Center for Demography and Ecology

University of Wisconsin-Madison

Geo-Demographic Methods for Improving Small-Area Population Estimates

Marc Perry

Paul R. Voss

CDE Working Paper No. 96-15

USING GEO-DEMOGRAPHIC METHODS FOR IMPROVING SMALL-AREA POPULATION ESTIMATES

Marc J. Perry
Department of Sociology
University of Wisconsin-Madison

Paul R. Voss
Department of Sociology
University of Wisconsin-Madison

This research is supported in part by a pre-doctoral training grant from the National Institute on Child Health and Human Development (5T32HD07014), as well as the College of Agriculture and Life Sciences, University of Wisconsin-Madison, and the U.S. Department of Agriculture, through a grant from the North Central Regional Center for Rural Development (NCRCRD Contract No. 94-4A). Research was carried out using facilities of the Center for Demography and Ecology at the University of Wisconsin-Madison, which receives core support from a center grant from the National Institute on Child Health and Human Development (P30-HD058760). Portions of this paper were presented at the 1996 annual meeting of the Population Association of America, New Orleans, LA. Correspondence should be directed to Marc Perry, Department of Sociology, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, Wisconsin 53706 or to perry@ssc.wisc.edu.

Using Geo-Demographic Methods for Improving Small-Area Population Estimates[1]

Marc Perry and Paul R. Voss
Applied Population Laboratory
Department of Rural Sociology
University of Wisconsin-Madison

This paper introduces, tests and demonstrates the efficacy of a methodology for improving post-censal population estimates for Census Block Groups.[2] While it is a methodology based on a new application of small-area geodemographic clustering systems, its conceptual and intellectual roots lie in a methodology -- called "social area analysis" -- proposed several years ago but apparently little used by those actually making population estimates (Goldsmith, Jackson and Shambaugh, 1982).

Small-Area Population Estimation

What is meant by a geographic "small area" has undergone considerable downward revision during the past five decades. In the 1950s and 1960s, a period of intensive innovation in the development of small-area population estimation models, the term "small area" typically referred to states, metropolitan areas, counties and large cities (for example, see Schmitt, 1952; Siegel, Shryock and Greenberg, 1954; Zitter and Shryock, 1964; Gurley, White, and Tarver, 1965). Passage in 1972 of the State and Local Fiscal Assistance Act (P.L. 92-512, commonly referred to as the Federal General Revenue Sharing Act), and similar statutes in several states, caused demographers to shift downward their notion of "small" areas. Small-area demography generally came to mean developing estimation or projection models for the 39,000 units of general purpose government in the U.S. Most of these governmental units are rural villages, towns and townships, and typically are smaller than the average census tract found in the nation's urbanized areas. The critical developments in small-area demography during this time were mostly taking place at the U.S. Census Bureau and in 400-plus additional settings in the public sector -- mostly by demographers at major public universities, in a growing number of state demographic units, and in local planning councils (U.S. Bureau of the Census, 1978).

Also in the 1970s, several companies sprang up the private sector that developed models for producing post-censal estimates and projections largely for commercial use (Russell, 1984; Merrick and Tordella, 1988). Since areas for which population estimates or projections were desired in the private sector typically were not standard units of political geography,[3] it became necessary for the commercial data vendors to devise ways of producing population estimates and projections for very small areas that could serve as the building blocks for larger, nonstandard areas. In metropolitan urbanized areas, these small building blocks typically were census tracts. In areas outside the officially recognized urbanized areas, minor civil divisions (MCDs) usually were the “update” units of analysis.[4] Apparently, in urbanized areas, tract-level estimates occasionally were disaggregated to census block groups on the basis of proportions observed in the most recent census. But tracts remained the basic geographic units used by commercial data companies throughout the 1980s to produce demographic[5] estimates and projections both for standard and nonstandard geographic areas.

As the term "small area" has come to mean smaller and smaller geographic units, demographers over the past decade have even begun to talk with excitement about a time soon when geographic information system technologies, coupled with extensive administrative records databases, hold the promise of making reasonably reliable estimates for Census Blocks, and even individual "block faces" (Edmondston and Schultze, 1995:156-77). Truly large-scale success on this front likely is at least a decade away, however. By 2000, the technology and the digital map -- augmented by the Census Bureau's planned nationwide Master Address File (MAF) -- likely will be in place to support such an effort. What certainly is yet farther off in time is the development of large and uniform administrative records databases that are appropriate for population estimation support, sufficiently reliable in their application, and (if ever) publicly accessible. Therefore, it seems prudent to continue to explore methodologies for improving current demographic estimation models operating at very small geographic levels.

Current Methods for Making Population Estimates for Block Groups[6]

Efforts among the private data vendors to produce population estimates for block groups generally are limited to rather naive extrapolation or step-down techniques. Such simple estimates (even if we admit extrapolation to be a valid estimation methodology[7]), typically are brought into alignment with population estimates for larger areas by proportional adjustment of the block group estimates. Through this technique, naive estimates for very small areas at least benefit from being forced to sum to an estimate for a parent level of geography where an independent population estimate has first been produced (usually by the U.S. Census Bureau or by a state demographic unit) using standard[8] estimation methodologies (Rives, *et al*., 1995). At least one data vendor also employs indicators of local residential change derived from new (and retired) listings from telephone books. It is not clear to us, however, whether any of the large data vendors currently are making use of the U.S. Postal Service's National Change of Address file in their update models. Most, if not all vendors, apparently do update their small-area databases using the results of special enumerations conducted by the Census Bureau or selected state agencies.

It is not generally known how good these commercially produced estimates for block groups are. To our knowledge, only the results of two somewhat limited studies have been reported. The first (Chapman, 1987) was a study, carried out by the International Council of Shopping Centers, of the statistical precision of estimates produced by six data vendors. Since our focus is on estimation accuracy, this report does not interest us here, except to say in passing that the estimates reported for Zip-code areas, which were the geographic areas in that report most likely to be estimated by aggregating tract-level estimates, had exceptionally high statistical coefficients of variation suggesting, by implication, that at least some of the individual estimates suffered from very large errors. The second study (Pitkin, 1992) does speak to accuracy since it compared 1990 population estimates from five different vendors with 1990 Census results. The reported results are somewhat difficult to interpret. Furthermore, the study, having been carried out only in one portion of southern California likely is difficult to generalize from.

