Journal of Official Statistics
Vol. 5. No. 2, 1989, pp. 171–182
© Statistics Sweden

On Finding and Delineating Area Sample Units in the Field

Alan Sunter[1]

Abstract: Multistage population samples almost invariably use area sampling, the areal units at one stage of sampling frequently being the enumeration areas defined by the most recent census. However, problems in identifying the selected EA's in the field can lead, particularly in developing countries, to serious errors of coverage in the survey. This article describes the types of problem that can occur and suggests procedures for ensuring that they are detected and resolved.

Key words: Area sampling; survey mapping; non-sampling errors.

1. Introduction

Multistage population samples using the household or dwelling as their ultimate sampling unit almost invariably use area sampling at one or more stages of selection. The unit used at the last, or next to last, stage of area sampling is usually the census enumeration area (EA), defined in the most recent population census and typically containing (at that time) a population of the order of 500 to 1500. Typically the statistical office will maintain, in one form or another, a set of EA files each containing an EA sketch map, description, some summary information from the census enumeration, and perhaps the census household listing itself. These EA files will usually be organized by some larger unit in the administrative hierarchy, say the Census District, for which there will be a file containing an index map for all the EA's within it. Excellent sources on the organization, administration, and techniques of census cartography are the training manuals published by the U.S. Bureau of the Census (1978, 1979).

However, since there are something like 1000 EA's per million of population, and they will have been defined, described, and sketched over a relatively brief period of one or two years in the pre-census mapping operation, it should not be surprising that the quality of the identification information in the files is sometimes low. This may not have been too serious in its implications for the census itself, since all of the EA's in a census district are being enumerated simultaneously. Under these circumstances little damage will be done if the decisions of the district supervisor as to which households are assigned to which EA are more or less

[1] President, A.B. Sunter Research Design and Analysis Inc., 63 Fifth Ave., Ottawa, Canada K1S 2M3.

Acknowledgements: The author would like to thank the referees and an editor for their helpful suggestions for some areas of the article in which his own description of problem detection and resolution procedures left ambiguities.

arbitrary, so long as each household is enumerated once and once only. Such arbitrariness is not permissible, however, in subsequent surveys in which isolated EA's have been selected in a probability sample.

The problem is compounded in many developing countries where

> The maps (topographic, cadastral, route, administrative, etc.) on which EA identification and control would normally be based are likely to be badly out of date or even non-existent.
>
> Both the survey enumerators and the local administrators with whom they must liaise are relatively unfamiliar with the map as a device for explaining and defining locations, relationships, and affiliations; in other words, where there is a general shortage of even minimal cartographic skills.
>
> There is a multiplicity, particularly in Africa, of spoken languages and practical difficulty in reducing many of them to the written form in which they must appear on a map or list, with consequent ambiguity in the identification of places by their place names.
>
> Physical features, particularly man-made features such as dwellings, roads, buildings, and even whole hamlets or villages tend to be rather less permanently located in the landscape than in developed countries.

Under these circumstances, mere exhortations to perform area mapping and subsequent relocation carefully and conscientiously may not be very helpful in resolving the problems that arise. Indeed, in at least one African country in which the author has worked such exhortations may have been part of the problem; the enumerators strove to demonstrate "care and conscientiousness" by always including doubtful areas in their sample EA's. In these cases substantial overenumeration would have occurred had the problem not been corrected by procedures along the lines of those described in this article.

2. Examples

Figures 1 and 2, based on real cases but simplified for the purposes of this exposition, illustrate a situation commonly occurring in rural areas.

Figure 1 represents dwelling clusters, or hamlets, and EA "boundaries" for a Census Enumeration District (ED), superimposed on a topographical map. The hatched areas are hamlets, with their approximate populations given for all hamlets except the one whose boundary is represented by a broken line. This exception is a hamlet which either did not exist at the time of the census or was overlooked by the mapping team. (In any event, whatever the reason for its omission, it now exists but is not represented on the ED index map or any of the EA sketches.) The census mapping team represented the locations of the hamlets as best they could (without survey instruments and with limited training) and then, since the instructions (let us suppose) were to form EA's of approximately 1000 population, sketched in EA boundaries as shown. These bear little relation, it should be understood, to land boundaries as they might be shown on a formal plan of survey. They are not marked on the ground and cannot be relocated by measured distances and bearings from permanent reference marks.

