SUPPRESSION VS. RANDOM ROUNDING

DISCLOSURE-AVOIDANCE ALTERNATIVES FOR THE 1980 CENSUS

Paul T. Zeisset

Data User Services Division

The purpose of this paper is to stimulate and support broad discussion within the Bureau of alternative disclosure-avoidance techniques for summary data from the 1980 census, specifically: suppression, random rounding, and other forms of introducing random variation.

The degree of protection required against statistical disclosure is necessarily a matter of judgment. The recently published Statistical Policy Working Paper 2 "Report on Disclosure and Disclosure-Avoidance Techniques" makes the important point that it is unreasonable to attempt to absolutely prevent any disclosure. For example, a statistic indicating that a block is 100 % Black discloses a characteristic of every resident of that block. Disclosure can be probabilistic as well, e.g., a figure of 90% discloses an individual's characteristic with high probability. Some level of disclosure must be tolerated in order to achieve important societal benefits. The task that this paper addresses is which technique most effectively maximizes utility of the data while reducing the risk of disclosure about individuals to an acceptable level.

One may note at this point that the use of the term suppression in this paper should not be confused with the practice of withholding certain derived statistics (e.g., medians or percents) if there were fewer than 100 persons, families or households in the distribution; or the practice of showing race/ethnic detail in the Census Tracts reports only if there were 400 or more persons in the race/ethnic category in a particular tract. In these cases the motives were statistical or practical and not related to the avoidance of disclosures about individuals.

This paper is generally organized as follows:

* Suppression rules used in 1970
* Inadequacies of the 1970 suppression scheme
* Two types of alternatives
  - combining areas
  - disturbing the data (including random rounding)
* User comments on the various alternative techniques
* Recommendations

originally prepared 1978

A. DISCLOSURE-AVOIDANCE PRACTICES EMPLOYED IN THE 1970 CENSUS

1. Rule of Five

In the 1970 census a system of "table suppression" was employed: if the number of entities in a particular critical universe failed to meet a particular criterion, then characteristics of the entities were withheld or "suppressed." This principle always allowed the publication of total population and housing counts.

In complete-count data the criterion was five. If there were fewer than five housing units in an area, then no housing characteristics were shown. The same area could, however, have five or more persons, resulting in the publication of population characteristics even while all housing data were suppressed. Population characteristics cross-classified by race were subjected to an additional level of scrutiny: there must have been five or more persons in a racial category before data (e.g., an age distribution) were shown for that race. For complete-count housing data, the rule of five similarly applied to each race-of-head category, to each tenure category, to race-tenure combinations, and to a few specialized universes dealing with the scope of rent and value.

The fact that the rule of five was applied only to certain critical universes is important. A table on household relationship of persons 65 years old and over could be shown even if there were only one such person -- as long as there were five or more persons in the total area -- since "persons 65 years old and over" was not considered a critical universe.

2. Adjustments for Sample Data

For sample data the suppression criterion was inflated by the average weight in the sample. For 20-percent sample data the rule of five became a rule of twenty-five in the estimate (representing roughly five sample cases), and twenty-five was used as the minimum number of persons or housing units in a critical universe for further data to be shown. Correspondingly, the cutoff was 33 for 15-percent data and 100 for 5-percent data.

The typically greater complexity of data tables made available from sample data was not considered, nor, on the other hand, was the fact that sampling substantially reduces disclosure potential relative to a particular individual who may or may not have been included in the sample.

A.2. continued

There are no corresponding suppression rules for sample surveys. Data cells are typically rounded to the nearest hundred or thousand so as to imply the imprecision of the data. In one case, where subcity data were made available from the Annual Housing Survey, the rule of five extrapolated by the average weight was applied.

3. Complementary Suppression

Complementary disclosure may occur where a distribution is published for both a total (e.g., all occupied units) and a subset of the total (e.g., owners). A distribution for the remainder (e.g., renters) can be obtained by subtraction. Complementary suppression was employed in relatively few types of cases in the 1970 census, and primarily in housing tables by tenure.1/ There was no complementary suppression in data by race (e.g., if data were shown for total and Black, there was no check to see that there were sufficient non-Blacks). There also was no attempt to avoid the suppression of only one ED within a small place, or one block within a block group or tract, even though the suppressed numbers in such cases could have been derived by subtraction.