Proposed Refinement

Our proposed refinement for improving small-area population estimates rests on two empirical questions. If the first question is answered in the negative, then the study ends there, and the refinement fails. If, however, the first question can be answered affirmatively, then the first hurdle is successfully passed, and the study may proceed to the second question. If it, too, can be shown to have an affirmative answer, then we will have demonstrated that there does, in fact, exist a methodology for improving population estimates for very small geographic areas. The questions are these:

Question 1: Do the geodemographic clusters that emerge from a standard cluster analysis of census data for block groups share a common demographic *dynamic* in ways that parallel the fact that they share common demographic *attributes*? That is, is there less within-cluster variation in terms population growth rates than there is between-cluster variation in growth? Or, put another way, in recent years, does it appear that neighborhoods that share common demographic attributes with similar neighborhoods elsewhere tend to also share a common rate of population change?

Question 2: If the first question is answered in the affirmative, can it then be further demonstrated that detailed, intensive knowledge regarding postcensal demographic change for clusters in a specific area can be used to improve estimates in other areas for which we have no such knowledge? That is, does special intelligence about population change for cluster *i* in one area spill over in ways that likely tell us about change in cluster *i* areas elsewhere?

Regarding the second question, it is assumed here that state and local demographers do have (or can obtain), on an ongoing basis, a steady stream of specialized intelligence regarding specific areas in their states or localities. Such intelligence generally is gathered in the process of carrying out other aspects of their work, for example, from detailed labor market studies, school district studies, neighborhood planning efforts, and special census enumerations. At the moment, for most state and local demographers, these opportunities arise in a rather unorganized and *ad hoc* manner. If the above two questions were to be answered affirmatively, then we assume that state and local demographers might establish a method for systematically storing and using such *ad hoc* inputs and even perhaps develop a specific case study initiative that would have as one explicit objective the gathering of population change information for geodemographic clusters that would then systematically feed into a population estimation system.

Test Methodology

We initially came to this study hoping that we might obtain from one of the nation's data vendors three pieces of information: (1) population counts for each cluster-coded Census Block Group in 1980; (2) population counts for the same block groups in 1990; (3) the vendor's 1990 population estimates for each block group. Armed with such information, our proposed test would have proceeded directly and fairly inexpensively. But no vendor stepped forward. We suspect the requisite data do not exist -- most especially the 1990 block group estimates. Admittedly, even had a vendor been willing to team up with us, the provision of such data would not have been easy. The introduction by the Census Bureau of a new nationwide system of Census Blocks and Block Groups in support of the 1990 Census was a change in census geography that made 1980-90 geographic comparability at the block group level virtually impossible. At best, comparability partially obtained in urbanized areas, but even here the block and block group numbering changes generally made this a daunting task.

We therefore proceeded to conduct our own cluster analysis of 1990 Census block groups. We added some interesting innovation at this stage of the research, but the fact that we did not have comparable clusters from 1980 -- nor 1980 Census counts associated with each clustered block group -- presented us with our first problem that somehow needed a "work-around" solution. More about this solution later.

A second problem emerged: We did not have any 1980-based estimates for 1990 block groups. This problem at first seemed to be an insurmountable hurdle. It's obvious: To test a set of estimates, one *needs* the estimates. But again we developed a "work-around" solution. We will discuss both of these work-arounds below. For now we simply will say that they both introduce some undesirable artificiality in our test that ideally should not be there. We believe, however, that our final conclusion and recommendation stand on their merits despite these two "work-around" short cuts.

Developing a Block Group Clustering Scheme

Cluster Analysis Methodology

Cluster analysis can be thought of as a multivariate statistical procedure that “starts with a data set containing information about a sample of entities and attempts to reorganize these entities into relatively homogeneous groups” (Aldenderfer & Blashfield, 1984). For the present study of block groups in the Midwest, cluster analysis helps us to answer the following question:

> Taking into account variation on demographic, economic, and agricultural data for the roughly 60,000 block groups in the Midwest, how would these block groups be reclassified into a specified number of clusters--in our case, 30--that maximize within-cluster homogeneity with respect to the demographic, economic, and agricultural variables under consideration?
>
> That is, using statistical procedures, how would we sort all of the block groups so that each cluster contains block groups that are *most* similar with each other and as different as possible from block groups in other clusters?

Most analysis of data at the block group level are univariate--they focus on only one variable at a time, such as income or race. This study incorporates a large number of variables into a multi-variate approach to anser the question: “Which block groups are most similar with respect to income ***and*** household types ***and*** age composition ***and***....?”

Several caveats are in order. First, while cluster analysis is seemingly a structure-seeking statistical procedure, it is, in actuality, very much a structure-imposing method. All objects (in our case, census block groups) will *always* be assigned to a cluster, as Aldenderfer & Blashfield (1984) observe. In addition, cluster analysis performs within parameters that are established by the researcher. For example, the clustering procedure used in this analysis, SAS FASTCLUS, requires the specification of either: (1) the maximum number of clusters allowed or (2) parameters for creation of new clusters. Different assumptions and clustering methods may produce different clustering results.

Data set Construction

Two different sources of data were utilized for this project. From the 1987 Census of Agriculture, we initially selected 43 variables on agricultural activity at the ZIP code level of geography, and from Summary Tape File 3A of the 1990 Census of Population and Housing we initially selected approximately 250 variables at the block group level of geography.

Agricultural Data: Because our unit of analysis in this study is the census block group, and the agricultural data was reported at the ZIP code level of geography, we used a polygon intersection routine from geographic information system software in order to attach agricultural data to block groups. For block groups wholly contained within a particular ZIP code, the values of the agricultural data for that ZIP code were attached to the block groups. In those cases in which a block group contained more than one ZIP code, we went down to the census blocks that formed the particular block group and selected the ZIP code of the census block group with the largest population. In some cases, a block group contained no rural population and was clearly urban in character, yet situated in a ZIP code in which agricultural activity occurred. Simple polygon overlay would have mistakenly attached agricultural characteristics to such a block group. In a few instances, a farm operator's farm was not within the same area as his or her mailed ZIP code (ie, mailing address). Because the 1987 Census of Agriculture employed a mailout/mailback enumeration method, these cases resulted in farms being enumerated in the mailing address ZIP code, not the ZIP code where the farms were located. To correct for these types of errors, we set agricultural variables to equal zero in those block groups which were 100 percent urban and therefore contained no rural population.

