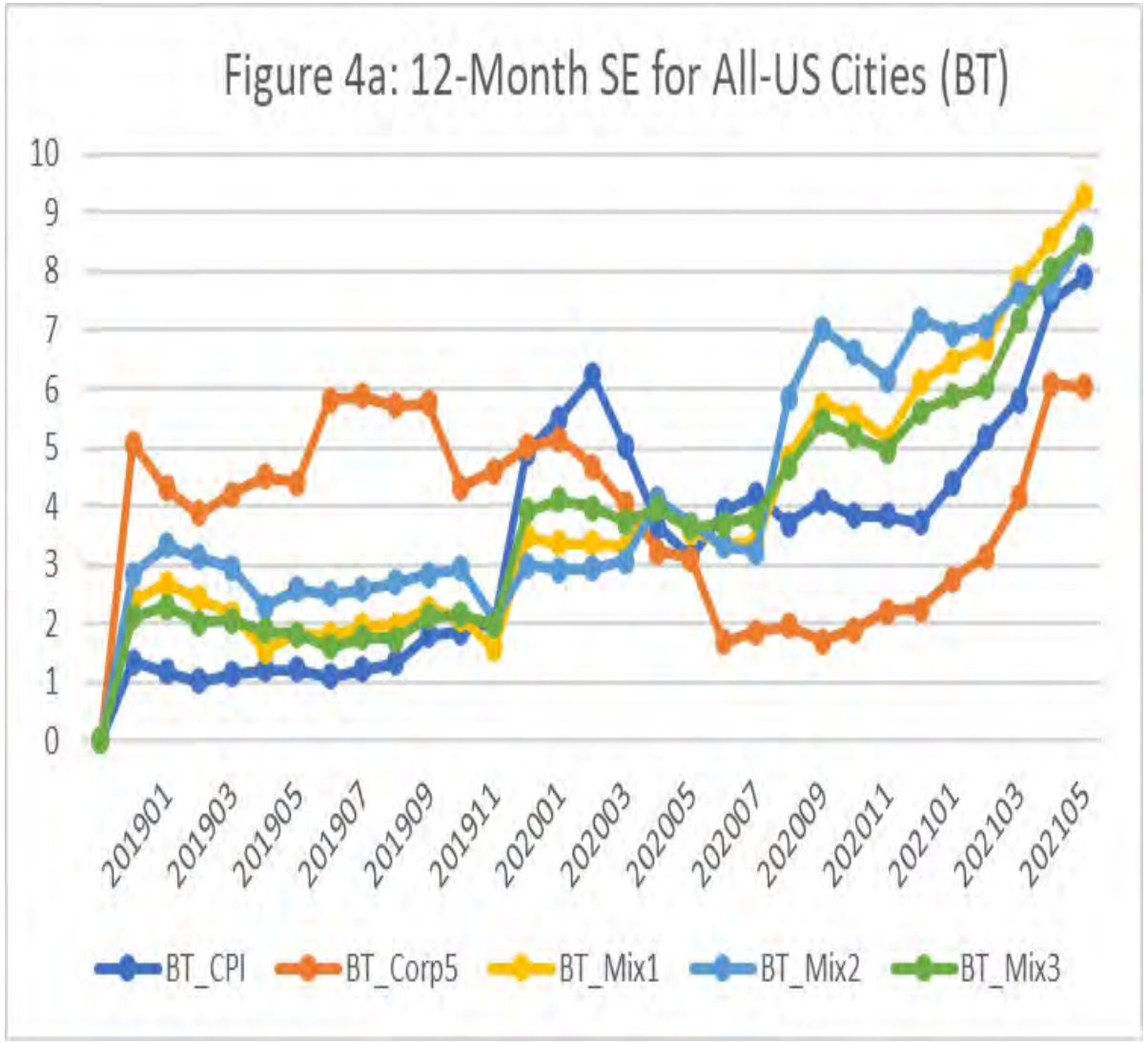
- Compare resulting values:
  - Among data groups
  - By variance method
  - And group by area:
    - All-US cities,
    - All Self-representing
    - All Non-self representing
- *Verdict:* The smaller their computed values the better
- NRMSE values = deciding factor

