Exploratory use of Spatial Statistics to Inform the Choice of Variables for Stratifying Non-Self Representing Cities for the Next CPI PSU Sample Selection.

William H. Johnson U.S. Bureau of Labor Statistics
2 Mass Ave., NE, Room 3655, Washington, DC 20212

Key Words: spatial statistics, stratification

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

Introduction:

Every decade after data from the decennial Census becomes available, a new sample of primary sampling units (PSUs) is drawn to support the Consumer Price Index (CPI). The primary sampling units are geographic areas which are unions of counties or minor civil divisions. Large primary sampling units, defined as those with a population greater than 1,500,000 are selected with certainty. Primary sampling units with populations less than 1,500,000 are grouped together into strata and one PSU is selected from each stratum.

As the sample of PSUs is to support the Consumer Price Index, it is desirable to stratify the PSUs so that strata are as homogeneous as possible in order to minimize the between PSU component of variance. Given the data available, previous research focused on identifying variables which model long term CPI change well in order to solve this problem. First PSUs are grouped by Census region and by whether they are metropolitan statistical areas (MSA) or not. Then they are stratified by the variables from the model of CPI change so that the strata have approximately equal populations.

As the design for the new sample of PSUs for the CPI is to be completed by the end of 2001 the author desired to use new tools to determine if additional insights could be gained which would inform the selection of variables

Data used:

The data used for this investigation are 12 month price changes in the all items Consumer Price Index for self representing PSUs. Indexes are not calculated for individual non-self representing PSUs, only for groups of them within Census defined regions and thus no indexes for non-self representing PSUs were used for this investigation. The 12 month price changes used covered the period from December 1987 through December 1999.

What was found earlier:

In 1983 a memo was written to summarize work to that point. One year price change ending in 1980 and 4 year price change ending in 1982 were modeled by variables obtained from the 1980 Census. This was done for the all items index as well as nine lower level indexes. The variables which produced the best R^2 were then reduced by determining which ones were highly correlated with one another and finally a set of seven variables was decided upon for creating strata for non-self representing PSUs.

In preparation for the 1998 CPI revision, further research was conducted in 1991-1992. It was noted that the information available at that time indicated that the component of variance due to PSUs was minimal and that if that information proved to be correct then the effort to stratify non-self representing PSUs would be ineffective. Research indicated that the variables chosen previously were describing the geographic region of the country and that Census region and the finer BLS region were better and simpler predictors of CPI change.

Further work compared a geographic model incorporating latitude, longitude, a normalized squared longitude and percent of population which is urban with previously used models and the best four variable model using variables from Census. This model compared favorably with the use of BLS regions and had an R^2 almost as high as a previously investigated 11 variable model based on Census data. Based on this, the four variable geographic model was used in stratifying non-self representing PSUs in three of four Census regions. In the fourth Census region the seven Census variables were used for stratifying in order to increase the overlap between old and new PSU samples.

Given the increase in computing capabilities, it was decided to revisit the previous work using more data since the 1992 work used only 1, 2, 3 and 4 year CPI change ending in January 1992. One through six year CPI changes calculated for the time from December 1986 through December 1998 were used, with all year to year CPI changes being from December to December. Additionally six month CPI changes from December to June and June to December were used over this time period. The conclusion as written in a memorandum was that the candidate models have not been good inflation predictors since 1992, although the models were found to be insignificant even for many time periods before 1992 for the one year CPI changes.

This last work cast doubt on the value of the models used thus far for estimating CPI change and thus the value of the variables in the models for stratifying non-self representing PSUs.

The investigation described in this paper is an attempt to explain why the models considered became poorer for estimating CPI change after 1992.

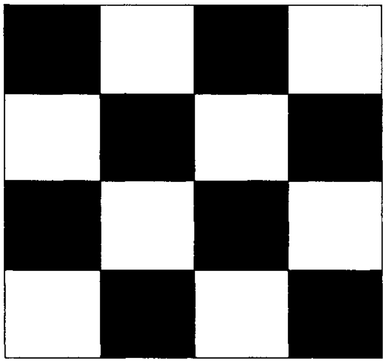
The current investigation:

The software used for this current investigation included SpaceStat version 1.90 and ArcView.