The initial variables in the 1987 Census of Agriculture were expressed in terms of three farm counts per category. For example, size of farm was originally expressed in terms of three variables: number of farms with land 1-49 acres, number of farms with land 50-999 acres, and number of farms with land 1000+ acres. These variables were recalculated in terms of percentages: percentage of farms in a ZIP code with land 1-49 acres, percentage of farms with land 50-999 acres, and percentage of farms with land 1000+ acres.

Data on crop cultivation at the ZIP code level was partly dependent on the State in which the ZIP code was located; crop information is only reported for the top three crops cultivated in each state. A total of 43 variables on agricultural activity at the ZIP code level of geography were obtained from the 1987 Census of Agriculture and are listed on the following page. All percentages are based on the complete count of farms in the particular ZIP code.

1 number of farms
2 percent of farms with land 1-49 acres
3 percent of farms with land 1000+ acres
4 percent of farms with market value of agricultural products sold less than $10k
5 percent of farms with market value of agricultural products sold greater than $100k
6 percent of farms with cattle/calf inventory, 1-49
7 percent of farms with cattle/calf inventory, 200+
8 percent of farms with hog/pig inventory, 1-49
9 percent of farms with hog/pig inventory, 200+
10 percent of farms with cropland harvested, 1-49 acres
11 percent of farms with cropland harvested, 200+ acres
12 percent of farms with corn for grain, 1-49 acres
13 percent of farms with corn for grain, 250+ acres
14 percent of farms with corn for silage, 1-49 acres
15 percent of farms with corn for silage, 250+ acres
16 percent of farms with sorghum for grain, 1-49 acres
17 percent of farms with sorghum for grain, 250+ acres
18 percent of farms with wheat for grain, 1-49 acres
19 percent of farms with wheat for grain, 250+ acres
20 percent of farms with barley for grain, 1-49 acres
21 percent of farms with barley for grain, 250+ acres
22 percent of farms with rice, 1-49 acres
23 percent of farms with rice, 250+ acres
24 percent of farms with sunflower seed, 1-49 acres
25 percent of farms with sunflower seed, 250+ acres
26 percent of farms with cotton, 1-49 acres
27 percent of farms with cotton, 250+ acres
28 percent of farms with tobacco, 0.1 to 4.9 acres
29 percent of farms with tobacco, 10+ acres
30 percent of farms with soybeans, 1-49 acres
31 percent of farms with soybeans, 250+ acres
32 percent of farms with potatoes, 0.1 to 14.9 acres
33 percent of farms with potatoes, 100+ acres
34 percent of farms with sugarcane, 1-49 acres
35 percent of farms with sugarcane, 250+ acres
36 percent of farms with peanuts, 1-49 acres
37 percent of farms with peanuts, 250+ acres
38 percent of farms with hay, 1-49 acres
39 percent of farms with hay, 250+ acres
40 percent of farms with vegetables, 0.1-14.9 acres
41 percent of farms with vegetables, 100+ acres
42 percent of farms with orchards, 0.1 -14.9 acres
43 percent of farms with orchards, 100+ acres

From this initial field of 43 variables, we deleted those crops (peanuts, sugarcane, potatoes, cotton, tobacco, corn for silage, and rice) which are not reported for any of the 12 Midwestern states. This left us with 29 variables to incorporate into the clustering scheme.

Social, Demographic, and Economic Data: From the 1990 Census of Population and Housing, Summary Tape File 3A, we obtained approximately 240 variables at the block group level of geography. These variables contained information on race & ethnicity, age structure, household types and relationships, educational attainment, labor force status, industry and occupation, income, poverty status, housing unit occupancy, and place of residence five years earlier. The Census Block Group is the smallest level of geography for which data from the census "long form" questions are tabulated.

Creating the Final Data set--Variable Selection

After assembling an initial data set with 267 variables for the 60,368 block groups in the twelve Midwestern states, we needed to create a subset of variables on which to perform the cluster analysis. The goal here was to identify a limited subset of variables that would best provide a full portrait of social, demographic, and economic activity for each block group.

All 267 variables were first standardized, since variables with large variances tend to have more effect on the resulting clusters than variables with small variances. Then, a matrix of zero-order correlation coefficients was examined in order to identify variables which were highly correlated with one another. In cluster analysis, including highly correlated variables has the undesired effect of assigning extra influence to those variables. For example, our initial data set contained 18 separate measures of poverty status and receipt of public assistance. If we had included all 18 variables in the cluster analysis, it would have served to significantly weight the impact of poverty status and receipt of public assistance as clustered variables in the model. Accordingly, we sought to minimize the effect of multicollinearity by paring our model down to slightly over 100 variables.

After initial clustering runs, we dropped 56 variables with small values of RSQ/(1-RSQ) from the model. This is the ratio of between-cluster variance to within-cluster variance, and is essentially one indication of whether a variable is helping to define one or more particular clusters. For these 56 variables, a block group's value for the variable was not highly associated with the block group's cluster. We also dropped all agricultural crop variables; the reporting of only the top three crops for each state introduced undesirable "border effects" between states with different types of crops reported. This left us with 36 variables in our final model. They are listed below.