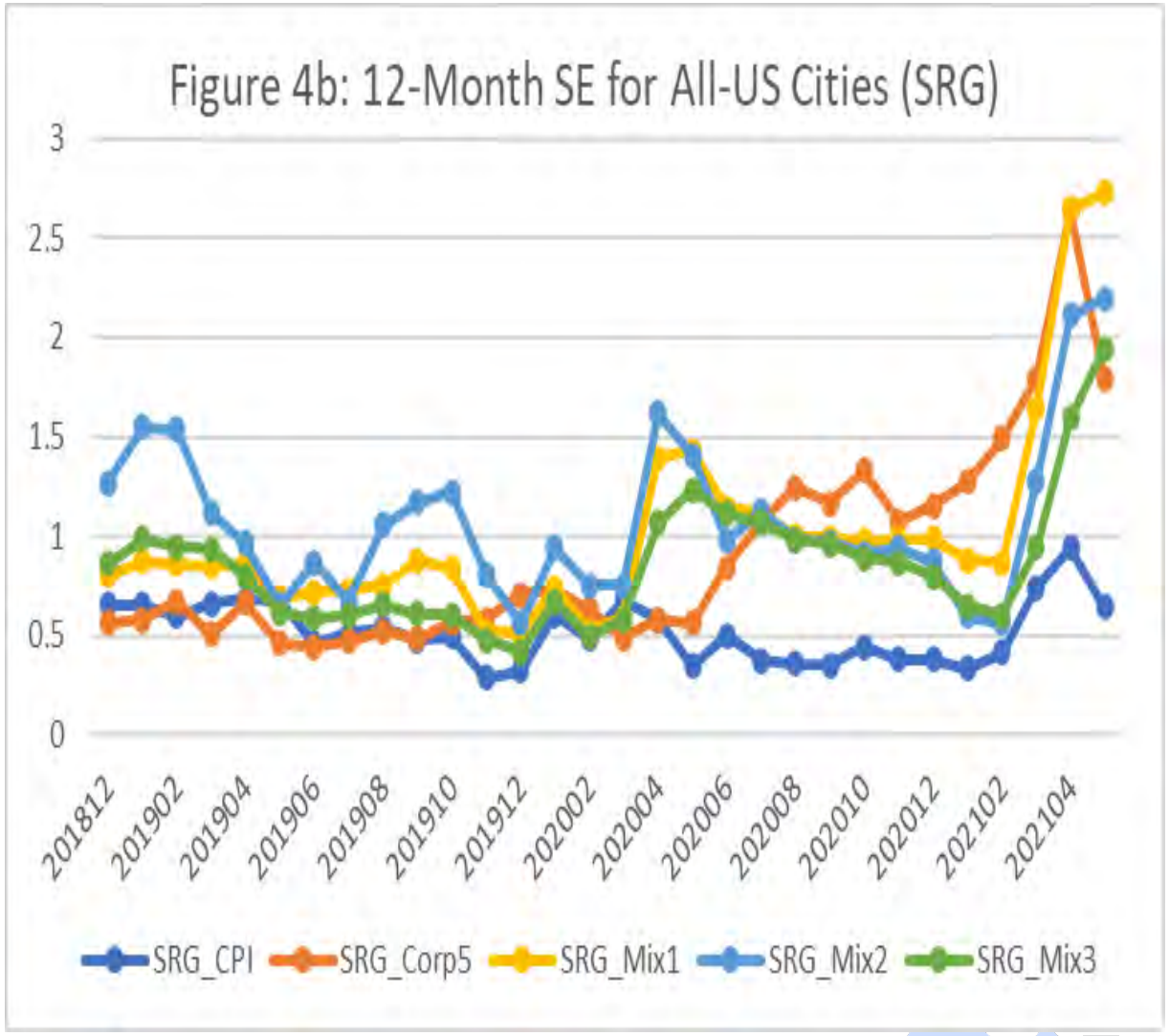
Results: One-Monthly SEs for All – US Cities by Data Group (SRG & BT Methods)