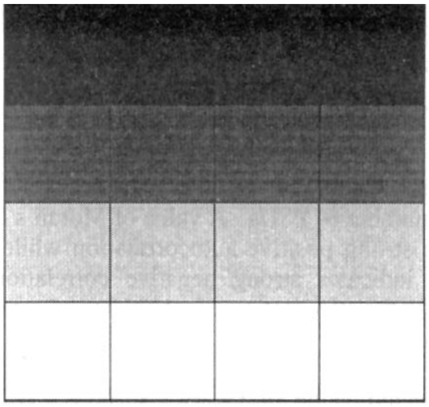
Components of variance were estimated as part of the effort to optimize the design of the sample for the Commodities and Services component of the Consumer Price Index using data from 1993 through 1997. It was found that less than 5% of total variance was attributable to the variation across PSUs within a group of PSUs known as an index area. See Shoemaker (1999) for further details on the estimation of components of variance for CPI change. The percent of variance attributable to variation across PSUs was less for most groups of item strata and was higher only for gasoline and food away from home. This supports earlier work that the component of variance due to PSUs is quite small and thus efforts to stratify the non-self representing PSUs may be of limited efficacy in reducing the variance of the CPI based on the resulting sample. As this work was based on data from 1993 through 1997 it gives some hint that geography was not playing too important a role in CPI change during this period.

In the last decade there has been increasing access to geographic information that can be combined with CPI data, and thus it was desired to determine what could be found about CPI change in this time period using exploratory spatial data analysis.

Spatial autocorrelation of CPI change:



The first concept used for examining one year changes in the CPI is that of spatial autocorrelation. Spatial autocorrelation is a means of describing the relationship of values to values at nearby points. Data is positively spatially autocorrelated if values at a given point are similar to values at nearby points. Data is negatively spatially autocorrelated if values at points are dissimilar to values at nearby points. No spatial autocorrelation indicates that the values are distributed randomly throughout space.



The checkerboard pattern in the picture on the left below is an example of negative spatial autocorrelation with values (color) being surrounded by dissimilar values. The picture above is an example of positive spatial autocorrelation as the value of squares is similar to that of neighboring squares.

As the sample of self representing PSUs for the CPI is not contiguous, there is considerable freedom in the choice of a spatial weights matrix. The spatial weights matrix contains information on what are the neighboring points for each point and what is their relative influence. By convention, a point is not a neighbor to itself.

For the analyses conducted, two different spatial weights matrices were used. The first set of weights used is based on the inverse distance squared. Thus every PSU was considered a neighbor of every other PSU. The spatial weights matrix is row standardized and thus the distance scale used factors out. The rationale for the inverse square distance weights is a model that the influence of the value at a point on other points decays with distance like gravity. The other set of weights used is defined by the three nearest neighbors rule. That is, the three PSUs closest to a PSU are considered to be its neighbors and to have equal influence upon it and all other PSUs are considered not to be neighboring.

Two different measures of spatial association were used, Moran's I and Geary's C.
Moran's I is defined as

$$I = \frac{N}{\sum_i\sum_j w_{ij}} \frac{\sum_i\sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2};$$

Geary's C is defined as

$$C = \frac{N-1}{2\sum_i\sum_j w_{ij}} \frac{\sum_i\sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2}$$

where x_i is the value at the i^{th} PSU and w_{ij} is the ij^{th} element of the spatial weights matrix.

The expected value of the Moran's I is –1/(N-1) where N is the number of PSUs. A value of Moran's I near +1 indicates strong positive autocorrelation while a value near –1 indicates strong negative correlation and a value near the expectation of –1/(N-1) indicates spatial randomness.

The expected value of Geary's C is 1. A value of Geary's C near 0 indicates strong positive autocorrelation while a value near 2 indicates strong negative autocorrelation and a value near 1 indicates spatial randomness.

SpaceStat software was used to calculate these measures for 12-month CPI change ending in each month from December 1987 through December 1999. The median values found are

	Moran's I	
Year	Inverse distance squared weights	Three nearest neighbor weights
Overall	0.1311	0.1555
1987	0.1664	0.2027
1988	0.1237	0.1932
1989	0.1212	0.1950
1990	0.1130	0.1965
1991	0.1203	0.0938
1992	0.1482	0.2013
1993	0.0446	0.0349
1994	0.0965	0.0479
1995	0.1726	0.1914
1996	0.1670	0.2039
1997	0.2194	0.0275
1998	0.0554	0.0863
1999	0.0842	0.0888

	Geary's C	
Year	Inverse distance squared weights	Three nearest neighbor weights
Overall	0.7431	0.7370
1987	0.6084	0.5922
1988	0.6500	0.6116
1989	0.6986	0.6392
1990	0.7755	0.7160
1991	0.6541	0.6789
1992	0.7550	0.6929
1993	0.8884	0.8289
1994	0.8114	0.8381
1995	0.7771	0.7066
1996	0.7182	0.7167
1997	0.6717	0.9221
1998	0.9135	0.8231
1999	0.7693	0.8883

Z-values were calculated for these statistics by SpaceStat under three different assumptions. It turned out that the majority of the time that the values were not significant at the 0.05 level however there were more months in which significance was achieved prior to 1992. On the basis of the observed values there is some slight evidence for positive spatial autocorrelation in 12-month CPI change and that this degree of spatial autocorrelation has decreased in recent years with 1996 being somewhat anomalous.

In examining the values by months, the greatest evidence of positive spatial autocorrelation occurs during March. Geary's C gives almost as much indication of positive spatial autocorrelation for January and December, but possibly slightly more in January. Moran's I however shows slight evidence of positive spatial autocorrelation during December and even less during January. Previous work utilized CPI changes ending in either January or December. Thus any influence that spatial autocorrelation may have on the predictive ability of the candidate models studied would have affected these studies more than if they had used CPI changes ending in other months.

In order to determine if there might be local clustering of CPI change, a local indicator of spatial association, the local Moran was also used. A local indicator of spatial association (LISA) gives an indication at each point of the extent to which there is significant clustering of similar values at neighboring points and the sum of LISAs for all points is a global indicator of spatial association.

The local Moran for a point i is defined as $I_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2} \sum_j w_{ij}(x_j - \bar{x})$, where x_i is the value of x at point i.

Between 10% and 15% percent of PSUs displayed local spatial correlation significant at the 0.05 level with no evidence of difference across years or across months. No PSU tested as having significant local spatial correlation in more than 16% of months from December 1987 through December 1999. This line of evidence suggests that there may be some local clusters of spatial association but the effect is somewhat weak.

Overall, the evidence for positive or negative spatial autocorrelation indicates that there may be some positive spatial autocorrelation. Also that it has slightly declined over the time in question. The weakness of the evidence for spatial autocorrelation may be due to the small number of self representing PSUs in the CPI. This result may also be due to the fact that the all items CPI for a self representing PSU is a highly aggregated number, representing a large number of prices collected for many different items from outlets spread across the PSU, which is typically composed of multiple counties.

A look at the candidate regression models:

Three sets of variables have been considered and compared in the past for use as stratification variables for non-self representing PSUs. The models are as follows:

Model 1)
Percent black consumer units
Percent electrically heated housing units
Percent owner occupied housing units
Mean interest and dividend income per housing unit
Mean wage and salary income per housing unit
Percent fuel oil heated housing units
Percent of housing units with retired persons

Model 2)
Percent white consumer units
Percent black consumer units
Average family size
Percent two or more wage earner consumer units
Percent wage and clerical consumer units
Mean contract rent
Mean gas bill for housing units with a gas bill
Percent housing units with an electric bill
Percent electrically heated housing units
Percent fuel oil heated housing units
Percent gas heated housing units

Model 3)
Normalized latitude
Normalized longitude
Normalized longitude squared
(number approaches +1 at the east and west coasts)
Percent urban

The models were treated slightly differently than in the past. The Anchorage, Alaska and Honolulu, Hawaii PSUs were kept for this analysis although they have been excluded in the past as they are outliers with respect to latitude, longitude and prices compared with other PSUs. They were kept this time in order to have the most data available as there are relatively few self representing PSUs. Given the nature of the spatial weights, either the inverse distance squared weights or the three nearest neighbor weights, their influence should be very small on PSUs in the 48 contiguous states.

Using an ordinary least squares approach, model 1 was significant at the 0.05 level in 11% of the months examined. Model 1 achieved significance in 7 out of 13 months from December 1987 through December 1988 but was rarely significant after that. Using the Breusch-Pagan test, model 1 only displayed evidence of heteroskedasticity in 7% of months, none of them during the time frame 1987 through 1988 when the model had the best predictive power.

Model 2 was significant at the 0.05 level in 40% of the months examined. This model was significant in 10 of 13 months from December 1987 through December 1988 and was rarely significant from 1992 through 1995 and 1997. Model 2 was significant at the 0.05 level in 8 of 13 months from December 1998 through December 1999 so the ability of this model to predict price change may be improving. Model 2 displayed evidence of heteroskedasticity in 8% of months examined with over half of those months being in 1993 and 1994.

Model 3 was significant at the 0.05 level in 24% of the months examined. Model 3 is noteworthy as it was frequently significant during the 1987-1988 time period, after which its predictive power dropped considerably and then tests significant in all 13 months from December 1998 through December 1999. Model 3 displayed evidence of heteroskedasticity in 10% of months examined, but did not display any such evidence from December 1998 through December 1999. Thus it appears that the ability of model 3 to predict one year CPI change improved considerably during the time period from December 1998 through December 1999.

Next these models were modified to determine if there was evidence supporting a spatial autoregressive or spatial error structure. A spatial autoregressive model is of the form $y = \rho Wy + X\beta + \varepsilon$. Here W is the spatial weights matrix and Wy is the spatially lagged version of the dependent variable y.

ρ is the spatial autoregressive coefficient and the null hypothesis of no autocorrelation corresponds to

H_0: $\rho = 0$

If there is autocorrelation which is ignored then the OLS estimates will be biased. In this case one should try to determine what variables are missing from the model that account for the autocorrelation.

For a spatial error model it is assumed that only the error term has spatial dependence and is of the form

$$y = X\beta + \varepsilon$$

$$\varepsilon = \lambda W\varepsilon + \xi$$

If spatial error dependence exists and is ignored then the OLS estimate is still unbiased however indications of model fit may be incorrect as the errors are correlated.

The following table summarizes how often among the months from December 1987 through December 1999 there was evidence of significant spatial autocorrelation or spatial error.

	Spatial Autoregressive	Spatial Error
Model 1	5.2%	9.7%
Model 2	6.7%	42.5%
Model 3	9.7%	19.4%

There appears to be some evidence that a spatial error model would be appropriate for model 2 or model 3. However, during the most recent time period, from December 1998 through December 1999 the results are as follows:

	Spatial Autoregressive	Spatial Error
Model 1	0.0%	7.7%
Model 2	0.0%	76.9%
Model 3	15.4%	0.0%

Thus it appears that during the recent time period during which models 2 and 3 have greater predictive power for CPI change, model 3 shows little evidence of spatial error dependence and slight evidence of being better suited to a spatial autoregressive model while model 2 seems to display strong evidence of a spatial error dependence.

Conclusion:

There is some evidence for a decrease in the level of spatial autocorrelation which can be measured in 12-month CPI change for the all items index for self representing PSUs after 1992, a point when it was found that the predictive capacity of models used in the past decreased. However it appears that the primarily geographic model, model 3, which was used for stratifying non-self representing PSUs for the 1998 CPI revision sample has become much better starting with December 1998 and displays little evidence of misspecification that would be corrected by a spatial autoregressive or spatial error model.

If model 3 continues to perform as well in 2000 as it did in 1999 then the variables are quite usable for stratifying non-self representing PSUs even if there is no evidence that stratifying by them will greatly reduce the variance of the CPI calculated using a new sample stratified with these variables.

References:

Anselin, Luc (1995), SpaceStat User's Guide

Arlinghaus, Sandra L. Ed. (1996), *Practical Handbook of Spatial Statistics*, CRC Press, Boca Raton, NY

Shoemaker, O. and Johnson, W. (1999), "Estimation of Variance Components for the U.S. Consumer Price Index", *Proceedings of the Government Statistics Section, American Statistical Association*, to appear.