Final 36 variables used in cluster analysis, by category

Characteristics of Farms--size and crop yield

1 Percent (of farms) with land 1-49 acres
2 Percent with land 1000+ acres
3 Percent with market value of agricultural products sold less than $10k
4 Percent with market value of agricultural products sold greater than $100k

Race & Ethnicity

5 Percent of population Hispanic
6 Percent of population Non Hispanic Black
7 Percent of population Non Hispanic Other (includes American Indian, Eskimo, Aleut, Asian or Pacific Islander, and Other)

Age of Population

8 Percent of population aged 0 to 4
9 Percent of population aged 18-24
10 Percent of population aged 25-49

Education of persons aged 25 and over

11 Percent of adult population without a high school diploma
12 Percent of adult population with a Bachelor's degree
13 Percent of adult population with graduate or professional degree

Industry and Occupation of employed persons

14 Percent of employed population in blue-collar (farming, skilled/unskilled labor) occupations

15 Percent of employed population in white collar (professional and technical) occupations
16 Percent of employed population in Primary Industries (Agriculture, Forestries, Fisheries, and Mining)
17 Percent of employed population in Service Industries
18 Percent of population unemployed

Household Type and Relationship

19 Percent of population in family households
20 Percent of population living alone
21 Percent of population in Other Nonfamily Households
22 Percent of population in group quarters

Income and Poverty Status

23 Percent of households earning less than $10k
24 Percent of households earning $50k-$100k
25 Percent of households earning greater than $100k
26 Percent of population receiving public assistance
27 Per capita income
28 Per capita income for Blacks
29 Percent of population in poverty

Housing Unit Characteristics

30 Housing unit density
31 Rental units as a percentage of occupied housing units
32 Percent of housing units in large, multi-unit (20+) buildings

Language Ability

33 Percent of population ages 5 and over that speaks English “not well” or “not at all”

Foreign Born Status

34 Percent of population foreign born

Urbanization/Farm location

35 Percent of population in urban areas
36 Percent of population living on a farm

Performing Cluster Analysis

Using SAS FASTCLUS

SAS FASTCLUS is the statistical program that was used in this project. Unlike some cluster analysis programs that require the calculation of an N X N matrix of similarities between observations, FASTCLUS works directly upon the raw data. Therefore, it can be used successfully with very large datasets. In this project, our final data set matrix contained roughly 60,000 block groups X 36 variables.

SAS FASTCLUS is an iterative partitioning method of cluster analysis. It requires the partitioning of the data into a *specified* number of clusters--in this case, 30. It begins by making an initial pass through the data set, assigning 30 cluster "seeds".   A seed is essentially the first guess of the means of the clusters. Each block group is assigned to its nearest seed, so that temporary clusters are formed. FASTCLUS repeats this process through the data set until no more block groups are reassigned. A block group can belong to one and only one cluster.

From initial clustering runs, six individual block groups were identified as extreme outlier cases (they consistently resulted in clusters with only one observation each) and were omitted from further clustering runs. In the end, we identified 30 clusters for the remaining 60,362 block groups in the twelve Midwestern states. These results are reported in Table 1.

Results of Cluster Analysis

We see in Table 1 that the 30 clusters vary widely in size. Clusters 16 and 20 both contain over 7,000 observations each. Together, they contain about one quarter of all block groups in the Midwest. Two clusters contain less than 100 observations each.

"Closest Cluster" indicates the cluster with the centroid (the mean for all variables clustered) closest to that of a given cluster centroid. It can be thought of as the cluster most similar to a given cluster with respect to **all** of the variables in the clustering model. Knowing which cluster is closest to a given cluster, we can better understand the relationships among the 30 clusters with respect to their centroids' similarities.

"Distance Between Cluster Centroids" gives an indication of *how* close the two cluster centroids are. For example, in Table 1, cluster 6 is the cluster that is closest to cluster 1, and the distance between them is 6.796. But notice that this is not necessarily a reciprocal relationship: cluster 1 is **not** the cluster closest to cluster 6--cluster 7 is even closer (a distance of only 5.149).

It should be emphasized here that the word 'cluster' as used in this report does not imply a geographic relationship of any kind among block groups. What is being clustered are the means of all 36 variables in each block group. These *centroids* are grouped together to form clusters. The block groups in a particular cluster, then, are not necessarily contiguous.

What are the key results of the cluster analysis?

The thirty clusters produced by the cluster analysis present a portrait of wide variation in types of block groups in the Midwest. While clusters 2, 9, 17, and 20 are all highly agricultural, rural clusters, there are large differences among them. Clusters 2 and 17 contain large-scale farming, both in acreage and market value of agricultural products sold, while clusters 9 and 20 are much more likely to contain small scale farms.

In metropolitan areas, clusters 12, 13, 15, and 25 are the most urban. Block groups that are racially and ethnically distinct emerge (6,7,12, and 22), as do areas with unique populations such as college students , foreign born persons (8,13,15,21,23,25, and 27), and group quarters persons (14, 24, and 29). By utilizing census block groups as our unit of analysis, our results go beyond the standard central city/suburb dichotomy. We have a much finer detailed portrait of a metropolitan area.

Recall that the purpose of this research was not to simply create another descriptive tool for commercial marketers. Our aim was to create a set of social and economic clusters that demonstrate potential research and analytic applications in rural community and economic development. The results could offer a method of comparing small-area geography across the rural Midwest.

Once the clustering system was finalized, we were prepared to ask the first question underlying this test: Does each cluster tend to have its own characteristic growth rate? This is one of the crucial intellectual questions prompting this research. But, as indicated above, in order to answer this question we were faced with an associated, and very practical, question: Without 1980 counts for each block group, how do we measure change over the decade for the block groups? This dilemma led to work-around solution number one.

We elected to use a proxy measure for population change between 1980 and 1990 based on the number of housing units added to each block group during the intercensal period. We were able to do this solely on the basis of 1990 Census information derived from long-form Q.H17 Our measure of change, then, is simply is the percentage increase in housing stock during the 1980-90 time period. This is a much preferred proxy than, say, the alternative of taking the proportion of the population who moved into the block group since 1985 (long-form Q.14b). Some block groups -- such as those in which college students who reside in rented housing predominate -- can have very high proportions of residents who were not present five-years ago, but in reality show little numerical population change.

Creating Block Group "Population Estimates" for 1990

Lacking 1980 population counts for our 1990 block groups also meant that we couldn't prepare 1980-based estimates of the 1990 population, as normally would be done for a traditional test of an estimation methodology. This shortcoming required that we devise a proxy system of 1990 estimates: work-around solution number two.