The final step in the mapping phase of the census was to prepare EA sketches and descriptions for insertion into the individual EA files. The sketch for EA/01 looked something like Fig. 2.

In practice "A" and "B" will be the local

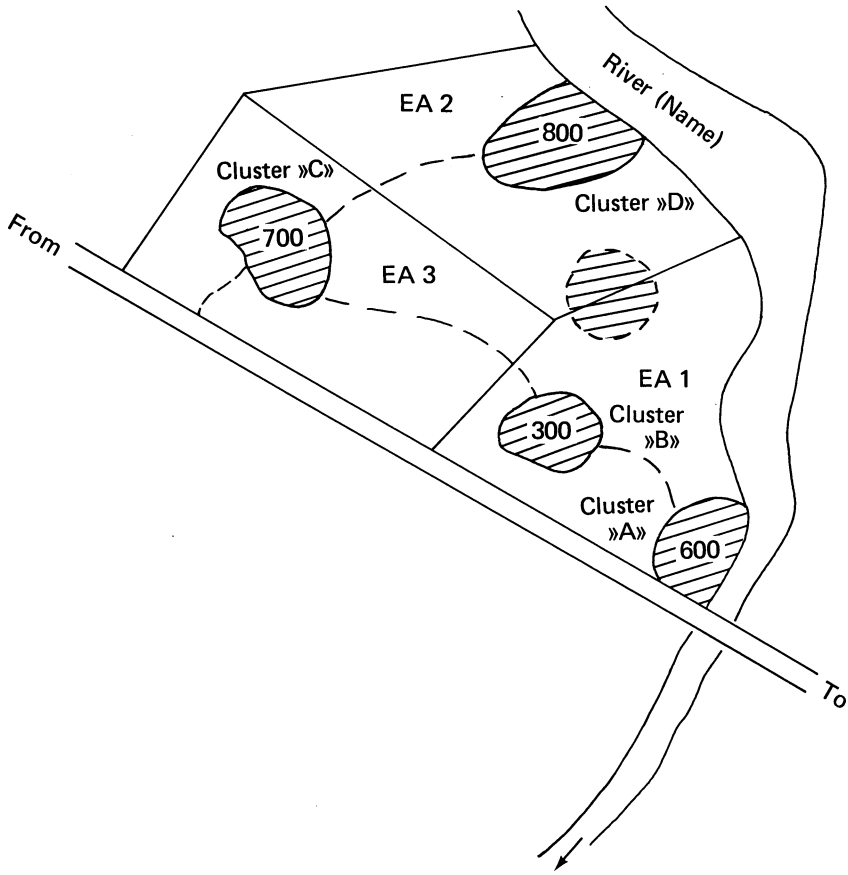


Fig. 1. Part of an Enumeration District Control Map

names by which these hamlets are known, or at least the best approximation that the cartographer is able to make of names for which there may be no agreed or customary written representation.

Now suppose that EA/01 is selected in a sample, drawn perhaps some years after the census (and even more years after preparation of the census maps) and that an enumeration team goes to the field equipped with only the EA file for this EA. Suppose further that it succeeds in finding and correctly identifying hamlets A and B but finds, in addition, the hamlet omitted in (or constructed since) the census. We show the boundary between EA/01 and EA/02 as passing through the omitted hamlet not as a supposedly correct representation (since the boundary was only an arbitrarily drawn line, not a surveyed and relocatable boundary) but as an indication that the enumerators think it could belong to either EA. Should this cluster be enumerated as part of the selected EA or not? It will be clear that the enumerators cannot resolve, or necessarily even identify, the problem by reference to the materials that they have at hand. Some further information and some methodology are required.

Figures 3 and 4, again adapted from an actual case, illustrate an urban area. Figure 3 shows the outline of selected EA/330 and its relationship, as best it could be determined from the mapping files, to its