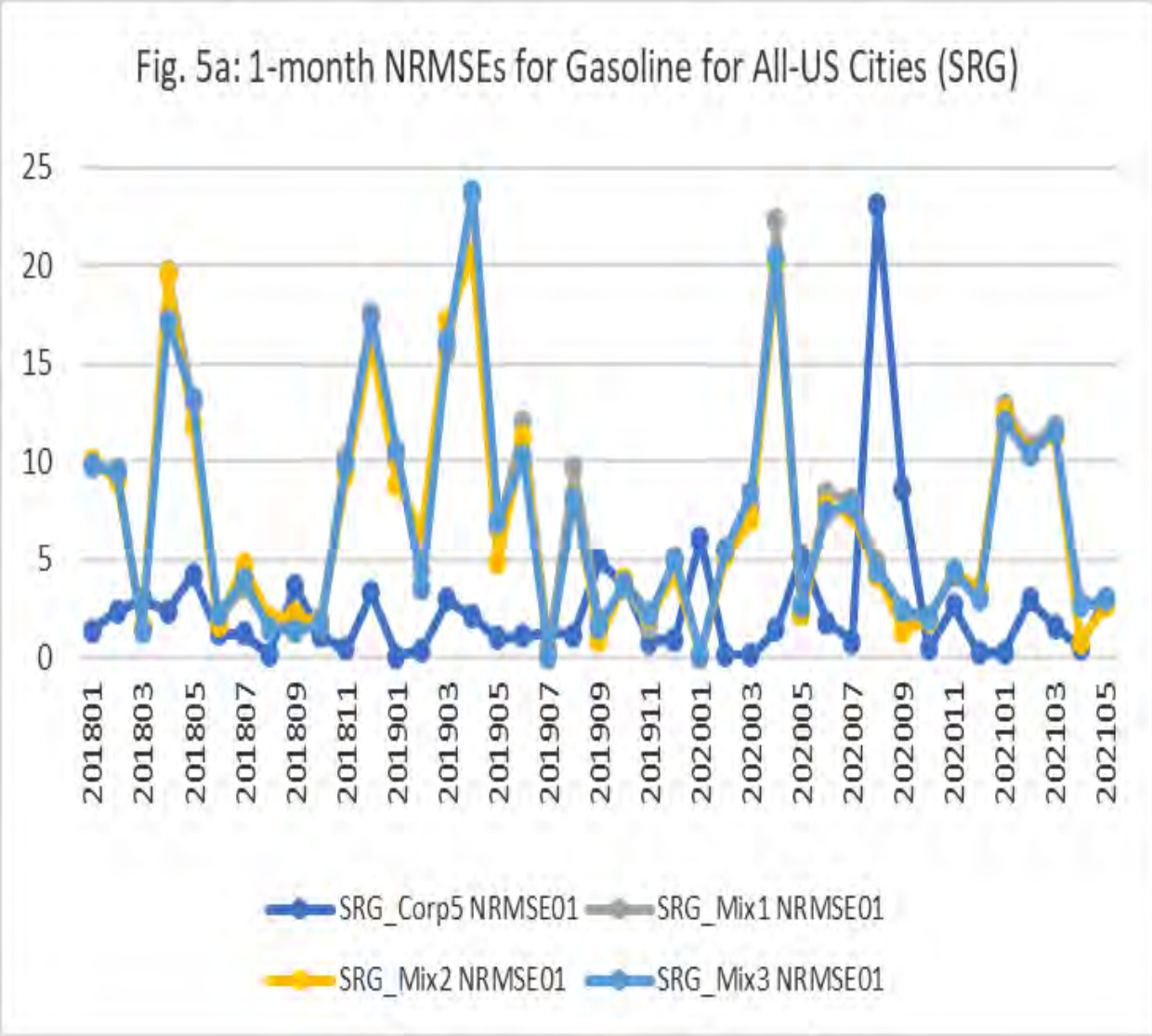


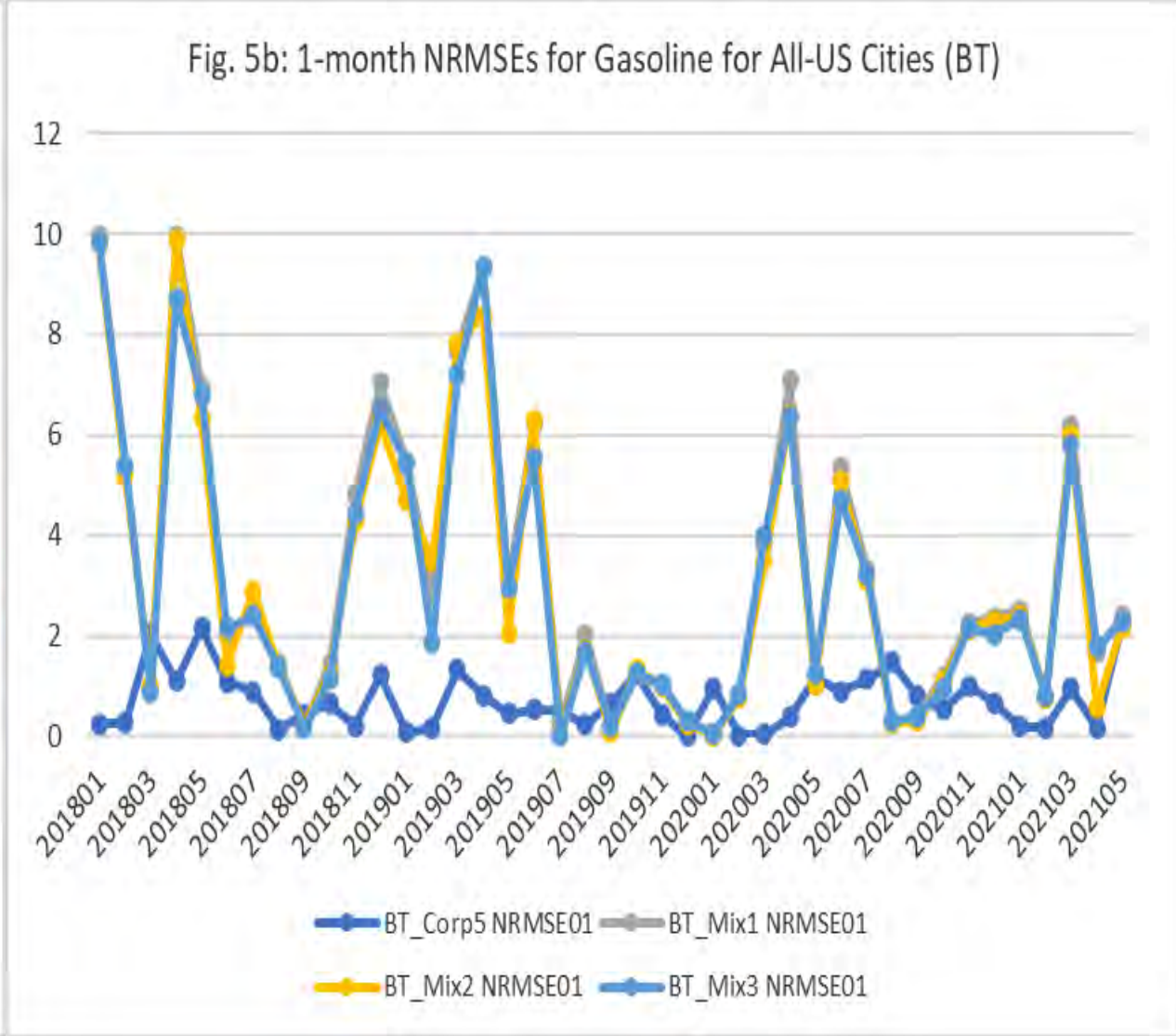
Results: : 12-Month SEs for All – US Cities by Data Groups (BT & SRG Method)