4. User Information

These disclosure rules were not initially explained to users, on the assumption that revealing them would make actual disclosure more likely. However, in December 1974 the suppression rules were published in Data Access Description No. 36 "1970 Census Fifth Count for Zip Codes, Counties, and Smaller Areas," given that a number of sophisticated users had by then figured out the essential points, the data were becoming increasingly out of date and less likely to be critically compared with observation, and suppression was quite prevalent in sample data being made available at the enumeration district and block group level in the Fifth Count.

B. INADEQUACIES OF THE 1970 SUPPRESSION SCHEME

The suppression scheme used in the 1970 census had two major types of drawbacks: 1) it didn't always prevent undesirable disclosure and 2) data users, especially those dealing with summary tapes, found it very difficult to deal with.

1. Inadequate Protection Against Disclosure

a. Complementary disclosure

Section A.3 above describes two types of complementary disclosure (among race categories and among areas) not

prevented by the 1970 suppression rules. Certainly such inadequacies could be addressed by making the scheme more sophisticated. The complexity of doing so, however, should not be underestimated. Census geography is not strictly heirarchical, with places and congressional districts crossing tract and other boundaries. Further, not all data are produced in the same series, and the system would ideally take into account data produced in other series, including special tabulations.

b. Noncritical universes

From a disclosure-avoidance point of view, the rule of five might apply to every universe. The fact that the rule of five was applied only to certain "critical" universes simplified suppression considerably for both the producer and the user. In population data, once it is established that data for Blacks can be shown, then all data with the same sampling rate can be shown -- including cross-tabulations. On the other hand, this can easily yield disclosures, as in the case where there is only one person 65 or over in an area and his/her relationship is shown in a relationship-by-age table, or in the case where there is only one employed person in an area (but assuming a sample estimated total population of 25 or more) -- that person's occupation, hours worked, and other labor force characteristics would be shown.

Another limitation of the critical universe approach for 1970 population data was that data were frequently reported in terms of families and households, although the criteria were in terms of persons. Thus, for example, family income, welfare recipiency and other sensitive characteristics could be reported for a single family of five or more persons if no other families lived in the same enumeration district.

Still another variation on the same theme could occur in a single one-dimensional table providing single years of age. Let's assume that there were dozens of persons represented in the table for a tract, all but one of whom were under 65. By reporting that one person in the tract was 83 years of age, it would obviously be disclosed that the one elderly person in the neighborhood was 83.

2. User Difficulties

Any time that desired data are withheld there will naturally be some frustration or inconvenience for the user. That inconvenience was accentuated in the 1970's with the issuance of unprecedented amounts of small-area data and their use in computerized form.

a. Programming difficulties

Programming with 1970 census summary tapes was made especially complicated by the fact that one had to allow for negative numbers representing suppression in the same fields as actual data values. The fact that there were several "types" of suppression to keep track of made programming still more complex, although, if the programmer understood them, certain suppression types allowed short-cuts for the computer. The fact that distributions of zero universes were suppressed on tape (even though shown as all zeroes in print) undoubtedly led to many uncharitable mutterings about the Census Bureau.

The programmer's task was made more difficult by his or her incomplete knowledge of our criteria. The documentation did not discuss the suppression of zeroes (later documented in Small-Area Data Notes and Census User Bulletin articles), nor was it documented that any suppressed number was predictably less than 5 (or 25, 33, or 100 in the case of sample data).

b. Area aggregation

The tremendous detail of the 1970 census was consumed so voraciously by data users not so much because data were needed in such detail for the typical application, but more because the detail allowed flexibility in aggregating the data to areal units more meaningful in one or another context. But, in aggregating summary data to "neighborhood" levels less prone to disclosure than the block or ED statistics used in the calculation, the frequent suppressions at the lower levels sometimes prevented the derivation of useable aggregates. Some users did the calculations anyway, accepting the erratic downward bias in their various statistics.

3. Exceptions

The incompatibility of suppression with certain uses in which flexible area aggregation is essential is illustrated by one exception to the normal suppression rules made for the journey-to-work special tabulation package performed for 121 urbanized areas. For a traffic-zone-of-residence by traffic-zone-of-work by mode-of-transit matrix no suppression was applied, although traffic zones were occasionally very small in terms of population. This decision, and subsequent decisions relating to other travel-to-work characteristics in the AHS, articulated the principle that travel-to-work was not a personal characteristic, was highly changeable over time, and therefore was not likely to disclose any information linkable to a particular individual at a later time. Such an argument would seem to apply equally to hours worked last week and perhaps other characteristics and might therefore be an unfortunate precedent for other loopholes.

