The Impact of High Variances at the Lowest Aggregate Levels on the CPI's All-US–All-Items Variance October 2009

Owen J. Shoemaker
U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Room 3655
Washington, DC 20212
shoemaker_o@bls.gov

Abstract

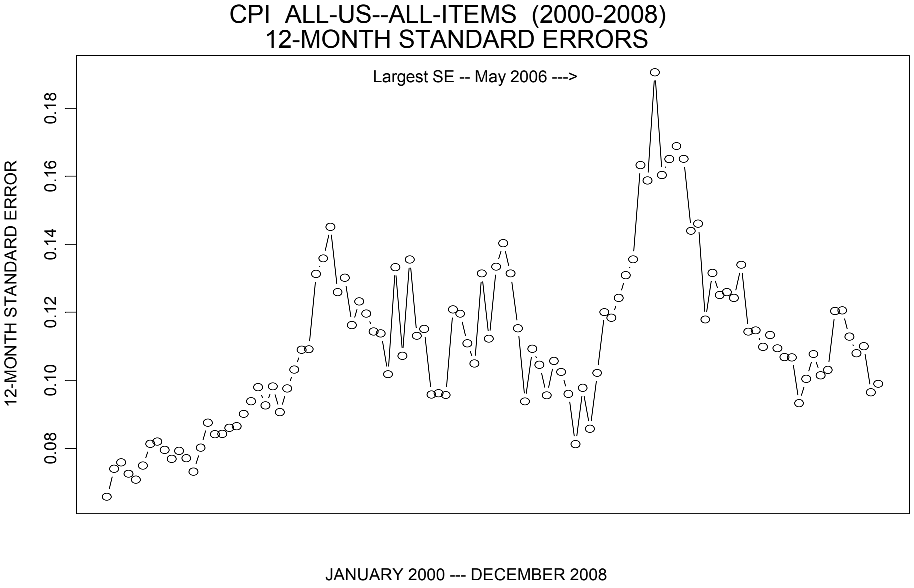
In 2006, the CPI's All-US–All-Items 12-month standard errors increased by more than 50% over the previous year's median average, returning to regular pre-2006 levels in 2007 and 2008. Since overall sample size had not been appreciably reduced in the 2006 time period, the analysis had to look elsewhere for an explanation for this rather significant rise in the All-US–All-Items 12-month standard errors. Perhaps one or more of the individual (replicate) variance pieces was contributing an inordinately high amount of variance to the overall variance. A decomposition analysis of the Stratified Random Groups variance calculation system was produced, and the results showed one or two major contributors at the lowest aggregate level producing as much as half of the entire All-US–All-Items variance. This paper will investigate the nature and genesis of these anomalies, their impact on the overall CPI variance, and then compare how different variance methodologies would have handled these anomalies.

Key Words: Stratified Random Groups

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

1. Introduction.

One of the CPI program's main performance goals is to produce a 12-Month All-US–All-Items percent price change with a standard error no greater than 0.25. Generally, since the 1997 Revision and the concomitant implementation of a Stratified Random Groups variance estimation system, this standard has been easily met each month. However, in 2006, the CPI's All-US–All-Items 12-month standard errors increased by more than 50% over the previous year's median average, with one anomalous month (May '06) producing a 12-month standard error of 0.19. Through 2007 and 2008 these 12-month standard errors returned to their pre-2006 levels. The graph below tracks the 12-month standard errors for the 108 months from Jan '00 through Dec '08.



In the first of these two years, 2000-2001, the standard errors held below 0.09. For the next four years, 2002-2005, the standard errors rose to a level that stayed below 0.12. After the substantial rise in 2006, the 12-month standard errors leveled below the 0.12 range. In a simpler tabular form, we can see these annual median 12-month standard errors as they progressed from 2000 through 2008.