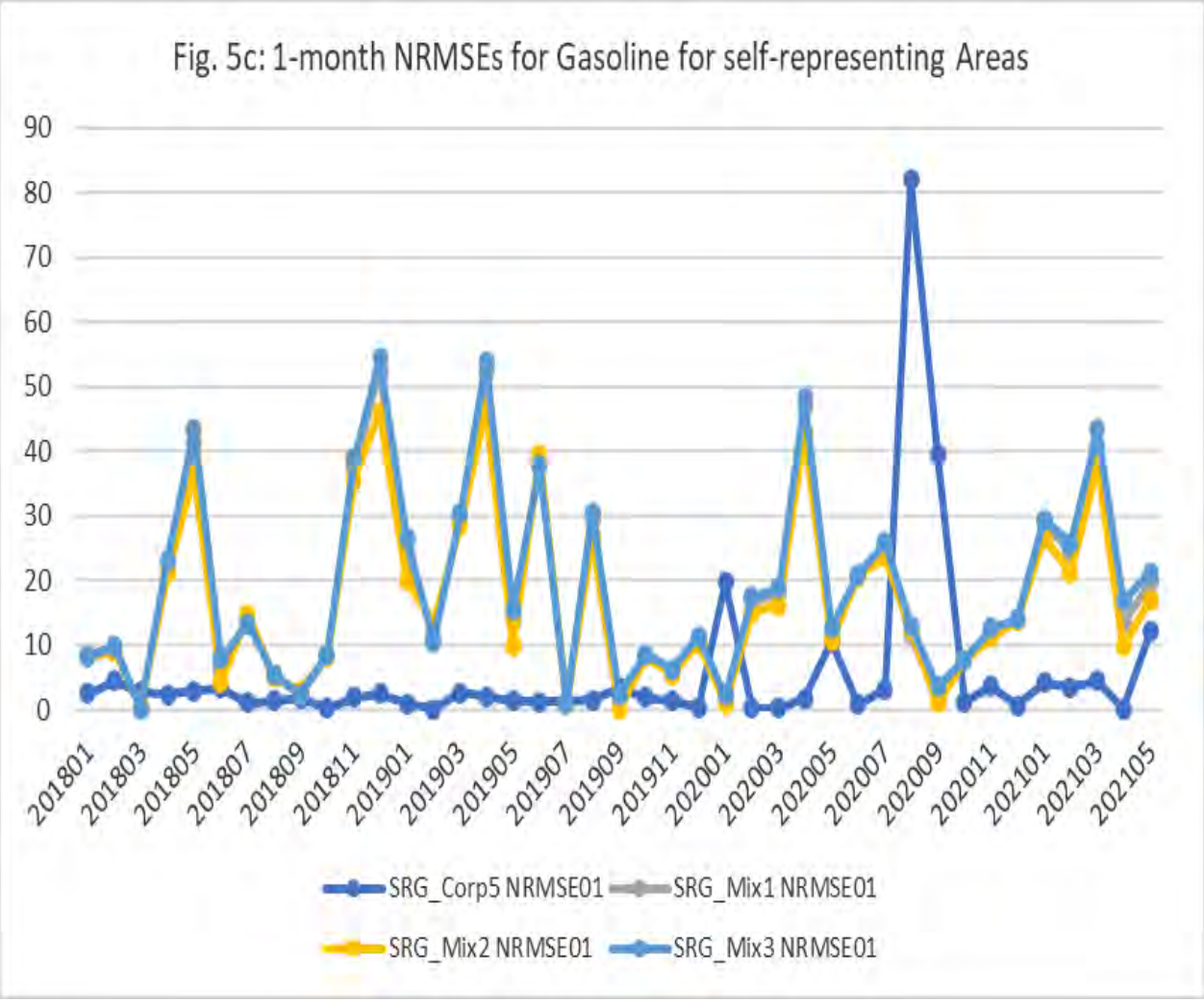


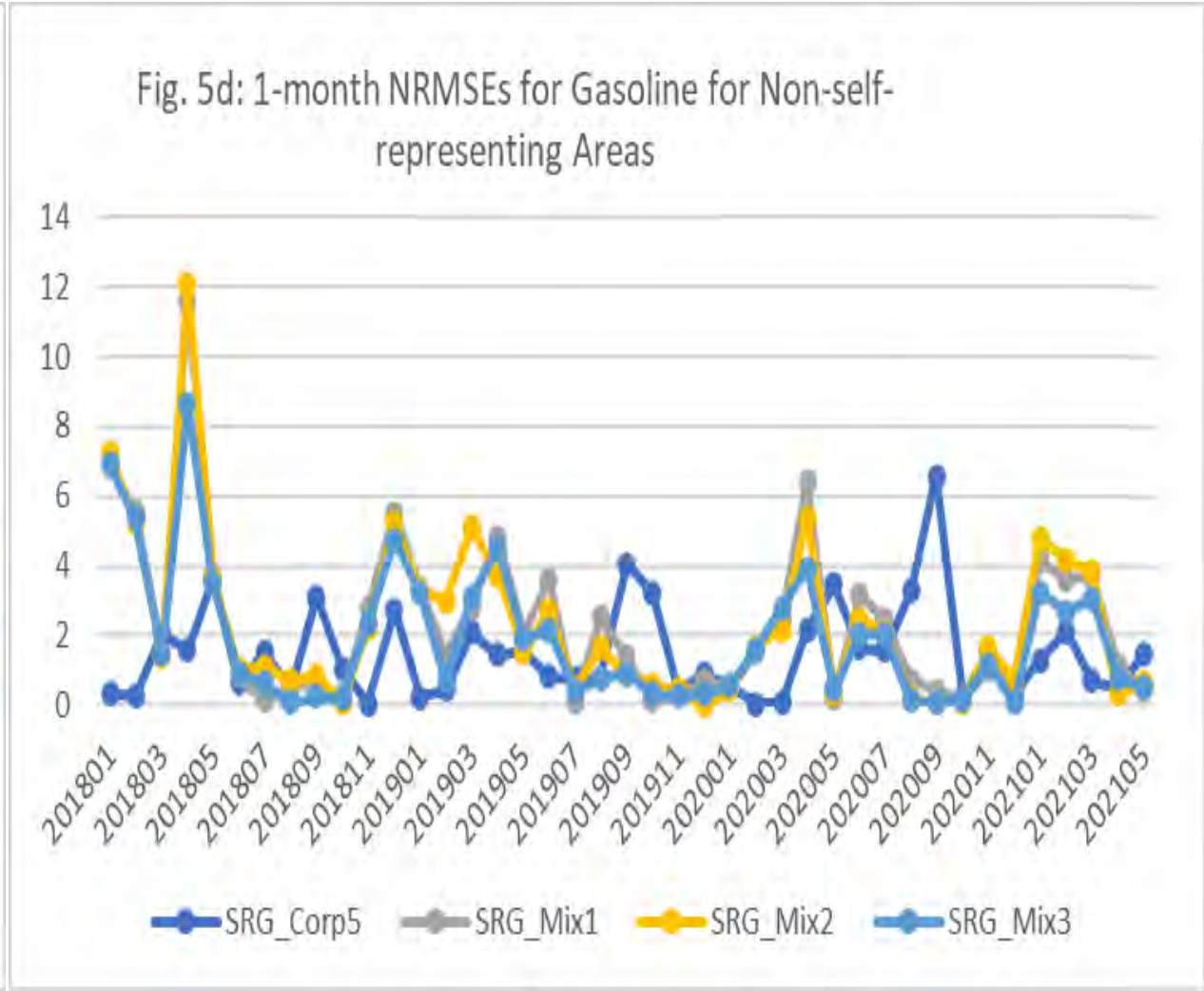
Results: One-Month NRMSEs for All – US Cities by Data Group (SRG & BT Method)