C. CONSOLIDATED AREAS -- A VARIANT OF SUPPRESSION

Users have occaionally suggested that suppression should be avoided by combining an area to be suppressed with another so as to create a base population large enough to meet whatever disclosure criteria we have. This alleviates the programming difficulties associated with missing data, and provides the user with complete matrices useable in area aggregation -- except where the two areas combined straddle the boundary of the user's desired area. Since suppressed numbers are always very small (or were so in the 1970 scheme), the impact on the second area into which the suppressed values are combined is not very substantial.

Despite its intuitive appeal there are several difficulties associated with implementing such a scheme. In the 1970 scheme, multiple criteria were used -- e.g., data for Blacks may have needed suppression while the total distribution could be shown, or data for owners may have been acceptable while data for renters were not. Either a single criterion would have to be used (e.g., 5 households) to govern the combining of areas, or else any one of several criteria might be used to force consolidation of the area, an alternative likely to lead to too much suppression if race categories are part of the criteria.

Other problems include the technical difficulty of selecting an appropriate second area for data from the first "disclosure-prone" area to be combined into. Alternatives range from selecting the area with the next higher number (presumably adjacent on tape) to tests of geographic adjacency using a geographic base file.

D. DISTURBING THE DATA AS AN ALTERNATIVE TO SUPPRESSION

Any unreliability in a data base incidentally reduces disclosure probabilities. If respondents lie to us, our summary data won't disclose their characteristics. Allocations, substitutions and processing errors all help the respondent retain anonymity. These errors do, however, degrade the general utility of the data, particulary where they introduce bias, and no one is advocating their intentional use for disclosure avoidance. Other forms of error, if unbiased and limited in size, may constitute significant disclosure-avoidance techniques, and may be employed as alternatives to suppression. Our census-taking colleagues in Canada and Britain have implemented various techniques of introducing limited random variation, which I term noise, into summary statistics.

1. Ordinary Rounding

Conventional rounding is the simplest example of disturbing data. Figures in a table, for example, may be rounded to the nearest multiple of 5. Where the figures involved are relatively large, this has little or no effect on the information value of the tables.

Ordinary rounding to multiples of 5 was used for most tables involving cross-tabulations for large areas in the 1971 Census of Population in Great Britain. Values were calculated from unrounded data and then rounded. Percentage figures were computed using rounded data and therefore did not necessarily add to 100%. The same technique was used for both 100% and 10% sample data.

Ordinary rounding obviously produces a variety of inconsistencies in tables. Only rarely do totals of component figures add to the total shown, and where total figures are analyzed in different ways in different tables the figures obtained by adding the respective component data cells will usually be different.

2. Random Rounding

"Random rounding" is a technique invented by Statistics Canada for use in all tabulations (100% and 33 1/3%) from its 1971 and 1976 censuses of population and housing.

In random rounding, each figure is rounded to a multiple of some integer, usually 5, but not necessarily to the nearer one. Whether a figure is rounded up or down is determined by chance, but with overall probabilities defined as follows, assuming a base of 5:

Final digit of number	Probability of rounding up
0 or 5	0
1 or 6	1/5
2 or 7	2/5
3 or 8	3/5
4 or 9	4/5

Thus, for example, 126 would be rounded up to 130 with probability of 1/5, or down to 125 with probability of 4/5. In conventional rounding 126 would always be rounded to 125.

Murphy discusses two advantages of random rounding relative to ordinary rounding:

> First, it insures against the possibility of deriving the original figures by comparing cells in a table against the independent rounded totals. Second, and most important, it makes the sum of the rounded numbers an unbiased estimate of the sum of the original numbers. This will not be the case in conventional rounding unless there is an even distribution of last digits. In census data there tends to be a preponderance of small last digits, and if conventional rounding were used, the sum of the rounded numbers would tend to underestimate the totals of the original numbers.[27]

Fellegi has discussed a mechanism for controlling the random rounding to assure that the totals would be subject to only the minimum rounding error at some predetermined higher geographic level,[37] but that idea has not subsequently been implemented.

