Analysis of the Impact of Secondary Source Data on the CPI's Gasoline Standard Errors October 2022

Jenny FitzGerald
U.S. Bureau of Labor Statistics
2 Massachusetts Avenue, NE, Room 3655 Washington, D.C. 20212 U.S.A.
fitzgerald.jenny@bls.gov

Abstract
The Consumer Price Index (CPI) estimates the change in prices over time of the goods and services U.S. consumers buy for day-to-day living. In the past, the program estimated its price change estimates for gasoline using about 4,000 price quotes selected from probability samples each month. In June 2021, the CPI replaced its sampled price data for its gasoline index with data from a secondary source. Currently, the secondary source provides the CPI with millions of gasoline prices each month. Given the substantial increase in the number of prices, one would expect to see a significant decrease in the gasoline standard errors (SEs). However, the CPI has not yet seen such a decrease. This study investigates why the gasoline SEs did not decrease with the increased number of prices from the secondary source. Specifically, the impact of increasing the number of replicate samples (or variance PSUs) assigned to the CPI Index Areas (variance strata) on the gasoline SEs from the secondary source data is investigated.

Key Words: Variance estimation, non-probability sample, stratified random groups

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

1. Introduction

The goal of the Consumer Price Index (CPI) is to estimate the change in prices over time of the goods and services U.S. consumers buy for day-to-day living. To estimate the change in prices, the Bureau of Labor Statistics (BLS) collects prices on goods and services in-person at sampled outlets or from websites for the sampled outlets. Unfortunately, response rates have been decreasing over the past few years and the cost of traditional manual collection is high. Consequently, the BLS has been exploring alternative sources of price data for several of its goods and services. One of the basic items that the BLS investigated for an alternative data source in the CPI and was successfully able to incorporate into the index was gasoline.

In June 2021, the BLS replaced the price data that it collects for the CPI's gasoline index with data from a secondary source that the agency refers to as CORP5. Currently, CORP5 provides the BLS with millions of gasoline prices each month. In the past, the BLS collected about 4,000 price quotes each month for gasoline. Given the substantial increase in the number of prices, one would expect to see a significant decrease in the item's standard errors (SEs). However, the BLS has only noticed a minimal decrease in the gasoline SEs from implementing the CORP5 data. One theory as to why there was such a

small decrease is because the BLS kept the same number of replicate samples for its gasoline index when it replaced its agency collected data with the CORP5 data.

This paper investigates the impact of increasing the number of replicates assigned to the CPI index areas on the gasoline SEs from CORP5 using the CPI's current method of variance estimation, Stratified Random Groups (SRG). Section 2 (Sampling and Index Estimation for Commodities and Services in the CPI) of this paper gives brief overviews of the CPI sample design and index estimation process. Section 3 (Methodology) describes the methodology of this study. Section 4 (Results) describes the impact that increasing the number of replicates had on the gasoline SEs from the CORP5 data. Finally, Section 6 (Conclusion) summarizes the findings of this study.

2. Sampling and Index Estimation for Commodities and Services in the CPI

2.1 Sampling for Commodities and Services

The CPI is calculated from a sample of price quotes, which are the ultimate outcome of several interrelated probability samples. First, the BLS selects a sample of geographic areas, which are the primary sampling units (PSUs) for the CPI (Bureau of Labor Statistics, 2020). The BLS traditionally updates its CPI area sample once every ten years. The area definitions for the PSUs are currently based on the 2013 Office of Management and Budget's (OMB) Core Based Statistical Areas (CBSAs). The BLS first classifies each PSU into one of the nine Census divisions and by its size. A PSU with a population greater than 2.5 million is a self-representing PSU, while a PSU with population less than 2.5 million is a non-self-representing PSU. Self-representing PSUs consist of only metropolitan CBSAs, and non-self-representing PSUs can be either metropolitan or micropolitan CBSAs.

After each PSU is mapped to its Census division and identified as a self-representing or non-self-representing area, the BLS stratifies the PSUs in each division-class size into strata of similar PSUs. Self-representing PSUs are placed in a stratum by themselves; non-self-representing PSUs are stratified based on geographic variables correlated with price change and expenditure level. A program selects one PSU per stratum using controlled selection to ensure that the selected PSUs are well-distributed across states and to maximize the number of old PSUs selected in the new area sample. Currently, the area sample for the CPI has 23 self-representing PSUs and 52 non-self-representing PSUs that make up the index's 32 index areas.

Within each sampled PSU, the BLS selects a sample of outlets where consumers shop using the data collected via the Consumer Expenditure (CE) Survey. The CE Survey consists of two surveys: the Interview Survey and the Diary Survey. The U.S. Census Bureau conducts both the Diary and Interview surveys for the BLS. The Quarterly Interview Survey collects data on large recurring expenditures such as rent and utilities, and the Diary Survey collects data on small, frequently purchased items such as food and clothing. Together the data from the two surveys cover the complete range of consumers' expenditures. The reported outlets form the frame of outlets that the BLS uses to select its sample for the CPI. The BLS selects its sample of outlets from the frame independently for each PSU and narrowly defined item category known as a rotation category using a systematic probability proportional to size (PPS) sample design, where each outlet's measure of size (MOS) is its reported expenditure in the item category.

The outlet sample is then merged to an independent sample of entry level items (ELIs) that consumers buy. Specifically, the BLS selects a systematic PPS sample of ELIs for each PSU from the expenditure data collected by the Consumer Expenditure (CE) survey, which is aggregated by item stratum and region. An ELI's MOS is its expenditure total for the region compared to the region's total expenditure value for the item stratum. The CPI outlet sample and ELI sample is updated each year for 25 percent of the item strata in each PSU.