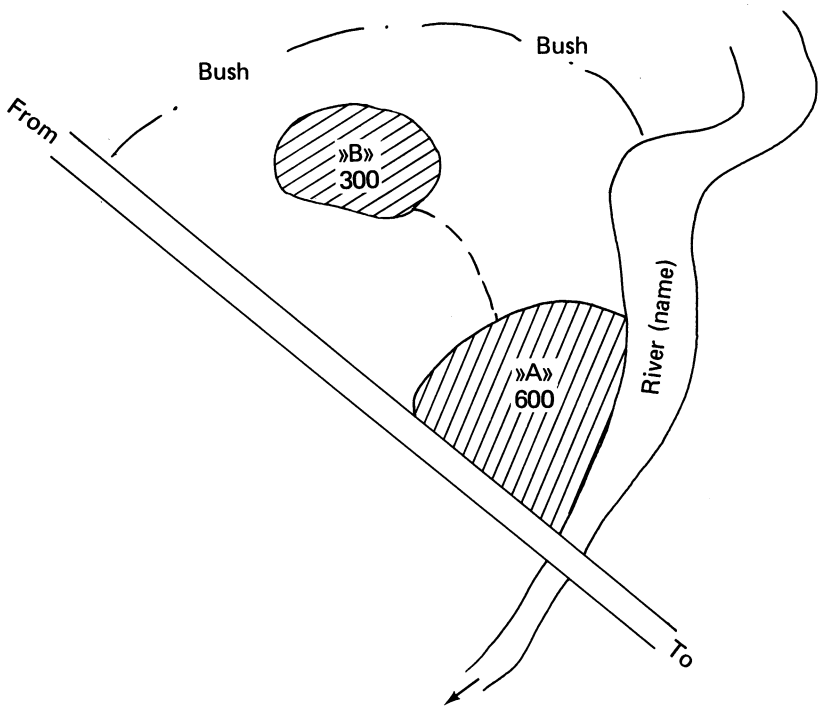
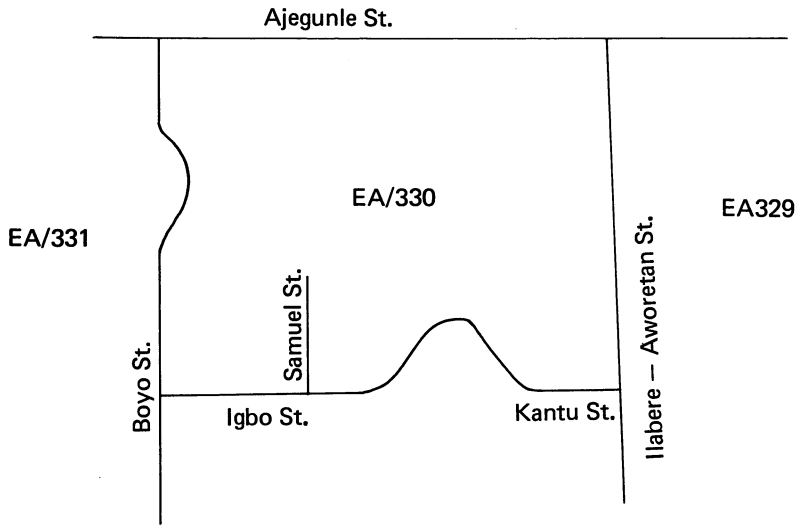


Fig. 2. A typical rural EA sketch (simplified)



Note: Contradiction: EA/330 and EA/329:
EA/330 shows Kantu running into Ilabere
EA/329 shows Kantu parallel to Ilabere
Both show Samuel as part of EA