First, because of data limitations, we were compelled to make estimates of housing units rather that population, *per se*. Fortunately, the 1990 Census question on "year housing unit was built" again came to our rescue. We had already created approximate 1980 "counts" of housing units in each block group as part of our first work-around solution by simply removing from the 1990 housing stock all housing added subsequent to 1980.[9]

Second, we chose the trended share method (Pittenger,1976) as our means of deriving an estimate of housing units in each block group for 1990.[10] To implement this method, we first determined the share of a county's housing stock standing in each of the block groups in the county in 1970.[11] We then calculated the same ratio for 1980. For each county in 1990, we calculated a projected ratio by assuming that the 1970 to 1980 change in the block group to county ratio of housing stock would continue between 1980 and 1990, but at only half the rate.[12] We then applied this estimated 1990 ratio to the county's observed housing stock to derive estimates of block group housing.[13]

Third, we compared our estimates of 1990 housing units in each block group with the observed 1990 housing unit counts. Based on these comparisons, we prepared a standard report of estimate errors.

Fourth, came the crucial step in our methodology. Recall that we wish to know whether or not we can improve on our small-area estimates by incorporating special intelligence derived for a sample of areas. To simulate such intelligence, we drew a random sample of block groups from each of the 30 clusters. From clusters with more than 500 block groups, we randomly sampled, on average, 21 block groups per cluster. For clusters with less than 500 block groups, we randomly sampled approximately 6% of the block groups in each cluster. The result was a total sample of 607 block groups. In these sample block groups we *assumed we knew* the actual 1990 housing stock, and calculated the 1980-90 growth rate of housing units based on the two censuses. We then assigned these block groups to their 1990 clusters and determined average growth rates for each cluster based only on these 607 block groups.

Fifth, using the cluster-type rates of housing unit change between 1980 and 1990, determined in step four from our limited sample of block groups, we re-estimated 1990 housing units for each block group in the Midwest assuming that all block groups of cluster-type *i* grew at the same rate as the sample of block groups of cluster-type *i*.

Sixth, we compared these re-estimates of 1990 housing units in each block group with the observed 1990 housing unit counts. And, again, based on these comparisons, we prepared a standard report of estimate errors.

Finally, as often is done to improve estimate systems, we took our initial estimates based on the trended share methodology and our estimates based on the geodemographic method and averaged the two to obtain a third set of estimates. Again, we prepared a standard report of estimate errors.

Findings

Regarding the question of whether geodemographic clusters tend to be characterized by a particular growth dynamic, Table 2 suggests that, indeed, they do. While the average percent change in housing unit stock, 1980-90, for the full 60,362 block groups was 19 percent, the values for individual clusters ranged from 4 to 84 percent, with smaller standard deviations than the overall value of 96 percent for the full dataset. Clearly, there were different housing stock growth patterns occurring in the different clusters during the time period 1980-1990.

Having answered this first question in the affirmative, we were prepared to test the implications of this finding by incorporating geodemographic techniques into a system of estimates of housing units. Table 3 shows the results of all three estimation techniques, by cluster: (1) our initial trended share estimation technique, uninformed by geodemographic information, (2) the geodemographic method, and (3) the estimates derived by simply averaging the two estimates. Table 4 presents Mean Percent Error (MPE) values for each of the estimation methods, by cluster, while Table 5 contains Mean Absolute Percent Error (MAPE) values by cluster.[14] The overall values of MPE and MAPE are shown in Table 6.

We see from Table 3 that for many of the 30 clusters, the three different techniques appear to produce somewhat similar estimates, at least in the aggregate. Certainly there exist some quite substantial variations. For example, in cluster 10, which, on average, experienced substantial growth in housing stock during the decade 1980-1990, the Trended Share method substantially *over*estimated 1990 housing stock, while the Geodemographic method substantially *under*estimated 1990 housing stock. For cluster 13, an urban cluster consisting mainly of high-rise apartments and condominiums, the Trended Share method underestimated 1990 housing stock, while the Geodemographic method's estimate was roughly three times the actual size of the housing stock. The Geodemographic estimate was based on a sample size of five block groups (out of eighty), and happened to contain several block groups which had grown enormously in the time period, thereby inflating the estimated growth rate for the full cluster.

Looking at how well these different techniques performed in estimating the housing stock for individual block groups, we turn toTable 4, which presents the Mean Percent Error (MPE) values, by cluster, for each of the three methods. The MPE is calculated by taking the difference between the estimate and the true value, averaged over all observations, and expressing that difference as a percentage of the true value. Ideally, we would hope for the MPE to approach zero, indicating that estimates are equally likely to both over and underestimate the true count.

We see in Table 4 that the GeoDemographic method has the lowest MPE in 8 of 30 clusters, while the Trended Share method has the smallest MPE in 11 clusters, and the Averaged Method has the smallest MPE in 11 clusters. Although Mean Percent Error can give some measure of bias in an estimator, by itself it doesn't indicate the accuracy of individual estimates. Accordingly, we calculated the Mean Absolute Percent Error for each of the three estimates. The MAPE is calculated by averaging the absolute value of the percent errors. Thus, although it can't indicate whether the model is over or underestimating the true value, it does give an indication of the size of individual errors. Results are shown in Table 5.

One of the more important things to note from Table 5 is the MAPE value for the Averaged Method. To reiterate, we are not so much comparing the *geo-demographic* method to the trended share method as we are comparing the *averaged* method to the trended share method. That is, our aim is to demonstrate that we can improve an existing set of estimates by including information for only a few areas. The averaged method has the smallest MAPE in 20 of 30 clusters. In addition, the MAPE for the averaged method is smaller than the Trended Share MAPE in 27 of 30 clusters.

Table 6 presents the overall MPE and MAPE values for each of the three methods. We see that although the Trended Share method has the lowest overall MPE, the standard deviation of the Trended Share's MPE is much larger than the standard deviations for the other two methods' MPEs. We would expect the Trended Share's MAPE, then, to be fairly large, then, since that method's larger standard deviation is indicative of wider "spread" in the errors. This is, indeed, the case: the Averaged Method has the smallest average MAPE.

