Survey Methodology, June 1997
Vol. 23, No. 1, pp. 61-71
Statistics Canada

Geographic-Based Oversampling in Demographic Surveys of the United States

JOSEPH WAKSBERG, DAVID JUDKINS and JAMES T. MASSEY[1]

ABSTRACT

Often one of the key objectives of multi-purpose demographic surveys in the U.S. is to produce estimates for small domains of the population such as race, ethnicity, and income. Geographic-based oversampling is one of the techniques often considered for improving the reliability of the small domain statistics using block or block group information from the Bureau of the Census to identify areas where the small domains are concentrated. This paper reviews the issues involved in oversampling geographical areas in conjunction with household screening to improve the precision of small domain estimates. The results from an empirical evaluation of the variance reduction from geographic-based oversampling are given along with an assessment of the robustness of the sampling efficiency over time as information for stratification becomes out of date. The simultaneous oversampling of several small domains is also discussed.

KEY WORDS: Sample design; Stratification; Rare populations.

1. INTRODUCTION

The sponsors of many broad multi-purpose demographic surveys require separate analyses of domains defined by race, ethnicity and income. Equal probability samples generally do not provide sufficient sample sizes for some of these domains to yield the precision needed, making some form of oversampling necessary. This requirement poses interesting methodological problems since there is no registry of the U.S. population from which samples stratified by these domains can be drawn. Housing lists containing identifiers for these domains are maintained at the Bureau of the Census, but they are not available to researchers outside of the Bureau. For surveys requiring face-to-face interviews, outside researchers are thus forced to use area sampling techniques. Even within the Bureau, geography is sometimes used as the basis of oversampling since the lists are only updated once every ten years. This paper describes efficient methods for oversampling the aforementioned domains in the context of area sampling.

Data from the U.S. Decennial Census on concentrations of various demographic domains are publicly available for small geographic units; race and ethnicity are reported for every block and income for every block group. (A "block" is an area bounded on all sides by roads and not transected by any roads. Block groups are combinations of several neighbouring blocks.) These data may be used to inexpensively improve the precision of statistics about rare domains by oversampling blocks or block groups that contain higher than average concentration of members of rare domains and then dropping or subsampling screened persons not in the targeted rare domains. The general theory for this type of sample design was worked out by Kish (1965, Section 4.5). An independent presentation of the theory with examples from the 1960 Decennial Census was given by Waksberg (1973). Further examples and a discussion of alternative methods are given by Kalton and Anderson (1986) and by Kalton writing for the United Nations (1993). In this paper, we extend prior illustrations to cover more domains, update results to 1990, and evaluate empirically the robustness of these methods over time.

We first briefly review the issues involved with screening and subsampling persons not in the targeted domains. Then we review the theory for optimal allocation where the strata are defined in terms of the density of rare populations and apply this theory to several rare populations. The main part of the paper is an empirical evaluation of the reduction in variance reduction from the geographic oversampling of various minority and other rare populations as well as how robust the variance reductions are over time. We also discuss the special problems involved with simultaneous targeting of several rare populations before summarizing our conclusions.

2. SURVEY COST STRUCTURE AND THE SCREENING DECISION

Let U stand for some target universe such as persons or households for which a sampling frame exists. Let D stand for some small domain of particular interest such as black persons that cannot be separately identified from the balance of U at the time of sampling. Let $\boldsymbol{Y}$ be a vector of characteristics of interest such as annual income, employment status, and number of doctors' visits in the last year. In some surveys, the only objective is estimation of the distribution of $\boldsymbol{Y}$ on D. In such surveys, members of U-D that are discovered in the course of screening sampled members of U will be dropped from the sample. A general inexpensive interview

[1] Joseph Waksberg, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; David Judkins, Research Triangle Institute, 5901-B Peachtree-Dunwoody Road, Suite 500, Atlanta, GA 30325, U.S.A.; James T. Massey, formerly of Westat Inc., now deceased.

questionnaire is used for the screening to determine who is eligible for a full questionnaire.

In other surveys, estimation of the distribution of Y on D and on U are both important objectives. For such a survey, at least some of the members of U-D that are discovered in the course of screening interviews will be retained for full interviews. If geographic-based oversampling is used, the initial sample will contain an oversample of those members of U-D who happen to reside in areas with heavy concentrations of D. Even when U-D is of interest, this oversampling of U-D in areas with high concentrations of D is usually undesirable since resulting variation in probabilities of selection for U-D leads to unnecessarily large design effects for statistics both about U and about U-D. These larger design effects mean that the extra sample size for U-D will usually result in only a trivial decrease in variances for statistics about U-D. Generally, the funds expended on the extra interviews with U-D would be better spent on increasing the total initial sample size.

It is fairly easy to set up subsampling procedures that result in an equi-probability sample of U-D. The subsampling can be done centrally after the completion of the entire screening operation, or it can be done by the interviewer while still in the sample household after obtaining data on household composition. Techniques have been developed that make the subsampling process very easy for the interviewer (Waksberg and Mohadjer 1991). Interviewers do not need to be trained to carry out random draws. With paper and pencil survey instruments, interviewers are given house-by-house pre-interview instructions about which domains can be interviewed at which households. These instructions are randomized centrally prior to screening to yield the desired sampling rates. Alternatively, with CAPI, the subsampling can be programmed and carried out automatically in the laptop computer used for CAPI; the computer notifies the interviewer which households are to be retained for the full interview and which ones to reject as a result of subsampling.

