INTRACLASS CORRELATIONS USING A SAMPLE OF 1980 CENSUS DATA

Kathryn F. Thomas, Deborah A. Harner and Robert Fay, U.S. Bureau of the Census

I. INTRODUCTION

Clustering of sampling units versus not clustering units is a prime consideration in developing a sample survey. While a simple random sample (of housing units, for instance) is an appealing design, in practice it is not always practical to implement a simple random sample. Factors such as travel costs and preparation of sample lists may make a cluster sample a more economical approach. However, cluster samples are usually less efficient in terms of variance than simple random samples of the same size. The selection of a cluster sample generally causes an increase in the sampling variance due to the homogeneity of neighboring units. Standard texts in sampling theory discuss the intraclass correlation coefficient as a measure of this homogeneity and its effect on the variance.

Specifically, the intraclass correlation coefficient, δ, is a measure of the degree of homogeneity of the clusters relative to the total variability. Theoretically this coefficient can vary from -1.0 to +1.0. In actuality, the coefficient can only achieve a value of -1.0 in the special case where there is an average of 2 cases within each cluster. Negative values for the intraclass correlation in samples of housing units discussed here are unusual, as attributes which are substantially less homogeneous than would be expected by chance are rarely encountered. Normally, values of the intraclass correlation coefficient for population and housing characteristics run between 0 and 1. While references to high positive and low positive values are often seen, there is no simple probabilistic interpretation of these values.

This measure of homogeneity, δ, is dependent upon both the between and within cluster variances. When units within the clusters are homogeneous, that is, they are highly correlated with respect to the characteristic under study, the variability within cluster is very small and the between cluster variance would account for most of the total variability. In this case the intraclass correlation coefficient, δ, would be high positive, close to +1. But if the within cluster variation were large, that is, units were heterogenous with respect to the characteristic under study, the between cluster variation would account for a small part of the total variability and the intraclass correlation coefficient would be small positive, possibly negative. When the sampling variance from a clustered sample is exactly that of an unclustered sample of the same size, the intraclass correlation will be zero, and the clusters would be approximately as homogeneous as might be expected by chance.

Past studies have shown that as the size of the cluster increases, the measure of homogeneity usually decreases. Small clusters or groupings exhibit a higher degree of homogeneity than larger clusters, indicating that units which are closer together are more similar than units which are further apart. However, the rate of decrease in homogeneity ordinarily is much slower than the rate of increase in cluster size. The national estimates of the intraclass correlation coefficients which have been computed for this paper uphold this fact. The attached table and graphs present the intraclass correlation coefficients for different cluster sizes and geographic disaggregations for selected characteristics collected from all households in the 1980 Census. This large bank of data contained no real surprises.

The intraclass correlation coefficients in this paper can serve as a guide when designing household sample surveys. Along with discussions of methodology for computing these measures, the results obtained and potential applications are discussed.

II. RESULTS

Estimates of national intraclass correlation coefficients were computed for a variety of characteristics collected from each household in the 1980 census. Table 1 displays correlations δ_1 and δ_2 (defined later) for several characteristics and cluster sizes at the U.S. level. Additional results are shown graphically in the attachments; they were computed for clusters of sizes 2, 4, 8, 16, and 32; one can interpolate for sizes in between. Each graph represents one characteristic and shows the values of δ_1 for a given cluster size and geographic area. The δ_2 values, which are lower, are not shown graphically but follow the same trends. Each graph is accompanied by a legend which indicates the geographic area being considered and the range of standard errors associated with that area.

Standard Metropolitan Statistical Areas (SMSA) were used as a reference in developing these geographic disaggregations. An SMSA is a unit which includes a large population nucleus and nearby communities whose activities form an integrated social and economic system.

Relative to SMSAs, the following geographic areas are defined:

- Metropolitan (M) - the entire SMSA; also divided into:
  - Central City (C) - an incorporated or Census-defined place recognized as part of the nucleus of the SMSA;
  - SMSA Balance (B) - within an SMSA but not the Central City
- Non-metropolitan (N) - not in an SMSA; also divided into:
  - Non-SMSA Urban (U) - in places or areas satisfying population density requirements, but outside an SMSA;
  - Non-SMSA Rural (R) - neither urban nor in an SMSA
- Total (T) - the total coverage of the census, composed of the preceding areas.

Further clarifications of SMSA, central city, urban and rural are given in many of the 1980 Census publications including the PC80-1-B series.

All data in the table and represented on the graphs are on a household basis, not a person basis. Some of the characteristics (e.g., number

of persons 65+ and number of black persons) are identified by the number of persons in the household possessing the characteristic. The remaining characteristics (occupancy status and owner occupied with value > 90K) are tabulated by the presence or absence of the characteristic.

Table 2 identifies all characteristics for which intraclass correlations have now been computed. Values for δ_1 and δ_2 at various geographic and regional levels for all these characteristics are available upon request from the authors.

III. METHODOLOGY

A. Sample Design

Data from the 1980 Decennial Census were used to derive the measures of intraclass correlation for this study. The sample used was comprised of EDs sampled at the second stage of selection of the "Enumeration" or "E-sample" from the 1980 Post Enumeration Program (PEP). The design of the PEP was based on the first stage of selection of primary sampling units (PSUs) consisting of counties or groups of counties (townships and MCDs in New England and Hawaii) used in the full sample (A/B/C/D/E) of the Current Population Survey (CPS) of April 1980. For the E-sample, census enumeration districts (EDs) were sampled with probability proportional to estimated size as the next stage of selection, clustered by the four digit ED code. That second stage of selection was performed in two waves; the first according to a preliminary measure of size, and the second as a supplement to the first for EDs with probabilities proportional to the increase in size, if any, of revised estimates of size relative to the preliminary measures. To simplify estimation, double hits were allowed and doubly weighted. The weights were the inverses of the products of the first-stage probabilities and the expected number of hits over the two waves of second stage sampling, regardless of whether the hit was in the first or second wave. This estimation procedure provides (design) unbiased estimates of total.

The E-sample for PEP involved a third stage of selection of housing units and persons in group quarters within the selected EDs. This third stage of selection was not considered here; computations are based on all enumeration within the sampled EDs. Variances have been estimated based on this PEP sampling design.

B. Cluster Formation

In forming clusters, housing units with all characteristics imputed during census processing from a neighboring unit were omitted from the computation, since inclusion of these cases would have resulted in an upward bias in the computed correlation. For all characteristics shown, the clusters were based on housing units regardless of occupancy status. In attempting to form clusters of units which were geographically contiguous, housing units were grouped on the basis of order of enumeration in the census (i.e., serial number order) into equal-sized clusters of sizes 2, 4, 8, 16 and 32. Block boundaries were ignored. Whenever the total number of households in the ED universe was not an exact multiple of these cluster sizes, the remaining households were dropped from the analysis.

C. Computations

Two formulae for computing intraclass correlation were used and are defined below. The first, δ_1, measures the intraclass correlation over all clusters in the stated universe and therefore includes the variation between EDs. This measure specifically indicates the correlation evident in housing units of varying cluster sizes and characteristics if the clusters were randomly chosen, independent of ED. The second, δ_2, provides a measure of intraclass correlation reflecting stratification by ED. Computations were made at the ED level and weighted over EDs. These correlations therefore are appropriate for samples chosen within EDs or for highly stratified samples in which the effect of between ED variance is effectively removed.

Intraclass correlations were computed for specific geographic areas by the methodology described below. Defining,

- x_{ijk} as the observation in the i^{th} household, j^{th} cluster and k^{th} ED,
- n as the cluster size, or number of households in each cluster,
- m_k as the number of clusters in the k^{th} ED,
- m as the total number of clusters in the specified geographic area, and
- p as the total number of EDs in the specified geographic area,

the following computations were made:

1. For each cluster in each ED, $x_{.jk}$ (the aggregate for the j^{th} cluster and k^{th} ED) and $x_{.}{}^{2}{}_{jk}$ were determined as:

$$x_{.jk} = \sum_{i=1}^{n} x_{ijk}$$

$$x_{.}{}^{2}{}_{jk} = \sum_{i=1}^{n} x_{ijk}{}^{2}$$

2. For each ED, $x_{..k}$ (the aggregate for the k^{th} ED), $x_{..}{}^{2}{}_{k}$ and $x_{.}{}^{2}{}_{k}$ were determined as:

$$x_{..k} = \sum_{j=1}^{m_k} x_{.jk}$$

$$x_{..}{}^{2}{}_{k} = \sum_{j=1}^{m_k} (x_{.jk})^{2}$$

$$x_{.}{}^{2}{}_{.k} = \sum_{j=1}^{m_k} x_{.}{}^{2}{}_{jk}$$

3. For each geographic area, $x_{...}$ (the aggregate for the geographic area), $x_{..}{}^{2}{}_{.}$ and $x_{.}{}^{2}{}_{..}$ were determined as:

$$x_{...} = \sum_{k=1}^{p} x_{..k}$$

$$x_{..}{}^{2}{}_{.} = \sum_{k=1}^{p} x_{..}{}^{2}{}_{k}$$

$$x_{.}{}^{2}{}_{..} = \sum_{k=1}^{p} x_{.}{}^{2}{}_{.k}$$

4. Non-stratified measures of intraclass correlation for the specified geographic area were determined as:

$$\delta_1 = \left[\frac{x_{..}{}^{2}{}_{.} - (x_{...})^{2}/m}{x_{.}{}^{2}{}_{..} - (x_{...})^{2}/n \cdot m} - 1 \right] \Big/ (n-1)$$

Measures at state (or higher geographic) levels were computed as:

$$\delta_1 = \left[\frac{(\sum x_{..}{}^{2}{}_{.}) - (\sum x_{...})^{2}/\sum m}{(\sum x_{.}{}^{2}{}_{..}) - (\sum x_{...})^{2}/n \sum m} - 1 \right] \Big/ (n-1)$$

where the summation was over all geographic areas at this level and $\sum m$ was the total number of clusters at this level.

5. Stratified measures of intraclass correlation for the specified geographic area were determined by computing:

$$\delta_2 = \left[\frac{\sigma_c^2}{\sigma_n^2} - 1 \right] \Big/ (n-1) \qquad \text{where}$$

$$\sigma_c^2 = \sum_{k=1}^{p} \left[x_{..}{}^{2}{}_{k} - (x_{..k})^{2} \Big/ m_k \right] \qquad \text{and}$$

$$\sigma_n^2 = \sum_{k=1}^{p} \left[x_{.}{}^{2}{}_{.k} - (x_{..k})^{2} \Big/ n \cdot m_k \right]$$

("c" for compact and "n" for non-compact).

Measures at state (or higher geographic) levels were computed as:

$$\delta_2 = \left[\frac{\sum \sigma_c^2}{\sum \sigma_n^2} - 1 \right] \Big/ (n-1)$$

where the summation was over all geographic areas at this level.

D. Limitations

In attempting to form clusters of housing units which were geographically contiguous, the units were grouped on the basis of their order of enumeration in the census. This was done using the census serial number as the basis for grouping. It was assumed that adjacent serial numbers and nearest housing units were synonymous. In fact, this is not always true. For example, when a housing unit was added to the address listing book, it was usually added at the end of the block in which it was located and the next available serial number in that enumeration district (ED) was assigned. This tends to slightly bias the data in that a small number of the clusters which were formed may not have been nearest neighbors.

Since the calculations were based on a sample of EDs, the estimates are affected by sampling error. Standard errors for national coefficients are generally no more than .02 or .03, but errors for geographic detail or characteristics defined for minority groups are somewhat higher.

IV. APPLICATIONS

A. Sample Design

Various sample designs should be considered in making a decision on the most cost effective sampling methodology. Cluster sampling will usually result in an increase in variance over simple random sampling although travel and enumeration costs associated with clustering will usually be less. Relative efficiencies of cluster designs are dependent on:

1. the degree of homogeneity within a cluster (δ),
2. the average cluster size (n), and
3. the cost associated with data collection.

B. Design Effect

For a specified cluster size (n) the variance for cluster sampling can be obtained by multiplying the variance under simple random sampling by the design effect, $\Psi = 1 + \delta(n-1)$, whenever clusters are equal in size. If the sample size is fixed and δ is greater than 0, two extreme situations exist when n=1 and n=N. In the first instance, Ψ collapses to 1 and the variance for sampling clusters of size 1 is the same as the variance for simple random sampling. When n=N, or the entire primary sampling unit (psu) is included in the sample, the variance reaches its highest level. Any increase from n=1 will result in an increase in the variance for a fixed sample size.

The data in Table 1 can be used to estimate the design effect, and thus the increase in the variance due to the use of specific cluster sample designs. An example of this application appears below.

EXAMPLE

Suppose a survey is being designed to measure the value of owner occupied homes. Suppose also that the characteristic of "owner-occupied, value $90,000 or more" from the 1980 census is chosen as the key variable for purposes of design. The effect of clustering will depend on how the clusters are selected. If the universe is, in effect, divided into clusters of approximately size 32 but no stratification is performed, then the value of $\delta_1 = .433$ for $n = 32$ (from Table 1) implies a design effect of $\psi = 1 + \delta(n-1)$ equal

to 14.42. If, instead, the survey designer is able to draw a multi-stage sample by selecting EDs with probability proportionate to size according to a highly stratified design and then, by selecting compact clusters of size 32 within these EDs, the best outcome that could be expected would be based on $\delta_2 = .161$ (from Table 1) implying $\psi = 6.0$. In practice, a somewhat higher ψ would probably be encountered.

In contrast, clusters of size 4 would do no worse than $\psi = 2.47$ based on $\delta_1 = .493$ (from Table 1) and could do as well as $\psi = 1.72$ with $\delta_2 = .240$ with careful stratification. Thus, cluster size and effect of stratification both have important effects on the reliability of the estimates from a sample of fixed size.

V. Conclusions

In summary, it is hoped that these intraclass correlations calculated using 1980 census data will prove helpful to users who are formulating sample designs. They can be used to help determine whether clustering of sampling units should be considered. Once it is decided to cluster, they can also be used to help determine the appropriate size of cluster.

****

BIBLIOGRAPHY

[1] Hansen, Morris H., William N. Hurwitz and William G. Madow, Sample Survey Methods and Theory, Volume I, Methods and Applications, New York: John Wiley & Sons, Inc., 1953.

[2] Blalock, Hubert M. Jr., Social Statistics, Second Edition, McGraw-Hill Book Company, 1972.

****

Table 1: Intraclass Correlations by Cluster Size - U.S. Level

Characteristic	Correlation	Size of Cluster 2	4	8	16	32
Number of Persons, 65+	δ_1	.126	.120	.111	.103	.096
	δ_2	.068	.061	.053	.045	.037
Number of Black Persons	δ_1	.527	.510	.492	.474	.455
	δ_2	.228	.201	.173	.146	.118
Number of Hispanic Persons	δ_1	.350	.332	.317	.301	.287
	δ_2	.136	.114	.094	.074	.058
Occupancy Status	δ_1	.303	.283	.263	.242	.222
	δ_2	.176	.153	.130	.107	.086
Owner Occupied, value > 90K	δ_1	.508	.493	.476	.456	.433
	δ_2	.260	.240	.216	.190	.161
Black Renter Occupied, rent > 100	δ_1	.396	.377	.354	.331	.310
	δ_2	.219	.196	.168	.142	.118

Table 2: Available Data

Age/race/household size characteristics

- Number of persons in occupied households
- Number of black persons *
- Number of hispanic persons *
- Number of persons, 55+
- Number of persons, 65+ *
- Number of black persons, 65+
- Number of hispanic persons, 65+
- Number of children, 5-17
- Number of black children, 5-17
- Number of hispanic children, 5-17
- Number of persons, 16-21
- Household has > 1 person age 65+
- Household has 6+ persons
- Household has 1+ persons per room

Income-related characteristics

- Household lacks plumbing
- Household is renter occupied, rent < 100
- Household is black renter occupied, rent < 100*
- Household is renter occupied, rent 100-149
- Household is renter occupied, rent 170-179
- Household is renter occupied, rent 200-224
- Household is owner occupied, value < 20K
- Household is black owner occupied, value < 20K
- Household is owner occupied, value 20 - 30K
- Household is owner occupied, value > 90K *
- Household is owner occupied, value > 100K

Other housing characteristics

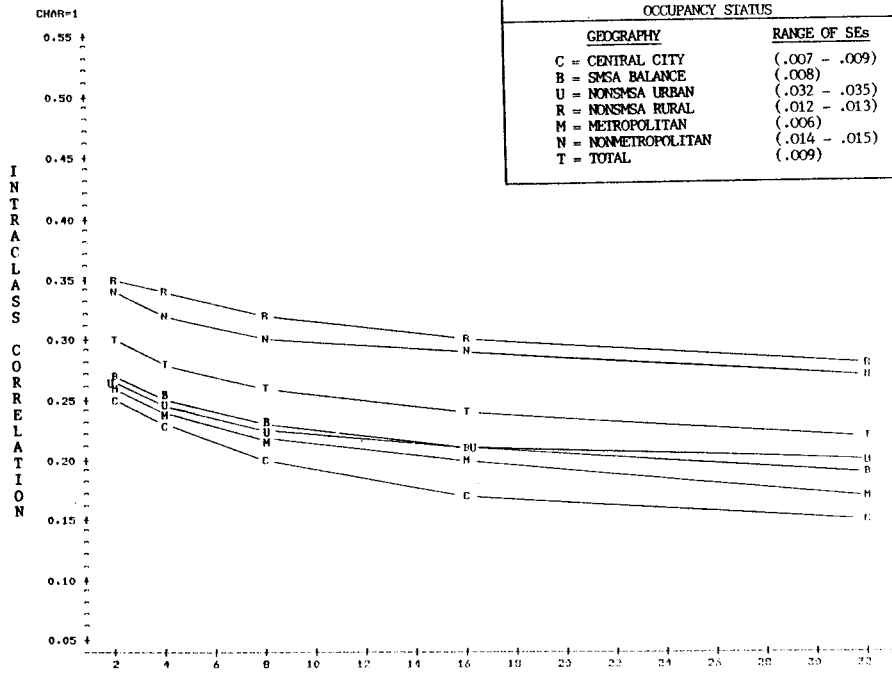
- Occupancy status *
- Household is renter occupied
- Household is owner occupied
- Household is occupied condominium
- Household is part of multiunit structure (10+ units)
- Household is year-round occupied
- Household is mobile home or trailer

* Data provided on graphs and/or in table 1.

MEASURES OF HOMOGENEITY (δ_1) FOR VARYING CLUSTER SIZES AND GEOGRAPHIC AREAS

OCCUPANCY STATUS

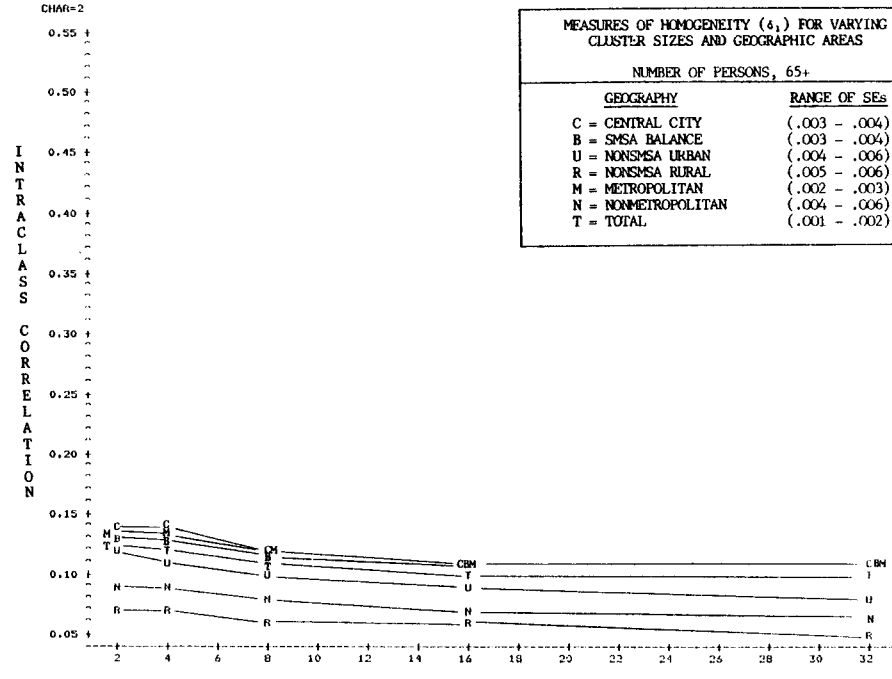
GEOGRAPHY	RANGE OF SEs
C = CENTRAL CITY	(.007 - .009)
B = SMSA BALANCE	(.008)
U = NONSMSA URBAN	(.032 - .035)
R = NONSMSA RURAL	(.012 - .013)
M = METROPOLITAN	(.006)
N = NONMETROPOLITAN	(.014 - .015)
T = TOTAL	(.009)

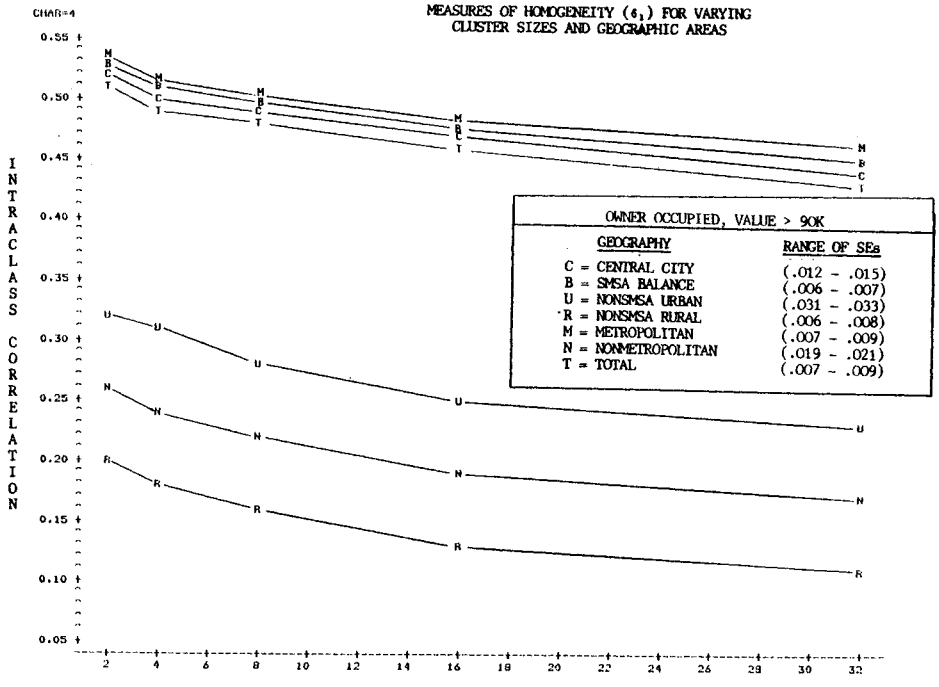


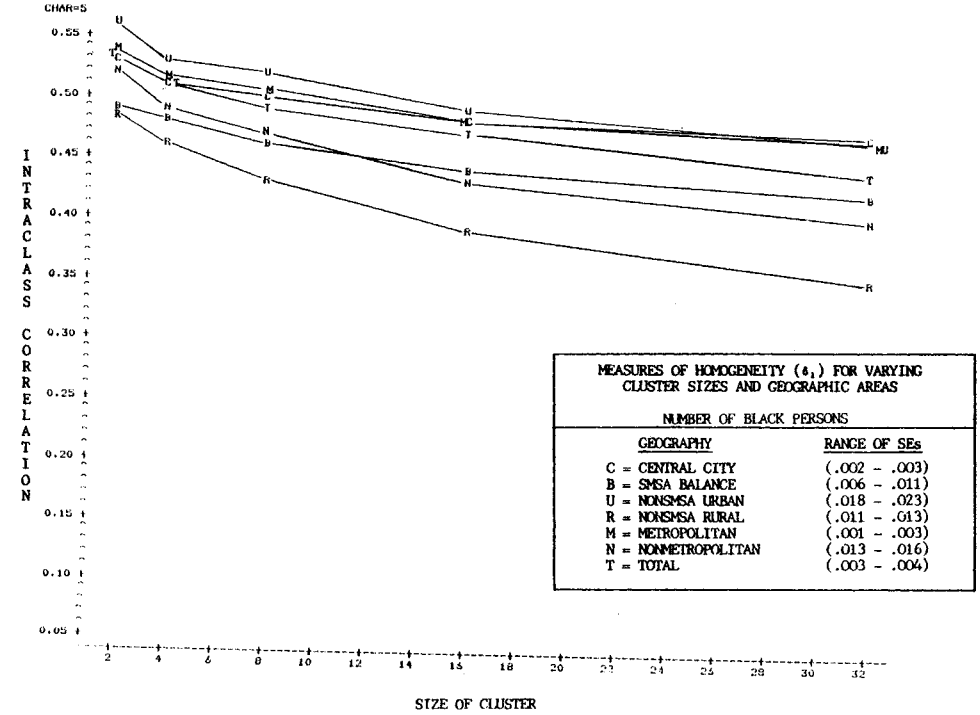
MEASURES OF HOMOGENEITY (δ_1) FOR VARYING CLUSTER SIZES AND GEOGRAPHIC AREAS

NUMBER OF PERSONS, 65+

GEOGRAPHY	RANGE OF SEs
C = CENTRAL CITY	(.003 - .004)
B = SMSA BALANCE	(.003 - .004)
U = NONSMSA URBAN	(.004 - .006)
R = NONSMSA RURAL	(.005 - .006)
M = METROPOLITAN	(.002 - .003)
N = NONMETROPOLITAN	(.004 - .006)
T = TOTAL	(.001 - .002)




CHAR=4
MEASURES OF HOMOGENEITY (δ_i) FOR VARYING CLUSTER SIZES AND GEOGRAPHIC AREAS
INTRACLASS CORRELATION
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
SIZE OF CLUSTER
OWNER OCCUPIED, VALUE > 90K
GEOGRAPHY — RANGE OF SEs
C = CENTRAL CITY (.012 - .015)
B = SMSA BALANCE (.006 - .007)
U = NONSMSA URBAN (.031 - .033)
R = NONSMSA RURAL (.006 - .008)
M = METROPOLITAN (.007 - .009)
N = NONMETROPOLITAN (.019 - .021)
T = TOTAL (.007 - .009)


CHAR=5
INTRACLASS CORRELATION
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
SIZE OF CLUSTER
MEASURES OF HOMOGENEITY (δ_i) FOR VARYING CLUSTER SIZES AND GEOGRAPHIC AREAS
NUMBER OF BLACK PERSONS
GEOGRAPHY — RANGE OF SEs
C = CENTRAL CITY (.002 - .003)
B = SMSA BALANCE (.006 - .011)
U = NONSMSA URBAN (.018 - .023)
R = NONSMSA RURAL (.011 - .013)
M = METROPOLITAN (.001 - .003)
N = NONMETROPOLITAN (.013 - .016)
T = TOTAL (.003 - .004)