Conclusions and Implications

Considering the breathtaking pace of technological change in recent decades, and the increasing calls for better information to inform decision-making at all levels of government and business, small-area population estimation methodology hasn't advanced much. Indeed most methods for preparing small-area estimates have been known at least since the 1960s. There are a number of reasons why this has occurred, but certainly data limitations are one obvious hurdle, as most small-area estimation methods continue to rely upon symptomatic data geocoded at the minor civil division or higher. As a result, such methods currently are unable to generate, with any sophistication, estimates for units of geography smaller than minor civil divisions.

Our geodemographic method offers a fundamentally different approach to leaping this hurdle. We began this study by asking two basic questions: (1) Do clusters have unique rates of population (or in our study, housing unit) growth? (2) Can we improve upon estimates by incorporating cluster type and ground truth? Based on the initial findings from our research, we can answer both questions in the affirmative. Clusters *do* appear to have distinctive rates of housing unit growth, and we *can* apply knowledge to all of the block groups within a particular cluster.

A few final words about our approach. As currently constructed, our model assumes that we only have ground truth for a small number of block groups. In this example, we produced estimates for 60,362 block groups based upon a sample of only 607 block groups--less than one percent of the total. Increasing the sample size would no doubt improve upon the accuracy of our method; we designed our model to operate with a bare minimum of ground truth from each cluster. Several of the smallest clusters have only five block groups in their sample. In reality, for many clusters there may be a great deal more ground truth available.

In addition, our clustering model does *not* include any variables that would be directly related to housing unit growth--such as percent of persons who moved within last five years, percent of housing units built in previous decade, etc. Including such variables would no doubt add greatly to the between-cluster differences in housing unit growth, thereby increasing accuracy. We opted for the most conservative approach; alternative clustering models could certainly incorporate any number of growth-sensitive measures into their clustering algorithms.

Another advantage in using a geodemographic method such as this one is the flexibility of the unit of geography. The Census Block Group is perhaps the ideal unit of geography to utilize in this type of analysis. It's large enough to contain data from the Census Questionnaire's Long Form, yet small enough to be used as a building block for constructing whatever unit of geography--census tract, school district, market area--we wish to estimate.

In summary, preliminary findings from our study suggest a definite potential for geodemographic methods in improving how we estimate housing units, and, presumably, populations. This study has demonstrated that we can improve upon existing methods by averaging with the geodemographic method. Further research in this arena, addressing questions such as, What types of clusters work best in a geodemographic method? How much "ground truth" is necessary to achieve reliable estimates for clusters? To what extent do clusters retain the same block groups over time?, could prove valuable in advancing our methods of small-area estimation.

References

Aldenderfer, Mark S. & Roger K. Blashfield. 1984. *Cluster Analysis*. Beverly Hills: Sage Publications.

Chapman, John. 1987. “Cast a Critical Eye.” *American Demographics* 9 (February):30-33.

Everitt, Brian. 1974. *Cluster Analysis*. New York: Wiley.

Goldsmith, Harold F., David J. Jackson, and J. Philip Shambaugh. 1982. "A Social Area Approach." Pp. 169-90 in *Population Estimates: Methods for Small Area Analysis*, edited by Everett S. Lee and Harold F. Goldsmith. Beverly Hills: Sage Publications.

Gurley, William R., David White, and James D. Tarver. 1965. "The Accuracy of Selected Methods of Preparing Postcensal County Population Estimates." *Estadistica* 23(86):70-90.

Merrick, Thomas W., and Stephen J. Tordella. 1988. *Demographics: People and Markets*, Population Bulletin 43,1. Washington, D.C.: Population Reference Bureau, Inc.

Pitkin, John R. 1992. “A Comparison of Vendor Estimates of Population and Households With 1990 Census Counts in California.” In *Applied Demography* 7:1: Spring, 1992.

Pittenger, Donald B. 1976. Projecting State and Local Popoulations. Cambridge, MA: Ballinger Publishing Co.

Rives, Norfleet W., Jr., William J. Serow, Anne S. Lee, Harold F. Goldsmith, and Paul R. Voss. 1995. *Basic Methods for Preparing Small-Area Population Estimates*. Madison, WI: Applied Population Laboratory, University of Wisconsin-Madison.

Romesburg, H. Charles. 1984. *Cluster Analysis for Researchers*. Belmont, CA: Lifetime Learning Publications.

Russell, Cheryll. 1984. *The Business of Demographics*, Population Bulletin 39,3. Washington, D.C.: Population reference Bureau, Inc.

Schmitt, Robert C. 1952. "Short-Cut Methods of Estimating County Population." *Journal of the American Statistical Association* 47(258):232-38.

Siegel, Jacob S., Henry S. Shryock, Jr., and Benjamin Greenberg. 1954. "Accuracy of Postcensal Estimates of Population for States and Cities." *American Sociological Review* 49(4):440-46.

U.S. Bureau of the Census. 1978. *State and Local Agencies preparing Population Estimates and Projections: Survey of 1975-76*, Current Population Reports, Series P-25, No. 723. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

U.S. Bureau of the Census. 1991. *Census Geography -- Concepts and Products*, CFF No. 8 (Rev.). Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Zitter, Meyer, and Henry S. Shryock, Jr. 1964. "Accuracy of Methods of Preparing Postcensal Population Estimates for States and Local Areas." *Demography* 1(1):227-41.

Endnotes

1. Support for this study was provided by the College of Agricultural and Life Sciences, University of Wisconsin-Madison, the Center for Demography and Ecology, University of Wisconsin-Madison through a grant from the Center for Population Research of the National Institute of Child Health and Human Development, and by the U.S. Department of Agriculture through a grant from the North Central Regional Center for Rural Development (NCRCRD Contract No. 94-4A). The authors wish to acknowledge with appreciation the assistance of Holly L. Hughes, Erin Kehoe and Swami Rajagopalan in various stages of the data analysis.

2. Census block groups (BGs) are subdivisions of census tracts and block numbering areas (BNAs). Block groups comprise all blocks with the same first digit in each census tract/BNA (U.S. Bureau of the Census, 1991).