ANNUAL MEDIAN 12-MONTH STANDARD ERRORS (ALL-US–ALL-ITEMS)

2000	2001	2002	2003	2004	2005	2006	2007	2008
0.0764	0.0871	0.1197	0.1134	0.1174	0.1023	0.1595	0.1145	0.1054

2. The Variance Decomposition

In order to determine more exactly where the new additional variance in 2006 was emanating from, we decomposed the total variance down into its 573 elemental pieces, extracting the individual pieces from the CPI's official variance estimator:

$$Var(A,I,t,t-k) = \sum_{a\in S_A} \frac{1}{N_a(N_a-1)} \sum_{i\in I}\sum_{r\in R_a} \left(PC(a,i,r,t,t-k) - PC(A,I,t,t-k)\right)^2$$
$$+ \sum_{a\in N_A} \frac{1}{N_a(N_a-1)} \sum_{r\in R_a} \left(PC(a,r,t,t-k) - PC(A,I,t,t-k)\right)^2$$

where $r \epsilon R_a$ refers to the set of replicates (REP > 0) in AREA = *a*, $i \epsilon I$ refers to the set of intermediate item aggregates in ITEM = *I*, N_a is the number of variance replicates in AREA = *a*, S_A is the set of self-representing index areas in AREA = *a*, and N_A is the set of non-self-representing index areas in AREA = *A*.

The CPI All-US–All-Items variance is calculated using a Stratified Random Groups method. Each AREA (broken down further by 8 Major Item Groups in the Self-Representing, or “A”, AREAS) is a random group, which consists of at least two replicates. All the “A” AREAS have two replicates, except for NYC, Chicago, and LA County, each of which have four. The four “X” AREAS have at least four replicates per AREA and the three current “D” AREAS have two apiece. (The “A” AREAS are the larger US cities, the “X” AREAS are groups of medium-sized US cities, and the “D” AREAS are groups of smaller-sized US cities.) The total number of replicate random groups comes out to be 573, with each of these 573 least-squares’ calculations factor-adjusted by the number of replicates in the random group. Each piece is the squared difference between the full-sample percent price change value and a replicate percent price change value, adjusted by a 1 /(N * (N–1)) factor, with N being the number of replicates. Each piece is then 1 / (N * (N–1)) • [$PC_{REP} – PC_{FullSample}$]^2, with the sum of these pieces equaling the total variance. Thus, a simple decomposition of the variance calculation down to its constituent replicate-level parts will give us a list of variance pieces which we can sort by their contributing amount of variance. (Note, PC_t = (CW_t / CW_{t-12} – 1) * 100. PC is a Price Change, CW is a Cost Weight, with CW = $INDEX_t$ * AGGREGATE_WEIGHT.)

In our analysis we will be concentrating our attention on just the three years of 2005, 2006 and 2007, mainly because our target month is May 2006. We first want to view some representative decompositions, including of course May 2006, and then try to explain why this surprisingly high May 2006 variance occurred and finally look at how some other variance methodologies might have handled the situation.

The table on the following page gives a representative sampling of top ten variance replicate values, including (on purpose) the May ‘06 set of values, where the single highest variance replicate value (A111-Housing) eats up almost 49% of the total All-US–All-Items variance, with all the rest of the 572 variance replicate values contributing the remaining 51%. If we throw in the second replicate value from A111-Housing (4.9%) we find that A111-Housing is carrying a full 53.6% of the total variance. Clearly something anomalous is going in the Housing sector in A111. (A111 is the New Jersey suburbs of NYC. To better understand the naming conventions, note that the second character in an AREA name stands for its region: “1” = East, “2” = Midwest, “3” = South, “4” = West. For instance, A433 is Denver and X300 is the X-sized cities in the South.)