Fig. 3. Urban EA index sketch

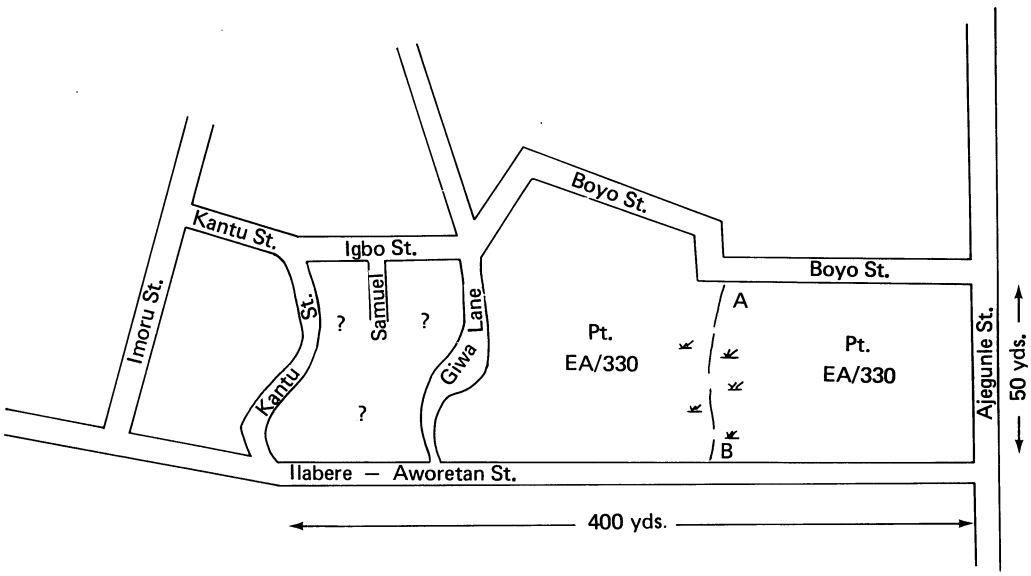


Fig. 4. Remapping of an urban EA

adjoining EA's. It also notes an apparent contradiction in the EA sketches for EA/330 and EA/329. We may observe, in passing, that the contradiction would not have been discovered had not the rules (see below) prescribed the examination of sketches not only for the selected EA but also for all adjoining EA's.

Examination in the field and the more careful mapping feasible for a relatively small sample survey, as opposed to a census, revealed the situation illustrated in Fig. 4. It was clear that the area marked by "???" in Fig. 3 had been included in the sketched areas for both EA/330 and EA/329—although it is probable (but not certain) that the households contained in this area (if any at the time) were enumerated only once.

As was the case for the rural example discussed earlier, it will be clear that the enumerators cannot resolve the problem by reference to the materials that they have at hand. A methodology is required.

These problems are only two of the many that can arise and whose resolution does not depend only on cartography. In fact, neatly drawn maps and sketches of apparently mutually exclusive enumeration areas, although pleasing to the eye, may tend to obscure the problem by giving an impression of precision that is seldom matched by what is found on the ground or by examination of the EA descriptions, lists of villages and hamlets, etc., that accompany these maps and sketches. Such examination will typically reveal one or more of the following problems:

1. Omissions: an existing urban block or rural hamlet cannot be found in any EA sketch or description. It was either missed by the previous cartography/listing operation, or has been constructed since that operation or in the case of a rural hamlet, has "migrated" from another location.

The current local name of the omission differs from that recorded in the earlier operation and the two cannot be reconciled. This may result from a real change in the local name or from the written representation of a spoken local name.

In many of these cases it will be found that the block or hamlet cannot be unambiguously related, either by its position vis-a-vis the EA "boundaries" as drawn or by its current local name, to any particular EA.

2. Contradictions: a block or hamlet appears to be shown on more than one EA sketch or listed in accompanying documentation as belonging to more than one EA. This may happen when adjoining EA's were mapped or listed by different teams, each team considering the block or hamlet as belonging to its EA. It may also result from giving the same name to different hamlets (in which case there may be a complementary omission).

3. Severely inadequate mapping or description: the documentation is so inadequate as to provide little basis for EA reconstruction in the field.

Useful discussions of these problems, in African rural contexts, will be found in Olsson (1984), UNECA (1984), and Kiregyera (1987).

In usual practice the enumeration team will not be carrying the documentation relating to all neighbouring EA's, so that its ability to detect omissions, contradictions, and ambiguities is severely limited. To the extent that it does detect them, there may be a tendency to resolve problem cases by deciding that the households in question must belong to some other EA, with resulting underenumeration. As noted in the Introduction, however, there may be the opposite tendency to always include ambiguous cases, with resulting overenumeration.

The key to the solution of the problem lies in the specification of an approach to field work that will ensure that all clusters that may belong to a selected EA will be found; in other words that all inherent ambiguities will in fact be detected. These ambiguities may then be resolved either by deterministic or by probabilistic rules and procedures.

3. Problem Detection

An application of the procedures outlined in this section, and of the deterministic approach (see below) to the resolution of ambiguities, will be found in Sunter (1981), a cartographer's manual prepared for the Nigerian Fertility Survey. The following procedures are adapted from the Nigerian Manual's section on the identification of rural EA's, the characteristic feature of which is considered to be that they consist of one or more distinct villages or hamlets.