3. Introduction of Noise Other Than Random Rounding

In their 1971 census, the British employed two types of disclosure-avoidance techniques: first, the ordinary rounding described above in cross-tabulations published for large areas, and second, a combination of random noise and suppression for enumeration district data, discussed here. At the ED level the British provide a number of one-dimensional distributions and two-dimensional cross-tabulations of data detail comparable to our own ED data. (Neither the British nor the Canadians, however, provide block statistics.)

Each data cell for an enumeration district was modified by +1, 0 or -1, in the ratio of 1, 2, 1 (e.g., 50% of the time there is no change). For every such adjustment there was a compensating adjustment in a second ED with thich the first had been paired, such that the sums of ED tabulations generally agreed with ward or parish totals as long as there were an even number of ED's within the area. Within an ED, totals were derived only from the adjusted data. This technique was not considered sufficient for all potential disclosure situations, and was supplemented by the suppression of all data for any enumeration district with less than 25 persons or less than 8 households.[47]

Since the British system does not relieve the necessity for suppression, it fails to achieve one of the most important potential benefits of noise introduction. On the other hand, it is obvious that larger amounts of noise could be added in order to avoid the necessity of suppression. One such example would be the introduction of errors from -4 to +4 in proportions dictated by a binomial distribution. The effect would be similar to that of random rounding except that final digits would not be constrained to 0 or 5.

The Swedish Statistical Bureau has proposed another variant based on the assumption that any value of one represents a disclosure. Their proposal would round a data value of one down to zero with a probability of 2/3, and up to 3 with a probability of 1/3.

4. Disturbing the Underlying Microdata

In each of the previously described methods, including suppression, the data are tabulated in the conventional fashion before any noise is added or any figures suppressed. With these methods, the unmodified tabulations may be retained by the producer for internal use or further manipulation. Another possible approach is to introduce noise into the underlying microdata, e.g., modifying an age or race at random. This technique would avoid inconsistencies in tabulations, but is fraught with a number of problems, including the inability to gauge its impact on a wide variety of tabulations. This issue is further discussed in a paper by Dalenius.[5]

5. Advantages and Disadvantages of Noise Introduction

a. Effectiveness in disclosure avoidance

In general the various forms of introduced noise result in figures which are insufficiently exact to disclose information about individual cases. A small number in a tabulation cannot be precisely associated with specific individual(s) since the user has no assurance that the number is correct. Further, many disclosure-prone values are changed to numbers which do not suggest disclosure (e.g., when a one is changed to a 5).

With the exception of the Swedish proposal where only small numbers are affected, these various techniques also protect against complementary disclosures or disclosure-by-subtraction. A disclosure-prone statistics A cannot be obtained by subtracting published statistic B from a published sum of A and B since neither B nor A + B are reported exactly (e.g., instead of 437 - 436 = 1, one might have, with random rounding, 440 - 435 = 5 or 435 - 435 = 0, or even 435 - 440 = -5).

b. Area aggregation

With noise introduction there are no missing values in summary tables, and the user can combine small area summaries into larger aggregates without having to worry about the downward bias created by the removal of values by suppression.

Aggregation of statistics which have been random-rounded yields a sum which is an unbiased estimate of the sums of the unrounded numbers. The variance added by rounding is a function of the number of statistics being summed and is unaffected by the magnitude of the numbers being summed. (Specifically the variance of random rounding error is 4N, where N is the number of data cells used to produce the total.)[6] If controlled random rounding were introduced, as proposed by Fellegi, the variances for certain are aggregations would be reduced. Neither of the British systems produce unbiased sums.

Neither of the British systems produce unbiased sums. Sums of conventionally rounded numbers will be biased if there is a preponderance of small last digits in the numbers being summed. In the introduction of noise not constrained to multiples of a particular number, negative numbers are possible results. If negative results are disallowed, as in the British system, an upward bias is introduced.

c. Design of data tables

On 1970 census summary tapes very few totals were given, since the user could always compute them from the cells shown. Where noise is introduced into the cells it is highly desirable to have independently adjusted totals so that the user is assured that the noise is within a certain limited range. Thus, if noise is introduced, the size of summary tape matrices should be increased to allow for marginal totals and at least some subtotals (e.g., data for persons 65+ in a more detailed age distribution).