Whether it is better to keep all sampled members of U-D or to subsample them depends on the relative sizes of U and U-D, the precision requirements for both and on the relative costs of full interviews and the shorter screening interviews. Let $c\star$ be the variable cost associated with sampling a single member of U and collecting and processing all data of interest about that member. Let c' be the variable cost associated with sampling, screening, and then dropping a single member of U. Let $c = c\star/c'$, be the ratio of the cost of a full interview to the cost of a screening interview. If c is much greater than 1, then subsampling should be considered for the survey that has interest in U-D even though subsampling of U-D will introduce some additional complexity into survey operations. Given that the full interview is by definition longer that the screening interview, it should always be the case that c is at least slightly greater than 1. On panel and longitudinal surveys, the cost of all follow-back interviews should be counted as part of $c\star$, typically making the cost of a full interview many times larger than the cost of a screening interview; *i.e.*, $c >> 1$. The same will be true of surveys that involve the collection of physical specima requiring expensive laboratory work and of surveys that require expensive experts (such as medical doctors) to participate in the primary data collection. For such surveys, we would highly recommend that geographic-based oversampling not be employed by itself, but rather, in conjunction with screening and subsampling. For a door-to-door survey with a single interview by a standard grade interviewer (trained to ask questions and record answers but not to make any technical or anthropological assessments), c is frequently in the range of 3 to 5. This is large enough in many applications to justify the complication of subsampling U-D in oversampled areas.

3. FORMING THE STRATA

We assume that even though D cannot be separated from U at the time of sampling, there is some information available about the distribution of D and U across a set of geographically defined entities. In the United States, the natural entities are blocks or block groups (BGs) and information for these entities is supplied by the decennial census. (Prior to the 1990 decennial census, blocks were not defined in rural areas; larger entities called "enumeration districts" were used for oversampling.) The U.S. Bureau of the Census makes data on the racial and ethnic composition of blocks publicly available along with mapping information so that these blocks can be identified years later by any survey organization. Income data are only made available at the BG level.

Standard practice calls for the stratification of the blocks or BGs by the local concentration of D. Thus, all blocks where D constitutes less that 10 percent of the block's total population might constitute one stratum. Further cutpoints for defining the strata might be 30 percent, and 60 percent, yielding a total of four strata. There has been little empirical study of the optimal number of strata nor of the optimal cutpoints. In general, more strata will yield more efficient designs, but, at some point, the operational complexities of a large number of strata outweigh the gains in efficiency. Conventional wisdom dating back to Kish (1965) holds that a fairly small number of strata will achieve most of the gains attainable through stratification.

4. OPTIMAL ALLOCATION FOR A SINGLE DOMAIN

Our objective is to adapt the general formulas for optimum allocation of a stratified sample to apply to the reduction in variance due to geographic-based oversampling. The derivations are essentially those given by Kish (1965) using the notation of Kalton in United Nations (1993). Let the population be divided into a number of strata as discussed above. Let N be the size of the total population and N_h be the size of the total population within the h-th stratum. Let P_h be the proportion of the h-th stratum that consists of members of D. Let P be the overall proportion of the population that belongs to D. We may use the prior decennial census to estimate P_h and P, or we may use some more recent large survey that carried block and/or BG codes for every sample household/person so that matching to the last decennial census will yield the stratum identification for every sample household/person.

We assume that c is constant across the strata even though this may sometimes not be very accurate. For example, interviewing in blocks with high concentrations of American Indians, Eskimos or Aleuts almost always means interviewing in remote locations with difficult transportation issues. However, estimation of even a national average for c is difficult for most survey operations. It will not generally be possible to get estimates by stratum.

We also assume that the distribution of $\boldsymbol{Y}$ on D is constant across the strata. More specifically, we assume that

$$E(\boldsymbol{Y}\mid D \text{ and } h) \equiv E(\boldsymbol{Y}\mid D) \quad \text{and that}$$

$$\operatorname{Var}(\boldsymbol{Y}\mid D \text{ and } h) \equiv \operatorname{Var}(\boldsymbol{Y}\mid D),$$

where the expected value and variance are with respect to the population, not the sample design. This is usually not a very good assumption, but given a vector of characteristics of interest, the components of the vector will usually behave differently across the strata so there is no point in trying to be more exact. Lastly, we assume that the sampling fractions are small enough in all the strata to make the finite population correction factors ignorable.

Given these assumptions, the optimal sampling fraction for the h-th stratum for a survey where all screened members of U-D are dropped is

$$f_h = k\sqrt{\frac{P_h}{P_h(c-1)+1}}, \tag{1}$$

where k is a constant determined by either precision requirements or budget constraints. (For a proof of (1), see either of the sources referenced above. This allocation rule is an application of Neyman allocation.) If $c=1$, (*i.e.*, screening is as expensive as interviewing), then this proportionality reduces to $f_h \propto \sqrt{P_h}$, which can yield allocations quite different from an equi-probability sample across strata. However, if the cost of screening is far less than the cost of interviewing (*i.e.*, $c>>1$) and D is not extremely rare (*i.e.*, P_h is not close to zero), then this relationship results in close to a flat set of sampling intervals, which is equivalent to allocation in proportion to total population.

Given a fixed budget of B, k is determined by the cost equation

$$B = \sum_h N_h f_h c'\left[P_h c + (1-P_h)\right]. \tag{2}$$

To obtain a simple random sample of size n from domain D would require selecting a screening sample of size n/P, resulting in a total cost of

$$B = ncc' + \left(\frac{n}{P} - n\right)c'. \tag{3}$$

By equating these two costs, we can solve for the constant of proportionality in (1) and get:

$$k = \frac{n\left(c-1+\frac{1}{P}\right)}{\sum_h N_h P_h \sqrt{c-1+\frac{1}{P_h}}}. \tag{4}$$

To calculate the benefits of this allocation realistically, it is necessary to acknowledge the fact that the estimates of P_h that are used to guide the allocation will be somewhat out of date by the time that the survey is actually conducted. Let A_h be the proportion of D actually to be found within the h-th stratum at the time of sampling and data collection. It is assumed that P is unchanged even though the distribution across strata changes according to A_h. By letting $NP = N_D$ and $N_D\, A_h = N_{Dh}$ it can readily be shown that the actual sample size, n_D, that will be achieved on D is given by

$$n_D = \sum_h NPA_h f_h. \tag{5}$$

From Kish (1965), this sample will have higher variance than a simple random sample of the same size on D. The variance inflation factor or design effect associated with the differential sampling rates across strata is the well-known

$$deff = \left(\sum_h A_h f_h\right)\left(\sum_h A_h \big/ f_h\right). \tag{6}$$

Thus, the *effective* sample size associated with the geographic-based oversampling is

$$\frac{n_D}{deff} = \frac{NP}{\left(\sum_h A_h \big/ f_h\right)}. \tag{7}$$

Substitution of formulae (1) and (4) into (7) yields

$$\frac{n_D}{deff} = \frac{n\left(c-1+\frac{1}{P}\right)}{\left(\sum_h A_h\sqrt{c-1+\frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP}\sqrt{c-1+\frac{1}{P_h}}\right)}. \tag{8}$$

This formula allows us to compare the variance for an arbitrary statistic on domain D given geographic-based oversampling with the variance for the same statistic given a simple random sample of D of the same total cost B. Formula (8) can be rewritten algebraically such that the proportion of simple random sample variance that is eliminated by the geographic-based oversampling is given by

$$\frac{\frac{\sigma^2}{n} - \frac{\sigma^2 deff}{n_D}}{\frac{\sigma^2}{n}} = 1 - \frac{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}{\left(c - 1 + \frac{1}{P}\right)}. \quad (9)$$

It is definitely possible for this reduction to be negative, meaning that a simple random sample would have provided lower variance for the same cost. This is most likely to happen when there exists a stratum for which $NPA_h >> N_h P_h$, meaning that there exists a stratum which was thought to have a very small portion of D but, in fact, has quite a significant portion of D. Note that if $P_h \equiv P$, then no variance reduction can be expected from geographic-based oversampling. Also, as c goes to infinity for fixed P (equivalent to screening becoming cheaper and cheaper relative to full interviews), the variance reduction approaches zero. Given the extra complication of a stratified sample, this means that for large c and moderate P, the sample designer should consider drawing a simple random sample instead of a stratified sample. Geographic-based oversampling increases in value as P approaches zero, c approaches 1, and D becomes more concentrated in a single stratum. As the small domain of interest, D, becomes more concentrated in a single stratum the sample becomes more efficient, since there are fewer cases from D in the remaining strata with large differential. The potential reductions in variance due to geographic-based oversampling under a number of conditions are shown empirically for several demographic domains in the section below.

5. EMPIRICAL EVALUATION

Equation (9) is quite difficult to evaluate for domains of interest. Data on P_h can be obtained from summary tapes from the decennial censuses that are published at the block, block group, and enumeration district levels by the Bureau of the Census. This allows one to define reasonable strata and to evaluate equations (1) through (4). If one were to assume that the P_h are static over time, then the rest of the equations could also be evaluated. However, Americans tend to move frequently, and the racial and ethnic composition of many blocks change in that process (Judkins, Massey and Waksberg 1992). To the extent that members of D move into areas where they were previously not common, the benefits of the geographic-based oversampling diminish. Not wishing to overstate the benefits of the procedure, we searched for some method to get reasonable estimates of the A_h at postcensal time points. Matching block- or BG-level data for two consecutive censuses might appear to be a good solution but is not possible. Up to now, blocks have been defined and labelled independently from census to census with no attempt to preserve definitions for longitudinal. Thus, alternate information sources are required to estimate A_h.

For the analysis of the benefits of geographic-based oversampling for the black and Hispanic populations, micro-level data from current household surveys conducted by the Census Bureau turned out to be a good source of information on the A_h. Specifically, we used data from the 1988 National Health Interview Survey (NHIS). Staff at the Census Bureau prepared a special tape for us that gave the 1980 block group or enumeration district code for almost all households interviewed in the 1988 NHIS in residences built prior to 1980. (Residences constructed during the 1980s would have been sampled for the NHIS from building permits rather than by area sampling. Due to technical difficulties, block and block group labels are not attached to such sample dwellings.) We then matched the 1988 NHIS against 1980 Census summary files by block group or enumeration district in order to classify NHIS households into strata defined by concentrations of blacks and Hispanics in 1980. Using survey weights, we were then able to estimate the distribution of various domains across those strata. (Housing built during the 1980s was assumed to be in the stratum with the lowest concentration of the rare domains.) Similar operations could have been carried out for Asians, Pacific Islanders, American Indians, Eskimos, Aleuts, and persons with low income but were not.

Tables and charts in the balance of the paper will refer to data at several points in time and from several sources. It is useful to bear in mind that the data used to form the strata do not have to be the same as the data used to allocate the sample, and that the data used to evaluate the sample may be from a third point in time or source. We have the following combinations in this paper:

Label	Source of stratification data	Source of allocation data	Source of evaluation data
80/80/80 BG	1980 Census (BG level)	1980 Census	1980 Census
80/80/88 BG	1980 Census (BG level)	1980 Census	1988 NHIS
80/88/88 BG	1980 Census (BG level)	1988 NHIS	1988 NHIS
90/90/90 BG	1990 Census (BG level)	1990 Census	1990 Census
90/90/90 blk	1990 Census (block level)	1990 Census	1990 Census