Finally, BLS data collectors visit the sampled outlets and select individual items for each sampled ELI to be priced each month (or every other month) through a multistage probability sampling technique known as disaggregation. The selection of a single unique item is referred to as a price quote. Regarding the CPI's gasoline index, data collectors would visit each sampled gasoline station each month and collect prices for the three grades of gasoline: regular, mid-grade, and premium. Obviously, obtaining gasoline prices in-person is extremely expensive for the BLS, and that cost limits the number of prices the agency can collect. The CORP5 data gives the BLS access to the millions of gasoline prices reported daily to the company's app in the CPI's 75 sampled geographical areas.

2.2. Index Estimation for Commodities and Services

Each month, the BLS calculates price relatives for all monthly and on-cycle bi-monthly elementary indexes for the CPI. An elementary index is an item stratum and index area combination. There are 243 item strata and 32 index areas in the CPI. Thus, the CPI consists of 7,776 elementary indexes (243×32).

Most elementary indexes use an expenditure-share-weighted geometric average $PRX^{G}_{t,t-1}$ for price relative calculation; other elementary indexes use the Laspeyres formula average $PRX^{L}_{t,t-1}$ (Bureau of Labor Statistics, 2020). Before using the CORP5 data in production, the BLS used the expenditure-share-weighted geometric average to calculate its price relatives for gasoline. The formulas for $PRX^{G}_{t,t-1}$ and $PRX^{L}_{t,t-1}$ are as follows for each index area *a* and item stratum *i* combination:

$$PRX^{G}_{t,t-1} = \prod_{j \varepsilon a,i} \left(\frac{P_{j,t}}{P_{j,t-1}} \right)^{\frac{W_{j,b}}{\sum_{k \in a,i} W_{k,b}}} \quad (1)$$

$$PRX^{L}_{t,t-1} = \frac{\sum_{j \varepsilon a,i} \left(\frac{W_{j,b}}{P_{j,b}} \right) P_{j,t}}{\sum_{j \varepsilon a,i} \left(\frac{W_{j,b}}{P_{j,b}} \right) P_{j,t-1}} \quad (2)$$

Where:
$P_{j,t} =$ the price of the *j*th observed item in month t for area-item combination a, i;
$P_{j,t-1} =$ the price of the *j*th observed item in month $t-1$ for area-item combination a, i;
$P_{j,b} =$ item *j*'s price in base period b; and
$W_{j,b} =$ item *j*'s weight in base period b.

An elementary index value for area *a* and item stratum *i* is calculated by multiplying the previous month's index ($IX_{a,i,t-1}$) by the price relative for area *a* and item stratum *i* in month *t* ($PRX_{a,i,t}$):

$$IX_{a,i,t} = IX_{a,i,t-1} \times PRX_{a.i,t} \text{ (3)}$$

In the base month (where $t = 0$), the index for area *a* and item stratum *i* is set equal to 100.

$$IX_{a,i,t=0} = 100 \text{ (4)}$$

The CPI item structure has four levels of classification. That is, the CPI's 243 elementary indexes feed into 70 expenditure classes (ECs); the 70 ECs make up eight major groups; and the eight major groups make up the entire CPI. To calculate the aggregated indexes, the 7,776 elementary indexes are multiplied by an aggregation weight derived from tabulated CE data; the product is called a cost weight ($CW_{a,i,t}$). These cost weights are then aggregated to calculate the indexes for the three levels above the elementary index level. For example, equation five gives the formula to calculate an index for a basic item *i* for all U.S. cities at time *t*:

$$IX_{\text{All U.S Cities.}i,t} = IX_{All\ U.S.Cities,i,t-1} \times \frac{\sum_{a\varepsilon \text{All U.S Cities.},i} CW_{a,i,t}}{\sum_{a\in \text{All U.S Cities.},i} CW_{a,i,t-1}} \text{ (5)}$$

Where:
$IX_{\text{All U.S.},i,t}$ = Aggregate Index for *A*ll U.S. Cities for item i at time t;
$IX_{\text{All U.S.},i,t-1}$ = Aggregate Index for *A*ll U.S. Cities for item i at time $t-1$
$\text{CW}_{\text{a,i,t}}$ = Cost weight CW for area a for item stratum i at time t; and
$\text{CW}_{\text{a,i,t-1}}$ = Cost weight CW for area a for item stratum i at time $t-1$.

2.3 Price Relative Calculation for CORP5

To use the CORP5 data in production, the BLS had to modify the CPI's traditional method of index calculation because it is not practical for estimating an average price relative from multiple prices for one item. For manual collection, the BLS collects a single price quote for a specific item each month. The CORP5 data, however, provides daily prices for each grade of gasoline for each reported station every month. The three grades of gasoline (regular, midgrade, and premium) are the entry level items (ELIs) that make up the CPI's gasoline index TB01. To estimate the price change at the station-ELI level, the BLS first calculates the arithmetic mean of the price of each grade of gasoline for every reported station each month. The price relative for each county within an Index PSU is then estimated using an unweighted geometric mean formula, which is also known as the Jevons price index formula ($PRX^{J}_{C,ELI,t,t-1}$). The International CPI Manual (IMF, et al, 2020) recommends using the unweighted mean formula when expenditure information is unavailable. Equation (6) provides the Jevons formula for calculating the price relative for county *c* for a gasoline ELI *i* at time *t*:

$$PRX^{J}_{c,i,t,t-1} = \prod_{g\varepsilon c,i} \left(\frac{\bar{P}_{g,\text{i},t}}{\bar{P}_{g,\text{i},t-1}} \right)^{\frac{1}{\text{n}}} \text{ (6)}$$

Where:
$\bar{P}_{g,i,t}$ = the average price of gasoline for ELI i from station g in county c in month t ;

$\bar{P}_{g,i,t-1} =$ the average price of gasoline for ELI i from station g in county c in month $t-1$; and
$n =$ number of average prices in month t and t-1 for ELI i in county c used in estimation[1].

