Simulated Statistics for the Proposed By-Division Design In the Consumer Price Index

John F Schilp
U.S. Bureau of Labor Statistics, Office of Prices and Living Conditions
2 Massachusetts Avenue NE, Washington, DC 20212
Schilp.John@bls.gov

Key Words: Consumer Price Index, Simulated Statistics, Design Verification

Abstract

This paper illustrates the research done in determining if by-Census Division stratification will give results similar to the existing Census Region city-size stratification results in the Consumer Price Index (CPI). Motivating this project is the proposed CPI switch from region city-size stratification to by-division stratification. Simulated by-division indexes for all Census Regions, as well as the non-self-representing part of the all U.S. CPI were completed by adjusting existing weights with respect to Census Division population share. In turn, these index time series are compared to hybrid by Census Region city-size index time series for the same areas and time period. A conclusion is given with comparison of results for CPI time series, 12-month percent changes and standard errors for the 12-month percent changes.

Background

While developing the specifications for the new geographic area design, the redesign team decided to stratify all non-self-representing geographic PSU by Census Division. This diverged from tradition as previous PSU designs were stratified by Census Region and City-Size. A city-size are either a large metropolitan or small micropolitan "Core Based Statistical Area" as defined by Census based on population. This Census Division stratification was in a response to requests for state level CPI indexes and Division level indexes come closer to this goal. There are nine Census Divisions. Of which, each of the four Census regions have two divisions except for the South which has three.

The by-division stratification was examined briefly by the redesign team in the stratification stage via calculating and examining trace(W) statistics (defined below) for different stratification groups. Simply, trace(W) is the sum of the major diagonal of the within strata variance-covariance matrix. Trace(W) statistics were used to determine the similarity of PSU within a strata for chosen statistics obtained from the American Community Survey. These stratification statistics were chosen to be Median Property Value, Median Household Income, Longitude and Latitude. Trace(W) is the sum of these PSU variables' variances within strata. Due to drastic differences in magnitude in these variables, they were normally standardized with mean equal to zero and standard deviation equal to one. A low trace(W) statistic of ACS variables was thought to be correlated with similarities in CPI 12-month percent change. Division is superior to Region – Size Class with regards to the size of trace(W). The tables below are taken from the BLS CPI-Statistical Method Division memo titled, "Changing the basic index areas for PSU stratification from Census Region to Census Division." They serve to illustrate the superiority of the by-Division stratification over Region – Size Class stratification via lower trace(W) statistics.

Here, W represents within-group dispersion matrix and the trace function is the sum of the major diagonal. The variables m = 1 to g is the stratum group and l = 1 to n_m is the number of PSU in stratum group, m.

$$W = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \bar{x}_m)(x_{ml} - \bar{x}_m)^T$$

In addition to W, the within-group dispersion matrix there is T, the total dispersion matrix and B, the between-group dispersion matrix. Both defined below:

$$T = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \bar{x})(x_{ml} - \bar{x})^T$$

$$B = \sum_{m=1}^{g} n_m (\bar{x}_m - \bar{x})\,(\bar{x}_m - \bar{x})^T$$

The matrixes W and B necessarily sum to T. When attempting to minimize W, the process is equivalent to the maximization of B.

While the most commonly used criteria is the minimization of trace(W) and that is what was done here, it does have its limitations. First it is scale dependent and all variables need to be standardized to arrive at consistent answers. The other limitation is that the use of this criterion may impose a 'spherical' structure on the clusters even when the 'natural' clusters in the data are of other shapes.

In future redesigns, it may be important to looks at the determinant of W because the minimization of the determinant of W may lead to finding these 'natural' clusters. Large values of det(T)/det(W) indicate that the group mean vectors differ. In maximizing this ratio, since for all partitions of n individuals into g groups, det(T) remains the same, a minimization of the det(W) may have led to better clustering. (Everitt, et al., pp. 115-116)

Regardless, when comparing the by-division stratification trace(W) versus the trace(W) for the region-city-size stratification the conclusion is that by-division stratification would reduce within-group variation and give the best homogeneity of strata for PSU. This merits more research with respect to resulting indexes, 12-month percent changes and percent change standard errors.

Census Region (by-Division)	Economic Variables Trace(W)	Geographic Variables Trace(W)	Total Trace(W)
1 North East	52.9	82.3	135.2
2 Midwest	198.4	188.8	387.2
3 South	245.1	176.5	421.7
4 West	92.5	74.0	166.5
		Total:	1110.6

Census Region (Existing)	Economic Variables Trace(W)	Geographic Variables Trace(W)	Total Trace(W)
1 North East	56.0	85.0	141.0
2 Midwest	192.8	231.4	424.2
3 South	321.6	246.2	567.8
4 West	164.0	96.1	260.1
		Total:	1393.0

Given the same data, one can expect some differences in the answer one gets for the CPI for non-self-representing areas in a region based on how the area is split into basic index areas. There are the Large and Small Midwest index areas, coded B200 and C200 respectively, or Division 3 and Division 4, which are both divisions in the Midwest region 2 – East North Central and West North Central, respectively. The reason for this is that a geometric mean formula is used to aggregate quotes within a basic index area and a Laspeyres' formula is used to aggregate across basic index areas within an aggregate index area. One would expect that the higher the percentage of data aggregated with a geometric mean formula, the lower the index will be.