When we isolate the May ’06 A111-Housing variance replicate value, we first note that its actual value is more than three times greater than the largest variance replicate value anywhere else in these decomposition tables. The Housing major group is by far the largest major group, maintaining a relative importance in the All-US–All-Items CPI of 40%+. This major group includes both Rent and Owner’s Equivalent Rent (REQ) in it, but as it turns out nearly the entirety of the May ’06 A111-Housing’s variance contribution comes from a much smaller Item-Stratum, Other Lodging Away From Home (SEHB02, which is commonly referred to as Hotels & Motels).

Top Ten Variance Replicate Values (a selection)

(All-US—All-Items 12-Month Price Change Variances)

200505

AREA GROUP	REP	VAR	PCT
A111-Housing	2	0.00117	12.7
X100	2	0.00052	5.6
A433-Housing	1	0.00038	4.1
A316-Apparel	1	0.00032	3.5
A111-Housing	1	0.00026	2.8
X200	3	0.00023	2.5
A110-Apparel	1	0.00020	2.2
A320-Apparel	1	0.00019	2.1
X300	3	0.00018	2.0
A318-Apparel	1	0.00017	1.9
TOT	**SE =**	**0.09601**	

200511

AREA GROUP	REP	VAR	PCT
A111-Housing	1	0.00379	27.0
A109-Apparel	2	0.00134	9.6
X300	4	0.00109	7.8
X499	3	0.00097	7.0
X300	6	0.00050	3.6
A103-Housing	2	0.00046	3.2
X300	2	0.00036	2.5
X300	1	0.00034	2.4
X100	1	0.00028	2.0
X100	3	0.00022	1.6
TOT	**SE =**	**0.11839**	

200605

AREA GROUP	REP	VAR	PCT
A111-Housing	1	0.01762	48.6
A109-Apparel	2	0.00518	14.3
A111-Housing	2	0.00178	4.9
X300	4	0.00135	3.7
X300	1	0.00098	2.7
A110-Housing	2	0.00088	2.4
X100	2	0.00077	2.1
A316-Apparel	1	0.00062	1.7
X300	8	0.00039	1.1
X100	1	0.00030	0.8
TOT	**SE =**	**0.19046**	

200611

AREA GROUP	REP	VAR	PCT
A109-Apparel	2	0.00598	28.0
A111-Housing	2	0.00213	10.0
X300	1	0.00119	5.6
X499	1	0.00098	4.6
X300	4	0.00087	4.1
A433-Housing	1	0.00083	3.9
X499	3	0.00079	3.7
D300	2	0.00064	3.0
A103-Housing	2	0.00047	2.2
X300	11	0.00047	2.2
TOT	**SE =**	**0.14601**	

200705

AREA GROUP	REP	VAR	PCT
A111-Housing	2	0.00526	29.3
A111-Housing	1	0.00268	14.9
A109-Apparel	2	0.00116	6.5
A420-Housing	1	0.00070	3.9
A420-Housing	2	0.00057	3.2
A433-Housing	1	0.00040	2.2
X499	3	0.00038	2.1
X200	5	0.00035	1.9
A102-Housing	2	0.00026	1.4
X300	1	0.00022	1.2
TOT	**SE =**	**0.13394**	

200711

AREA GROUP	REP	VAR	PCT
A420-Housing	1	0.00125	11.0
X300	4	0.00125	10.9
A420-Housing	2	0.00118	10.3
X300	6	0.00103	9.0
X300	11	0.00073	6.4
X499	1	0.00052	4.6
X200	3	0.00025	2.2
D300	1	0.00025	2.2
A102-Apparel	2	0.00019	1.7
A312-Apparel	1	0.00016	1.4
TOT	**SE =**	**0.10679**	

We performed an exercise on the entire cost weight structure, in which we zeroed out the variance contribution of this one A111-SEHB02 cell and its two replicate variance values. The reduction in variance on the All-US–All-Items total was 52%, which roughly matches the 53.6% contribution of the cell when it is included. The culprit here was not Housing in New Jersey in general, but specifically Hotels & Motels in New Jersey-NYC. In May '06 the one relatively small cell of A111-SEHB02 was causing a 0.19 all-time high standard error for the All-US–All-Items CPI. When this cell gets zeroed out, the May '06 CPI standard error reduces to 0.13.