After the price relatives for each county and ELI are derived, the BLS then calculates the following:

1. Price relatives for each Index PSU-ELI using a weighted geometric average of the county-ELI price relatives,
2. Price Relatives for each Index Area-ELI using a weighted geometric average of the Index PSU-ELI price relatives, and
3. Price Relatives for the CPI's 32 Index Areas for all gasoline.

Once the BLS has price relatives for the 32 Index Areas for all gasoline, the agency can resume its traditional method of index estimation for the CPI.

3. Methodology

As mentioned earlier in the Introduction, when the BLS replaced its manually collected gasoline price data with the crowd sourced CORP5 data, the agency expected to see a significant decrease in its SEs for gasoline because of the substantial increase in the number of prices used in the index. However, when one-, two-, six-, and twelve-month percentage change (PC) SEs were calculated using the CORP5 data and compared to the CPI SEs from production from January 2018 – May 2021, the decrease was minimal. Specifically, the one-month SEs decreased on average for All U.S. Cities by about 0.0471, the two-month SEs decreased by 0.0199, the six-month SEs decreased by 0.0050, and the 12-month SEs decreased by only 0.0003. For context, Table 1 below provides the median one-, two-, six-, and twelve-month price changes and corresponding standard errors for Gasoline for All U.S. Cities from 2016 – 2021.

Table 1. Median Price Change and Median Price Change Standard Error for Gasoline (All Types) for All U.S. Cities for 1-, 2-, 6-, and 12-month intervals

	1-Month		2-Month		6-Month		12-Month	
Year	Med. Price Change	Med. Std. Error	Med. Price Change	Med. Std. Error	Med. Price Change	Med. Std. Error	Med. Price Change	Med. Std. Error
2016	1.79	0.16	-0.72	0.19	0.53	0.17	-14.57	0.15
2017	-0.14	0.13	2.04	0.13	5.53	0.16	12.54	0.18
2018	0.32	0.20	0.48	0.24	10.38	0.31	13.02	0.38
2019	-0.05	0.21	-2.57	0.27	-3.73	0.39	-4.34	0.34
2020	-0.51	0.20	-1.62	0.24	-8.96	0.33	-17.39	0.32
2021	2.64	0.13	6.37	0.24	19.68	0.42	43.87	0.70

Also stated above, one theory as to why there was such a small decrease is because the BLS kept the same number of replicate samples for its gasoline index when it replaced its agency collected data with the CORP5 data. To see if increasing the number of replicates might further decrease the gasoline SEs from the CORP5 data, one-, two-, six-, and twelve-

[1] The BLS imputes prices for gasoline station – ELI combinations in month t for up to four months.

month PC SEs were calculated twice using the CPI's stratified random groups (SRG) variance methodology from January 2018 – May 2021. In short, the SRG variance for Area *A* – Item *I* is equal to:

$$V(A, I, t, t\text{-}12) = \sum_{a \in A} \frac{1}{N_a(N_a-1)} \sum_{r \in R_a} (PC[(A, I, 00) - (a, I, 00) + (a, I, r), t, t-12] - PC(A, I, t, t-12))^2 \quad (7)$$

Where:
$r \in R_a$ refers to the set of replicates in AREA a,
$a \in A$ refers to the 23 self-representing and 9 non-self-representing index areas in A,
I = All Gasoline Types, and
N_a is the number of variance replicates in AREA a.

The first set of SEs used the same number of replicates as in CPI production. The second set of SEs used fourfold the number of replicates in CPI production for the self-representing areas and twice the number of replicates for the non-self-representing areas[2]. Table 2 below provides the current number of replicates assigned to each index area and the new number of replicates assigned to each index area for this study.

Table 2. Number of Replicates Assigned to Each CPI Index Area

Area	Area Description	Number of Replicates from Production	New Number of Replicates
N110	New England - Size Class B/C	2	2
N120	Middle Atlantic - Size Class B/C	2	4
N230	East North Central - Size Class B/C	4	8
N240	West North Central - Size Class B/C	2	4
N350	South Atlantic - Size Class B/C	6	12
N360	East South Central - Size Class B/C	3	6
N370	West South Central - Size Class B/C	4	8
N480	Mountain - Size Class B/C	2	4
N499	Pacific Size Class B/C Other Than Urban Hawaii and Urban Alaska	2	4
S11A	Boston-Cambridge-Newton, MA-NH	2	8
S12A	New York-Newark-Jersey City, NY-NJ-PA	8	32
S12B	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	2	8
S23A	Chicago-Naperville-Elgin, IL-IN-WI	4	16
S23B	Detroit-Warren-Dearborn, MI	2	8
S24A	Minneapolis-St. Paul-Bloomington, MN-WI	2	8
S24B	St. Louis, MO-IL	2	8
S35A	Washington-Arlington-Alexandria, DC-VA-MD-WV	2	8
S35B	Miami-Fort Lauderdale-West Palm Beach, FL	2	8
S35C	Atlanta-Sandy Springs-Roswell, GA	2	8

[2] The new number of replicates assigned to Area N110 is equal to the number of replicates in production because that area only contains two Index PSUs.