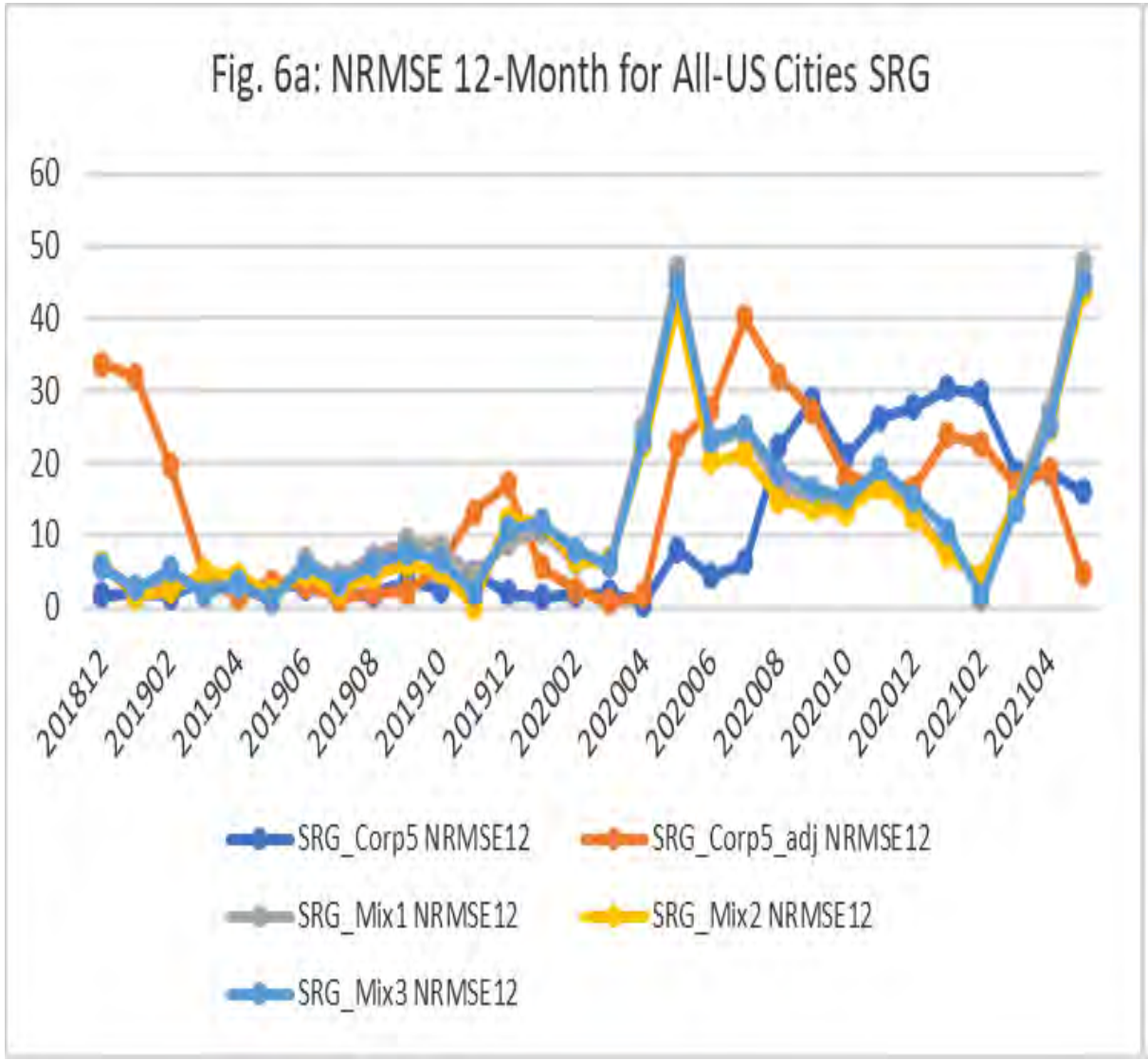


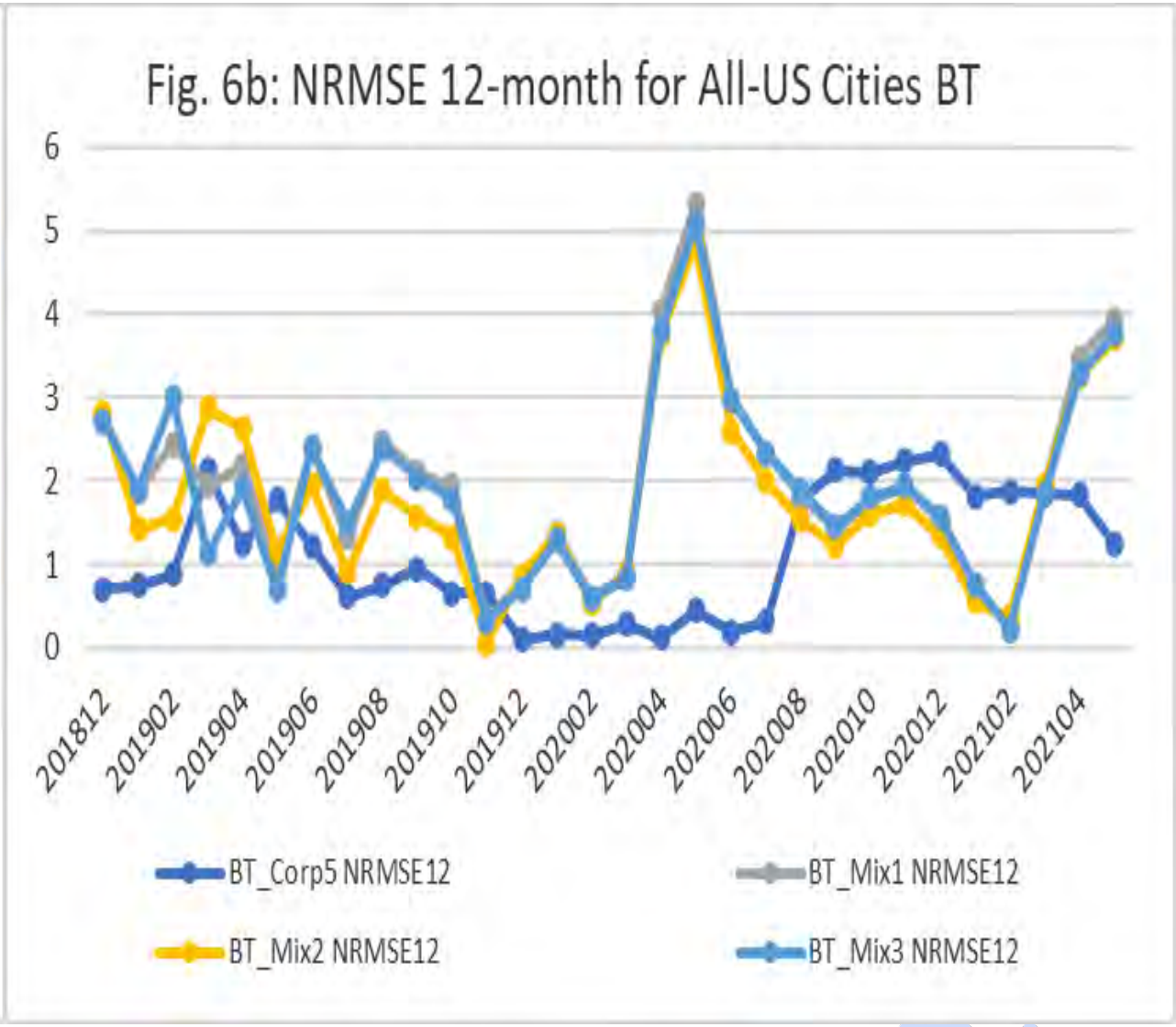
Results: One-Monthly NRMSEs for Gasoline for All Self & Non-self-representing Areas