Methodology

In order to more thoroughly determine the merits of this by-division stratification design, simulated division based indexes, 12-month percent changes and standard errors were calculated from existing CPI price quotes and weights with some adjustments. These weight adjustments are illustrated below.

The by-division Midwest indexes were combined to create region indexes, called N200. The two division indexes are combined in order to compare them with properties of the existing region index time series, B+C200. So, this division index time series was then compared to a hybrid region city-size index time series. These exclude the self-representing PSUs, and are also calculated from existing price quotes and weights.

Self-representing PSU were excluded from calculation in order to bring out the differences in these simulated index time series. If self-representing PSU were included in the calculation the results would be more similar.

In order to produce the N200 index time series, the primary adjustment was to the weights. The strata population of Census Region 2, the Midwest can be broken into 2 divisions. These divisions are East North Central and West North Central, or Division 3 and Division 4 respectively. For instance, for the large PSU in the Midwest stratum B222, 57% of the population was found in Division 3, 41% of the population was found in Division 4 and the remaining 2% is in Division 6, the East South Central division of the South region. This was done for each large and small stratum to see how the population falls in each division. These proportions were used to adjust the weights that feed into the simulated values. The simulation program then calculates by-division indexes using only the percent of weight found in that particular division for each PSU.

In our example, while large and small Midwest index areas B200 and C200 have pre-existing replicates and replicate indexes. The simulated Division 3 and Division 4 index areas had to be subdivided into replicates containing two or more PSUs, with at least one PSU from each pricing cycle. The replicates were not as balanced as if the sample had

been designed with division replicates in mind. Illustration below. The results in standard error may not look good because the by Census Division replicates were not optimally constructed as shown below.

PSU	Current Replicate	Computation Cycle	Div3	Div4	Div3 rep #	Div4 rep #
B218	5	2	0.39	0.61	1	1
B220	1	2	1	0	2	
B222	3	2	0.57	0.41	3	2
B224	1	3	1	0	1	
B226	4	3	1	0	2	
B228	2	3	1	0	3	
B230	3	3	0.95	0.05	4	3
B232	4	2	0.90	0.07	4	1
B234	2	2	0.61	0.39	5	3
B236	5	3	0	1		1
B356	2	3	0.09	0	6	
B372	8	2	0.11	0	5	
C212	2	2	0.75	0.25	6	2
C216	2	3	0.19	0.81	6	2
C218	1	2	0.21	0.79	1	3
C222	1	3	0.83	0.17	5	3
C328	2	2	0.003	0	6	
C332	1	3	0.005	0	6	

The imbalances in this replicate assignment are illustrated here. First, Division 4 has 6 replicates on even month cycle 2 while there are 4 replicates on cycle 3, or the odd month cycle. Also, Division 4 cycle 2 contains the weight of 2.527 PSU while cycle 3 contains the weight of 2.022 PSU. This is compared to Division 3, which has the weight of 4.543 PSU on the even cycle 2 and the weight of 5.065 PSU on the odd cycle 3. The Division 3 also has 9 replicates on an even cycle 2 and 8 replicates on odd cycle 3.

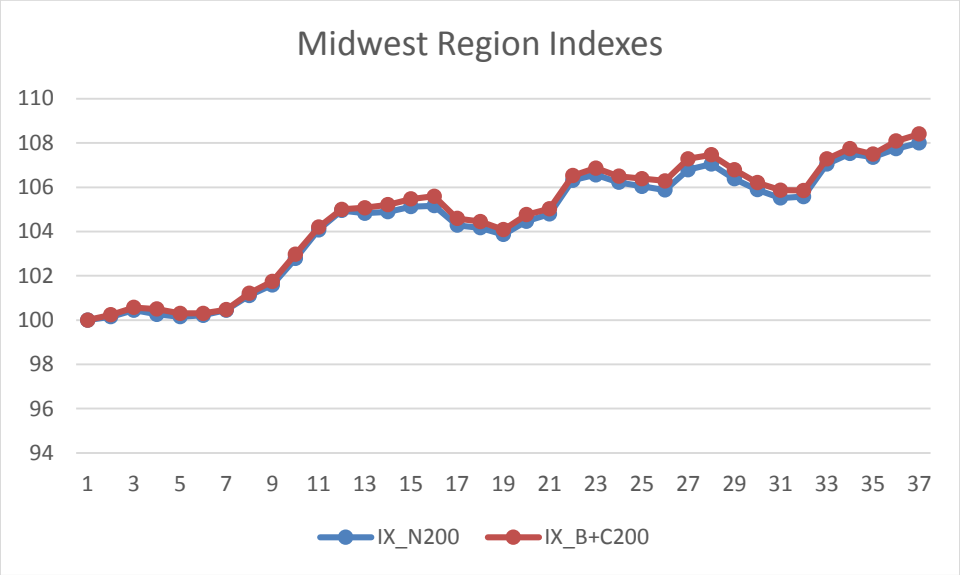
If the replicate assignments were optimized in production to achieve more balance, then the standard errors would be closer for proposed N200 and the existing B+C200. Here, the standard error for N200 is a "close estimate" for what is expected in production, when the replicate assignments are optimized for a by-division design.