3. Sampling History of A111-SEHD02

The New Jersey-NYC–Hotels & Motels cell (A111-SEHD02) has an unfortunate sampling history in this three year time frame (2005–2007). The CPI program replenishes the entire C&S (Commodities and Services, less Rent and REQ) set of ITEMS and AREAS every four years, by rotating in 1/8 of the ITEMS and AREAS every six months. When A111-SEHB02 was rotated in 2001, it was scheduled to be refreshed completely in 2005. However, a newly revised rotation matrix and schedule delayed that next rotation date until 2007. So, A111-SEHB02 had to hold onto its sample a full two years longer than the ideal rotation system would have called for. Moreover, and more crucially, the Feb '01 new Sample Design allocated a much smaller number of Outlet and Item hits to A111-SEHB02 than it had gotten previously, due to a misplaced notion at the time that SEHB02 would not be carrying the weight for Vacation Home Rentals as well as for Hotels & Motels. This substantially underweighted the A111-SEHB02 cell for the Feb '01 Sample Design and resulted in an allocation of only 6 Outlets with only 3 Quotes per Outlet. Then, after initiation of these outlets and quotes, two of these outlets were not included in the new sample and so for the next 6 years, A111-SEHB02 consisted of only 4 outlets with 3 quotes apiece, for a grand total of 12 quotes for the entire cell. Replicate 1 consisted of 2 of these outlets, Replicate 2 included the other 2 outlets. By comparison, the previous sample size for the entire A111-SEHB02 cell was 40 quotes, with 20 apiece for each its two replicates. (When the later Feb 2007 Sample Design was finally introduced in late 2007, the highly augmented sample size was set to over 80 quotes. Note that A111-Housing drops out of the Top Ten in the 200711 list.)

Further complicating matters, in Replicate 1 of A111-SEHB02, one of its two outlets dropped out of the sample during more than half the recorded months of its existence and was taken out of the sample altogether in early 2006, leaving Replicate 1 with only one outlet and with only three quotes. This sample-weakened replicate was clearly the main reason that this cell was causing the variance problems attributed to it.

One other complication and contributing factor to higher variability in this cell was the fact that its four outlets represented seasonally different motels. Replicate 1 consisted of a motel in Florida, with its unsteady partner being a motel in Upstate New York. The former was seasonally higher-priced during the winter months, while the latter was seasonally higher-priced in the summer months --- in fact, only available for pricing at all during the summer months. By contrast Replicate 2 consisted of a high summer rental motel in Maryland, along with an odd spring- and fall-seasonal rental hotel in Washington, DC. Nothing *but* highly contrasting ups and downs in pricing could occur in this volatile and low-sample mix.

A final contributor to the elevated variance in A111-SEHB02 was the fact that not only did this cell have a relatively large relative importance, but the two sets of quote-level weights for the two replicates were at a ratio of 2 to 1, which then rose even more dramatically (4 to 1) when the Lake George outlet went away altogether in early 2006.

All in all, just about everything bad, from a sampling, representative, weighting and seasonality point of view that could be present for causing higher variances did in fact occur in this case. To summarize:

- An under-weighted sample was initially allocated to A111-SEHB02 in 2001
- The sample size in the cell dropped from 40 to 12 (and often to 9)
- The volatile mixture of a cell with a relatively large relative importance with a set of very seasonal quotes and a too small sample size
- Higher quote-level weights being assigned to the replicate with the fewer (and more volatile) number of quotes

4. Variance Comparison

BLS computes standard errors down to the smallest ITEM–AREA cell, so we can also directly compare the 12-month standard errors for the A111-SEHB02 cell with All-US–All-Items around this crucial May '06 time period.