Table 2. Number of Replicates Assigned to Each CPI Index Area			
Area	**Area Description**	**Number of Replicates from Production**	**New Number of Replicates**
S35D	Tampa-St. Petersburg-Clearwater, FL	2	8
S35E	Baltimore-Columbia-Towson, MD	2	8
S37A	Dallas-Fort Worth-Arlington, TX	2	8
S37B	Houston-The Woodlands-Sugar Land, TX	2	8
S48A	Phoenix-Mesa-Scottsdale, AZ	2	8
S48B	Denver-Aurora-Lakewood, CO	2	8
S49A	Los Angeles-Long Beach-Anaheim, CA	4	16
S49B	San Francisco-Oakland-Hayward, CA	2	8
S49C	Riverside-San Bernardino-Ontario, CA	2	8
S49D	Seattle-Tacoma-Bellevue, WA	2	8
S49E	San Diego-Carlsbad, CA	2	8
S49F	Urban Hawaii	2	8
S49G	Urban Alaska	2	8
Total Number of Replicates		**83**	**276**

3.1 Replicate Assignments

The replicate samples for the variance estimates in the CPI are formed using two methods depending on whether an index area is a self-representing index area or a non-self-representing index area. Self-representing index areas consist of only one Index PSU that the BLS selected with certainty for its geographical area sample as discussed in Section 2.1, and non-self-representing index areas are comprised of two or more Index PSUs that the program selected with probability. For the CORP5 variance estimates, the BLS creates replicate samples for the self-representing index areas by randomly assigning each gasoline station along with all its prices to a replicate. The stations keep the same replicate assignments throughout their entire existence in the sample. In the non-self-representing index areas, the BLS assigns all the stations from an Index PSU to a single replicate. For example, Index Area N120 (Middle Atlantic - Size Class B/C) has four design Index PSUs:

1. Pittsburgh, PA (N12C),
2. Buffalo, NY (N12D),
3. Rochester, NY (N12E), and
4. Reading, PA (N12F).

In production, the Index PSUs Pittsburgh, PA (N12C) and Reading, PA (N12F) are assigned to replicate 41, while Buffalo, NY (N12D) and Rochester, NY (N12E) are assigned to replicate 42. The BLS assigns the Index PSUs from the non-self-representing index areas to replicates based on whether the bi-monthly items in the Index PSUs are priced in even months or odd months. Each replicate should have an equal number of "even" and "odd" Index PSUs so that the replicate sample has collected price data every month. CORP5 provides price data every month for all the sampled PSUs. Therefore, each Index PSU from the non-self-representing index areas can be assigned to its own replicate. Furthermore, by doubling the number of replicates in the non-self-representing areas, each index PSU ends up assigned to its own replicate.

3.2 Significance Tests

Wilcoxon signed-rank tests were conducted to see if the differences between the SEs from CORP5 with the current number of replicates and with the increased number of replicates were significant. The Wilcoxon signed rank test is a nonparametric alternative to the paired student's t-test. To be precise, data must be normally distributed for a paired t-test but not for the Wilcoxon signed rank test. From Hollander and Wolfe (1999), the signed-rank test is designed for analyses in which the primary interest is centered on the median of a population. It was hypothesized that there would be no difference between the SEs calculated from the current number of replicates and the SEs calculated from the increased number of replicates. Thus, the median of the differences was expected to be zero.

4. Results

4.1 Differences between Gasoline SEs for All U.S. Cities

Table 3 below provides the median differences between the gasoline one-, two-, six-, and twelve-month SEs "For All U.S. Cities (0000)" calculated from the CORP5 data with the current number of replicates in CPI production and with the increased number of replicates from January 2018-May 2021. The differences are equal to the CORP5 SEs with the increased number of replicates less the CORP5 SEs using the current number of replicates. Table 3 also provides the results of the Wilcoxon signed rank tests applied to the differences.

The signed rank test results clearly show that there are significant differences between the CORP5 SEs calculated with the two sets of replicates (at the $\alpha = 0.05$ level), and that increasing the number of replicates does appear to decrease the SEs. For example, the median differences between the one-month and twelve-month SEs from the CORP5 data with the two sets of replicates are 0.0388 and 0.0703, respectively, with the CORP5 SEs using the number of replicates in production being greater. This decrease, however, is not as large as the BLS expected when it replaced its collected data with the secondary source data. Figures 1 and 2 below graph the one-month and twelve-month gasoline (TB01) SEs for All U.S. Cities from CORP5 with the two sets of replicates 201801-202105. The graphs clearly illustrate the decrease in the SEs after increasing the number of replicates.

Table 3. Wilcoxon Signed Rank Test for Gasoline SEs from CORP5 with Increased Reps. vs. CORP5 w/Prod. Number of Reps.