Data

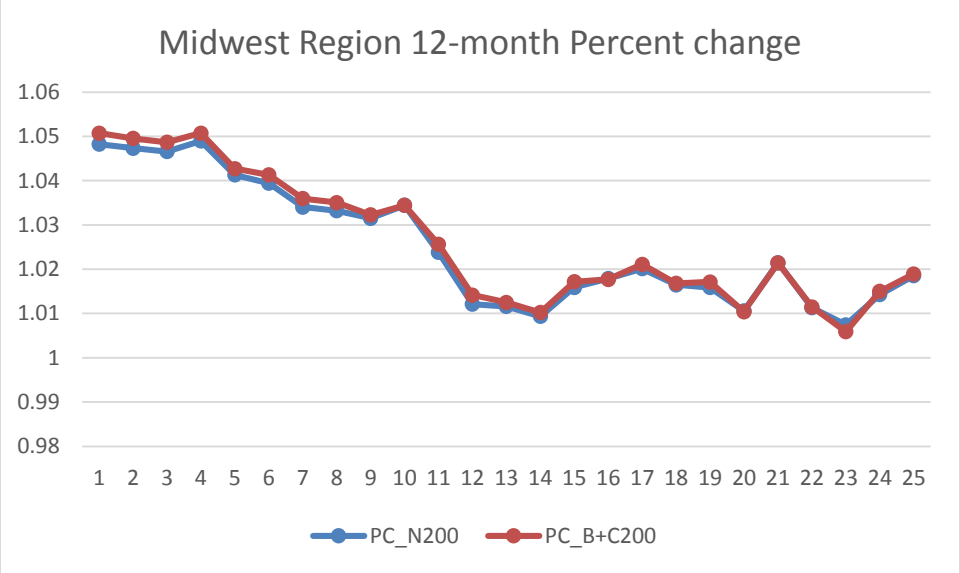
The time period examined starts in June 2010 and continues monthly until June 2013. This gives 3 years of simulated data or 2 years of 12-month percent changes. Due to Medical Insurance, Rent and Owner's Equivalent Rent being calculated at the region level, and fed into the simulator for final calculation, these item-area prices were excluded from the examined time series' data. This series is informally called SA0CS within BLS as this aggregate is not an officially produced aggregate.

Results

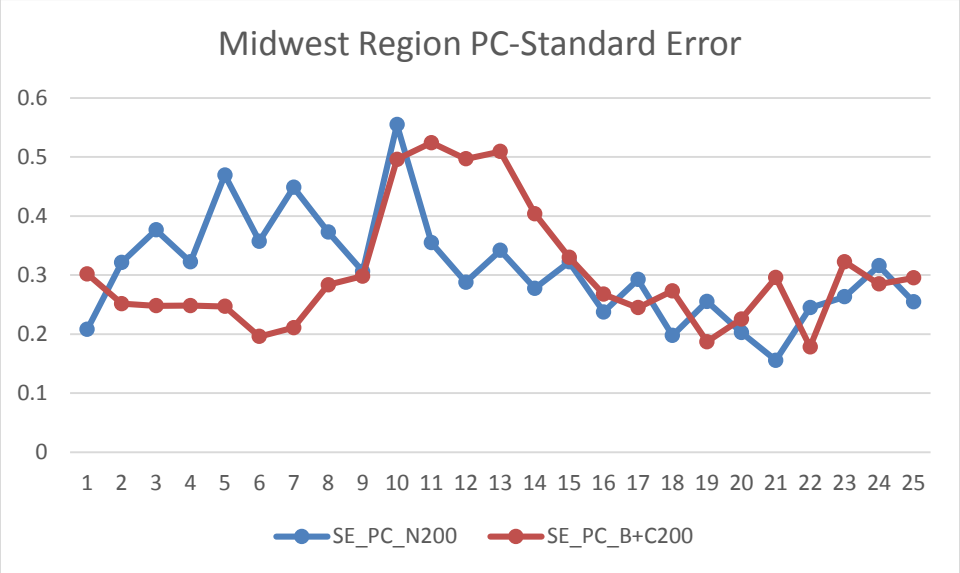
Below are the plots for Region 2 Indexes, Percent Changes and Standard Errors. The remainder of the plots for the other regions and the all US are included in the appendix. In the Midwest Region Index plot both time series begin at 100 for the base period. Both continue closely over the 37 months investigated and the final distance between N200 and the existing B+C200 is .3949.