d. Computation of derived statistics

In the Canadian and British rounding schemes, totals are derived independently. Percentages, on the other hand, are derived from the rounded data so as not to reveal the underlying values. Means can be calculated from the original data. Aggregates, on the other hand, are much less amenable to rounding and need to be manufactured by multiplying the actual mean by the rounded number of cases in the universe upon which the mean was based.

e. Cross-checking for errors

One of the concerns voiced around the Bureau is that with the introduction of noise, real data errors could go undetected. In random rounding, for example, the same value might be

rounded up in one case and down in another yielding an apparent discrepancy (e.g., 436 rounded to 435 or 440), while on the other hand, two values could differ by as much as 8 and go undetected (e.g., 436 and 444 could both be rounded to 440, although by chance a discrepancy of 8 would be masked less than 1% of the time). Of greater consequence would be the uncertainty in dealing with a series of numbers summed. Clearly, checking for programming errors would need to be done using intermediate tapes with unrounded data.

f. User misunderstanding

A direct corollary to Bureau problems in checking for errors would be the user's difficulty or uncertainty in checking for his/her own programming errors. Further, the fact that figures may not add up right has occasionally been the key to user realization of the distinction between household and family counts or to the discovery of other errors of interpretation. Addition of random noise, on the one hand, hides some real discrepancies, but on the other hand, and perhaps more significantly, may lull the user into discounting real discrepancies as attributable to rounding error.

Among the various forms of error introduction, ordinary and random rounding to a base of 5 or 10 have one distinct advantage over other forms which allow the full range of final digits. When every frequency count ends in 5 or 0 the user cannot escape noticing the fact that the data have been modified.

This user awareness that each data item is subject to some error can have beneficial side-effects. Too many unsophisticated users apply what may be called an "accountant mentality" to census data, taking each number as exact and being oblivious to the various nonsampling and sampling errors that affect the data. Quite a few users have been introduced to sampling variability in seeking an explanation for discrepancies between complete-count and sample reports.

Whether discrepancies are immediately obvious or not, the introduction of noise to summary statistics cannot help but lead to discrepancies among data tables. As long as there are many distributions in which independent figures can be derived for owner-occupied units, persons under 18, families, and so forth,

discrepancies from one table to another are inevitable. Only if it is the underlying microdata, rather than summary data, which have been disturbed can absolute consistency among tables be maintained.

g. Conflict with demands for precise counts

It is not immediately clear whether courts, Congress, or other legal authorities would contend that the intentional disturbance of census counts would violate the need for precise data in Congressional redistricting. Further, the process of local review will require unmodified counts at least at that stage.

Therefore, one can conceive of the provision of precise counts for total population and total housing units in each area down to the block level, but with all other numbers subjected to random adjustment. This has, however, a serious drawback where the actual number of persons or housing units is very small. For example, take the case of a block with a population of one, subjected to random rounding: on the average four-fifths of all data tables would have only zeroes, but one-fifth of those tables would show a value of five in one category -- each category with a value of 5 could be deduced as a characteristic of the one resident, and one-fifth of his or her characteristics would be disclosed. Regardless of this acknowledged danger, Statistics Canada does provide exact population and housing counts along with random-rounded characteristics, although the problem of extremely small populations is less prevalent there since no block statistics are provided and the smallest reporting area is the enumeration district.

6. Choosing Among These Techniques of Noise Introduction

If one is to choose among the three major alternative methods for introducing noise for disclosure avoidance -- ordinary rounding, random rounding, and the British system for small areas -- the choice centers around three factors: understandability to data users, the degree and kind of bias introduced by the various techniques, and the amount of noise necessary to avoid inferences of confidential information.

Ordinary rounding to the base 5 has the advantage of being familiar and understandable to all users. Its primary drawback is that, since census figures typically have a slight predominance of small

last digits, a sum of rounded numbers does not produce an unbiased estimate of the sum of the unrounded numbers. There are also a number of combinations of values and tables where the underlying distribution can be deduced exactly.

The British system of compensating errors for paired small areas within larger areas does not affect the data in such a way that it is immediately obvious that noise has been introduced (a disadvantage shared by rounding to bases other than 5 or 10). Discrepancies can be avoided in the summing of enumeration areas to a larger area total, however this does not preclude other discrepancies, such as differing values for the number of persons 65 years old or over in different tables for the same area. Secondly, if negative numbers are disallowed (e.g., zero is never reported as -1 even if required to compensate for the addition of +1 to a corresponding statistic in the other paired enumeration district) a small bias is introduced. The most serious drawback of the British system would seem to be that it does not alleviate the need for suppression, since the noise added (in the range -1 to +1) is not sufficient to mask small disclosure-prone values. Errors taken from a larger distribution would be necessary to avoid suppression.

Among the alternative noise-introduction techniques, the one of choice would seem to be random rounding. Assuming a base of 5 or 10 is used, the modification cannot go unnoticed by the user. Sums of random rounded numbers are unbiased estimates of their unrounded counterparts. Finally, the amount of noise added should be sufficient to preclude direct and indirect disclosure.

E. USER COMMENTS ON THE VARIOUS ALTERNATIVE TECHNIQUES

1. Data User Services Division Contacts

The Data User Services Division and the organizational components that preceded it have had a significant amount of contact with data users regarding their experience with current disclosure-avoidance practices. Unfortunately, there is no effective documentation of most of the user comments since most of the contacts were at conferences or by telephone. A number of generalizations are possible, however, based on the experience of key DUSD staff, and have been discussed earlier in this paper, especially:

- Frustration with the technical complexities of dealing with the suppression indicators on 1970 summary tapes.

- o Frustration with the high proportion of data suppressed on the various tapes, particularly in data for blocks, ED's and BG's.
- o Frustration with the impediments to area aggregation engendered by suppression.

2. 1980 Local Public Meetings and Correspondence

A much better level of documentation exists with regard to input received at the 73 Local Public Meetings and the 16 State Agency Meetings, as well as the various letters filed and indexed over the last few years by Decennial Census Division. It should be noted that data dissemination issues were not particularly stressed in the two series of meetings and the frequency of comments on suppression was not high. Nonetheless, at three meetings and in three letters, suppression was denounced in very strong terms such as describing the resultant summaries as "nearly worthless." Three of the meetings yielded unprompted recommendations that the Bureau adopt random rounding in lieu of suppression, an idea also supported by four letters. Two meetings and two letters espoused the idea that small area data to be suppressed should somehow be combined with data for other areas to avoid suppression. Two meetings and four letters supported various technical reforms to suppression. Suppression was mentioned in other meetings and correspondence, but without particular emphasis or recommendations.

3. Summary Tape Processing Center Conferences

Two conferences were held late in 1977 for representatives of Summary Tape Processing Centers and other users of Census Bureau data on computer tape. The introduction of random noise in lieu of suppression was specifically mentioned as a possibility in the background paper for the conference but was not discussed extensively prior to the deliberations of the working groups which generated the conference recommendations. Each of the 4 relevant working groups and one of the other groups chose to comment on the subject. Two specifically recommended the adoption of random rounding, two more advocated that the Bureau consider random rounding as an option, and one declined to advocate random rounding over suppression primarily because it did not have enough information about random rounding.

The users represented at these conferences were, of course, not a cross section of census data users, most being interested primarily in computerized data products. Incidentally, at the August 1978 meeting of the Urban and Regional Information Systems Association, it was noted that the Bureau was seriously considering alternatives to suppression in the 1980 census. That announcement was greeted with cheers from the audience.

4. Users of Canadian and British Data

An important group to be heard from are the users of those data bases which have already been released in "disturbed" form. Unfortunately, documentation in this area is almost totally lacking. No surveys have been taken and no papers have been written on the subject. The author has queried a number of relevant Statistics Canada officials, most of whom are unaware of substantial complaints about random rounding. However, inquiry clerks in the Toronto and Montreal regional offices indicate that, for the inquirer with no previous experience with census data, random rounding causes trouble and confusion. At minimum it simply requires time for explanation.

One paper on the British techniques does mention complaints about discrepancies between published large area figures and corresponding figures produced by aggregating small area data, even though those discrepancies are controlled to be quite small. The magnitude of those complaints was not characterized.

5. Conclusions

From the foregoing it can be inferred that there is not presently any significant opposition among users to systems which introduce controlled noise in the data to protect against disclosure. At the same time there is within the Bureau significant skepticism about the ability of unsophisticated users to deal with the repeated discrepancies which are inevitable with random rounding or other noise introduction techniques. The Bureau may well wish to further explore user reactions through user surveys, conferences such as those which have been held for STPC's, and papers at professional conventions.