Standard Error	Area	Med. SE from 83 Replicates	Med. SE from 276 Replicates	Med. Diff. btw. CORP5 SEs using the Two Sets of Replicates	n	W	Z-Value	PVAL
SE01	0000	0.1828	0.1286	-0.0388	41	70	-4.67	0.0000
SE02	0000	0.2444	0.1852	-0.0576	40	101	-4.15	0.0000
SE06	0000	0.3641	0.2648	-0.0776	35	14	-4.93	0.0000
SE12	0000	0.3525	0.2568	-0.0703	29	136	-1.76	0.0390

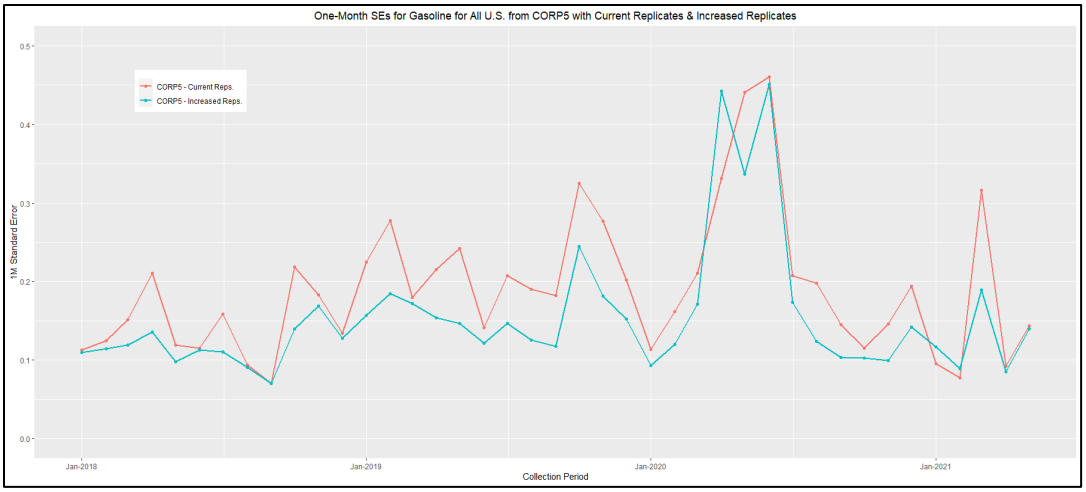

Figure 1. One-Month SEs for All Cities for Gasoline from JAN 2018 – May 2021

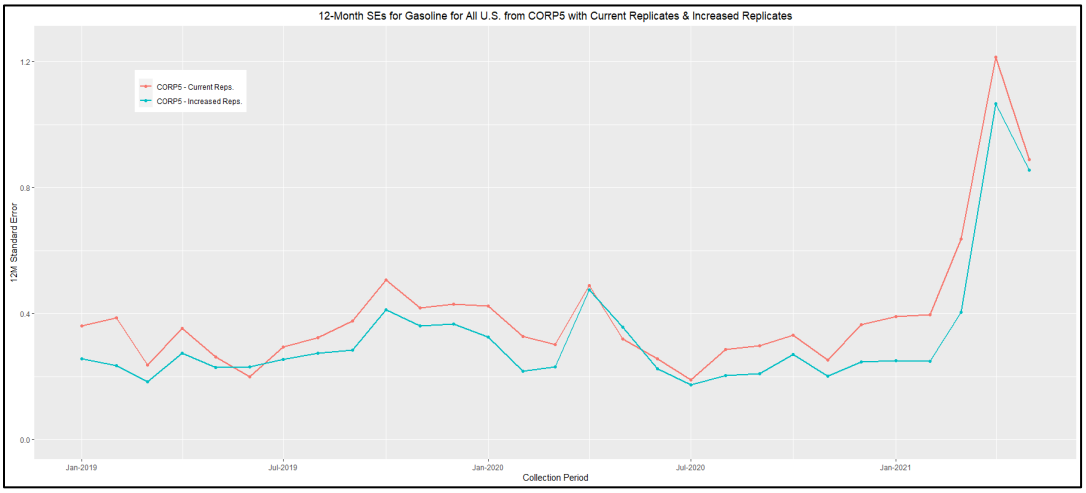

Figure 2. Twelve-Month SEs for All Cities for Gasoline from JAN 2019 – May 2021

In addition to calculating the PCs and corresponding SEs for "All U.S. Cities," the BLS also calculates PCs and SEs for the CPI at lower levels of aggregation. As described in Section 2, there are two types of index areas in the CPI: self-representing index areas and non-self-representing index areas. The following two subsections investigate how increasing the number of replicates impacted the gasoline SEs from CORP5 for "All Self-Representing Index Areas (S000)" and for "All Non-Self-Representing Index Areas (N000)."

4.2 Differences between Gasoline SEs for All Self-Representing Index Areas

Table 4 below provides the median differences between the gasoline one-, two-, six-, and twelve-month PC SEs for all "Self-Representing Areas (S000)" calculated from the CORP5 data with the current number of replicates in CPI production and with the increased number of replicates from 201801-202105. The differences are again equal to the CORP5 SEs with the increased number of replicates less the CORP5 SEs using the current number of replicates. Table 4 also provides the results of the Wilcoxon signed rank tests applied to the differences.

The signed rank test results show that the one-, six-, and twelve-month SEs from the CORP5 data are not significantly different. This outcome might be due to the fact that there was such a large decrease in the gasoline SEs for S000 when the BLS replaced the agency collected data with the CORP5 data. Specifically, the median differences between the one-month and twelve-month SEs from CORP5 with the current number of replicates and CPI production were -0.1332 and -0.2372, respectively. Figures 3 and 4 below provide line graphs of the one-month and twelve-month gasoline (TB01) SEs for "All Self-Representing Areas (S000)" from CORP5 with the two sets of replicates from 201801-202105.