3. These areas typically were highly localized trade areas, service territories and labor market areas.

4. The nature of census geography prior to 1990 dictated that fairly large enumeration districts (or small towns and villages, where Census Bureau EDs were larger than political geographic areas) were the building blocks of commercial estimates and projections.

5. The term "demographic" is used here (rather than "population") to signify that these commercial estimates and projections almost always included efforts to go well beyond mere estimates of current or future population. Typically estimates and projections are produced for several population and household demographic, social and economic attributes.

6. We fully admit our rather substantial lack of understanding of the details and intricacies of the population estimation models used by various demographic data vendors. These models, and tests concerning the accuracy of the estimates derived from them, are largely proprietary. This is fine. Our intent is not to critique these models but, rather, to suggest a methodology they might benefit from. Moreover, our proposed methodology is general enough that it should be possible to incorporate this refinement into the work of public agencies as well. Indeed, as will be explained, it likely is in the public sector that this proposed methodology holds the greatest promise because of the need to feed it with intensively obtained "ground truth" in selected areas. We have selected the Census Block Group as the small area with which to test our proposed procedures.

7. It has become commonplace among applied demographers to reserve the term "estimate" for attempts to model current or past populations using data (frequently called "symptomatic data") that is roughly contemporaneous with the estimate date. Pure extrapolation of a historical trend to a current date does not meet the requirements of this definition.

8. By this we mean estimates produced using one or more of several estimation models that are fed by current "symptomatic indicators" of population change for the areas being estimated or models that depend on counts of registered activities (component estimation models and

variations of the housing unit estimation model).

9. While defensible, this approach has several weaknesses. One obvious weakness is that the question on "year unit built" is a long-form question. As such, it was obtained only from the 1990 census sample and thus suffers from sampling error. Moreover, our approach must assume that the person completing the sample form *knew* when the unit was built and then filled out the form correctly and with fidelity. The extent of non-sampling errors introduced in violation of our assumption cannot be known. The best one can hope for is that the errors are small. Finally, our approach ignores diminution of the housing stock between 1980 and 1990 resulting from housing demolitions, accidental loss from fire or natural disaster, etc.

10. The estimation method used at this stage of the research is not particularly critical. Almost any method would have sufficed. But given the various constraints that confronted us we were highly limited in our choice of methodologies.

11. We used the 1990 Census question on "year housing unit was built" to calculate the number of housing units in each block group for 1970 by removing from the 1990 housing stock all housing units added subsequent to 1970.

12. The rationale here is that block groups with rapidly increasing housing stock would probably face shortages of land, thereby tempering additional growth in the housing stock in the following decade. Additionally, for block groups with stagnant or dwindling numbers of housing units, halving the 1970 to 1980 change for the period 1980 to 1990 prevents block groups from reaching the unlikely ratio of zero for 1990.

13. The discerning reader will object at this point. Using 1990 Census counts of housing units for counties is cheating and likely makes the block group estimates better than they should be. Ideally, in a defensible test of a set of small-area estimates, what we would want to use is a set of 1990 *estimates* of housing stock for each county. Unfortunately, to our knowledge, none exists. This difficulty is hardly a "fatal flaw" in the research, however. Recall, our objective is not to test the utility of a specific set of estimates. Rather, our goal is to show that those estimates (even if a bit artificial) can be *improved upon* using geodemographic techniques. Crudeness in the original estimates should not matter much for the real goal of our test.

14. These are typical measures for evaluating estimation errors. Ideally one hopes for a MPE close to zero (indicating little bias in the estimation model) and a small MAPE (indicating the absence of large errors).

Table 1: Results of Cluster Analysis

Cluster Number	Number of Block Groups	RMS Standard Deviation	Closest Cluster	Distance Between Cluster Centroids
1	393	1.27	6	6.8
2	118	1.14	27	11.3
3	3639	0.58	16	2.8
4	1046	0.95	30	5.1
5	6079	0.52	20	2.8
6	462	0.93	7	5.1
7	2198	0.63	12	3.4
8	500	0.85	23	5.2
9	4972	0.55	20	2.5
10	3303	0.62	11	3.7
11	5073	0.45	16	2.7
12	1786	0.55	7	3.4
13	80	1.49	4	12.4
14	805	0.74	16	4.2
15	368	1	8	7.4
16	7766	0.45	11	2.7
17	2860	0.75	9	5
18	2001	0.69	30	3.3
19	401	1.13	4	8.9
20	8418	0.48	9	2.5
21	55	1.45	27	8.9
22	861	0.88	7	3.7
23	699	0.78	3	4.8
24	306	1.15	14	10.8
25	1137	0.62	11	3.2
26	820	0.89	10	6
27	172	1.23	8	7.1
28	475	0.97	10	5.8
29	250	1.3	14	8.4
30	3319	0.57	11	2.8

Table 2: Percent Change in Housing Stock, 1980-1990, by Cluster

Cluster Number	Number of Block Groups	Mean % Change 1980-1990	Standard Deviation
1	393	24	52
2	118	42	66
3	3639	7	13
4	1046	84	480
5	6079	29	36
6	462	26	141
7	2198	5	18
8	500	3	9
9	4972	13	14
10	3303	79	448
11	5073	22	113
12	1786	4	39
13	80	46	173
14	805	17	39
15	368	7	19
16	7766	8	53
17	2860	11	11
18	2001	21	91
19	401	25	51
20	8418	20	16
21	55	23	114
22	861	7	18
23	699	4	9
24	306	18	52
25	1137	9	28
26	820	59	291
27	172	11	36
28	475	65	301
29	250	51	285
30	3319	32	195
Total:	60,362	19	96