F. RECOMMENDATION -- TWO ALTERNATIVES

From the foregoing it may be evident that the author favors the institution of random rounding as the disclosure mechanism for the 1980 census. The continuation of suppression is, however, also discussed in this section, including a number of reforms or improvements which address some of the inadequacies of the system used in 1970.

1. Random Rounding With Area Consolidation

The basic proposal is to use random rounding with a base of 5. One major variation from the system used by Statistics Canada is proposed,

however, given the substantial desirability of providing unaltered population and housing counts for each area, yet given the dangers discussed in section D.5.g (that in an area known to have only one person or housing unit, characteristics rounded to 5 are disclosed as characteristics of the person or unit).

It is proposed that any area with more than zero but less than 10 persons be combined with another area for the purpose of reporting characteristics. (In 1970, about 6% of blocks and 1.8% of ED's had 1 to 9 persons.) This combination could not affect area mapping; but would be represented in reports something like this: "Block 101+102". A supplementary table would be published giving the exact population and housing counts for each area before combination, an idea consistent with current plans to provide supplementary tables for blocks and enumeration districts with zero population.

Selection of the most appropriate algorithm for area consolidation will require research. The simplest alternative would be to select the next area in the sequence of areas on an internal unsuppressed summary data file. This could be satisfactory insofar as blocks or ED's have been numbered in a serpentine fashion such that areas with adjacent identifying numbers are usually physically adjacent. Some modification would be necessary to avoid combinations that cross block group, place or other higher level boundaries. A second alternative would be to select for combination the smallest area which is adjacent to the first area with under 10 population (as long as it is within the same higher level geography). This would be highly desirable from the user's point of view but would also be technically difficult, since adjacency information would have to be obtained from a GBF/DIME file or generated clerically outside of GBF coverage.

Some question may be raised regarding the adequacy of random rounding to multiples of 5 for 1-in-6 sample data, since any number reported as 5 has a high probability of representing a single case (probably over 95% probability). One alternative might be to round sample data to a base of 10 instead of 5, an operation which would also call user attention to the fact that those data are based on a sample. The need for doing so, however, should be evaluated in light of the fact that sampling actually reduces disclosure potential in most situation.

Area consolidation in sample data should be consistent with that done for complete count data. Thus, if two ED's are consolidated in complete count tabulations, the same two areas should be consolidated in sample data (i.e., sample estimates would not be considered).

2. Suppression Reformed

Any suppression scheme seriously proposed for 1980 must be more conservative generally than that used in 1970, given the inadequacies of the 1970 system in protecting against certain kinds of disclosure as discussed in section B.1.(pages 3-4).

a. Basic methodology

First, it is suggested that no characteristics be shown for an area with fewer than 5 occupied housing units, regardless of the number of persons. This would help erase 1970 inconsistencies between population and housing suppression (e.g., a single 5-person family had all personal characteristics reported but no housing characteristics), and alleviate some problems with data being shown for noncritical universes of persons of under 5. This rule does not address treatment for an area with many vacant housing units or many persons in group quarters but without the required number of occupied units.

Retention of race and tenure as critical universes, to which the minimum of five households is also applied, appears inevitable since these variables are so frequent as stratifiers. Nonetheless, the basic logic of disclosure-avoidance suggests that every cross-tabulation should be scrutinized to see that no marginal total is too small. (For example, an age-by-marital-status table could disclose the marital status of the only elderly person in a block.) There are, on the other hand, compelling reasons for not making the disclosure mechanism more complex than necessary. (In fact, most disclosure literature deals with distributions relating to one or two or three units. The suppression criterion of 5 households therefore provides some leeway.)

Full complementary suppression should be implemented. Not only does that mean complementary suppression between owners and renters, as done in 1970, but also complementary suppression by race and by area. For example, if a place contains three ED's and data for one must be suppressed, the suppressed data should not be derivable by subtracting the other two from place

totals; presumably the smaller of the remaining ED's would also be suppressed. Such a system will no doubt be expensive to implement.

For sample data, the inflation of the suppression criterion by the average weight (e.g., 30 for a 1-in-6 sample) may be more conservative than necessary. A criterion of 10 in the weighted count of occupied units would seem adequate, which would in effect eliminate only areas with a single household falling in the sample. The result would also not be far different from the 1970 criterion of 25 persons.

The advisability of continuing the special exception for journey to work data described in section B.3. can be debated separately.

b. Technical reforms in the preparation of data on tape

(1) Documentation

Whatever method is selected it must be possible to provide the user with full documentation of the techniques, complete with pointers on how to anticipate what data are suppressed.

(2) Type 2 suppression

Type 2 suppression was an awkward practice in 1970 necessitated by the absence of a data cell for total population on certain summary tapes. The addition of one cell to the data matrices will alleviate the need for it.

(3) Suppression indicators

Data fields should contain data only (i.e., no "-1" for suppression). Suppression indicators should be reserved to separate "flag" fields. In 1970 only 17 one-character flags would have been needed, one for each critical universe observed. Suppressed data cells would be blank. Actual zero cells would contain zeroes and therefore would be distinguishable from the blank suppressed cells in visual inspection of printouts.

(4) No suppression of zero distributions

There should be no zero suppression, i.e., superfluous suppression of the zeroes in a distribution known to have no cases. (This occurred only on tape -- zeroes were not suppressed in 1970 reports.)

(5) File of suppression indicators

During the creation of user tapes a separate file of suppression flags with a record for each area should be created. That file would also be made available in some eye-readable form. Thus, a user interested in particular small areas could determine, prior to programming a retrieval, whether the desired data were unavailable because of suppression.

(6) Report on suppression impact

Some report should be prepared on the frequency of suppression, average values suppressed, impact on certain kinds of problems, etc.

3. Compromise Between Random Rounding and Suppression

Some observers have suggested that the best answer may be somewhere between suppression and random rounding. A working group at one of the STPC conferences in 1977 actually suggested the use of random rounding for tapes, but with suppression in reports.

It can be argued that the provision of precise unrounded counts at some high geographic level would not perceptibly undermine the effectiveness of random rounding for small areas. Unrounded State totals, for instance, might be useful to the user wishing to make sure that he/she knew just which figures were supposed to add up. Random rounding in National level reports would serve no purpose other than consistency with small area presentation. Whether exact counts could be preserved below the State level, for instance down to the SMSA and county level, can be determined only after further research. Nonetheless, inconsistency of disclosure techniques from table to table or report to report should be discouraged.

G. NOTES ON FURTHER RESEARCH NEEDED

Certain statistical research or other work may be appropriate during the discussion of alternative techniques. These include:

- o The most feasible method for area consolidation
- o How best to represent the consolidation of two or more areas on tape
- o Estimates of the impact of random rounding on various statistical applications
- o The effects and advisability of implementing controlled random rounding as proposed by Fellegi, or any other method for controlling rounding errors in such a way as to be compensating and to sum to higher level totals
- o Whether any methods are feasible for controlling rounding errors in such a way as to be compensating and to sum to higher level totals
- o Whether any changes to the rounding algorithm can reduce rounding variance
- o To estimate the impact of random rounding on various statistical applications
- o Modifications to the suppression scheme described on page 17 needed to account for the situations where there are few households but significant numbers of a) vacant housing units or b) persons in group quarters.
- o An efficient algorithm for complementary suppression if suppression is adopted.

FOOTNOTES

1/ As an interesting aside, such housing tables were not necessarily suppressed in the same way in reports and on tapes. In reports, owner and renter data were frequently published without a corresponding distribution for all occupied units. If renter data were suppressed owner data were not affected. However, on tape, there was typically an additional distribution for occupied units, requiring the complementary suppression of both the owner and renter data if either failed to meet the criterion. We are not aware that any user ever discovered that one could, in those few cases, derive the suppressed data by combining information from tape and reports.

2/ Michael Murphy, "Confidentiality and Its Effects on Census Data." Unpublished paper. Census Division, Statistics Canada, Ottawa.

3/ Ivan P. Fellegi, "Controlled Random Rounding." Survey Methodology 1:123-133. Statistics Canada, Ottawa 1975.

4/ Dennis Newman, "Techniques for Ensuring the Confidentiality of Census Information in Great Britain." Paper read at the 2nd Session of the International Association of Survey Statisticians, Warsaw, 1975.

5/ Tore Dalenius, "The Invasion of Privacy Problem and Statistical Production -- An Overview." Statistisk Tidskrift, National Central Bureau of Statistics: 213-225, Stockholm, 1974.

6/ J.G. Stinson, "Effects of Random Rounding on User Aggregated Data." February 1973. This paper also includes charts showing maximum absolute error and maximum percent error expected at 3 confidence levels as a function of the number of data cells.