The Use of Geocoding to Find Location of Outlets Outside of Sample Area Boundaries and to Determine Significant Areas of Commerce

December 2006

John Schilp and Fred Marsh III
U.S. Bureau of Labor Statistics, 2 Mass Ave., NE Room 3655, Washington, DC 20212
Schilp.John@bls.gov Marsh.Fred@bls.gov

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

Abstract:

This paper gives an introduction to the method of sampling retail outlets in the U.S. Consumer Price Index (CPI). Some of these retail outlets are located outside the primary sampling area boundary and not priced. The Office of Field Operations at the Bureau of Labor Statistics (BLS) has established a 25-mile/10 quote rule to reduce data collection costs. The 25-mile/10 quote rule is an established collection procedure that says BLS will price any outlet that is less than 25 miles from the primary sample area boundary or has more than 10 quotes in a cluster. This paper will examine the effects of eliminating these groups of outlets from the sample on the CPI index. This paper will also explain our definition of a spatial cluster and the new "significant area of commerce" as they pertain to this project.

Keywords:

Geocoding, Sampling, Index Calculation

1. Introduction:

This paper presents the results of an assignment by the Bureau of Labor Statistics' Office of Field Operations (OFO) to help reduce travel costs accrued while collecting data outside the PSU boundary. As it stands now, field economists are required to sometimes travel great distances to perform pricing for an outlet contained in the sample. OFO would like to limit the amount of traveling done by their field economists by eliminating, from our sample, outlets that are outside the PSU. However, we must learn how this will impact the CPI before we can eliminate these sample units.

In this project we considered three options for eliminating sample.
1) The current 25-mile/10 quote rule that says BLS will price any outlet in the sample that is less than 25 miles from primary sample area boundary or has 10 or more quotes in a cluster of outlets. (We give an operational definition of "cluster" in section 3.1 below.).
2) To eliminate all outlets not contained within the PSU boundary.
3) To eliminate any outlet outside the primary sample area that is not contained in a cluster of ten quotes or more regardless of the distance. This is the "significant area of commerce" option.

Each of these three options impacts the CPI indexes to different degrees.
1) The 25-mile/10 quote rule is vague, particularly with regards to the definition of a cluster. It has a moderate impact on the Indexes while eliminating a reasonable number of outlets that are not cost efficient to collect.
2) The second option of removing every outlet outside the PSU eliminates the most outlets and has the greatest impact on the CPI.
3) The significant area of commerce (SAC) option eliminates a reasonable number of outlets yet has moderate impact on the CPI.

For now, a cluster is a group of outlets that are contained in close proximity, e.g. a zip code area. A Significant Area of Commerce is a cluster of outlets that contain a considerable number of price quotes; specifics are given in section 3.

1.1 The Consumer Price Index

The Consumer Price Index is a measure of the average change over time in the prices of consumer items — goods and services that people buy for day-to-day living. The Consumer Price Index is broken down into several indexes including the All-USA, All-items index. There is an all items index for each of the four regions, as well as for each of the 28 self-representing PSUs, each major group, PSU combination etc. In all there are 12,926 lower level indexes that make up the main All USA-All items CPI. While the three adjustment scenarios above do not significantly impact the higher level indexthere are some changes to the lower level indexes.

1.2 Telephone Point of Purchase Survey

The Telephone Point of Purchase Survey (TPOPS) is conducted by the Census Bureau for BLS to obtain a sufficient outlet frame for consumer commodities and services that are to be priced in the CPI. In TPOPS the consumers from our primary sampling areas are asked via telephone where they purchased specific goods and services within given recall periods. Obviously a consumer is not limited to reporting only items they purchased inside the PSU definition. Sometimes they travel great distances to shop. BLS still needs to capture these outlets in the sample because the CPI focuses on the shopping habits of people inside the sampling area. This is a guiding philosophy of the CPI.

2. Statement of Problem

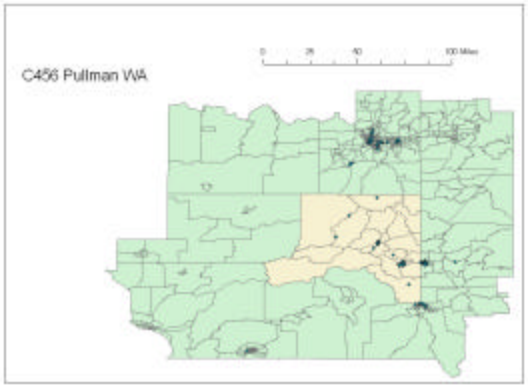
The existing 25 mile/10 quote policy was never properly enforced. This was due to a vague definition of a

cluster. OFO would like to solidify a set of rules to eliminate sample that does not satisfy distance and cluster size requirements. There are three scenarios that we examined to rectify this problem. 1) Eliminate sampled outlets that do not satisfy the 25 mile/10 quote rule 2) Eliminate all sampled outlets that are not located inside the PSU 3) eliminate everything that is not in a cluster of 10 or more quotes regardless of the distance from the PSU boundary. A cluster that contains 10 or more quotes will be called a significant area of commerce or SAC. OFO economic assistants will travel to theses SACs to collect prices from the sample.

3. Geocoding to Find Clusters

Geocoding is a process used to find the longitude and latitude location of an outlet or housing unit after being given its address and zip code. When we originally geocoded the set of outlets we tried to match the address to a TIGER data set from Census. TIGER data set is the first attempt from Census to map all of the streets and housing unit numbers in the US. These maps produce very accurate locations for the outlets. However, this produced a match rate of approximately 65%. This was not good enough because 35% of the sample would have an unknown location and not be counted in a cluster definition. On the next attempt matches were determined by zip code location, lowering the accuracy of the outlets' locations but producing a match rate of 97%. We therefore adopted zip code centroids as an important element in defining clusters.

Once the Outlets were geocoded and put onto maps they had to be looked at to determine what would constitute a cluster and more importantly a "Significant Area of Commerce." Some areas were easy to identify as a SAC. Consider Pullman, WA, C456 as an example. There are a large percentage of outlets for this PSU that are not contained inside the PSU boundary. Spokane to the north contains 36 Outlets and 154 quotes, and Moscow, ID to



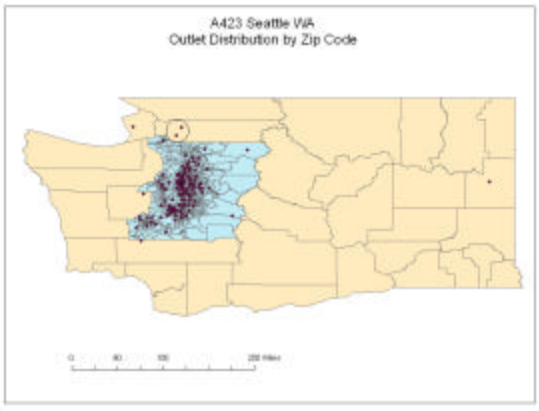
the east contains 28 Outlets and 348 quotes. These areas will surely be considered SACs for Pullma and will remain in the sample. However, Walla Walla WA, to the south west is 33 miles from the PSU boundary and contains only 1 outlet with 1 quote. This will most likely be dropped, but it is important to determine if not pricing this quote would have a significant impact on the CPI.

3.1 Definition of a Cluster/SAC

1) Any single outlet outside the PSU is considered a cluster unto itself.
2) Two or more outlets, outside the PSU, in the same zip code are considered a cluster.
3) Two outlets that are in the same town, within 10 miles of each other are considered in the same cluster. For three or more outlets in the same town the distance will be measured from the farthest two but the distance will be set at less than 20 miles to be a cluster.
4) Two addresses that maybe in different towns that are within 10 miles of each other are a cluster though they have different zips. For three or more outlets the distance will be measured from the farthest two but the distance is set at 20 miles.

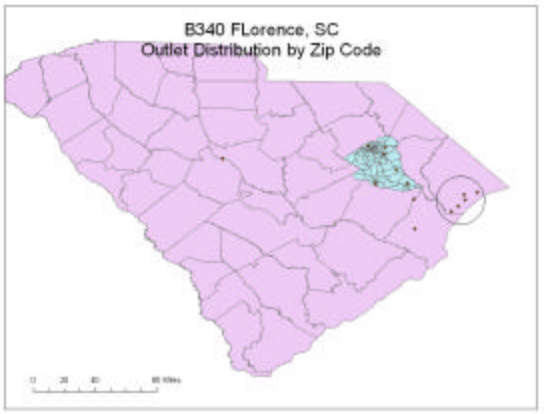
A significant area of commerce (SAC) is any cluster that contains more than 10 quotes regardless of number of outlets.