12-Month Standard Errors --- All-US–All-Items vs. NJ–Lodging

AREA-ITEM	200602	200603	200604	200605	200606	200607	200608
All-US–SA0	0.1355	0.1632	0.1587	**0.1905**	0.1603	0.1688	0.1651
A111–Hotels	34.02	49.49	54.99	**127.88**	64.93	84.75	62.19

Correlation = 0.891

The correlation between these two series for the entire year of 2006 was a somewhat lower 0.727, but the strength of the point remains. These very high variances at the lowest aggregate level are highly correlated with the All-US–All-Items results. Moreover, our main culprit, A111-SEHB02 in May '06, is indeed nearly the largest of all the 8,018 lowest cell level 12-month standard errors. Only two Apparel cells have slightly larger 12-month standard errors in May '06. (Girls' Apparel in Atlanta is tops at 133.45, and Women's Dresses in the D-sized West region is 129.16. But NJ-Lodging's relative importance is, respectively, 40 and 150 times bigger than these two higher variance cells. With a relative importance of 0.15%, the A111-SEHB02 cell's percentage variance contribution ends up more than ten times greater than either of these two cells.)

5. Three Other Variance Methods

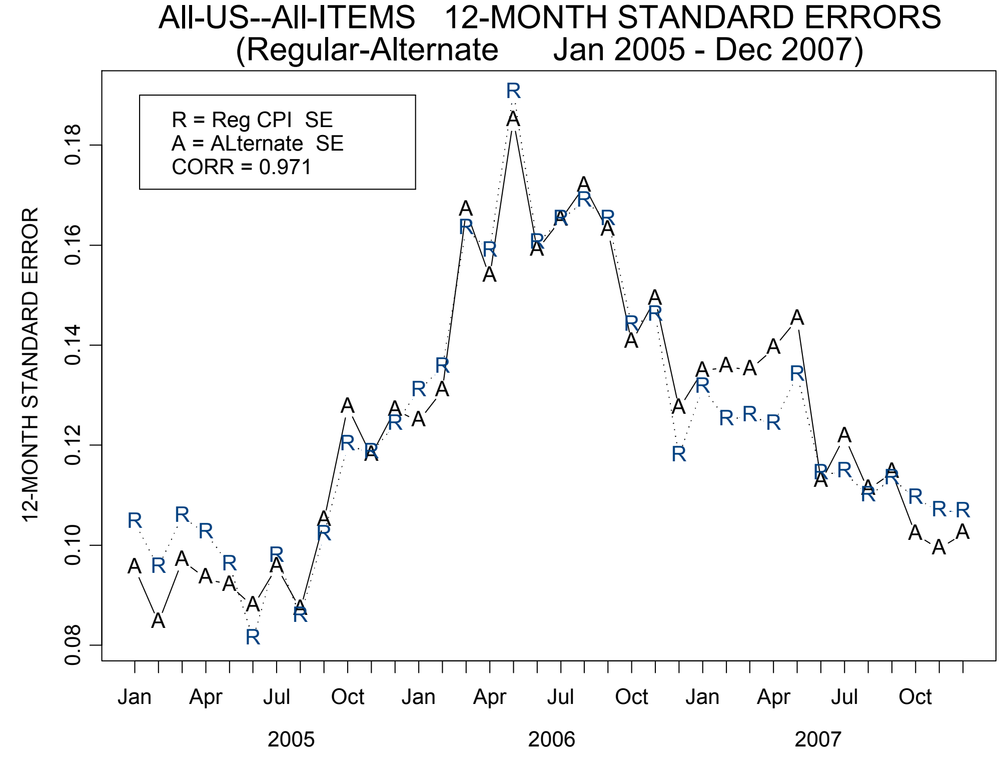
Having demonstrated that a single high variance AREA-ITEM cell (in this case, A111-SEHB02) can ramify up to the highest aggregate level (All-US–All-Items) rather dramatically and singularly, we would now like to investigate a few other variance methodologies in order to see how they might handle the impact of one of these lower level cells in the CPI on the upper level variance. The three methods we will explore are (1) CPI's own Alternate Stratified Random Group Method, which does *not* break out the random groups into Major Group by AREA categories, but uses all appropriate replicate values in each AREA; (2) a Stratified Jackknife Method, which uses the 38 Index AREAS as its strata to jackknife, but none of the replicate values; and (3) the Regular

Bootstrap Method, which, like the Jackknife uses only the 38 Index AREAS to produce its bootstrap estimates. All three of these other variance methods are viable methods for estimating CPI variances, and all three work from the same full-sample price changes (here only the 12-month price changes) and the same set of cost weights, albeit in different combinations.

By not breaking out the Stratified Random Group structuring, the Alternate SRG Method uses 97 replicates in its All-US–All-Items variance calculation as compared with CPI's current SRG method which utilizes 573 replicate categories.

$$Var_{alt}(A,I,t,t-k) = \sum_{a \in A} \frac{1}{N_a(N_a-1)} \sum_{r \in R_a} \left(PC\left(a,r,t,t-k\right) - PC\left(A,I,t,t-k\right)\right)^2$$

The Alternate SRG Method uses perforce the same 1-, 2-, 6- and 12-month price changes as CPI's current SRG Method (as will Jackknife and Bootstrap as well). The following graph tracks the two sets of 12-month standard errors for the 36 months from Jan '05 to Dec '07, with our particular attention drawn to the comparison at May '06.



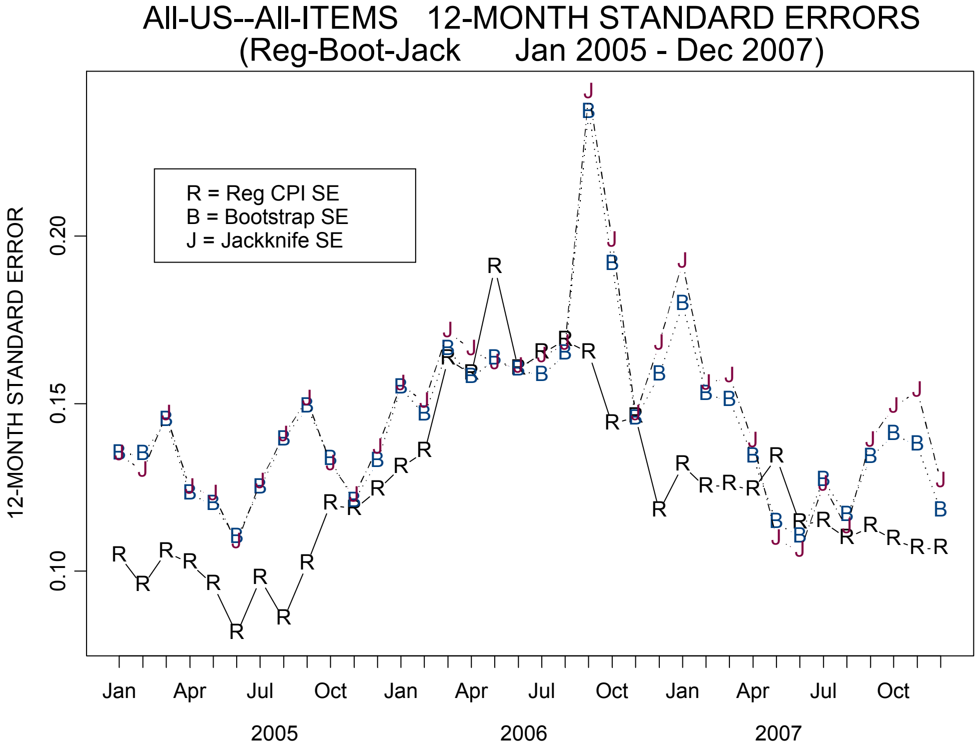
These two SRG methods seem to be producing very similar results. Note the very high correlation between these 36 months of 12-month standard errors: 0.971. Even the mean standard errors of the two methods are equal over these 36 months: Mean (CPI SRG) = 0.125 and Mean (Alt SRG) = 0.125. A paired comparison t-test between these two sets of standard errors produces a non-significant difference between the two sets (P-value = 0.872). At the month of interest (May '06), the Alternate SRG Method does produce a lower (3% smaller) standard error, but the difference is slight even negligible. This Alternate SRG Method certainly seems to be doing as good a job estimating CPI variances at the All-US–All-Items level, but it does not seem to be any less immune to the ramifications of a lower cell's very high variance impacting the highest level standard error.

The Bootstrap and Jackknife Methods provide a greater contrast for our study. Both bootstrap and jackknife use only the 38 full-sample Index AREA cost weight values to compute their variances and standard errors. The jackknife formula is

$$Var_{JK}[PC(A,I,t,t-k)] = \frac{N_A - 1}{N_A} \sum_{a \in A} \left(PC(A-a,I,t,t-k) - PC(A,I,t,t-k) \right)^2$$

where each PC(A–a, I, t, t–k) omits the ath t and t–k cost weights from the respective sums of full-sample cost weights, and then calculates the percent price change just as done in the SRG methods.

The bootstrap resamples these same 38 AREAS (i.e., the 38 row vectors of cost weights) applying 4000 resamplings for each time period. Instead of individual values of a vector being resampled, here the row vectors of a matrix of cost weights are resampled. Each bootstrap resample is then used to compute a percent price change, just as it is done in the SRG methods ($PC_t = (CW_t / CW_{t-12} - 1) * 100$). Each bootstrap standard error is the standard deviation of 4000 bootstrap percent price change resampling results.



The bootstrap and jackknife are producing nearly identical results month by month. Correlation between the bootstrap and jackknife values is 0.99, with the jackknife producing just slightly higher standard errors. Mean (Jack) = 0.147 while Mean (Boot) = 0.144. Compare those means with the SRG means of 0.125. The correlation between Bootstrap and SRG is 0.65, and the correlation between Jackknife and SRG is 0.62. At our point of interest, May ’06, both bootstrap and jackknife provide a significantly lower standard error than SRG. Bootstrap is 14.3% lower and Jackknife is 15% lower. However, as well as bootstrap and jackknife may be handling the lower-level variance issue in May ’06, in several other months, both bootstrap and jackknife are spiking dramatically higher than SRG (particularly in Sep ’06 and Oct ’06). Both bootstrap and jackknife are legitimate and serious candidates to be the CPI variance method of record. Their standard errors here are running 15% to 18% higher than SRG, but their own variability is as high as SRG’s. The bootstrap standard errors are just slightly *less* variable than the SRG standard errors, and the jackknife standards errors are just slightly *higher* then SRG. However, the tendency to spike seems to be just as strong or stronger in the bootstrap and jackknife methods, even though both do a better job of handling the standard error of interest at May ’06.

6. Conclusions

- High variability in just one of the lowest aggregate CPI cells can ramify up the aggregate structure to dramatically impact the All-US–All-Items variance itself.
- A near “perfect storm” of circumstances combined to produce a very low sample size and a very high variance in the May ’06 New Jersey-NYC–Hotels & Motels cell.
- The Alternate SRG Method provides a legitimate and worthy variance methodology to adopt if necessary, but it is also unable to mitigate the impact of a singularly high lower-level variance.
- Bootstrap and Jackknife provide legitimate and worthy alternatives to our current SRG method. Both were able to substantially mitigate the spiking effects of our target time (May ’06), but then both also produce even higher spikes of their own elsewhere in year. Bootstrap and Jackknife produce consistently higher standard errors, but their results are no more or less variable than SRG.