Table 3: Mean Housing Unit Estimates, by Method

Cluster Number	Number of Block Groups	Mean Number of Housing Units 1990 Census	Mean Housing Unit Estimate for Cluster		
			Trended Share	GeoDemographic Method	Averaged Method
1	390	369	401	364	383
2	117	343	339	434	386
3	3638	355	342	351	346
4	1020	667	715	728	722
5	6075	467	463	460	462
6	462	492	474	546	510
7	2197	307	278	310	294
8	500	497	439	504	471
9	4971	373	373	376	374
10	3220	475	540	391	465
11	5025	425	464	610	537
12	1785	330	300	337	318
13	80	1846	1617	5497	3557
14	803	389	375	375	375
15	367	422	361	448	405
16	7752	382	367	385	376
17	2857	201	195	194	195
18	1995	504	517	494	506
19	399	449	393	430	411
20	8416	396	395	384	389
21	54	481	459	453	456
22	857	281	260	296	278
23	697	324	293	335	314
24	267	266	255	272	263
25	1123	396	418	457	438
26	805	358	325	342	334
27	170	411	373	442	407
28	467	538	629	646	637
29	158	135	128	141	134
30	3302	550	572	539	556

Table 4: Mean Percent Error (MPE) for Methods, by Cluster

Cluster Number	Number of Block Groups	MPE and Standard Deviations for Clusters, by Method					
		Trended Share		GeoDemographic		Averaged	
		MPE	Std Dev	MPE	Std Dev	MPE	Std Dev
1	390	6.35	63.62	0.18	20.96	3.27	35.22
2	117	-0.55	24.13	27.81	22.17	13.62	19.9
3	3638	-4.33	17.16	-0.69	8.57	-2.51	9.97
4	1020	5.61	91.09	14.11	28.59	9.86	48.83
5	6075	-0.78	28.61	1.22	16.87	0.22	18.51
6	462	-3.44	42.18	9.05	19.91	2.8	25.04
7	2197	-9.61	11.05	0.98	7.61	-4.31	8.01
8	500	-12.14	11.62	1.64	6.72	-5.25	6.99
9	4971	-0.24	16.46	1.59	9.31	0.67	9.93
10	3220	10.26	129.61	-8.67	25.16	0.79	66.29
11	5025	7.27	129.74	49.93	25.25	28.6	65.93
12	1785	-9.93	14.86	2.02	6.32	-3.95	8.67
13	80	-12.06	26.42	191.24	69.83	89.59	44.99
14	803	-3.5	23.27	-2.06	13.95	-2.78	14.97
15	367	-14.38	13.4	6.56	10.31	-3.9	10.28
16	7752	-4.82	34.67	2.07	10.03	-1.38	17.94
17	2857	-3.19	11.8	-3.49	8.19	-3.34	8.73
18	1995	-0.36	45.48	0.28	16.36	-0.03	24.58
19	399	-11.41	21.68	-1.41	19.43	-6.41	18.28
20	8416	-0.37	12.91	-2.26	10.5	-1.31	9.59
21	54	-1.72	34.53	-3.88	15.75	-2.81	20.95
22	857	-7.52	17.96	5.36	10.58	-1.07	12.09
23	697	-10.03	11.52	3.73	6.98	-3.15	7.3
24	267	-4.04	40	4.53	17.6	0.24	23.77
25	1123	1.77	116.16	17.2	13.59	9.49	57.95
26	805	-7.51	60.22	0.84	23.9	-3.33	34.39
27	170	-9.78	16.31	6.35	10.87	-1.72	10.41
28	467	11.18	78.58	28.94	33.17	20.06	42.49
29	158	-9.87	26.5	9.25	24.14	-0.31	22.1
30	3302	0.76	71.04	4.44	20.72	2.6	36.97
TOTAL	60,362	-1.1	59.7	5.4	22.5	2.1	32.5

Table 5: Mean Absolute Percent Error (MAPE) for Methods, by Cluster

Cluster Number	Number of Block Groups	MAPE and Standard Deviations for Clusters, by Method					
		Trended Share		GeoDemographic		Averaged	
		MAPE	Std Dev	MAPE	Std Dev	MAPE	Std Dev
1	390	29.6	56.6	16.1	13.3	18.4	30.2
2	117	15.4	18.5	30.8	17.8	19.5	14.1
3	3638	9.5	14.9	5.9	6.2	5.6	8.6
4	1020	32.5	85.2	28.2	14.9	23.4	43.9
5	6075	15.8	23.9	13.3	10.4	12.1	14
6	462	22.4	35.9	19	10.8	14.7	20.4
7	2197	11.5	9	5	5.8	5.8	7
8	500	14.7	8	4.9	4.9	6.4	5.9
9	4971	8.2	14.2	7.4	5.9	6.2	7.8
10	3220	40.2	123.6	17.6	20	23	62.1
11	5025	30.8	126.2	53.2	17.3	32.6	64
12	1785	12.7	12.6	4.4	4.9	5.7	7.7
13	80	19.2	21.7	195.6	56.1	96.2	27.9
14	803	13.1	19.5	10.2	9.7	9.3	12
15	367	16.3	11	10.22	6.7	5.9	9.3
16	7752	12.5	32.7	7.6	6.8	6.6	16.7
17	2857	8.2	9	6.2	6.3	6.5	6.7
18	1995	18.5	41.5	12.3	10.7	11.9	21.5
19	399	17.7	16.9	14.6	12.8	12.9	14.4
20	8416	8.7	9.6	8.2	7	7	6.7
21	54	17.7	29.6	7.9	14.1	10.8	18.1
22	857	12.6	14.8	10	6.4	6.9	10
23	697	11.9	9.5	6.6	4.3	5	6.1
24	267	17.9	36	14.9	10.4	12.4	20.3
25	1123	30.1	112.2	20.7	7.1	14.4	57
26	805	25.1	55.3	18.1	15.6	15	31.1
27	170	14.2	12.6	11.1	6	6.1	8.6
28	467	36.2	70.6	40.1	18.2	30.2	36
29	158	16.7	22.8	21.4	14.4	13.6	17.4
30	3302	23.1	67.2	17.1	12.5	13.9	34.3
TOTAL	59,969	16.6	57.3	14.5	18.1	11.8	30.3

Table 6: OVERALL MPE and MAPE, by Method

Method	Average MPE	Standard Deviation	Average MAPE	Standard Deviation
Trended Share	-1.1	59.7	16.6	57.3
GeoDemographic	5.4	22.5	14.5	18.1
Averaged	2.1	32.5	11.8	30.3

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: perry@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu