CLUSTERING IN DEMOGRAPHIC SURVEYS: LONG-TERM RESULTS

Patrick J. Cantwell, Bureau of the Census, Washington, DC 20233*

KEY WORDS: equalizing characteristics, variance reduction, cluster stability.

1. Introduction and Summary

This paper evaluates several methods of forming clusters of housing units for sampling. We compare the variances of key estimates when clusters are formed compactly, systematically, and by methods which use census information obtained from the units. Through a study based on longitudinal survey records, the effectiveness of the several clustering techniques is compared as the census information becomes outdated.

The Census Bureau is planning the sample selection for the years 1994 through 2004 for the major household surveys it conducts, mostly for other government agencies. In this new design, several surveys will continue to sample housing units in "compact" clusters. They will select and interview four consecutive housing units from the 1990 census list. Compared to sampling isolated units, the decrease in travel costs is thought to outweigh the small increase in variances due to clustering.

This clustering method ignores responses to the census which are available for each unit on the list. Can we use this information to form clusters which will reduce the variance of estimators?

In a recent paper (Cantwell 1990) a new method of forming noncompact clusters of housing units was studied. This method, called equal characteristic clustering (ECC), tries to use information from decennial census records (the X variable(s)), such as the number of people in the household, or the sex and race of the householder. Through ECC, we hope to decrease the variance of target (Y) variables, such as household income or welfare recipiency.

That study investigated how well ECC performs with (then) current information using records from the 1984 Survey of Income and Program Participation (SIPP). We compared variance components under ECC and compact clustering. Although ECC showed measurable reductions in variance, the SIPP study did not answer our main concern with using 1990 census information. Through the years, as people move in and out of the housing units which will eventually be in sample, how stable are the ECC clusters formed from 1990 data?

We address this issue by testing ECC on a longitudinal file obtained from the American Housing Survey Metropolitan Sample (AHS-MS). For most records, responses were available for the same housing units for four survey years from 1974 through 1985. This allowed us to create ECC clusters based on 1974 responses. Think of these X's as census responses. We then compare the variances of target variables based on several clustering methods--compact, systematic (noncompact), and ECC--for each of the four survey years.

We summarize our results on the effectiveness of ECC as it was applied to AHS data from the Los Angeles metropolitan area. Note that these results are descriptive in nature.

- Soon after clustering (1974 and 1977), ECC performed moderately well. Compared with compact clustering, ECC reduced the variance of most Y variables, often by more than 10%.

- After 6 years (1980) ECC clusters lost some of their advantage. ECC beat compact clustering most of the time, but generally by smaller amounts. About half the time, systematic clustering now yielded smaller variance than ECC.

- After 11 years (1985) ECC was no better than other methods. ECC worked better than compact clustering for some variables, worse for others. Systematic clustering worked better than ECC for most of the variables we studied.

- Balancing clusters on two or three X variables (rather than one)--"multiple ECC"--to form clusters generally produced no further reduction in variance.

In light of these results, we see little chance for implementing ECC. We investigated clusters formed from responses 0 to 11 "years old." Surveys would actually sample from clusters originally formed from responses now 4 to 15 years old. Our results probably underestimate the instability of ECC clustering.

Census information cannot be kept current and accurate with a mobile population. Further, if a survey wants to cluster noncompactly, systematic clustering is simpler and works about as well over the sample decade.

2. Methods of Clustering

To assign sample for any survey in the 1990 Design, the Census Bureau first divides the United States into

many primary sampling units (PSUs). A PSU is typically the size of one or several counties. To simplify matters, a PSU could be divided into strings of housing units which we label "segments." A segment might consist of, for example, 40 housing units. From each segment, units or clusters can be assigned in any specified manner. For most surveys, the number of units or clusters assigned to a segment is the number of different samples needed by the survey over the ten years of the design, 1994 to 2004.

To determine its sample, a survey selects a number of PSUs by some appropriate method. Within any chosen PSU a simple random sample of segments is drawn. The Census Bureau actually selects segments systematically. However, we assume simple random sampling here to compare variances.

Within each chosen segment, one cluster or housing unit is selected for the current sample. A second cluster from this segment is chosen for the next sample, and so on. Therefore, within the PSU, the sample is selected in two stages--a number of segments are drawn first, and then one cluster from each segment. How the clusters are formed will affect the variance of the estimators.

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 1990 census can be used to cluster.

In drawing sample from census address lists, most of the major household surveys in the 1990 Design will select either single housing units or compact clusters, i.e., four consecutive housing units. Reasons why the different surveys select units or compact clusters are given in Cantwell (1990) and will not be repeated here.

Alternative methods of clustering usually fall between these two extremes. A noncompact cluster is a group of housing units from the same neighborhood, but not in consecutive order. A survey might want to separate its housing units slightly, to keep neighbors from being in sample simultaneously, or to lessen the impact from clustering on variances.

One option is a systematically noncompact cluster. For example, to form 10 clusters of size four from a segment of 40 housing units, one might combine the 1st, 11th, 21st, and 31st units into the first cluster. Nine other clusters are formed similarly. Under current plans, the Census Bureau will use systematically noncompact clustering only in its area frame.

Our method of ECC also forms noncompact clusters. However, unlike the two options just described, ECC tries to reduce the variance of estimators by using information which is available about the housing units on the census list.

3. Forming Clusters by Equalizing Characteristic Levels

Let Y represent a target variable, one we wish to estimate from the sample. In this paper, we address only the within-PSU variance of an estimator. 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 important characteristics.

Under the sampling scheme we described in the last section, 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 compact clusters of size four from strings of 40 units in the AHS data study, this component ranged from .08 to .55 of the total within-PSU variance for the characteristics we examined.

Usually, the larger 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.

Equal characteristic clustering (ECC) tries 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.

The chief problem with ECC 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_k) is available about the housing units from the census or another source. Possible X's 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.

We select one or more balancing variables which are known and highly correlated with Y. For example, suppose that we are forming 10 clusters from a segment of 40 units. If we know from the census that there are 90 people living in these 40 units, we try to form ten clusters with 9 people in each. Or if about 25% of the householders in this segment are owners, we try to put one owner and three renters in each cluster.

By forming clusters which have fairly constant means for one or several of the X's, we hope to make

the cluster means for Y more nearly equal. Whether this actually happens depends largely on the strength of the correlation between Y and X.

We applied equal characteristic clustering to data from Wave 7 of the 1984 panel of the Survey of Income and Program Participation (SIPP) (Cantwell 1990). For several Y variables, we compared variances under compact clustering and ECC. When the X values were completely up-to-date, ECC effectively reduced the variance of some Y's, but did not work well for all of them.

However, left unanswered in that study was how stable ECC clusters are over time. Clusters would originally be equalized based on accurate 1990 census values. These data will then be four or five years old by the time the sample is phased in, and close to fifteen years old when the last clusters are in sample. The chance is great that many X values will have changed between 1990 and the time of interviewing.

4. The AHS Data Study--the Longitudinal File

To analyze the long-term stability of ECC clusters--how well they reduce variances after several years--we obtained parts of a longitudinal file put together from the American Housing Survey Metropolitan Sample (AHS-MS). For the Los Angeles metropolitan statistical area, the file contains records for 1974, 1977, 1980, and 1985.

Each record represents one household, with all responses for the four years in sample. If someone in sample moves, AHS interviews the new residents. We can form ECC clusters based on 1974 X values, and monitor the variance reduction (compared to compact clustering) from 1974 to 1985. Think of 1974 as a census year--comparable to 1990--when the X information is obtained.

The sample file includes many recodes of the variables to make the responses equivalent across the four survey years. Depending on the variable and how the question was asked, the file contains responses for two, three, or all four years. For example, the number of people in the household is available for the years 1977, 1980, and 1985, but not for 1974. Household income, on the other hand, is available for all four survey years.

We chose as balancing (X) variables for ECC those from the AHS file which are also available on the census 100% Detail ("short form") File. These are shown in Tables 1 and 2. The 1974 values for the first two--number of people and number of people 16 years or older--were not on the file. There we used the 1977 values instead. Rent/home value is a categorical variable which assigns a value according to the rent (home value) if the unit is rented (owned).

We selected as target (Y) variables those which are not on the census short form, but which are of interest to one or more of our survey sponsors:

Household income (from persons related to the householder)
Households with income less than $6000
Households with income greater than $20,000
Householders on welfare
Householders on unemployment compensation
Unemployed householders
Employed householders

Out of 4480 records for the LA metropolitan area, 3643 remained with complete responses to all the fields we used. For ease in forming clusters, we treated the first 3600 records as one PSU. Segments were defined by taking consecutive strings of 40 housing units. Within these segments, we formed clusters in several ways.

5. The AHS Data Study--Forming Clusters

In this study, each segment had 40 housing units. We formed "compact" clusters by joining the first four units in the segment, the next four, etc., until we had 10 clusters. From the same segment, we also formed systematic (noncompact) clusters. The first cluster combines the 1st, the 11th, the 21st, and the 31st units. Nine more clusters were created similarly.

Finally, ECC was used to produce clusters. In this section, we discuss clustering which "equalizes" on only one X variable at a time. First the segment records (housing unit values) are ordered on the X value. The first and last (ordered) records are combined into a cluster of size two, the second and second last records are combined, etc. These clusters of size two are now ordered according to the cluster's total X value, and combined into clusters of size four: the first and last, the second and second last, etc.

This method of equalizing clusters is not the only way, and need not be optimal. However, it is simple, easy to automate, and appears to give results which are optimal or very close to it.

Recall that for all but two X's, we used only the value in 1974 to cluster. ECC clusters are "formed in 1974" (1977, for the two exceptions) and then followed for the four survey years. We can see how well ECC clusters work in 1974, when the X values are current, as well as in 1977, 1980, and 1985, by which time many changes will have occurred.

6. AHS Data Study--Longitudinal Results

In this section, we compare methods of clustering by presenting their variances in estimating Y as a percentage of the variance for compact clustering. Formulae for the estimates and their within-PSU variances are found in Cochran (1977).

Tables 1 and 2 focus on separate target (Y) variables: the number of households with income less than $6000, and total household income, respectively. ECC variances are given for each of the nine X's. For a given X and Y, we can observe how effective ECC clustering is over the four survey years. Note again that these observations are merely descriptive. If we had formed the clusters again starting at a different unit in the file, the resulting variance may be slightly higher or lower.

For each of the four sample years (i.e., down each column), we underlined the value of the smallest variance. For comparison, the variances obtained under systematic sampling are shown at the bottom of the tables.

In these tables and in general, there was a slight increase in the ECC variances from 1974 to 1985 compared to compact variances, reflecting some cluster instability. During this time, ECC variances also increased relative to systematic clustering. In 1974 and 1977, over all the X's and Y's we studied, ECC produced a smaller variance than systematic clustering in well over half the comparisons. However, ECC worked better than systematic in only about half the comparisons in 1980, and about 20% in 1985.

There was also no consistently good balancing variable. To see this consider Table 3. For the different Y variables in each year, we compare clusters formed compactly, systematically, and by ECC on the most successful of the nine X's. All of the nine X variables worked best for at least one Y in one year, and no X was best more than three times in the table. Even the best X's only beat systematic clustering in about half their comparisons.

Actually, ECC based on the number of people in the household and the number 16 years or older worked slightly better than systematic clustering. Yet because these two X's were not available on the research file, we formed clusters based on the 1977 responses to these questions. Their ECC clusters were only tested up to eight years (1985) after their formation (1977).

7. ECC With Two or Three X Variables

If one X variable works moderately well equalizing the Y values, why shouldn't two X's work even better? Multiple ECC tries to form clusters which have approximately the same value of several X variables. We hope that this balancing will carry over more strongly to the Y values.

We imposed a method of equalizing clusters where the X characteristics are taken according to their relative importance in the balancing operation. Suppose we balance on a set $X_1, X_2, ..., X_k$. The records are first ordered on X_1, as described in Section 5. 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. When two or more records have the same values of X_1 and X_2, we use X_3 to order the records. We continue in this way, as necessary.

The order in which the X's are introduced affects the cluster formation and the resulting variances. If X_1 is essentially a continuous variable, we may want to recode it. Otherwise, subsequent X's will have little or no influence on the balancing algorithm.

There are other ways to form clusters based on multiple X values. In the one we used, the first variable strongly influences the resultant clusters. The second or third variables generally have a much smaller effect, especially if X_1 takes many values.

For many Y variables we compared multiple ECC on different combination of X's with other clustering methods. Table 4 shows the within-PSU variances for just one example--estimating household income (Y) using ECC on tenure (X_1), tenure with either mobile home status or number of rooms (X_1, X_2), and tenure with each of these two variables (X_1, X_2, X_3), in either order. The results here are typical of what happened with other Y's and X's.

Note that a different ordering of the X's produces different clusterings and variances. To summarize our observations:

- There is no consistent improvement when we use a second or third X variable to help balance the clusters.
- Multiple ECC yields no consistent and sizable improvement over systematic clustering. Although the best combination of X's generally beats systematic clustering, we don't know ahead of time which X's to use.
- There is again an increase in the variance as the information becomes dated. In most cases, variances using multiple ECC are higher for the 1985 variables than for the earlier ones.

In all, this algorithm for multiple ECC requires no more effort or cost than ECC based on a single X. But there is no way to predict which combinations will yield the best results. One can generally select the best X, use it alone, and obtain similar results.

We also tried representing several variables X_1, ..., X_k by various functions, from simple sums to sums of standardized variables. The function value was then used in place of the X_i's to balance the cluster. Some functions have shown a small decrease in the variances for certain Y's. But even these have not worked well for a variety of targets.

8. Conclusions

Based on this research, we see little future for equal characteristic clustering in census surveys. Compact clustering is cheaper, and its variances are not much higher over the survey decade, 1994 to 2004.

This study is limited in several ways. First, we only looked at data from the Los Angeles metropolitan area. In addition, we treated the sample file as if consecutive records were next door to each other. This deficiency could affect the variance reductions that ECC produces. Finally, responses for some variables were not available on our research file for certain years of the survey.