The plot below is 2 years of 12-month Percent changes for both N200 and B+C200. These are very close to each other and follow the same trend.



The percent change standard errors are below. These values are more spread apart than the other 2 graphs. The average SE for N200 is 0.310 while the B+C200 average is .305. The difference between these two average SEs is .005.



Census Region	By-Division Average SEs	Region-City-Size Average SEs
National (All-US)	.108	.068
Northeast	.266	.273
Midwest	.310	.305
South	.166	.186
West	.208	.188

Conclusion

It is encouraging how similar indexes and percent changes are from by-division method versus by region-city-size method. There are little practical differences between the index values for both simulations. There are also no significant differences between percent change values for both simulations as well. Also, it appears that standard errors will be close with the by-division stratification than with the existing region-city-size stratification CPI has now.

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

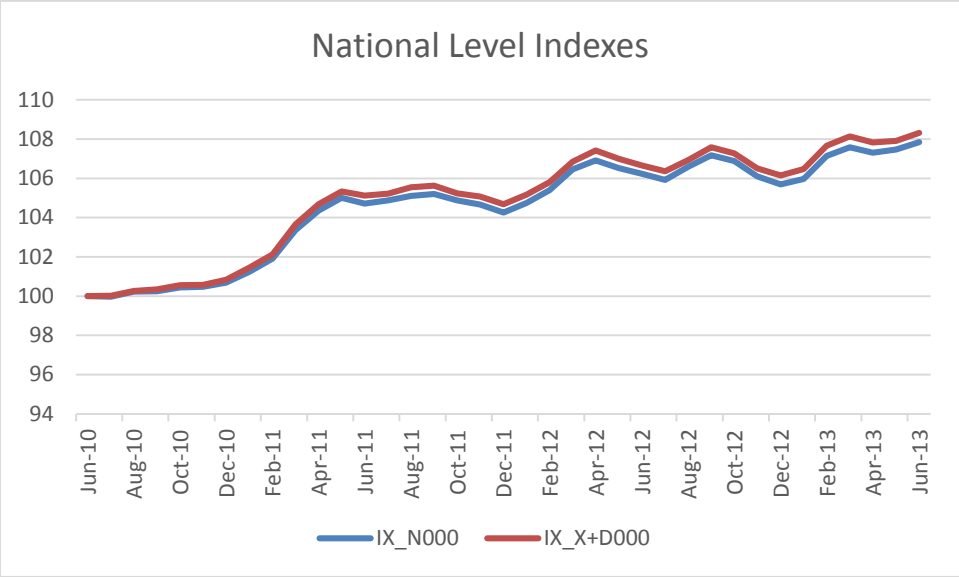
References

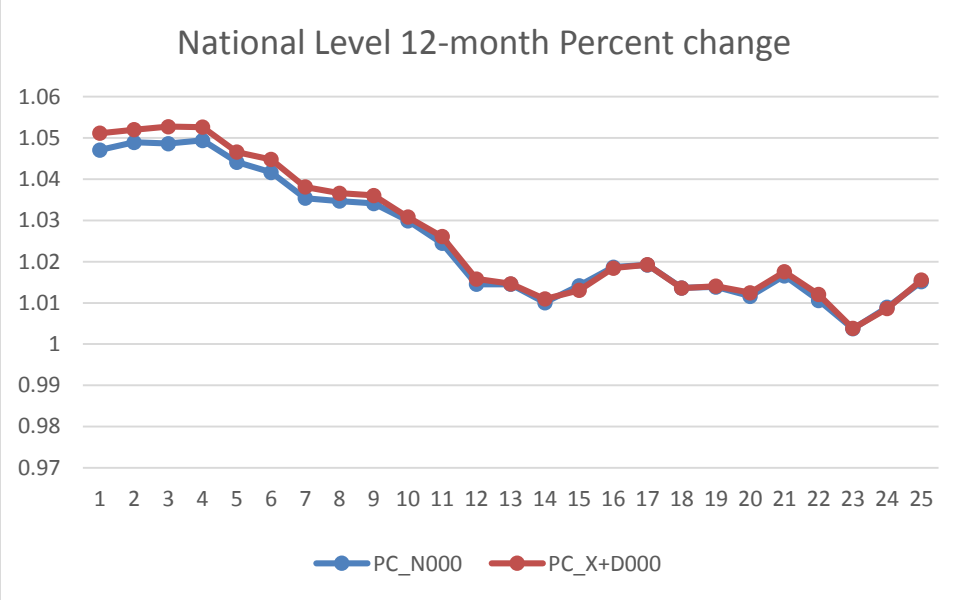
CPI/CE Area Redesign Team. 2011. Changing the basic index areas for PSU stratification from Census Region to Census Division. Statistical Methods Division Memorandum to BLS management. Washington, DC.

