EQUAL CHARACTERISTIC CLUSTERING

Patrick J. Cantwell, Bureau of the Census[1]
Bureau of the Census, SRD, Washington, DC 20233

KEY WORDS: Noncompact clusters, variance reduction, demographic surveys.

1. Introduction

In most government household surveys, clusters of nearby housing units are in sample at the same time. If the intracluster correlations for characteristics to be estimated are moderately high, such clustering can increase the variances of the estimates, compared with sampling isolated units. However, the resultant decrease in cost often makes clustering worthwhile on balance (Cochran 1977).

This paper does not revive the argument whether or not to cluster housing units. Instead, for surveys which cluster, we examine the manner in which clusters are formed.

The most common methods of constructing clusters are combining consecutive housing units, or taking a systematic sample of housing units in a small area. These operations are simple to implement, and make it easy for the field interviewer to locate the sample units. However, they ignore information which is available for each unit.

In most areas, characteristics from decennial census records can be obtained which exhibit correlations with the target variables, those to be estimated in surveys. By exploiting these correlations, one might hope to form clusters in which the average of the target variable is fairly constant from cluster to cluster. If this is done, an important component of the variance of the estimate can be decreased substantially, as will be described shortly. This method of combining housing units is called equal characteristic clustering (ECC). The variables which are known before the survey is taken are referred to as "balancing" variables; they are used to equalize the clusters.

The idea of equalizing clusters developed in the mid 1980's, as a more specialized method called equal person segmentation. By balancing the number of people in the clusters, it was hoped that variances could be reduced. An experiment was conducted where equal person segmentation was introduced into certain parts of the Current Population Survey (CPS) sample. As far as we can determine, the data from the experiment have not yet been analyzed. Recently, Gary Shapiro of the Census Bureau suggested forming clusters by balancing other characteristics. His ideas motivated this study on ECC.

This study investigates how well ECC performs using actual data from the 1984 Survey of Income and Program Participation. The entire sample was treated as a primary sampling unit. Several types of clusters were formed--compact, random, and others by ECC--and the variance components of estimates of target variables were computed for each. Various characteristics were used, some as balancing variables, others as target variables. Other factors, such as the size of the cluster or the length of the string of housing units from which to cluster, were varied. Ratio and non-ratio estimators were used in the appropriate settings.

From the results of the study, it appears that ECC has limited potential in the major household surveys. The variances components studied for some characteristics can be reduced moderately using certain balancing variables, sometimes by 15% to 25%, usually by at least 10%, under the proper conditions. But the most effective "balancing" variables are not always available on the census 100% ("short form") detail file, the file which contains responses from all households in the country.

A major drawback of the technique is how well it works when the information used to form the clusters is several years old. By the time the first samples from the 1990 redesign are phased in, information from the decennial census will be at least four years old. Later samples will use census data which are more outdated.

This problem was addressed through a stability simulation. Housing units were combined so that the numbers of people per cluster were as nearly equal as possible. Using longitudinal data from the American Housing Survey, the sizes of the housing units were projected seven years and fourteen years later. The variability of the cluster sizes was compared in the later years.

The results indicate a serious lack of stability over time in the values used to cluster the housing units. This implies a loss in much of the effectiveness of ECC.

Section 2 of the paper describes methods to form clusters of housing units used at the Census Bureau. The method of ECC is described in Section 3. Sections 4 through 6 provide the main results of this study. In 4 and 5, details of the SIPP data file, how clusters were constructed, and results using ECC methods are given.

In Section 6 the stability of cluster sizes over time is examined through a simulation on American Housing Survey data. Section 7 contains a summary, further areas for investigation, and some practical ideas for ECC.

2. Methods of Clustering Used by the Census Bureau

The most common method of clustering used by the Bureau is to interview a fixed number of consecutive housing units. These "compact" clusters are assembled from census address lists in several surveys conducted by the Bureau. CPS and the National Crime Survey sample clusters of four consecutive housing units. (The Census Bureau defines a household and a housing unit equivalently.) Some surveys, such as the Consumer Expenditure Surveys and the American Housing Survey, retain an unclustered design.

An alternative is a "noncompact" cluster, a group of housing units from the same block or neighborhood, but not in consecutive order. For example, to form systematic noncompact clusters of size four from a string of forty housing units, one might combine the first, eleventh, twenty-first and thirty-first units into the first cluster. Nine other

clusters are formed similarly.

Currently the Census Bureau uses systematic noncompact clusters or "measures" of size four where area listing and sampling is done. The National Health Interview Survey, which has mostly an area sample, combines two consecutive measures to create clusters of size eight. Because the intracluster correlation tends to decrease with the size of the string, variances for designs which use noncompact clusters typically fall between those which use compact clusters and unclustered units.

This paper deals only with forming clusters where sampling is done from census address lists. Not only are addresses available before interviewers go into the field, but data obtained in the decennial census can be used in clustering.

To assign sample for any survey, the United States is first divided into many primary sampling units (PSUs). In the Bureau's surveys, a PSU is typically the size of one or several counties, or a part of a large county. A number of these are selected appropriately for the sample.

To simplify matters, a PSU could be divided into strings of housing units which we label "segments" or hits. A segment might consist of, for example, 40 units. From each segment, compact or noncompact clusters can be formed in any specified manner. Typically, the length of the segment is determined by multiplying the number of units per cluster (for any one sample) by the number of different samples needed by the survey over the ten years of the design.

Within a PSU selected for sample, a simple random sample of segments is drawn. This sampling is typically done systematically at the Bureau. However, srs is assumed here to compare variance results.

From each chosen segment, one cluster is selected randomly for the first sample of the decade. A second cluster from this segment will be chosen for the next sample, and so on. Therefore, within the PSU, the sample is selected randomly in two stages--a number of segments are drawn first, and then one cluster from each segment is selected. How the clusters are formed will affect the variance of the estimators. It is important to realize that most of the clusters within any segment are not in sample at any point in time.

3. Forming Clusters by Equalizing Characteristic Levels

Let Y represent a target variable, one which is to be estimated from the sample. In this study, only the within-PSU variance of an estimator is studied. However, Banks and Shapiro (1971, p.43) have shown, at least for the CPS, "that the overwhelming component of variance is within-PSU variance rather than between-PSU or between-stratum variance." In their tables, within-PSU variance accounts for 90% to 99% of the total variance for most of the important characteristics.

Under the sampling scheme described, the within-PSU variance itself has two components. The first is the between-segments component, which is a function of the variability among segment means. For a given segmenting of the PSU, this component of the variance is the same for all clustering methods. It represents a lower bound below which no clustering method can decrease the variance, unless the segments are redefined. When we formed clusters of size four (two) from strings of 40 (20) units, however, this percentage was below 20% of the total within-PSU variance for all characteristics studied.

The more prominent part of the variance is the within-segments component, which depends on the variability of the cluster means. For each sample, one cluster is selected from each segment. The idea behind equal characteristic clustering (ECC) is to form clusters within segments so that the cluster means for Y are as nearly equal as possible. The within-segments component can be reduced, and with it, the entire variance.

One immediate problem is that Y is unknown for each unit. This makes it impossible to form clusters with equal y values. Instead, a "proxy" variable or group of variables must be used. Certain information, which we call the "balancing" variable(s) X (X_1, ..., X_m), is available about the housing units from the census or another source. Possible choices include the number of people in the housing unit at census time, the tenure of the unit--whether the residents own or rent it--or the race or sex of the householder.

The plan is to select one or more of these variables which are known and are highly correlated with Y. By forming clusters which have fairly constant means for one or several of the X's, one hopes that the cluster means for Y become more nearly equal. Whether this actually happens depends largely on the strength of the correlation between Y and X.

For example, suppose clusters of size four are used in estimating household income, and X is tenure (i.e., owner/renter). If it is known that about 75% of the people in the area are renters, we would try to create clusters containing three renters and one owner. It is hoped that equalizing the number of renters per cluster will reduce the variability of household income totals among the clusters.

4. The SIPP Data Study--Background

To investigate how well ECC performs, a sample data file was obtained. The answers to seven questions from all 14,722 housing units responding in Wave 7 of the 1984 panel of the Survey of Income and Program Participation (SIPP) were made available. Only the first 14,400 records were used in this study, to simplify breaking the PSU into segments of various sizes.

The variables for each household were:

PEOPLE: the number of people in the household,
TENURE: the tenure of the unit--whether it is owned, rented, or occupied (but not owned) with no payment,
COLLEGE: the number of people with some college education,
EMPLOYED: the number of people employed,
SOCSEC: the number of people who are on social security,
HHINCOME: the household monthly income, and
PPEARNINGS: the principal person monthly earnings.

Record entries for HHINCOME and PPEARN were truncated at $50,000 (per month). This forced only eight

records to be edited.

For the seven variables, the means, standard deviations, and correlations for all pairs are provided in Table 1. We were interested in estimating the PSU total for the Y variables EMPLOYED, SOCSEC, HHINCOME and PPEARNINGS. PEOPLE, TENURE and COLLEGE were used to equalize the clusters. PEOPLE and TENURE are available on the census short form; COLLEGE is not.

The 14,400 housing units in the SIPP file were treated as a PSU. Segments were defined by taking consecutive strings the length of the desired segment. Within these segments, clusters of size four (and later two) were created through a variety of clustering algorithms.

In this study, "compact" clusters were defined merely by joining the first four housing units, the next four, etc., until the segment was depleted. "Random" noncompact clusters were obtained by randomly selecting and combining four units from the segment, and then repeating this process among the remaining units in the segment. Although the noncompact clusters we study were formed randomly rather than systematically, as described in Section 2, the variances resulting from the two methods are probably relatively close, compared to those for compact or ECC clusters.

Finally, ECC was used to produce clusters of four housing units. The study looked at clusters formed by "equalizing" any one of several X variables, any two, or any three. When a single X is used, the segment records are ordered on that X value. Then the first and last records are combined into a cluster of size two, the second and second last are combined, etc. These cluster totals are now ordered on the X value, and clusters of size two are combined: the first and last, the second and second last, etc. If clusters of size two are desired, the second step is omitted.

If two X's are used for ECC, the clusters are first balanced on X_1 in the manner just described. Wherever two or more records in a segment have the same value of X_1, the records are then ordered on X_2. Equal values of X_1 are common when the set of X_1 responses is small, or when X_1 has been recoded. Three or more X's would be treated in an analogous fashion.

The order in which the X variables are introduced affects the cluster formation and the resulting variances. If X is essentially a continuous variable, X might have to be recoded. Otherwise, subsequent X's will have little or no influence on the balancing algorithm.

In a real PSU, many characteristics are somewhat homogeneous within neighborhoods. But the 1984 SIPP panel used for this study is a set of records from housing units across the country. Thus the "compact" clusters formed and analyzed here will generally not exhibit the higher intracluster correlations of genuine compact clusters formed from the census file. The "compact" variances we compute will underestimate the true variance of compact clusters. Accordingly, the results obtained will be conservative in evaluating the effectiveness of ECC. Using the census file instead of the SIPP sample file would not help much, because most of the characteristics to be estimated are not available on the census short form.

5. The SIPP Data Study--Results

To study ECC clustering, HHINCOME, PPEARNINGS, EMPLOYED, and SOCSEC were used as target (Y) variables. The balancing variables (X's) were PEOPLE, TENURE, and COLLEGE. Non-ratio (unbiased) estimates were used for measurements based on the number of households--HHINCOME and PPEARNINGS. Where estimates are based on the number of people, i.e., for EMPLOYED and SOCSEC, ratio estimates are substituted. Formulae for each type of estimator and their variances are found in Cochran (1977). The methods of clustering are compared by examining their variances in estimating Y as a percentage of the variance for compact clustering.

The 14400 records were split into 360 segments of 40 housing units. Within each segment, groups of four units were combined via compact and random clustering, and ECC on one of the X variables (three possibilities), ECC on any two X's (six possibilities) and ECC on any three X's (six possibilities). Recall that, when clusters are balanced on more than one variable, the order of the X's affects the variances.

In estimating HHINCOME, note first its correlations with the X variables: with PEOPLE, .234; with TENURE, -.204; and with COLLEGE, .385. It should be apparent that the sign of the correlation is not important in balancing clusters, only the strength of the correlation in absolute value.

"Random" clustering can lower the variance 11.8%. The results of the ECC method are shown in Table 2. Using PEOPLE, TENURE, or COLLEGE alone, the variance can be reduced 17.3%, 15.2%, or 20.1%, respectively. Balancing on either or both of the other two X's after the first brings out little or no improvement. Often the variance increases. Although COLLEGE is the most effective balancing variable for estimating HHINCOME, it is not on the census short form.

When the target variable is PPEARNINGS, the conclusions about ECC are similar. The correlations between the X's and PPEARNINGS are smaller than the corresponding ones between the X's and HHINCOME. It is not surprising then that variance reductions for PPEARNINGS under ECC are smaller than those attained for HHINCOME. Again, adding a second or third variable helps little or not at all. PEOPLE, TENURE, and COLLEGE produce variance reductions of 11.9%, 6.9%, and 13.9%, respectively, when acting alone. Using random clusters here reduces the variance by 7.4%.

Table 2 also contains results for estimating EMPLOYED and SOCSEC using ratio estimation. For the latter, ECC yields greater variance reduction--28.2% balancing on PEOPLE, 31.2% on PEOPLE and TENURE (in that order). Random clustering reduces variance by 8.4%.

The effect of ECC is smaller when the target is EMPLOYED. The variances are reduced from 11% to 17% under the various ECC combinations.

These tendencies were checked under other conditions. ECC works slightly better as the segment size increases. See Table 3 for a comparison of the variances with segment sizes

of 20, 40 and 80 units. ECC clusters of size two were also formed and examined. The trends seen before are repeated here. Corresponding variance reductions were smaller everywhere than with clusters of size four.

6. Stability Simulation

Perhaps the most serious concern with using ECC in the major household surveys is that of the stability of the clusters based on the X variables. An assumption intrinsic to the operation of ECC is that the X values used for balancing the clusters are accurate and current. It is unlikely that these conditions could be met if 1990 census data are the source of this information. Clusters would originally be equalized based on fairly accurate census values. However, these data will be four or five years old by the time the sample is phased in, and close to fifteen years old as the 1990 design winds down. The chance that many X variables will have changed between census time and sample phase-in is great.

To see the effect of using 1990 X values to help estimate Y characteristics in, say, the year 1998, one would like to observe the correlations between the pairs of variables eight years apart. Unfortunately, records for the same housing unit many years apart are not easily obtained. We used an indirect approach to get an indication of how stable the data are over time.

In an earlier study (U.S. Dept. of Commerce, Cahoon 1982) using the American Housing Survey (AHS) Longitudinal (abbreviated) File, Cahoon tabulated housing unit sizes for the same unit in 1973 and 1980. Data from the file included every housing unit interviewed from 1973 to 1980. Empty units were eliminated. After removing all noninterviews, supplements, special places, the later new construction, and the CENSUP units, there remained 35,034 units with good interviews in 1973 and 1980. In the tabulation, Cahoon truncated housing unit sizes at ten so that SPSS output would fit on a single page.

It should be mentioned that AHS follows addresses rather than individuals or families within houses. If the residents of Address A moved to Address B between 1973 and 1980, the housing unit size in 1980 is the number of people who are then living in Address A. This is consistent with CPS sampling procedure. On the other hand, the SIPP follows movers to their new residence if that place is within one hundred miles.

Associated with each housing unit in the AHS table are unit sizes, i.e., the number of people in the unit in 1973 and 1980. This provides not only marginal frequency distributions for the two years, but also conditional distributions for the 1980 unit size, given the size in 1973. In 1973, the mean number of people per unit from this table was 3.02, with a coefficient of variation of .566. (Recall that some units, including empty houses, were removed from the file.) In 1980, the mean number per unit dropped to 2.77, but the c.v. remained at .567. These numbers are compatible with the summary statistics from the 1984 SIPP file, Wave 7: the mean per unit is 2.70, with a c.v. of .565.

If the AHS file had been available, actual segments could have been formed from nearby units for this stability study. Our best alternative was to simulate segments of housing units from Cahoon's table. The segments generated here are then assembled from a composite of the whole country, rather than from within a PSU.

In the first part of the simulation, 10,000 segments of 40 housing units were randomly generated according to the 1973 distribution of unit sizes in the AHS table. The sizes of these housing units are assigned to "Year 0," corresponding to a census year. Within each segment, ten clusters were constructed in two ways: "randomly," that is, by combining any four units, and through ECC on the unit size.

Using the conditional frequency distributions for unit size in 1980 given that in 1973, as specified in the AHS table, subsequent housing units sizes were generated for each unit in "Year 7." The same distributions were used a second time to project a change in unit size to "Year 14." These years correspond to periods in the middle and near the end of the survey design. If unit sizes for the same units were available in 1987, they would have been used for Year 14 instead.

Each housing unit has three sizes associated with it, sizes for Years 0, 7 and 14. The clusters formed in Year 0 are followed in the subsequent years. No reclustering is done in Years 7 and 14.

For each of the three years, the standard deviation of the cluster size was compared for random and ECC clusters. A comparson of the ratio of the standard deviations--ECC to random--indicates how stable the ECC cluster formation remains after seven and fourteen years.

The results are displayed in Table 4. For segments of 40 housing units, the ratios in Years 0, 7 and 14 are .400, .864, and .964, respectively. After seven years, according to these data, much of the smaller variability of cluster sizes is lost; after fourteen years, almost all of it is gone.

One way to view these numbers is to note that in Year 0, the cluster size variability under ECC on PEOPLE is 40.0% that of random clustering. With current data from the 1984 SIPP data file, the variance reductions under ECC on PEOPLE, compared to random clusters (rather than compact clusters) are 5.5% in estimating HHINCOME, 0.9% for EMPLOYED, and 19.2% for SOCSEC. Seven years later, there is not much difference in cluster size variability between the two clustering methods. It seems reasonable to guess that much of the variance reduction from ECC would also be lost with old data. Fourteen years later, the variances under ECC would be almost the same as with the original random clusters.

The stability simulation was also run with segments of 20 and 80 housing units. The ratio of standard deviations for the cluster sizes comparing random to ECC are also included in Table 4. For segment sizes of 20, 40 and 80 units, the ratios in Year 0 are .503, .400 and .333, respectively. However, in later years the ratios are almost the same for different segment sizes: about .87 after seven years and .97 after fourteen years.

Clusters of two housing units were also examined for stability in size. The trends observed with clusters of size four were generally repeated here.

This stability study has several shortcomings. Foremost among them is that actual variances in later years based on

information obtained earlier cannot be computed. Secondly, only one variable is recorded for each unit in 1973 and 1980, the number of people in the housing unit (PEOPLE). But the results imply that much of the variance reduction obtained through ECC clustering on any variable will be lost as the cluster totals become more variable with time. If the sample is taken four to fifteen years after the X variables are recorded, the ability to balance clusters is greatly impaired.

7. Conclusions and Further Study

It appears that equal characteristic clustering can be effective in estimating certain variables, but does not work consistently well. Between some pairs of variables used in this study, moderate gains were observed. For example, estimation of SOCSEC responds particularly well to ECC. In the case of ratio estimates, part of the gains to be achieved through ECC may have already been attained through ratio estimation.

Longer segments generally produce better ECC variance reduction, but not in all cases. Similarly, the ECC technique performs better in clusters of size four than of size two. With more housing units to manipulate in a cluster and the segment, the X variables can usually be balanced better. However, this does not always equalize the target (Y) variable better.

The results of the stability simulation cast doubt on ECC's effectiveness when using data from the decennial census. As the information used for balancing the clusters becomes outdated, the clusters fail to retain their smaller variability based on the X variables. This will likely mean less variance reduction under ECC.

There are other details which should be investigated before ECC is fully evaluated. First, other data files should be run and other characteristics investigated. If the results already obtained are repeated, the conclusions about ECC will be confirmed. It would be useful to determine if particular characteristics almost always work well in ECC either as X or as Y variables.

We have obtained for further research a file with AHS responses for the years 1974, 1977, 1981, and 1985. In addition to some of the variables studied in this paper, the AHS file includes the sex, race, and ethnicity of the householder (all available on the census short form), unemployed status, welfare and unemployment recipiency, and others.

Because each household on the file has responses from some or all of the four years mentioned, we can also test our doubts about stability of the clusters over time. Clusters can be formed from census variables using 1974 data. Then variance reductions under ECC can be monitored as they apply to the data in years 1977, 1981, and 1985.

It was disappointing to observe that, after one variable was used to balance clusters, adding a second or third brought out little or no improvement. An interesting possibility is to use a function of several X variables instead. This could allow the second or third X variables to contribute more (perhaps equally) to the clustering algorithm.

If particular X variables prove to be consistently more effective, they could be weighted more heavily. Still, this would make the clustering more complicated. The original ECC scheme assumes nothing about the X's; it requires only their values in the segments selected for sample. Our initial attempts replacing the X's with various functions of the X's have produced slightly better results.

Finally, we express our thoughts on the future of ECC. It appears that this clustering technique can produce minor improvements in the major household surveys. Whether it is worth the added cost in sample programming and interviewer travel is questionable. However, ECC could prove very beneficial in other types of surveys which cluster.

Occasionally follow-up surveys are used on a fraction of the entire sample, or on retired sample. An initial set of responses may be available from a large sample. These data might be used to screen the original sample for members with particular traits. Once the target variables for the follow-up survey are determined, correlations between these and other responses from the original survey can be computed and can indicate which variables should be used for the clustering.

Advantages in this setting are numerous. The information available from the original survey should be much more extensive than that from the decennial census. From a greater assortment of characteristics, there should be some which have high correlations with the target variables. In addition, because the follow-up survey relies on the screening information being up-to-date, the responses from the original survey are likely to be fairly current. The variance reductions would be greater and more stable.

ACKNOWLEDGMENTS

I would like to thank Lynn Weidman for many beneficial discussions on ECC cluster formation, Todd Williams for preparing the SIPP data file used in this study, and Promod Chandhok for reviewing the paper and making helpful suggestions.

REFERENCES

BANKS, M.J. and SHAPIRO, G.M. (1971). "Variances of the Current Population Survey, Including Within- and Between-PSU Components and the Effect of the Different Stages of Estimation," Proceedings of the Social Statistics Section, American Statistical Association, p. 40-49.

COCHRAN, W. G. (1977). Sampling Techniques, 3rd Edition. John Wiley and Sons, New York, N.Y.

U.S. Department of Commerce, Bureau of the Census. Internal memorandum from Cahoon to Shapiro, "CPS Redesign - Stability of Household Size," October 28, 1982. ID# F1-22.

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

TABLE 1: THE VARIABLES

	PEO	TEN	COL	EMP	SOC	HHI	PPE
Mean:	2.70	1.39	0.77	1.26	0.39	2446	1317
Std. Dev.:	1.52	0.55	0.90	1.01	0.67	2614	2081

Correlation Matrix:

PEOPLE	1.00	-.14	.24	.51	-.20	.23	.17
TENURE	-.14	1.00	-.12	-.13	-.12	-.20	-.12
COLLEGE	.24	-.12	1.00	.40	-.20	.38	.29
EMPLOY	.51	-.13	.40	1.00	-.46	.41	.32
SOCSEC	-.20	-.12	-.20	-.46	1.00	-.14	-.30
HHINCOM	.23	-.20	.38	.41	-.14	1.00	.80
PPEARN	.17	-.12	.29	.32	-.30	.80	1.00

TABLE 2: VARIANCE COMPARISON

Each segment contains 10 clusters of size 4.

Within-PSU variances, as a fraction of that obtained under "compact" clustering.

	Y (target) Variable[1] :			
	HHINCO	PPEARN	EMPLOY	SOCSEC
"Random" Clustering:	.882	.926	.875	.916
ECC on X (balancing) Variable[2] :				
PEOPLE	.827	.881	.866	.718
TENURE	.848	.931	.877	.874
COLLEGE	.799	.861	.832	.869

NOTES:

1 Ratio estimates and their variances are compared for Y variables EMPLOYED and SOCSEC.

2 Variance results given here only where ECC was used on one X variable. Adding a second or third balancing variable after the first generally did not decrease the variance.

TABLE 3: SEVERAL SEGMENT SIZES

Variances for Y (target) variable: HHINCOME.

All clusters of size four.

Segment Size:	**20 Units**	**40 Units**	**80 Units**
"Random" Clustering:	.911	.822	.862
ECC on X (balancing) variable:			
PEOPLE	.874	.827	.796
TENURE	.872	.848	.831
COLLEGE	.840	.799	.752

TABLE 4: STABILITY SIMULATION RESULTS

All table entries are the ratio of:

The std. dev. of cluster size under ECC on PEOPLE,

to: The std. dev. of cluster size under "random" clustering

==========

Clusters with 4 housing units:

Segment Size:	**20 Units**	**40 Units**	**80 Units**
Year 0	.503	.400	.333
Year 7	.876	.864	.862
Year 14	.968	.964	.965

==========

Clusters with 2 housing units:

Segment Size:	**10 Units**	**20 Units**	**40 Units**
Year 0	.620	.534	.485
Year 7	.900	.881	.872
Year 14	.974	.967	.967