Nevertheless, with current, accurate information, ECC can reduce variances, often 10% below that of compact clustering. However, systematically noncompact clusters can work about half as well, and are simpler to implement.

As the census responses become outdated, as people move in and out of units which will be in sample, ECC loses much of its effectiveness. Within about six years after the balancing information is obtained--only a couple of years after the surveys start phasing in sample--systematic clustering appears to catch up to ECC. Through the remainder of the survey decade, ECC shows little or no improvement over systematic clusters, and only slight improvement over compact clusters.

Unfortunately, no one X variable balances clusters well for many Y's. This is not surprising, due to the different correlations with different Y's. Further, we could not obtain consistent improvement by using more than one X variable at a time. Both multiple ECC and ECC on functions of variables worked well only sporadically.

Until we show that ECC can reduce variances for several key target variables and retain its effectiveness over a decade, we would hesitate to recommend ECC. However, different forms of noncompact sampling with their accompanying variance reductions might be investigated further.

In particular, we feel that systematic clustering should be reconsidered as an alternative to compact clustering. There are several important reasons:

- The variances obtained under systematic clustering are consistently lower than those for compact clustering because of the smaller intracluster correlation.
- Systematic clustering does not make use of census information. Hence, there is no deterioration in the variance reductions (compared to compact clustering) over time, as there is with equal characteristic clustering.
- Currently, the Current Population Survey and the National Crime Survey use compact clusters, mainly to reduce travel time and costs. However, most of their interviews are conducted by telephone. For telephone interviews, there is no difference in cost between the two methods of clustering.

ACKNOWLEDGMENTS

I would like to thank Larry Ernst for reviewing the paper and making helpful comments.

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.

CANTWELL, P. (1990). "Equal Characteristic Clustering," Proceedings of the Section on Survey Research Methods, American Statistical Association, p. 231-236.

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

* 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: Variances[1], Estimating INCOME LESS THAN $6000 (Y)

Balancing (X) Variable	1974	1977	1980	1985
No. of People (1977)	....[2]	.859	.916	.914
No. Peo., 16+ (1977)	....	.849	.941	.937
Rent/Home Value	....	.874	.849	.935
Tenure (Own/Rent)	....	.895	.908	.962
Mobile Home	....	.893	.874	.971
No. of Rooms	....	.885	.903	.912
Female HH	....	.891	.933	.937
Black HH	....	.893	.897	.939
Hispanic HH	....	.874	.952	.887
Systematic Clusters	....	.895	.901	.920

TABLE 2: Variances[1], Estimating HOUSEHOLD INCOME (Y)

Balancing (X) Variable	1974	1977	1980	1985
No. of People (1977)	....[2]	.881	.879	.934
No. Peo., 16+ (1977)	....	.871	.885	.906
Rent/Home Value	.860	.891	.861	.889
Tenure (Own/Rent)	.848	.881	.883	.887
Mobile Home	.858	.895	.851	.883
No. of Rooms	.862	.858	.868	.905
Female HH	.858	.874	.900	.935
Black HH	.867	.891	.873	.935
Hispanic HH	.871	.867	.877	.929
Systematic Clusters	.863	.893	.890	.849

TABLE 3: Variances[1], Most Successful ECC

Target (Y) Variable	1974	1977	1980	1985
Household Income				
Systematic	.863	.893	.890	.849
Most succ. ECC	.848	.858	.851	.883
	Tenure	No.Rms	MobHm	MobHm
Income < $6000				
Systematic	....[2]	.895	.901	.920
Most succ. ECC	....	.849	.849	.887
		No. 16+	RtHmV	HspHH
Income > $20,000				
Sytematic	....	.919	.888	.941
Most succ. ECC	....	.869	.864	.939
		HspHH	MobHm	No. 16+
HHldrs on Welfare				
Systematic	.929	.893	.924	.972
Most succ. ECC	.900	.838	.902	.986
	BlkHH	No.Peo	RtHmV	FemHH
HHldrs on Unempl. Comp.				
Systematic	....	.945	1.084	1.056
Most succ. ECC	....	.881	1.028	1.028
		HspHH	No.Rms	RtHmV

TABLE 4: Multiple ECC Variances[1], Estimating HOUSEHOLD INCOME (Y)

Balancing (X) Variable(s)	1974	1977	1980	1985
No. of Rooms	.847	.875	.874	.895
↗				
Mobile Home	.847	.864	.868	.874
↗				
Tenure	.848	.881	.883	.887
↘				
No. of Rooms	.860	.905	.866	.880
↘				
Mobile Home	.884	.898	.867	.866

[1] Variances are expressed as a fraction of the variance under compact clustering for the same year.

[2] Data for some variables in some years were not available on these files.