Geocoding Procedure to find Geographic Identifiers in the Housing Component of the Consumer Price Index

August 2005

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

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

Abstract

This paper gives an introduction to the Housing component of the Consumer Price Index (CPI), outlines the steps taken to geocode housing units for use in the CPI and discusses some uses of geocoded housing units. The purpose of geocoding is to find a geographic identifier called a Census Block Group Number for housing units. Bureau of Labor Statistics Statisticians will use this identifier for weighting, sampling and identification purposes. The software used is ArcView 9 created by ESRI. ArcView 9 is GIS software that is used primarily by Geographers and is beginning to find prevalence in Statistical realms. Other software used is SAS and Map Loader, also created by ESRI.

Key Words: Geocoding, Spatial Statistics, Weighting

1. Introduction

The Consumer Price Index (CPI) is an economic indicator that measures inflation for the urban population of the United States. Items are clustered in seven major categories. The category with the highest relative importance is housing, which has a relative importance of 42 percent. Thus it is exceptionally important to measure inflation in this category as accurately as possible.

The housing component is broken down into 2 primary categories; Owners' equivalent rent and Rent, which have relative importances of 23 percent and 6 percent respectively. The remaining 13 percent is made up of many item strata including housing insurance and maintenance and repair.

Owners' equivalent rent is not easy to calculate because it is not directly measurable. Ideally we would like to find a rental house among a neighborhood of owned houses. Then we can use this information to project the change in rental value of similar houses.

1.1 Geographic Stratification

Based on 1990 census data we found the two best predictors of rent change are geographic location and rent level. Before the 1998 revision we did not take geography into account when stratifying PSUs for the CPI housing sample designs. It was not taken into account because geographic software was unavailable at the time.

The geographic stratification accomplishes five goals[i]:

1) It helps ensure sample coverage for the major characteristics (geography and rent level) that are correlated with rent change.
2) It is felt to be the best way to correlate renter-occupied units with owner-occupied units in the same neighborhood, in order to produce the 'rental equivalence' index.
3) Housing units constructed after 1990 can be located and assigned to the existing geographic strata.
4) Because goals 1 through 3 will be met, there should be a reduction in the sampling variance of the 'rent' and 'rental equivalence' indexes.
5) It sets up a stratification structure that will allow the rotation of Housing samples on a rolling basis, thereby distributing the introduction of future census samples over an extended period.

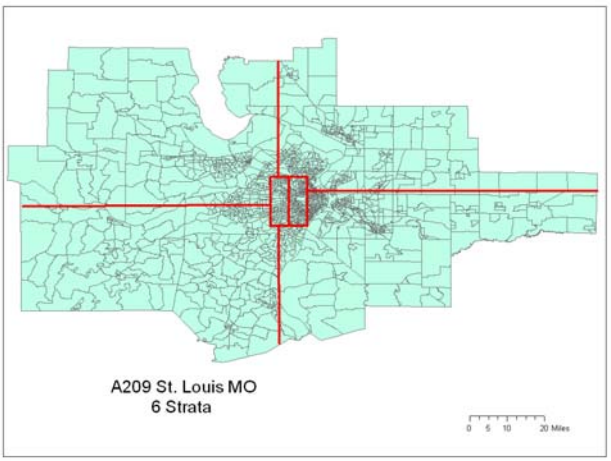
With the advent of Geographic software these goals can now be realized. To expedite this process every housing unit whether apartment or house is assigned a Census number that relates to its geographic location. This number is called a Census Block Group number. It consists of the State, county and tract numbers along with the 3 digit block group number. This 12 digit number defines the neighborhood in which the housing unit is contained.

These block group designations are also useful for selecting a sample. Across the country we select Urban and Metropolitan areas that are called Primary Sampling Units (PSUs). There are 87 PSUs currently in the CPI PSU sample. Each PSU is broken down into 6 strata. To do this:

1) First the smallest box containing a third of the PSU's total housing expenditure is determined.
2) This box is then split either horizontally or vertically into strata of nearly equal housing expenditure by rent level. The split that maximizes the difference in median rent levels is used to define strata 1 and 2.

3) The four non-central strata are found iteratively in a similar fashion to arrive at the 6 strata.

See graph I below. The little areas on the map below are Census Block Groups.



Segments are selected with probability proportional to size. The probability of selection is the block group expenditure divided by the stratum expenditure multiplied by the number of selections from the stratum. Block groups are ordered according to expenditure within strata and then a systematic sample is taken to capture a representative sample from each stratum. This guarantees that the sample is not composed exclusively of high rent segments or exclusively of low rent segments. Each segment has a weight, which is the reciprocal of the probability of selection.

2. Geocoding

Now that we understand the utility of Census Block Group identifiers we can now move on to the process of geocoding housing units to find these values. Geocoding is the process of determining the longitude and latitude of units based on initial information such as address or block group number. For example if one has the address of a housing unit and needs to find the block group number, first one has to geocode the address to create a point-shape file. A point-shape file is native to ArcView. It contains the information for the geocoded addresses as well as X-Y coordinates and the geometry of the shape, in this case a point. Once the point-shape file is created it can be placed on a map of block groups. Then it can be determined in which block group the unit lies. This is done with the new construction housing units in the CPI housing sample. When a house is built the address is supplied by the Census Bureau from building permits. The Bureau of Labor Statistics (BLS) must determine to which block group it belongs in order to include it in the sample.

First I had to download the street and block group information for all of the counties in all of our PSUs to set up this procedure. Once this was finished I had a complete record of all the streets and block groups in each PSU. The street record contains numerous fields that will be used in the geocoding service. Included are the following fields: From Address Left, To Address Left, From Address Right, To Address Right, Street Name, Zip Code Left, and Zip Code Right. There is also information on geometry and location.

The geocoding service is created from this street shape file. The geocoding service is produced in ArcCatalog and is the "program" that will find where the address' location geographically. Say we have an address at 123 Main St. 08011. It should be related to the 100 block of Main St. in the zip code area 08011. It can also be found to be on the left side of the street about one third of the way up depending on the range of addresses on that block. Once the address is geocoded it is now a point-shape file and can be place on the map of block groups or streets in ArcMap. It will show the precise location of the address with a point on the map. This example is for just one address. In actuality geocoding is done in a batch process and one can find the geographic locations for hundreds of addresses in a mater of moments.

To find the Block Group number an action is taken called a joins and relates. This is a process done in ArcMap. Once this is completed it will return a database file showing where an address point intersects a block group and will assign to that address its corresponding block group number.

If one wants to map the locations of the housing units, this is done in ArcMap. ArcMap and the aforementioned ArcCatalog are two application found in ArcView. In ArcMap the layers for block group and geocoded addresses (and streets if you wish) are put together to make valuable maps. The block group layer is added first because this is on the bottom and the point shape file for the addresses is added second to go on top of the block group layer. Other layers such as zip code boundaries, school districts or water can be added later to make the map more useful.

3. Mapping Project

ArcView is a program for creating and exploring geographic information. In the following project it was my job to determine the accuracy of a list of addresses that was obtained from a third party vendor of addresses.

First we started with a list of 442 addresses from the vendor. Several BLS employees then used American Fact Finder, located at factfinder.census.gov,

to verify if the provided block group numbers from the vendor denoted Census2000 and the FactFinder derived block group numbers agreed. At American Fact Finder anyone can type in an address and find Census information including block group number.

There were 29 addresses in the vendor list that did not match the block group number provided on American Fact Finder. This is a small percentage (6.5%) but as I have outlined above it is important that these identifying numbers be accurate for sampling and weighting purposes. So I had to do research to determine the nature of these discrepancies.

Twenty-three (23) of the 29 address could be geocoded with ArcView. Ten of the 23 addresses I could geocode were just on the opposite side of the street from their listed block group. See graph II below. The dark line is the street that defines the boundary of the block group and the light lines are streets. Listed are the block group numbers as well. When the street is not a major highway, as in this case, then it is safe to assume if we sample a house on the opposite side of the street it will be representative of the neighborhood we are trying to sample.

Of the remaining 13 the difference was a more significant than being on the opposite side of the street. See Graph III below, but they were all in adjacent block groups. Team leaders decided that if there was discrepancy in Census2000 designations it would be acceptable as long as the differing addresses were at least in adjacent block groups. This way the inaccurately labeled addresses would likely be representative of the neighborhood in which we are sampling.

4. Conclusion

Census block group numbers are a way to identify the "neighborhood" in which a residence or outlet is located. However BLS is not restricted to strict interpretations of these numbers. If a unit has different block group numbers according to the vendor but is located on the opposite side of a street or in adjacent block group to the FactFinder derived block group then it can still reasonably be considered in the same neighborhood. Thus the unit is likely to have similar characteristics to units in the correct block group. The 29 discrepancies out of 442 units are a small minority while the vast majority of units have precise block group designations.

Geocoding is a useful tool for BLS housing purposes. It has a long set up time while downloading the county information and creating the geocoding services but once these are created it is rather fast to geocode and analyze a list of addresses.

At the BLS Geographic Information Systems (GIS) is a relatively new tool and with time we will find other applications for this new technology.

Reference

[1] Ptacek and Baskin, "Revision of the CPI housing sample and estimators" Monthly Labor Review, December 1996 pg 32 – 33