3.2 Examples



The map above shows two circled outlets. These outlets are in different towns but they are close enough together to be considered a cluster. However, they have only 8 quotes between them so this cluster is not considered a SAC.

Conversely, the following map shows a large cluster that is considered to be a SAC. The outlets are in Myrtle Beach, SC and Garden City, SC. though not all in the same zip code. The distance between the two farthest points is 18 miles thus it is defined as a cluster. There are 10 outlets and 23 quotes in this cluster so it is considered a SAC.



All 87 PSUs were looked at in this way to determine significant areas of commerce.

4. Impact on the Indexes

As mentioned above, it is necessary to determine how these deletions will impact the various indexes once it is determined which outlets are candidates for removal from the sample.

To determine if something is being significantly impacted we will have to look at correction triggers to see if they are violated. We do not use standard errors (SE) and t-tests to tell if indexes are significantly different primarily because the SEs are difficult to calculate. The difficulty arises in the replicate structure, the use of imputation and in the substitution of quotes. Instead, correction triggers are established thresholds that if violated in normal production, these indexes will be examined for possible errors.

BLS looks at the 12-month price relative and not index values since the magnitude of an index is somewhat unimportant. A price relative is the index at time t +12 divided by the index at time t, minus 1. Essentially it is the percent change over the 12 month period of time.

$$PRC = \frac{IX_{t+12}}{IX_t} - 1$$

We were able to determine the production indexes and the reduced sample indexes by using a generic index calculator that was developed by BLS economist Craig Brown. This program produces two indexes for each aggregate index. They are called production (prod) and reduced. If the ratio of these two indexes is larger than a correction trigger a flag is set indicating that eliminating sample significantly impacts that aggregate index.

$$D = \frac{PRC_{reduced} - PRC_{\text{Prod}}}{PRC_{prod}}$$

There are three levels of correction triggers. There are 94 all items aggregate indexes with a correction trigger of .2%, 3402 major group and sub-group aggregate indexes with a correction trigger of .6% and 9430 item level aggregate indexes with a correction trigger of 1%.

If the absolute value of D is larger than the correction trigger for that Index area, it will be judged significantly impacted.

4.1 Results

The scenario of removing everything outside the PSU has the largest impact on the set of aggregate indexes. It eliminates 1,065 of 43,500 outlets (2.4%) and 3,994 of 115,420 quotes (3.4%). There were 1,072 aggregate indexes that violated their correction trigger. This is 8.29% of all indexes. This scenario will save the most fuel and time resources, but it impacts too many indexes.

The second scenario of keeping the 25-mile/10-quote rule in place has the least impact on the aggregate indexes. It eliminates 91 of 43,500 outlets (0.2%) and 211 of 115,420 quotes (0.2%). It affects only 7 of the 12,926 aggregate indexes (.05% of all aggregate indexes). However, this saves the least amount of time and fuel resources.

The option of determining significant areas of commerce has a moderate impact on the aggregate indexes. It eliminates 423 of 43,500 outlets (1%) and 1,079 of 115,420 quotes (1%). It affects 3.04% of the set of indexes or 393 out of 12,926 aggregate indexes.

4.2 Bias

A test for bias is needed in addition to looking at the number of aggregate indexes that were impacted. Some of the indexes increased, while others decreased. If the number of increases does not come close to equaling the number of decreases there may be a problem with index bias. A non-parametric sign test will be used to test if there is a directional bias in the index. The sign test is used due to the lack of independence in the aggregate indexes.

In the scenario where all sample outside the PSU is eliminated, there are 4,184 increases, and 7,304 decreases in aggregate indexes. The z-statistic is -29.1 (p-value < .0001), thus concluding that there is a significant decrease in the aggregate indexes under this scenario.

In the 25-mile/10-quote rule scenario there are 3,735 increases in aggregate indexes and 7,563 decreases. The z-stat is -36.05 (p-value < .0001), thus concluding that there is a significant decrease in the aggregate indexes under this scenario.

In the SAC scenario there are 1,116 increases in aggregate indexes and 1,416 decreases. The z-stat is -5.96 (p-value < .0001), thus concluding that there is a

significant decrease in the aggregate indexes under this scenario.

5. Conclusion

This project looked at only one month of sample. There were 43,500 outlets that month and they will change the following month. We believe that with more time and resources other months should be looked at for patterns of SACs. It could turn out that an identified SAC from one month is an anomaly and that in future months it may not be a SAC.