Table 4. Wilcoxon Signed Rank Test Results for Gasoline SEs from CORP5 with Increased Reps. vs. CORP5 w/Prod. Number of Reps. for All Self-Representing Areas

Standard Error	Area	Med. SE from 83 Replicates	Med. SE from 276 Replicates	Med. Diff. btw. CORP5 SEs using the Two Sets of Replicates	n	W	Z-Value	PVAL
SE01	S000	0.0129	0.0141	0.0007	41	321	-1.42	0.0780
SE02	S000	0.0167	0.0181	0.0009	40	204	-2.77	0.0028
SE06	S000	0.0223	0.0234	0.0014	35	219	-1.57	0.0579
SE12	S000	0.0232	0.0275	0.0019	29	209	-0.18	0.4271

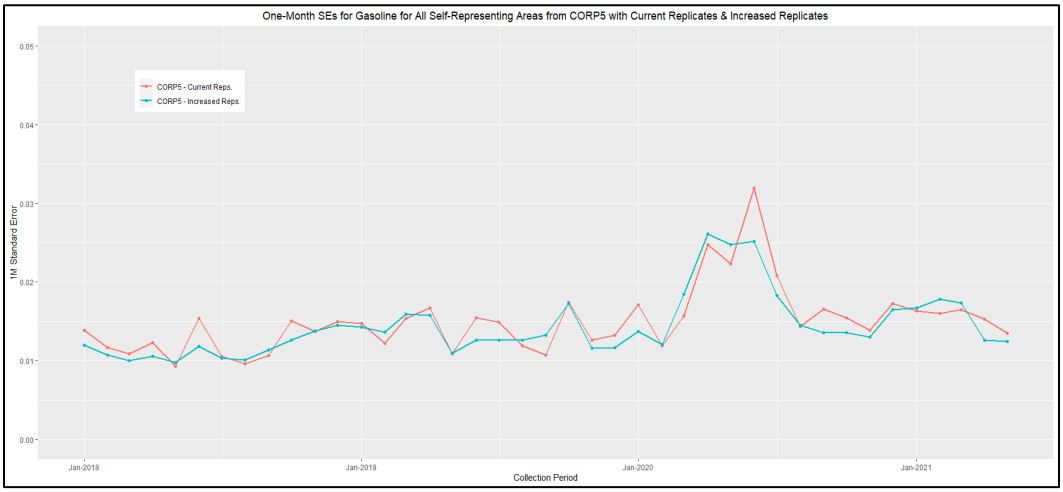

Figure 3. One-Month PC SEs for All Self-Representing Areas for Gasoline from JAN 2018-May 2021

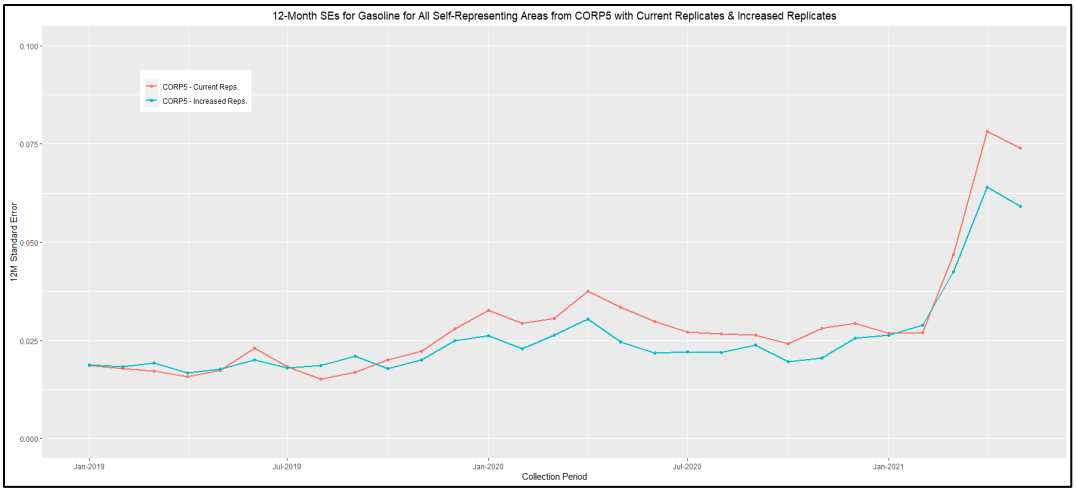

Figure 4. 12-Month Standard Errors for All Self-Representing Areas for Gasoline from JAN 2019 – May 2021

4.3 Differences between Gasoline SEs for All Non-Self-Representing Index Areas

Table 5 below provides the median differences between the gasoline one-, two-, six-, and twelve-month SEs for all "Non-Self-Representing Areas (N000)" calculated from the CORP5 data with the current number of replicates in CPI production and with the increased number of replicates from 201801-202105. The differences are again equal to the CORP5 SEs with the increased number of replicates less the CORP5 SEs using the current number of replicates. Table 5 also provides the results of the Wilcoxon signed rank tests. The signed rank test results show that the one-, two-, six-, and twelve-month PC SEs from the CORP5 data are significantly different at the $\alpha = 0.05$ level. These differences, however, are nowhere near the differences seen earlier between the SEs from CPI Production and the CORP5 data for S000. For example, the median differences between the one-month and twelve-month SEs from CPI-Production and CORP5 for S000 were 0.1332 and 0.2372, respectively with the SEs from CPI production being greater. The median differences between the one-month and twelve-month SEs from CORP5 with the two sets of replicates for N000 were only 0.0662 and 0.1172, respectively, with the SEs from the CORP5 data using current number of replicates in production being greater.