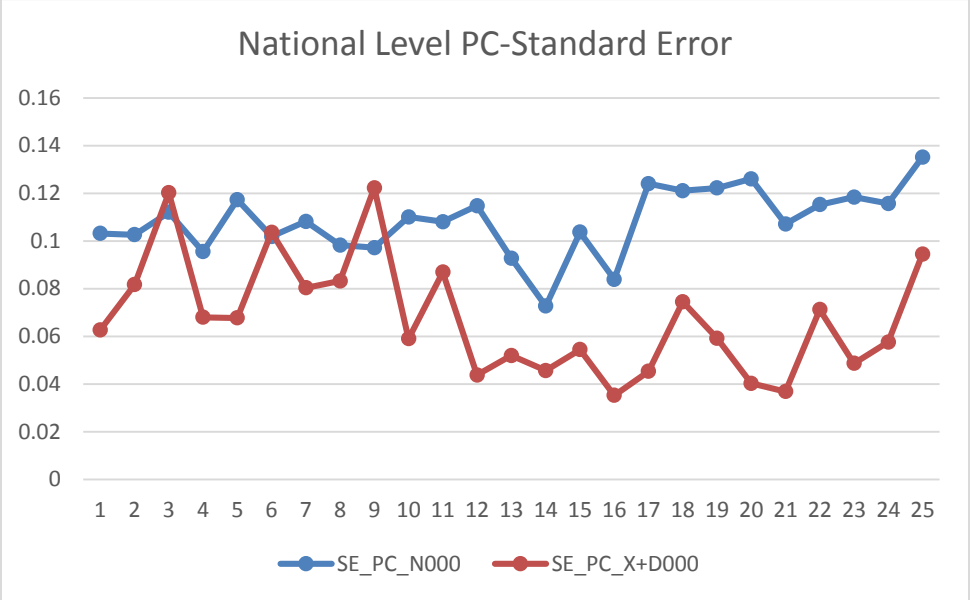
Everitt, Landau, Leesem and Stahl. 2011. Cluster Analysis, 5th Edition. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK. Wiley Publishing, pp. 113 – 116.

Appendix

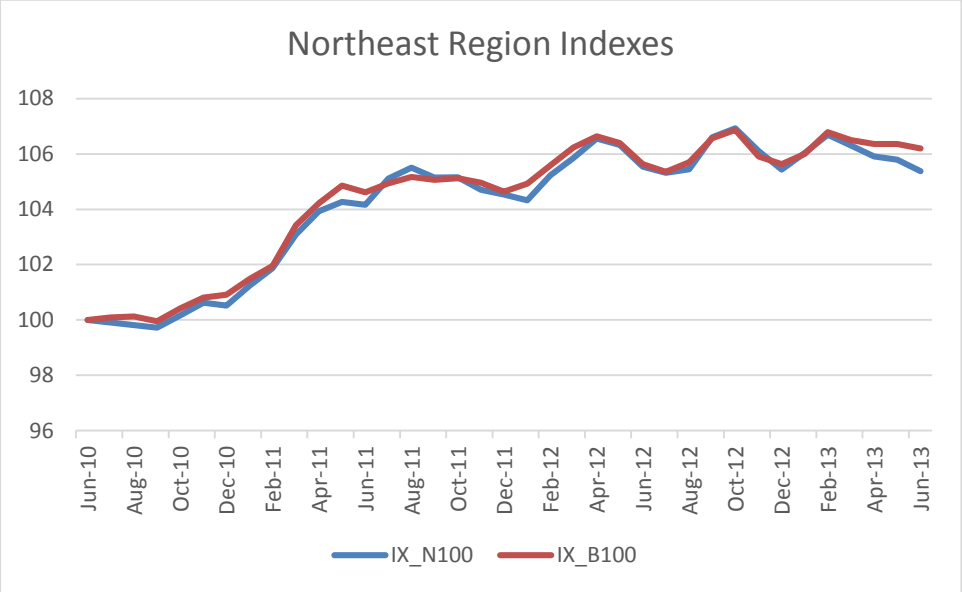
National Level, denoted area 000

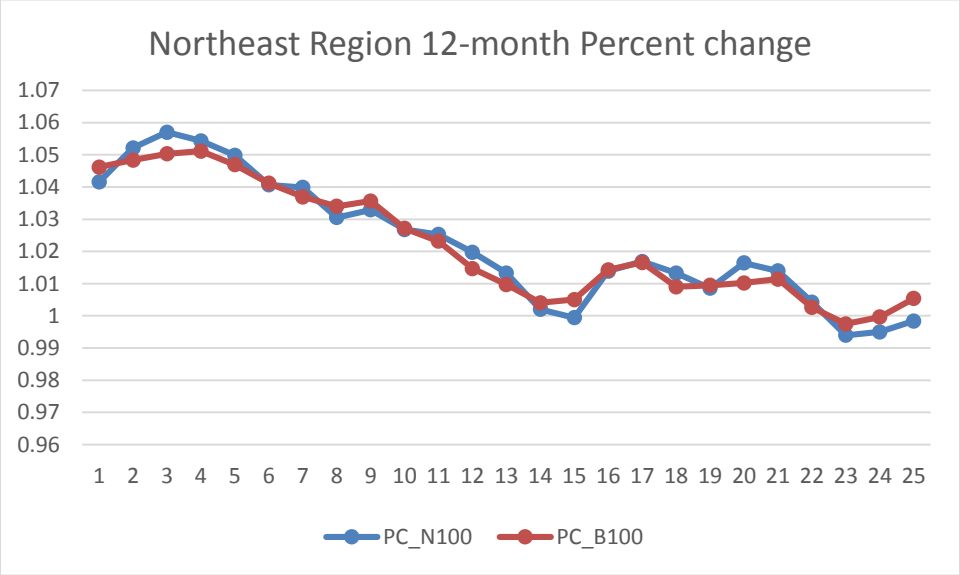


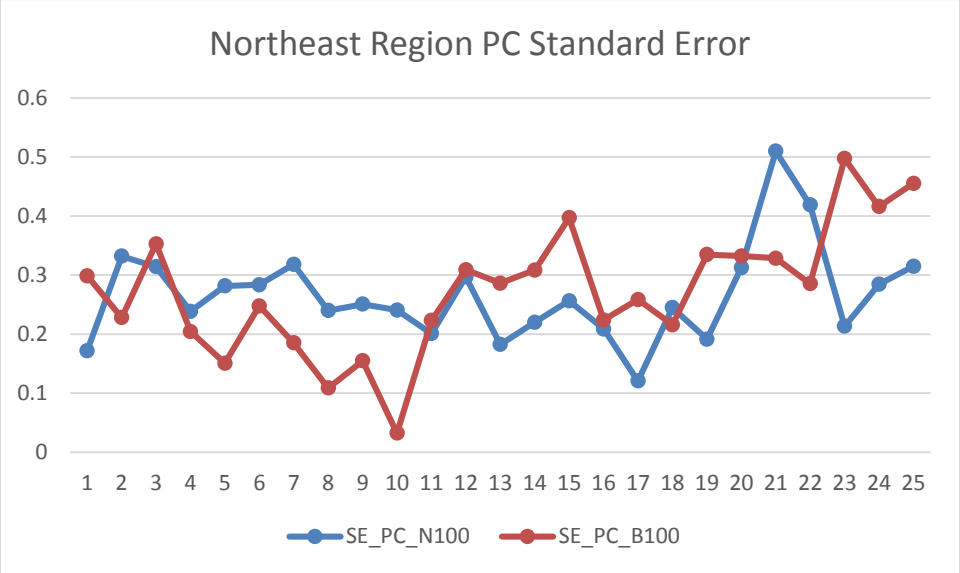



National Level PC-Standard Error
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
SE_PC_N000
SE_PC_X+D000

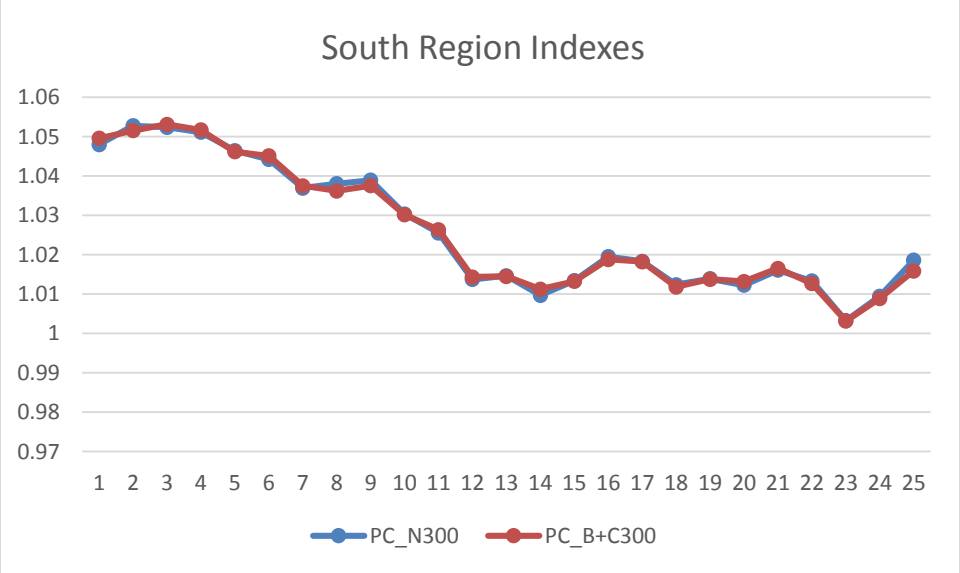
Census Region 1, Northeast. (note: there are no small C-sized PSU in Northeast)