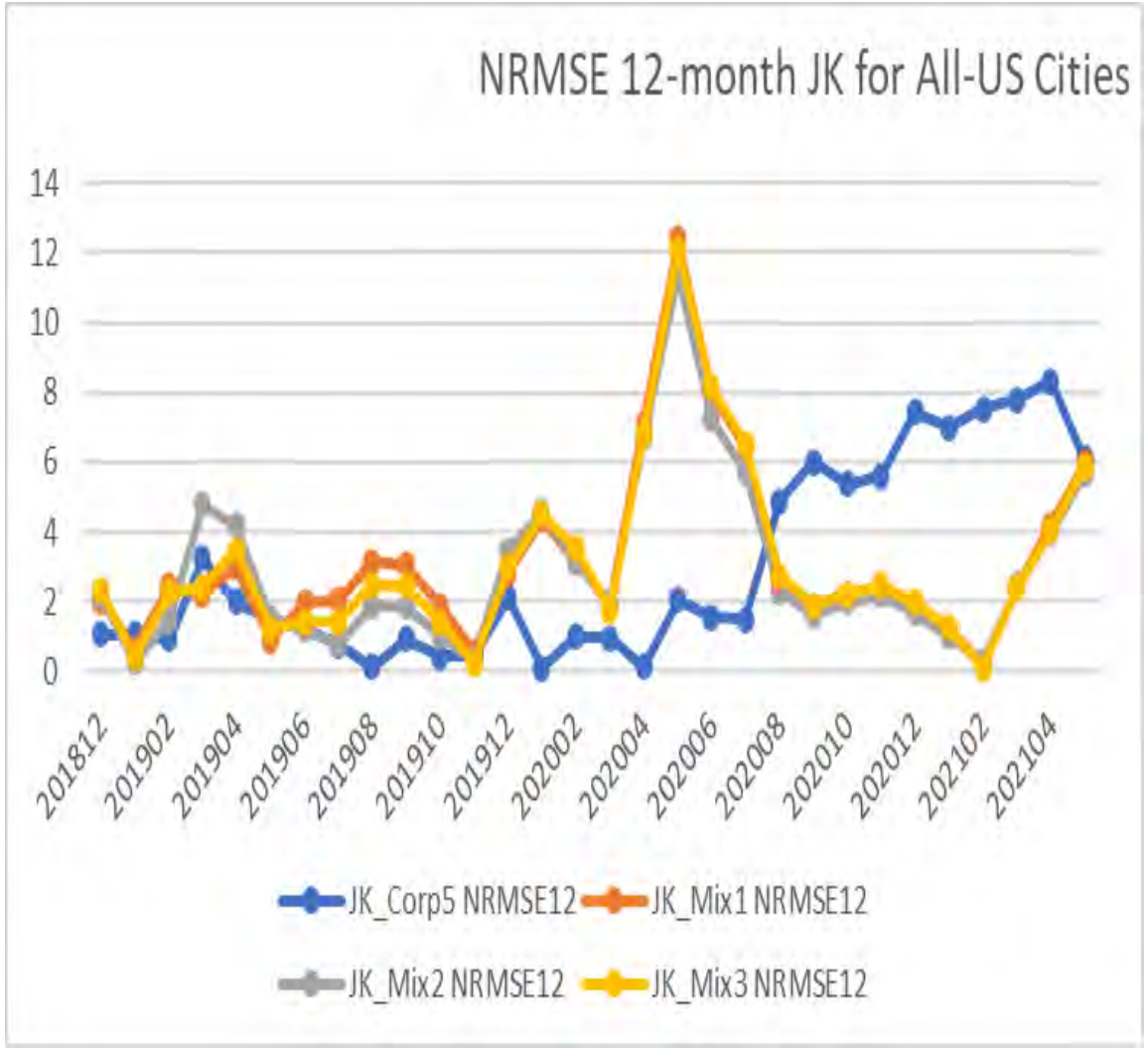


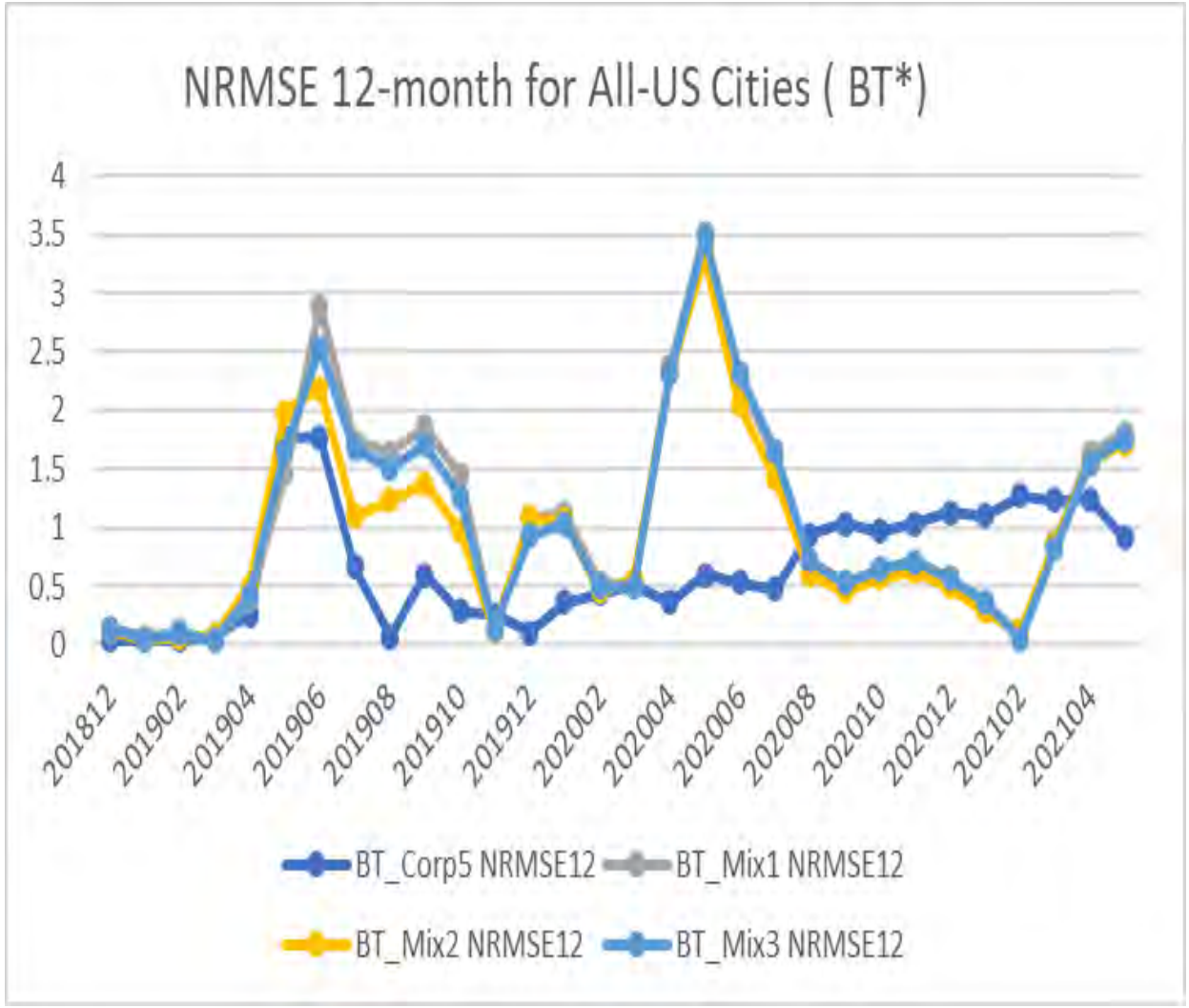
Results: 12-Month NRMSEs for All – US Cities by Data Groups (SRG & BT Method)





Results: 12-Month NRMSEs for All – US Cities for Data Groups (JK & BT Method)





Results: Summary of All Periods calculated NRMSEs for All US Cities for Data sets (12/2017 – 5/2021)

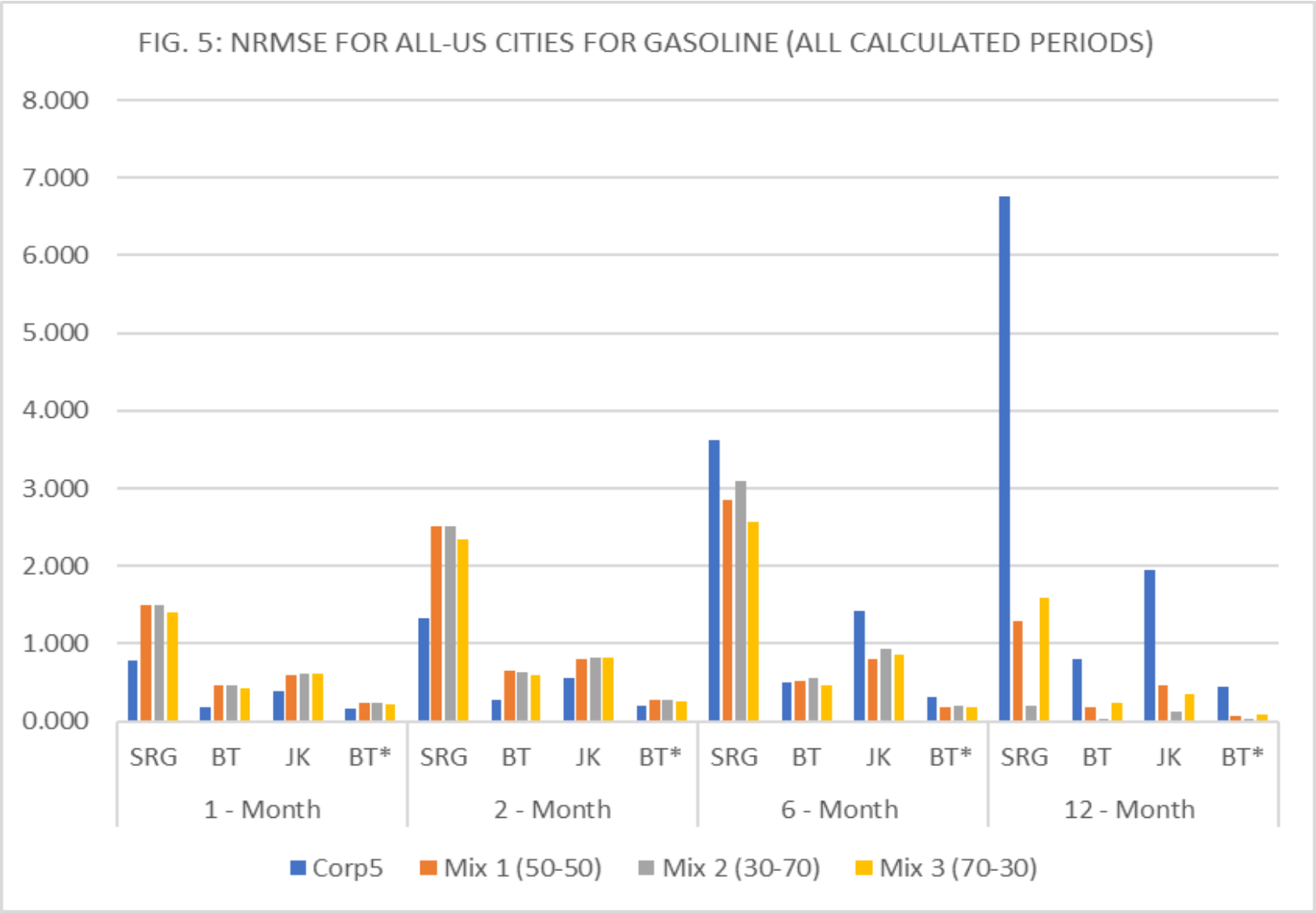


TABLE 1: NRMSE CALCULATION FOR ALL-US AREAS FOR GASOLINE

Period	Data Group	Corp5	Mix 1 (50-50)	Mix 2 (30-70)	Mix 3 (70-30)
1 - Month	SRG	0.774	1.496	1.494	1.396
	BT	0.179	0.467	0.464	0.432
	JK	0.397	0.587	0.610	0.604
	BT*	0.161	0.231	0.228	0.221
2 - Month	SRG	1.329	2.521	2.503	2.350
	BT	0.266	0.643	0.637	0.595
	JK	0.548	0.798	0.820	0.812
	BT*	0.196	0.270	0.266	0.259
6 - Month	SRG	3.623	2.850	3.091	2.567
	BT	0.495	0.519	0.549	0.463
	JK	1.419	0.808	0.934	0.861
	BT*	0.308	0.183	0.190	0.172
12 - Month	SRG	6.760	1.293	0.200	1.587
	BT	0.807	0.176	0.034	0.234
	JK	1.945	0.455	0.124	0.358
	BT*	0.446	0.076	0.033	0.085

Conclusion and Recommendation

- Study goal: Assess the possible advantage of blending the CPI survey data with alternatively sourced nonprobability data such as the corp5 data, and to explore the best way to do it.
- The result showed a mix picture especially for 1-month and 2-month estimates, and further exploration could give a clearer picture.
- Mixed data performed better at longer term (6-month and 12-month) periods.
- Weight assignment is an issue: What weight should be assigned to the mixed data – county or sample weight?
- With the current way of doing things, we think there is no much benefit in using mixed data except to give credence to the estimates.
- The study shows that combining CPI survey data with corp5 data could provide better enhancement when computing at ELI – Area level than at Item level.

Contact Information

Onimissi M. Sheidu
Mathematical Statistician
Statistical Method Division/CPI Survey
Research and Analysis Branch
www.bls.gov
202-691-6901
sheidu.onimissi@bls.gov