Table 1
Residential Clustering of Blacks

Density stratum (Blacks as a percent of the stratification unit in the year of stratification)	Percentage of blacks living in the stratum in the indicated year				Percentage of the total population living in the stratum in the indicated year			
Measurement year	1980	1988	1990	1990	1980	1988	1990	1990
Stratification year	1980	1980	1990	1990	1980	1980	1990	1990
Stratification unit	BG/ED	BG/ED	BG	Block	BG/ED	BG/ED	BG	Block
< 10%	9.7	20.5	12.0	8.5	78.2	81.4	75.7	77.5
10-30%	13.5	13.2	16.8	13.9	8.9	7.1	11.4	9.6
30-60%	18.9	20.4	20.3	16.2	5.1	5.1	5.7	4.5
60-100%	57.9	45.9	51.0	61.4	7.8	6.4	7.2	8.4
Total populations (1000s)	26,495	29,380	29.986	29,986	226,546	240,876	248,710	248,710
Blacks as percent of nation in measurement year	11.7	12.0	12.1	12.1				

Sources: 1980 Decennial Census (Westat tabulation)
1988 National Health Interview Survey (Westat tabulation)
1990 Decennial Census (Westat tabulation)

6. OVERSAMPLING THE BLACK POPULATION

Table 1 shows various aspects of residential segregation for the black population in the U.S. that are important to know about when designing a population survey. Although the percentage of blacks living in densely black (60+ percent) block groups declined between 1980 and 1990, it is clear that blacks were still strongly segregated. The columns about the population in 1988 are particularly important since they show the dynamics of the stratification data over time. By 1988, the percentage of the black population living in the block groups that were less than 10 percent black in 1980 had doubled, from just 9.7 percent of blacks to 20.5 percent. This has major implications for the efficacy of geographic-based oversampling as will be shown below. It is also interesting to note that the total population in the block groups that were densely black (*i.e.*, over 60% black) in 1980 actually declined by about 2 million persons between 1980 and 1988. At least part of this shift came from abandonment of some old housing and neighbourhoods. Concentration levels are sharper at the block level than at the block group level in 1990, as would be expected. (Block level data are not available for the whole nation from 1980.) Although sampling blocks is slightly more costly than sampling block groups (due to the larger number of blocks and the need to make provisions for blocks that have fewer inhabitants than the desired sample cluster size), it does allow sharper focus on the targeted domain.

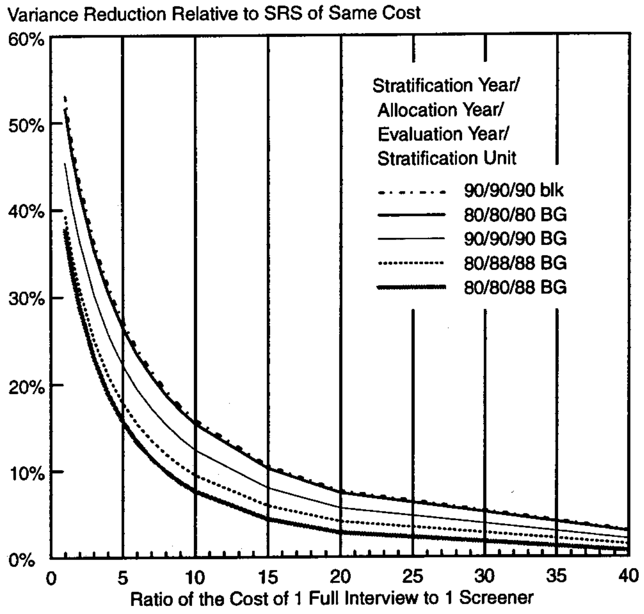

Figure 1. Variance Reduction from Geographic-based Oversampling for Blacks

Figure 1 summarizes the implications of the density data shown in Table 1 for oversampling blacks. This figure shows the substantial effect of *c* on the efficiency of geographic-based oversampling. For values of *c* beyond 20, the best way to sample the black population is probably just to screen an equi-probability sample.

The figure also illustrates the danger of relying upon the stratification data to evaluate the benefits of geographic-based oversampling. The 80/80/80 line shows the variance reductions that could be made if there were no change over time in the distribution of the black population across the density strata defined in terms of 1980 block group data. The 80/80/88 line shows the actual variance reductions that are possible in 1988 for the same strata and allocation. At $c = 5$, the variance reduction given a static distribution is 26 percent, while the variance reduction given observed changes in the distribution is just 16 percent. We examined whether allocating the sample across the old strata according to new distribution data could improve the actual variance reduction in 1988. The answer is yes, but not by much. The 80/88/88 shows the variance reductions that are possible using the 1988 distribution across the 1980 strata to guide the allocation for a survey conducted in 1988. At $c = 5$, the variance reduction given this allocation is 18 percent, a very modest improvement over the 16 percent variance reduction possible with the allocation guided by the old distribution. This led us to conclude that the major problem was the old stratification itself. By 1988, the extent of migration by the black population from block groups that were densely black in 1980 into block groups that had lower concentrations of black populations in 1980 was so great as to cut the variance reduction achievable through oversampling almost in half. The shift of the black population into block groups with lower concentrations of blacks in 1980 results in more sample blacks with large weights thus increasing the variability among weights which increases the variance. Nonetheless, the variance reductions indicated by the 80/80/88 line for $c < 10$ are certainly large enough to be useful.

Turning attention to the 1990 data in Figure 1, we observe that the 90/90/90 BG line is consistently several points below the 80/80/80 line, indicating that geographic oversampling at the block group level is likely to be slightly less useful during the 1990s than it was during the 1980s. This is a reflection of the slight reduction in segregation of the American black population in 1990 compared to 1980 noted above. On the other hand, the 90/90/90 blk line is almost exactly the same as the 80/80/80 line, indicating that the geographic oversampling at the block level can be expected to be as effective during the 1990s as it was at the block group level in the 1980s. Although data have not yet been collected on the distribution of the black population in the late 1990s across 1990 density strata, we would expect that migration has continued and that therefore the gains indicated by the 1990 lines should probably be reduced (along the general trend indicated by the 80/80/88 line) when projecting savings into the late 1990s and the first few years after 2000.

7. OVERSAMPLING HISPANICS

Table 2 shows various aspects of residential segregation for Hispanics in the U.S. that are important to know about when designing a population survey. Several points are interesting to note. First, it appears that Hispanics (unlike blacks) became slightly more segregated between 1980 and 1990. Other patterns, however, are similar for the black and Hispanic populations. In 1980, 30 percent of the Hispanic population lived in block groups that were 60 percent or more Hispanic. By 1988 these same block groups contained only

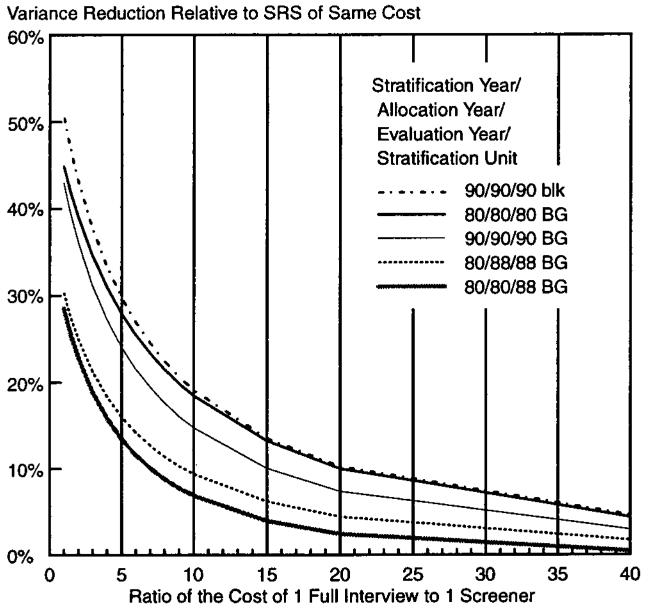

Figure 2. Variance Reduction from Geographic-based Oversampling for Hispanics

Table 2
Residential Clustering of Hispanics

Density stratum (Hispanics as a percent of the stratification unit in the year of stratification)	Percentage of Hispanics living in the stratum in the indicated year				Percentage of the total population living in the stratum in the indicated year			
Measurement year	1980	1988	1990	1990	1980	1988	1990	1990
Stratification year	1980	1980	1990	1990	1980	1980	1990	1990
Stratification unit	BG/ED	BG/ED	BG	Block	BG/ED	BG/ED	BG	Block
< 5%	14.8	29.3	10.6	6.6	76.8	79.8	68.4	68.9
5-10%	9.6	9.5	8.7	8.1	8.8	7.7	10.9	10.3
10-30%	22.6	21.2	22.8	22.1	8.5	7.4	11.8	11.5
30-60%	23.1	18.8	24.1	23.3	3.5	3.0	5.1	4.9
60-100%	30.0	21.2	33.9	39.8	2.4	2.0	3.8	4.4
Total populations (1000s)	14,609	19,393	22,354	22,354	226,546	240,876	248,710	248,710
Hispanics as percent of nation in measurement year	6.4	8.1	9.0	9.0				

Sources: 1980 Decennial Census (Westat tabulation)
1988 National Health Interview Survey (Westat tabulation)
1990 Decennial Census (Westat tabulation)

about 21 percent of the Hispanic population. In contrast, the percent of Hispanic population living in the 1980 block groups that were less than 5 percent Hispanic increased from 15 percent in 1980 to 29 percent in 1988. These changes reflect both a shift of the Hispanic between areas and the increase in the Hispanic population coming into the United States. The restratification of the Hispanic population using 1990 data shows patterns similar to the 1980 distribution patterns.

Figure 2 summarizes the implications of these segregation data on oversampling schemes. The curves show the same general patterns as the black curves. Geographic-based oversampling appears to be a useful tool for values of $c < 10$. Again though, it is important to be mindful of the effect of migration on the variance reduction. The gap between the 80/80/80 and 80/80/88 lines is greater for Hispanics than for blacks, particularly for $c < 5$. At present, we do not have a good basis for predicting whether this will be as true in the 1990s as it was in the 1980s.

8. OVERSAMPLING OTHER RACIAL MINORITIES

Tables 3 and 4 show segregation data for Asians and Pacific Islanders and for American Indians, Eskimos and Aleuts, respectively. Figures 3 and 4 show corresponding implications for oversampling these domains. Data from 1980 and 1988 were not tabulated for this work because the 1990 data are not encouraging for the inexpensive oversampling of these populations even with the use of stratification by density. The percent reductions in variance are quite large, greater than those for the black and Hispanic populations, since the amount of screening that would otherwise be required is much larger. However, the rarity of these populations in the U.S. means that very large screening samples are still required in order to get respectable interviewed sample sizes. For example, with a cost ratio of 3, even with geographic-based oversampling, it is necessary to screen 61,000 persons (or about 24,000 households) in order

Table 3
Residential Clustering of Asians and Pacific Islanders

Density stratum (Asians and Pacific Islanders as a percent of the 1990 block or block group in 1990)	Percentage of Asians and Pacific Islanders living in the stratum in 1990		Percentage of the total population living in the stratum in 1990	
Stratification unit:	BG	Block	BG	Block
< 5%	30.5	19.4	86.4	85.2
5-10%	17.2	17.7	7.2	7.4
10-30%	27.8	32.1	5.0	5.7
30-60%	14.6	18.0	1.0	1.3
60-100%	9.8	13.0	0.4	0.5
Total population (1000s)	6,968	6,968	248,710	248,710
Asians and Pacific Islanders as percent of nation in measurement year	2.8	2.8		

Sources: 1990 Decennial Census (Westat tabulation)

Table 4
Residential Clustering of American Indians, Eskimos and Aleuts

Density stratum (American Indians, Eskimos and Aleuts as a percent of the 1990 block or block group in 1990)	Percentage of American Indians, Eskimos and Aleuts living in the stratum in 1990		Percentage of the total population living in the stratum in 1990	
Stratification unit:	BG	Block	BG	Block
< 5%	50.3	34.6	98.3	97.4
5-10%	7.4	12.1	0.8	1.4
10-30%	12.4	15.9	0.6	0.8
30-60%	6.0	7.7	0.1	0.1
60-100%	23.8	29.6	0.2	0.2
Total population (1000s)	1,793	1,793	248,710	248,710
American Indians, Eskimos and Aleuts as percent of nation in measurement year	0.7	0.7		

Sources: 1990 Decennial Census (Westat tabulation) to obtain a sample of American Indians, Eskimos and Aleuts with precision equal to a (theoretical) simple random sample of 1,000 persons from this domain. (Of course, to successfully screen 24,000 households, more housing units would have to be selected to allow for vacants and nonresponse). The comparable number for Asians and Pacific Islanders is 18,000 persons or roughly 7,000 households.

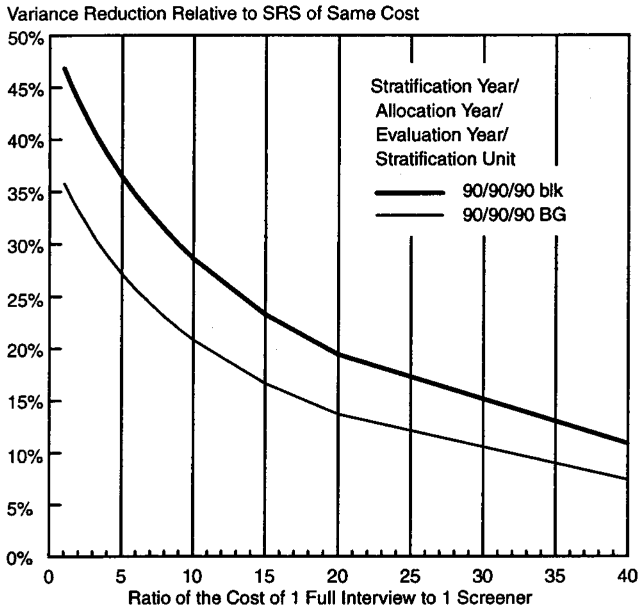

Figure 3. Variance Reduction from Geographic-based Oversampling for Asians and Pacific Islanders

9. OVERSAMPLING THE POOR

Table 5 shows the 1990 distribution of the low income population by block groups classified according to the proportion of low-income population in the BG. The BGs in each of the classes depends on the definition of low income. The figures shown in the table are the percentages of low-income persons in each class. Table 5 shows a rather flat distribution of low income among the classes for all three definitions in 1990. Data (not shown) from the 1970 decennial census and the Current Population Survey indicate that segregation of persons below the poverty level increased between 1970 and 1990 (Waksberg 1995), but the segregation is still far less than the segregation of racial and ethnic groups. The concentrations are somewhat greater for persons under 150 percent than for the other two definitions but, even for this group, it is considerably less than for racial and ethnic groups. As can be seen, with this definition, only about 25 percent of the poor live in BGs where 50 percent or more of the population is poor. The comparable percentages are 19 percent for persons below 125 percent of poverty and only 13 percent for persons below 100 percent of poverty. Such distributions imply that oversampling households in the strata with relatively high percentages of low-income persons will not be much better than oversampling and screening the entire sampling frame unless the full interview costs are only slightly higher than screening costs.

Figure 5 shows the ratio of the variance of the optimum sample to an SRS at the same cost, for statistics relating to the low-income populations. Interestingly, despite the greater concentration associated with the broadest definition of low

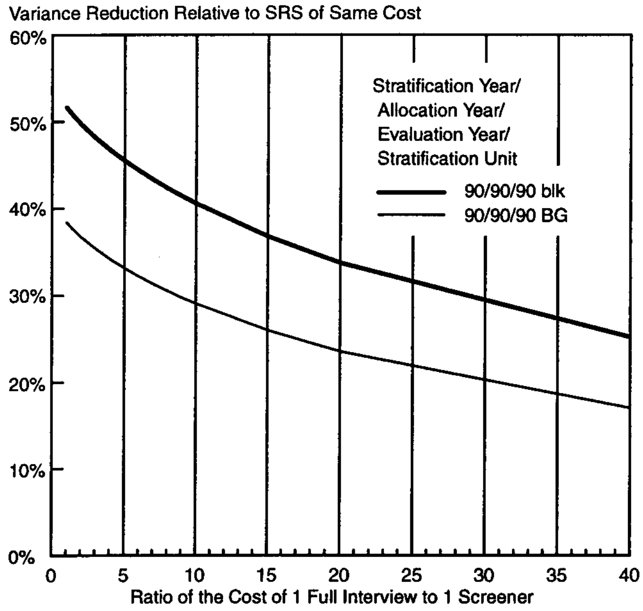

Figure 4. Variance Reduction from Geographic-based Oversampling for American Indians, Eskimos and Aleuts

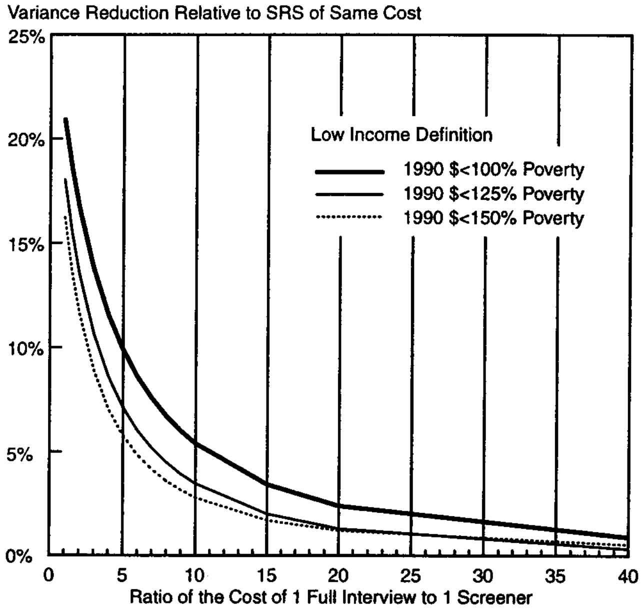

Figure 5. Variance Reduction from Geographic-based Oversampling for Persons with Low Income

Table 5
Residential Clustering of the Low Income Population

Density stratum (Persons with low income as a percent of 1990 block group in 1990 according to various definitions of low income)	Percentage of persons with low income living in the stratum in 1990			Percentage of the total population living in the stratum in 1990		
Low income definition:	$ < Poverty	$ < 125% of Poverty	$ < 150% of Poverty	$ < Poverty	$ < 125% of Poverty	$ < 150% of Poverty
< 5%	5.8	3.2	1.8	33.3	22.4	15.4
5-10%	12.3	8.3	5.7	22.1	19.7	16.7
10-20%	24.8	21.0	16.8	22.8	25.2	24.8
20-30%	19.8	20.2	19.2	10.7	14.4	16.8
30-40%	14.3	15.9	17.0	5.4	8.1	10.7
40-50%	10.0	12.2	13.7	2.9	4.8	6.7
50-100%	13.0	19.3	25.7	2.8	5.4	8.8
Total populations (1000s)	31,797	42,316	52,521	248,710	248,710	248,710
Persons with low income as percent of nation in measurement year	12.8	17.0	21.1			

Sources: 1990 Decennial Census (Westat tabulation of STF-3)

income, the reduction in variance for geographic-based oversampling is strongest for the narrowest definition because it requires more screening and thus has more to gain from a sampling strategy that reduces screening. For all three definitions, there appear to be moderate advantages to oversampling when *c* is under 3 or 4, about a 10 or 15 percent reduction in variances. When *c* is as large as 10, the gains are very slight, and there is virtually no advantage to oversampling BGs with high levels of poverty when *c* is 20 or larger. Of course, migration must be taken into account here as well, but we did not obtain the necessary data. Due to the effects of migration, the actual variance reductions will almost certainly be smaller than those shown in the chart. Furthermore, the income data in the 1990 Census are based on a one-sixth sample. The sample size in a typical block group was a little under 100 households. The classification of blocks according to percentage of low-income persons therefore has a fair amount of fuzziness to it, and many block groups will not be in the categories that Census data assign them, but in neighbouring classes, further weakening the variance reductions that can be achieved with geographic-based oversampling. As a result of these factors, it is unlikely that geographic-based oversampling will improve the efficiency. In fact, by mid-decade or later, it may actually result in an increase in variance. A related unpublished study by Waksberg in 1989 showed similar results when considering the possibility of merging ZIP-code level summary income data onto banks of telephone numbers used in RDD sampling. The gains achievable through stratification appear quite limited.

An examination of more detailed tables (not shown) indicates that the effectiveness is about the same for various types of geographic breakdowns, *e.g.*, states, large or small MSAs, central cities, suburban areas, and nonmetropolitan areas. Conclusions drawn from this analysis will thus approximately apply to subnational surveys.

However, geographic-based oversampling is an extremely effective tool for the low-income black and Hispanic populations. As shown in Table 6, blacks and Hispanics living in poverty are highly concentrated and others living in poverty are not. The left-hand side of Table 6 indicates the distribution of the poor black, Hispanic, and other populations across density strata defined in terms of poverty rates specific to the domain of interest. Interpreting one example from the left side, 32 percent of poor Hispanics lived in 1990 in block groups where the poverty rate for Hispanics was over 50 percent. The right hand side indicates the distribution of the poor black and Hispanic populations across density strata defined just in terms of the local concentrations of blacks or Hispanics without regard to income levels. Interpreting one example from the right side, 44.8 percent of poor Hispanics lived in 1990 in block groups where Hispanics constituted over 60 percent of the local population. From these numbers, we infer that over 90 percent of both poor blacks and poor Hispanics live in areas with above average concentrations of their respective racial/ethnic groups. This means that a sampling strategy that oversamples blocks with high black or Hispanic concentrations will automatically yield disproportionately large numbers of poor blacks and Hispanics. Furthermore, almost no poor blacks or poor Hispanics live in areas with low poverty rates for their groups. This stands in marked contrast to the patterns for poor people who are neither black nor Hispanic. It appears that many poor nonhispanic whites live in close proximity to more well-off whites, possibly because poverty tends to be a transitory phenomenon for them, or perhaps because they are retired and purchased their homes when they were in better circumstances.

Table 6
Residential Clustering of the Low Income Population by Race and Ethnicity

Density stratum (Poverty rate in 1990 for persons of the indicated race/ethnicity within the block group in 1990)	Percentage of persons with the indicated race/ethnicity and income below the poverty line living in the stratum in 1990			Density stratum (Indicated minority as a percent of 1990 block in 1990)	Percentage of persons with the indicated race/ethnicity and income below the poverty line living in the stratum in 1990		
	Domain				Domain		
	Blacks	Hispanics	Others		Blacks	Hispanics	Others
< 5%	0.6	0.6	10.4	< 5%	4.0	4.6	n/a
5-10%	2.2	2.4	19.6	5-10%	3.7	5.1	n/a
10-20%	8.8	11.0	32.6	10-30%	13.2	19.9	n/a
20-30%	13.8	17.0	18.1	30-60%	19.0	25.5	n/a
30-40%	17.0	19.3	9.0	60-100%	60.0	44.8	n/a
40-50%	17.3	17.7	4.6				
50-100%	40.4	32.0	5.6				
Total populations (1000s)	8,557	5,536	17,975	Total populations (1000s)	8,557	5,536	17,975

Sources: 1990 Decennial Census (Westat tabulation of STF-3)

10. SIMULTANEOUS OVERSAMPLING OF SEVERAL RACE-ETHNIC DOMAINS

In general, geographic-based oversampling can be used as easily and effectively for targeting multiple race-ethnic domains as for a single race-ethnic domain. In fact, the optimal sampling rates for the strata with high concentrations of each of the targeted domains will be about the same as if only it were being targeted. However, the overall level of screening will be increased since the number of areas with high sampling rates will increase with the number of targeted domains. Both these observations are due to the limited overlap between the highly segregated areas of the examined racial and ethnic minorities.

Table 7 presents some data on this subject from the 1990 Decennial Census. The only domains that overlap significantly in their concentrated areas are Hispanics and Asians and Pacific Islanders, and even that overlap only works one way. Since there are so many more Hispanics in the U.S. than Asians and Pacific Islanders, the proportion of Hispanics that live in blocks with Asian /Pacific Islander populations over 10 percent of the local population is only 13.7 percent while the percent of Asians and Pacific Islanders that live in blocks with Hispanic populations over 10 percent of the local population is a high 40.8 percent. The practical significance of this particular overlap is probably slight, however, since it would take such a large screening sample (both in and out of highly concentrated areas) to find enough Asians and Pacific Islanders to meet moderate precision requirements that such

Table 7
Residential Mixing of Minorities

Density stratum (Indicated minority as a percent of 1990 block in 1990)	Percentage of blacks living in the stratum in 1990			Percentage of Hispanics living in the stratum in 1990			Percentage of Asians and Pacific Islanders living in 1990			Percentage of American Indians, Eskimos and Aleuts living in 1990		
	Stratification domain			Stratification domain			Stratification domain			Stratification domain		
	Hispanic	Asian and Pacific Islander	American Indian, Eskimo and Aleut	Black	Asian and Pacific Islander	American Indian, Eskimo and Aleut	Black	Hispanic	American Indian, Eskimo and Aleut	Black	Hispanic	Asian and Pacific Islander
< 10%	79.2	95.4	99.6	73.4	86.3	99.1	78.9	59.2	99.6	85.9	81.4	95.1
10-30%	12.7	3.8	0.3	15.5	10.7	0.8	15.2	26.9	0.4	8.2	12.3	3.9
30-60%	5.8	0.7	0.0	7.4	2.5	0.1	4.2	10.8	0.0	3.3	4.5	0.8
60-100%	2.2	0.1	0.0	3.6	0.5	0.1	1.6	3.2	0.0	2.5	1.8	0.2

Sources: 1990 Decennial Census (Westat tabulation) a screening sample would probably find enough Hispanics without resorting to disproportionate allocation of the sample to blocks with higher concentrations of Hispanics.

11. CONCLUSIONS

For household surveys in the U.S., geographic-based oversampling using data from the most recent decennial census is a useful sampling strategy for improving the precision of statistics about the black and Hispanic populations provided that the cost of full interviews is less than 5 to 10 times the cost of screener interviews. It is also a useful strategy for improving the precision of statistics about the Asian/Pacific Islander and American Indian/Eskimo/Aleut populations, even at very high ratios of the cost of full interviews to the cost of screener interviews.

However, this does not mean that a survey of reasonable cost can be designed to simultaneously provide highly precise statistics about all these domains while maintaining desired precision levels for the total population. Most demographic surveys require reasonable precision for both targeted domains and for the total population. Shifting some portion of the full interviews from the white nonhispanic population to the other domains is bound to decrease the precision of statistics about the total population. It is generally useful to strike a balance between precision attained for subpopulations and the total population. The point of this observation is merely that geographic-based oversampling does not obviate the need to select very large samples and conduct many screening interviews when trying to obtain precise statistics about rare domains at the lowest possible cost. Furthermore, precise statistics about rare domains will continue to be expensive even when using geographic-based oversampling.

For surveys of low-income persons, only small gains are possible with geographic-based oversampling, and those only when the cost of a full interview is only a few times larger than the cost of screening and dropping a household. Most of these gains are likely to disappear when deterioration over time is taken into account. In fact, by the middle of a decade or later, when Census data become seriously outdated, there is the distinct possibility that geographic-based oversampling could reduce efficiency rather than improve it because of migration of the poor and sampling error in measuring poverty at the block group level. Geographic-based oversampling is a useful tool, however, when the focus of interest is on the black or Hispanic poor.

ACKNOWLEDGMENTS

This research was conducted by Westat Inc. under contract 200-89-7021 sponsored by the National Center for Health Statistics, Centers for Disease Control and Prevention. David Judkins and James Massey participated in the project while they were with Westat and NCHS, respectively. The authors would like to gratefully acknowledge the programming contributions of John Edmonds and Robert Dymowski of Westat and to thank the referees for their useful comments and suggestions on an earlier version of the paper.

REFERENCES

JUDKINS, D., MASSEY, J., and WAKSBERG, J. (1992). Patterns of residential concentration by race and Hispanic origin. *Proceedings of the Social Statistics Section, American Statistical Association,* 51-60.

KALTON, G., and ANDERSON, D.W. (1986) Sampling rare populations. *Journal of the Royal Statistical Society, Series A,* 149, 1, 65-82.

KISH, L. (1965). *Survey Sampling.* New York: Wiley.

MASSEY, J., JUDKINS, D., and WAKSBERG. J. (1993). Collecting health data on minority populations in a national survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 75-84.

UNITED NATIONS (1993). Sampling Rare and Elusive Populations. Department for Economic and Social Information and Policy Analysis, Statistical Division, National Household Survey Capability Programme. New York.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association,* 429-434.

WAKSBERG, J. (1995). Distribution of poverty in Census block groups (BG's) and implications for sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 497-502.

WAKSBERG, J., and MOHADJER, L. (1991). Automation of within-household sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 350-355.