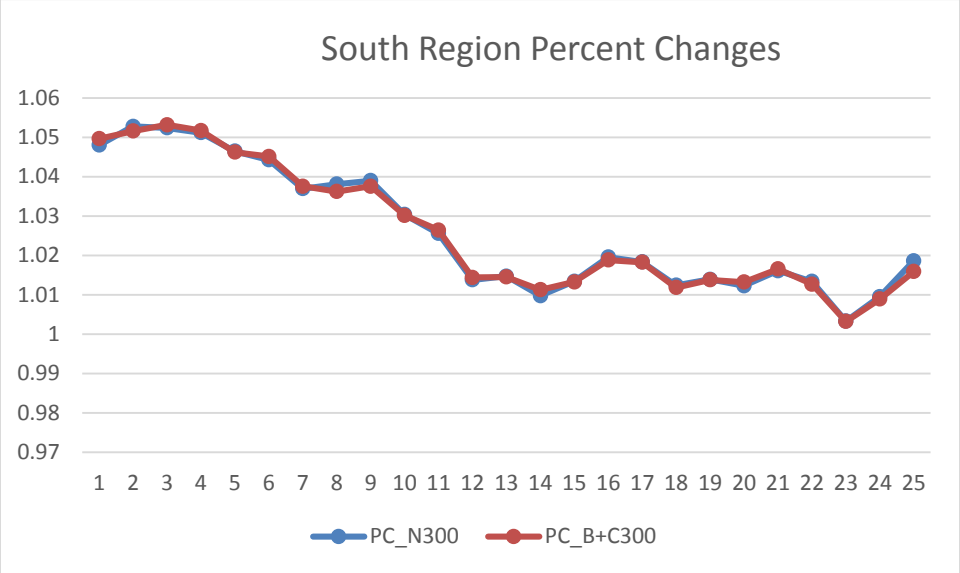


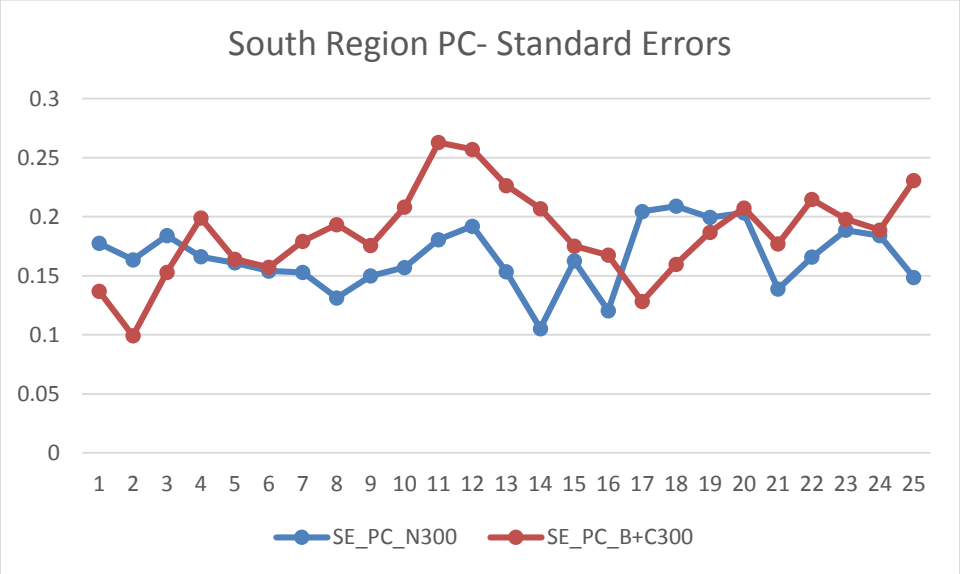




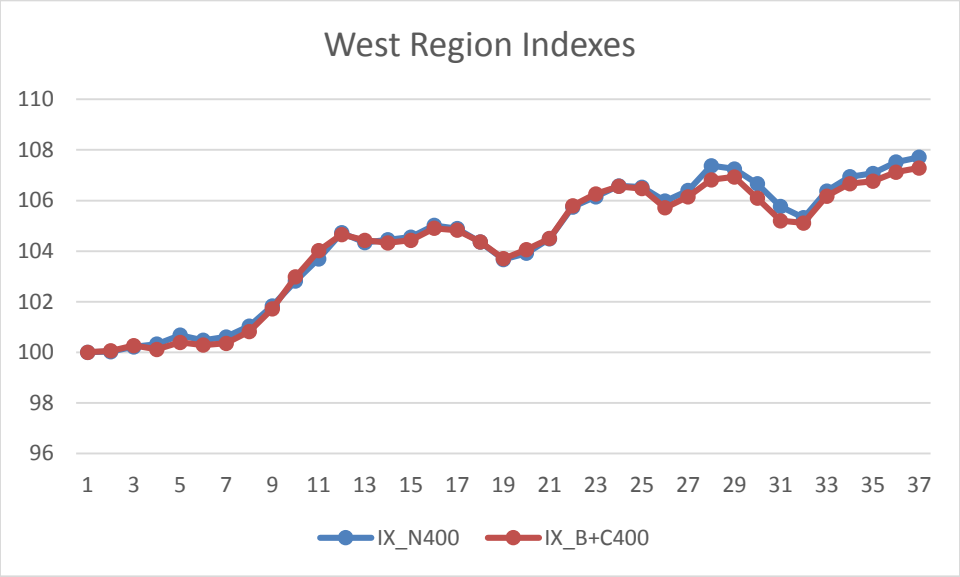
Census Region 3, South

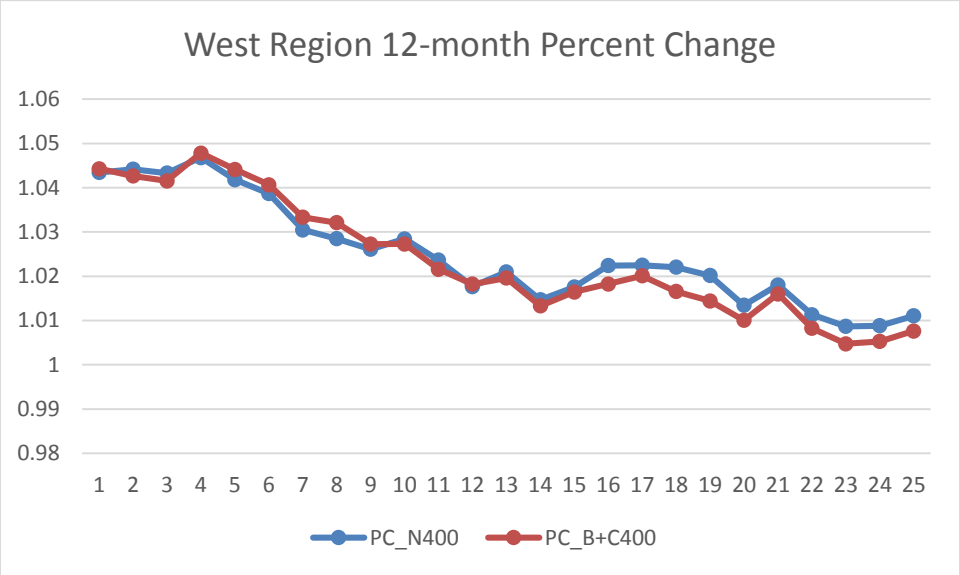


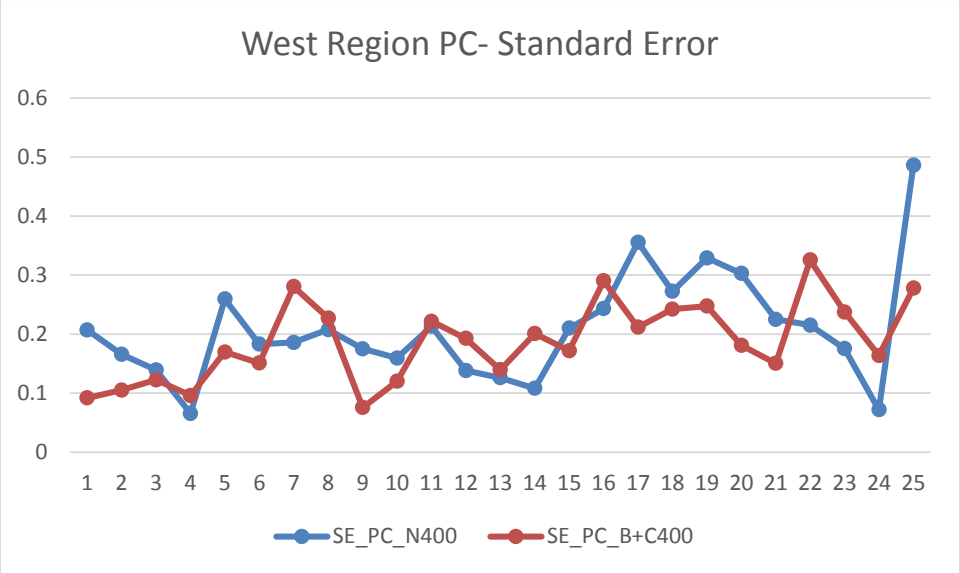

South Region Percent Changes
1.06
1.05
1.04
1.03
1.02
1.01
1
0.99
0.98
0.97
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
PC_N300
PC_B+C300


South Region PC- Standard Errors
0.3
0.25
0.2
0.15
0.1
0.05
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
SE_PC_N300
SE_PC_B+C300

Census Region 4, West






West Region PC- Standard Error
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
SE_PC_N400
SE_PC_B+C400