Table 5. Wilcoxon Signed Rank Test Results for Gasoline SEs from CORP5 with Increased Reps. vs. CORP5 w/Prod. Number of Reps. for All Non-Self-Representing Areas

Standard Error	Area	Med. SE from 83 Replicates	Med. SE from 276 Replicates	Med. Diff. btw. CORP5 SEs using the Two Sets of Replicates	n	W	Z-Value	PVAL
SE01	N000	0.3130	0.2210	-0.0662	41	70	-4.67	0.0000
SE02	N000	0.4189	03163	-0.0984	40	102	-4.14	0.0000
SE06	N000	0.6280	0.4531	-0.1329	35	14	-4.93	0.0000
SE12	N000	0.6012	0.4382	-0.1178	29	136	-1.76	0.0390

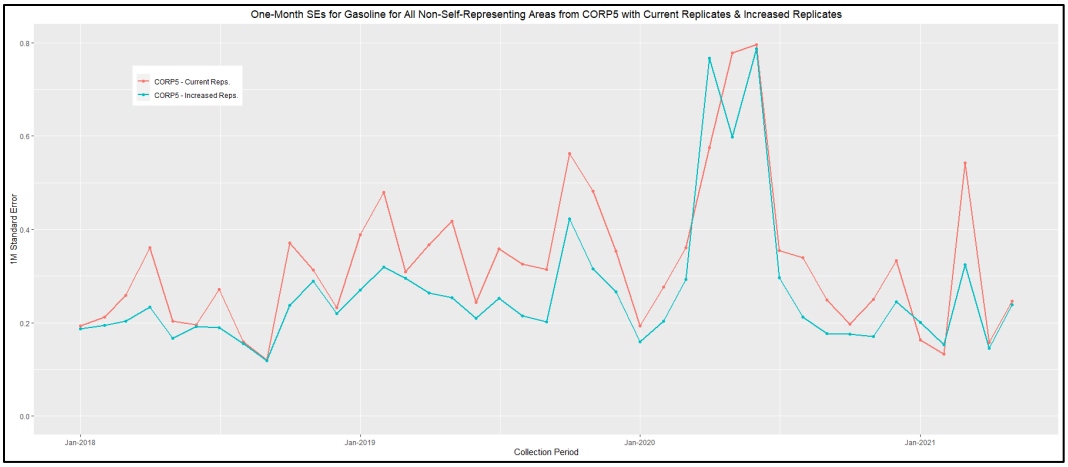

Figure 5. One-Month SEs for All Non-Self-Representing Areas for Gasoline from JAN 2018-May 2021

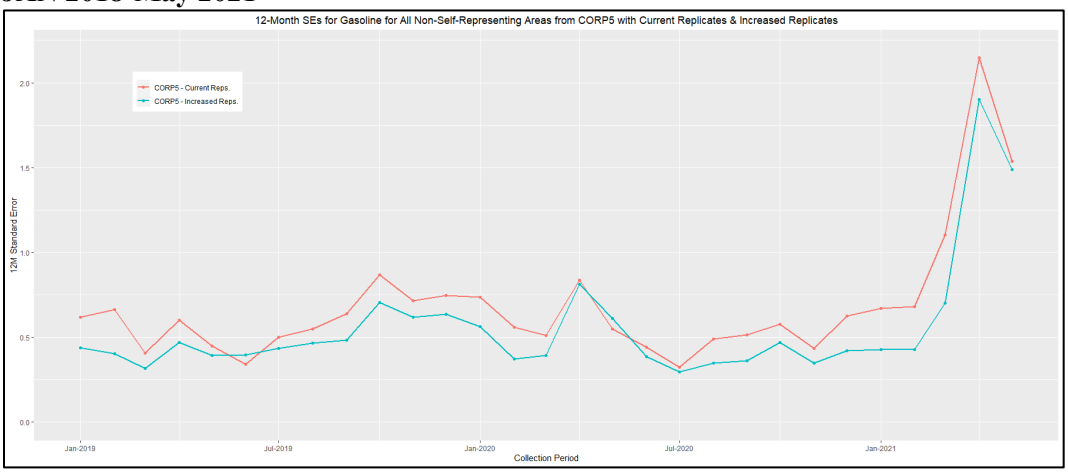

Figure 6. 12-Month SEs for All Non-Self-Representing Areas for Gasoline from JAN 2019-May 2021

Consequently, increasing the number of replicates did decrease the SEs for gasoline at the "All U.S. Level (0000)." That decrease, however, only occurred because the SEs for the aggregate area N000 decreased significantly. Significant differences were not found between the CORP5 one-, six-, and twelve-month gasoline SEs for S000 using the two sets of replicates. This is likely because the S000 SEs from CORP5 with the current number of replicates in production were already dramatically less than the gasoline SEs from CPI Production. Table 6 provides the median and mean one-month and twelve-month SEs for gasoline by aggregate area and year from 2018 – 2021[3] from CPI production and CORP5 with the current and increased number of replicates. The median and mean 12-month SEs for S000 pooled over January 2019-May 2021 from CPI Production were 0.2587 and 0.2618, respectively. During the same time, the median and mean SEs for S000 from CORP5 with the current number of replicates were 0.0232 and 0.0255, respectively. Given that the S000 SEs from CORP5 were already close to zero, decreasing the SEs in this area anymore would be extremely difficult. CORP5, however, did not have the same type of impact on the N000 SEs. By increasing the number of replicates, however, the median 12M

[3] The year 2021 only has data for January – May 2021.

SE for N000 decreased from 0.6012 to 0.4382, and the mean 12M SE decreased from 0.6831 to 0.5560, respectively.

Table 6. Median and Average Gasoline (TB01) SEs for SE01 and SE12 by Area and Year from 2018 - 20221

Year		CPI PRODUCTION				CORP5 with Current Number of Replicates				CORP5 with Current Increased Number of Replicates			
		SE01		SE12		SE01		SE12		SE01		SE12	
		MED	AVG	MED	AVG	MED	AVG	MED	AVG	MED	AVG	MED	AVG
'18	0000	0.20	0.20			0.13	0.14			0.11	0.12		
'19		0.21	0.23	0.34	0.34	0.21	0.22	0.36	0.35	0.15	0.16	0.27	0.28
'20		0.20	0.28	0.32	0.36	0.20	0.23	0.31	0.32	0.13	0.20	0.23	0.26
'21		0.26	0.25	0.61	0.62	0.10	0.15	0.64	0.71	0.12	0.12	0.40	0.56
'18	S000	0.13	0.14			0.01	0.01			0.01	0.01		
'19		0.17	0.17	0.17	0.18	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.02
'20		0.16	0.20	0.27	0.27	0.02	0.02	0.02	0.03	0.02	0.02	0.03	0.03
'21		0.12	0.12	0.42	0.45	0.01	0.01	0.03	0.04	0.02	0.02	0.04	0.05
'18	N000	0.33	0.33			0.22	0.24			0.19	0.20		
'19		0.35	0.37	0.56	0.57	0.36	0.38	0.61	0.59	0.27	0.27	0.45	0.48
'20		0.33	0.46	0.51	0.58	0.34	0.39	0.53	0.55	0.23	0.34	0.39	0.45
'21		0.43	0.42	1.00	1.02	0.16	0.25	1.10	1.23	0.20	0.21	0.69	0.99

5. Conclusion

In June 2021, the BLS replaced the price data that its collects for its gasoline index in the CPI with data from a secondary source that the program refers to as CORP5. Given the substantial increase in the number of prices from CORP5, the CPI expected to see a significant decrease in its standard errors (SEs) for gasoline. The program, however, has only seen a small decrease in its production SEs from the data.

One theory as to why there has been little to no decrease in the SEs is because the BLS did not change the number of replicates for its gasoline index when it replaced the agency collected data with the CORP5 data. To see if increasing the number of replicates assigned to each index area would decrease the gasoline SEs from CORP5, SEs for gasoline were calculated using the CORP5 data twice from January 2018-May 2021. For the first set of variances, the number of replicates were kept the same as in CPI production. For the second set of variances, the number of replicates in the self-representing areas were increased fourfold and the number of replicates in the non-self-representing areas were doubled.

Wilcoxon signed rank tests showed that significant differences exist between the CORP5 SEs calculated with the two sets of replicates at the $\alpha = 0.05$ level, and that increasing the number of replicates decreased the gasoline SEs for All U.S. Cities (0000). To be specific, the median differences between the one-month and twelve-month SEs from the CORP5 data with the two sets of replicates were 0.0388 and 0.0703, respectively, with the CORP5 SEs using the current number of replicates in CPI production being greater. This decrease,

however, is not as large as the CPI expected when it replaced its directly collected price data with its secondary source data.

After investigating the differences between the SEs from the two sets of replicates for "All Self-Representing Index Areas (S000)" and "All Non-Self-Representing Index Areas(N000)," it was apparent that the decrease in gasoline SEs for "All U.S. Cities" only occurred because the SEs for the aggregate area N000 decreased significantly. Significant differences were not found between the CORP5 one-, six-, and twelve-month gasoline SEs for S000 using the two sets of replicates. This was likely because the S000 SEs from CORP5 with the current number of replicates in production were already considerably less than the gasoline SEs from CPI Production. The median and mean 12-month SEs from January 2019 – May 2021 from CPI Production were 0.2587 and 0.2618, respectively. During the same period, the median and mean SEs for S000 from CORP5 with the current number of replicates were 0.0232 and 0.0255, respectively.

One possible reason why the SEs for gasoline in the non-self-representing areas did not decrease from the increased number of prices from CORP5 as much as is in the self-representing areas is because the data did not increase the number of Index PSUs within the non-self-representing index areas. Thus, the variance between the PSUs remains the same. The self-representing areas, in contrast, are made up of only one index PSU. Thus, there is no variance between PSUs in the self-representing Index Areas. This theory will be further investigated going forward.

References

Bureau of Labor Statistics (2020). "BLS Handbook of Methods: Consumer Price Indexes." Accessed February 4, 2022 from the Bureau of Labor Statistics https://www.bls.gov/opub/hom/cpi/ .

Hollander, Myles and Douglas A. Wolfe, (1999). Nonparametric Statistical Methods, Second Edition, New York: John Wiley, pp. 35-42.

International Monetary Fund (IMF), International Labor Organization (ILO), Statistical Office of the European Union, United Nations Economic Commission for Europe, Organization for the Economic Co-operation and Development, and the World Bank (2020). "Consumer Price Index Manual, Concepts and Methods." Accessed May 10, 2022 from the International Labor Organization Website https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication/wcms_761444.pdf.