1. Assemble the documentation (the EA files) relating not only to the selected EA but also to all neighbouring EA's. This documentation must be carefully examined before going to the field and all apparent contradictions noted. (If there is not already an index map, on which such ambiguities as apparent duplication of villages, hamlets, etc., may be noted, it will be convenient to construct one.) Some such contradictions may be resolved by appealing to other documentation (e.g., the census enumeration records), but most will require resolution in the field.
2. Begin work in the field in the usual way with a reconnaissance of the area, noting all routes (i.e., normally negotiable roads and paths and, in

some areas, navigable water routes) to and through it, and classifying dwelling clusters found in this step as

A—definitely belonging to the selected EA,

B—definitely belonging to another EA,

C—not apparently or unambiguously belonging to any EA; i.e., not shown on any EA sketch or included in any EA list of dwelling clusters,

D—apparently belonging to more than one EA sketch or included in more than one EA list of dwelling clusters,

All of these clusters are, of course, to be shown on a field sketch, the most important features of which are cluster names and approximate populations as well as inter-cluster routes and approximate distances.

3. From each cluster of class A, C, or D follow every route (and every branch of every route), noting and classifying every dwelling cluster encountered, until that route either
   i. terminates without reaching any other dwelling cluster; or
   ii. reaches an unambiguously identifiable physical EA boundary such as a main road, railway, or large stream; or
   iii. reaches a previously noted and classified dwelling cluster or joins a previously traversed route.

We will now have identified all relevant clusters as belonging to one of the four classes A to D, relative to the EA we are trying to identify. It remains to reassign the ambiguous cases of class C or D to either class A or B. Procedures for such reassignment are discussed in the next subsection.

Procedures will differ somewhat for urban EA's, characterized in the Nigerian Manual as follows:

i. They are comprised of one or more urban blocks; the boundaries are usually named streets; at least some of the buildings within the EA have commonly understood street addresses.
ii. Population density is high and at least some of the buildings in the EA may have two or more levels.
iii. Some of the buildings in the EA may not be dwellings.
iv. The adjacent EA's are also urban in character; in other words it requires more than one EA to cover the town or city of which this one is a part.

We do not deal here with the extreme case, though it is not uncommon in rapidly urbanizing societies, of urban reconstruction to the extent that the district of which the selected EA is a part is hardly recognizable as the same district shown by EA sketches. Nor do we deal with the case of extreme growth even in EA's whose boundaries are easily located. The first of these cases may call for remapping of a whole area and subsequent reselection of the EA sample within the remapped area; the second may call for mapping within the selected EA and statistical procedures appropriate to the modified selection probabilities of subselected "sub-EA's" or segments. The typical problem with which we do deal is that illustrated by Figures 3 and 4; recognizable but contradictory EA boundaries for the selected EA and its neighbours.

The detection of omissions, contradictions, and ambiguities in urban areas begins, as before, with the assembly of the EA sketches for the selected EA and its neighbours. It will be useful, if an index sketch (which may be based on a street map, if one covering the area is available) is prepared on the basis of this material. All problems

evident from the examination of these EA files should be noted before going to the field. Figure 3 illustrates the result of this process in one case.

The main task of the cartographic team is to prepare an accurate representation of the selected EA and as much of the adjoining EA's as will be necessary to illuminate and resolve the problem areas. In this process they will take advantage of the characteristics of urban EA's listed above, by showing street names, urban landmarks, street numbers for households at key points in the sketch map (e.g., at street intersections, adjacent to proposed segment boundaries), etc. (We have not demonstrated those features, however, in Fig. 4 although they were all exemplified in the actual sketch of which Fig. 4 is a simplified version.) In many cases, particularly in areas with regular street patterns, the process of boundary identification will be a straightforward matter of updating the original sketches, correcting obvious errors that may have occurred in showing landmarks in these sketches, and determining and mapping suitable boundaries (e.g., line A–B in Fig. 4) for any required segmentation. However, where omissions, contradictions, and ambiguities are discovered, either at the index sketch stage or in the field mapping stage, their nature must be shown clearly on the field sketch. This is illustrated by Fig. 4.

There are areas of course, particularly on the fringes on large towns and cities, which are somewhere between urban and rural in character and in which cartographic teams will have to adopt a synthesis of the procedures described here. This applies both to detection and to resolution of problem cases. At the extreme we have situations in which EA's that were rural or only semi-urban at the time of the original EA definition have now become completely urbanized and their layout has changed beyond recognition. Under these circumstances, remapping of a whole census district may be called for. A less extreme and fairly common situation is illustrated in Fig. 5.

The top sketch in Fig. 5 shows two adjoining EA's separated, at the time of census mapping, by an area of bush devoid of dwellings. The EA sketches themselves showed EA/36 and EA/37 as being comprised of the dwellings contained in the areas bounded by named roads and by sketched in "boundaries" representing the bush areas; descriptions and sketches that were quite adequate for the immediate census purposes. Some years later one of these EA's is elected in a sample survey and, on going to the field, the survey team finds the situation shown in the bottom sketch. The built up area has now expanded to fill in the area of bush that used to separate the EA's and the original boundaries, to the west of EA/37 and the east of EA/36 respectively, have now disappeared. No other EA's have a claim to the intervening area so that the extent of the ambiguity is quite clear.

4. Problem Resolution

4.1. Deterministic reassignment

The essential idea here is to determine a set of rules whose effect is to create a logical one to one correspondence between the old EA's (with, as we have seen, their omissions, contradictions, and ambiguities) and new ones (with none of these flaws). This correspondence must be replicable, in the sense that enumeration teams working independently in neighbouring EA's would reach the same conclusion as to the disposition of dwelling clusters to which more than one EA has a claim. This is necessary if we are to preserve

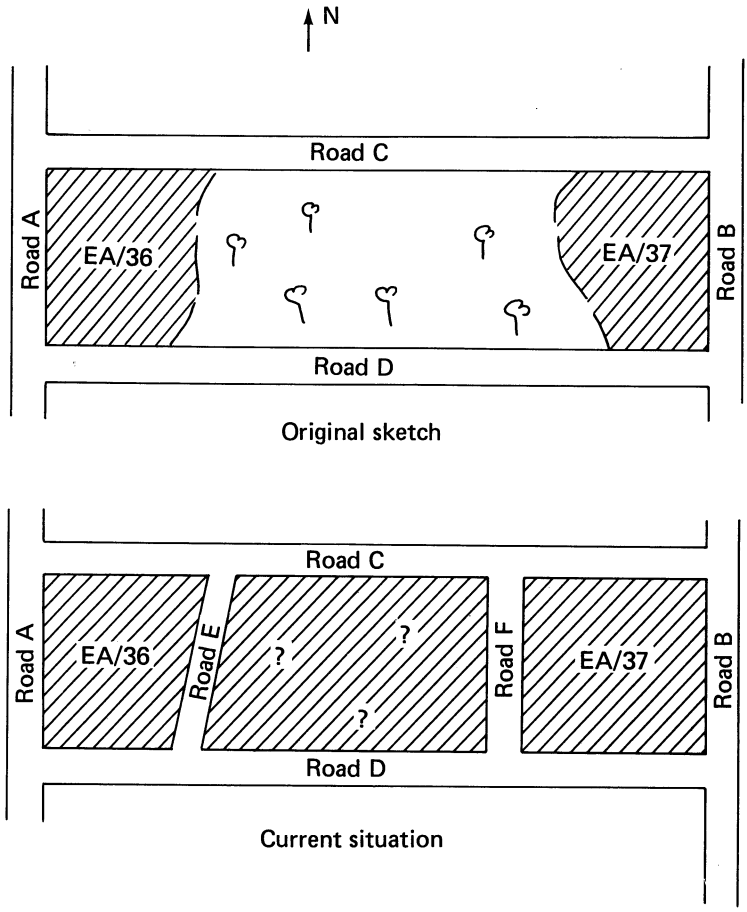


Fig. 5. Change in a semi-urban situation

the essential characteristic of a probability survey; namely, that every member of the population under study has a known non-zero probability of selection.

For rural EA's, one way of doing this is to assign each class C or D dwelling cluster to A or B depending on which cluster of one of these classes it is "closest" to. "Closeness" or "distance" may be defined in the usual physical sense but, since this presents difficulties in many cases, is more satisfactorily determined in terms of some hierarchical metric such as allegiance to the same village, language, kinship, market ties, church allegiance, ease and time of mutual access, etc. The rules given in the Nigerian Manual were:

"A village (or isolated dwelling) is associated with another if it satisfies one or more of the following criteria:

i. There is a road or path connecting the "new" village to the "old" one but to no other "old" one.
ii. The inhabitants acknowledge the same village head.
iii. The children go to school and/or the adults go to church there.
iv. The women sell and/or buy at the market there and at no other market.
v. It is much closer and easier to get to than any other village.

There may remain one or more villages whose association does not appear to you to be uniquely determined. In such cases they should be shown in your sketch map with the relevant symbol (???) and the relevant circumstances included in your report."

The intention, in the undetermined cases, was to "invent" any additional rules necessary for their resolution and, subsequently, to add these to the general rules. These rules may be relatively arbitrary, and may differ from one region to another or between urban and rural areas. The important thing is that they can be strictly applied, once determined, and their results replicable.

The notion of "closeness" breaks down in an urban setting. The problem area in Fig. 4, for example, is equally "close" to the two EA's with which it is associated. In Fig. 5 we can devise a "closeness rule" (which we would now add to our set of rules, to be applied in similar situations) that will assign the eastern and western blocks to EA/37 and EA/36 respectively (as shown in the figure), but this rule will not help to assign the area between Road E and Road F. For such cases we will have to adopt some relatively arbitrary rules for deterministic assignment, or we will have to turn to some form of probabilistic assignment. The possibilities for this are now examined.

4.2. *Probabilistic reassignment*

An alternative, or a supplement to be used when the deterministic rules fail, might be to assign ambiguous areas to one of the EA's that have a "claim" to them, under some appropriate probability structure. Suppose that dwelling cluster i is to be assigned to one of three EA's (which we index 1, 2, 3) that have a claim on it. If, say, EA/1 was the originally selected EA then cluster i is also to be selected with probability $\pi_{i|1}$, if EA/2 then with probability $\pi_{i|2}$, and if EA/3 then with probability $\pi_{i|3}$. It will obviously be convenient if, once assigned, it can be treated in subsequent subsampling and estimation procedures in the same way as other dwelling clusters in the EA; i.e., *as if* it has been selected with the probability of that EA. To see what this implies for the conditional assignment probabilities it is convenient to rewrite the usual estimator, based on EA totals, of the population total as

$$\hat{Y} = \sum^{N} t_\alpha Y_\alpha / \pi_\alpha, \tag{1}$$

where the summation now extends over the whole population, and the random variable

$$t_\alpha = \begin{cases} 1 \text{ if } \mathrm{EA}_\alpha \text{ is in the sample} \\ 0 \text{ otherwise} \end{cases}$$

Taking the expectation of (1), with respect to the probability distributions of the $\{t_\alpha\}$ we have

$$\begin{aligned} \mathrm{E}(\hat{Y}) &= \sum^{N} \pi_\alpha Y_\alpha / \pi_\alpha \\ &= \sum^{N} Y_\alpha \end{aligned}$$

as required. Thus, the expected contribution of each EA is its own population total. This is the condition we impose on the assignment of dwelling cluster i. Thus, denoting the population total for dwelling cluster i as Y_i, we require

$$\mathrm{E}(t_{i1} Y_i/\pi_1 + t_{i2} Y_i/\pi_2 + t_{i3} Y_i/\pi_3) = Y_i \tag{2}$$

where

$$t_{i\alpha} = \begin{cases} 1 & \text{if } \mathrm{EA}_\alpha \text{ is in the sample } \textit{and} \\ & \text{cluster } i \text{ is assigned to it} \\ 0 & \text{otherwise.} \end{cases}$$

Taking the expectation of (2) with respect to the $(t_{i\alpha})$, we have

$$\pi_1 \pi_{i|1} Y_i/\pi_1 + \pi_2 \pi_{i|2} Y_i/\pi_2 + \pi_3 \pi_{i|3} Y_i/\pi_3$$

$$= Y_i(\pi_{i|1} + \pi_{i|2} + \pi_{i|3})$$

$$= Y_i \text{ if and only if } (\pi_{i|1} + \pi_{i|2} + \pi_{i|3}) = 1.$$

Thus any consistently applied rule for assigning the conditional selection probabilities will do provided these probabilities sum to one. For example, one such rule might be to assign conditional selection probabilities proportional to the original EA selection probabilities. There are good reasons, however, to assign *all* of the conditional probability to the EA with the largest selection probability:

i. simplicity of application by the enumeration team;
ii. any variance estimator, valid in the absence of this complication, remains so;
iii. in self-weighting multi-stage sampling schemes, in which the within EA sampling ratio is inversely proportional to the EA selection probability, the additional sample take from the added dwelling clusters is minimized.

If this rule is adopted, the "probabilistic rule" is, of course, no longer probabilistic but has reverted to a "deterministic rule." It will be noted that any such deterministic rule will do provided that it can be consistently applied. The rule just given (i.e., assign to EA with largest selection probability) is one such rule. A more simply applied one, however, might be to assign to the "claimant EA" with the lowest EA number.

5. Conclusion

Problems of the type illustrated by the examples given in this article arise, at least in developing countries, in a large enough proportion of cases to jeopardize the credibility of the whole survey unless they are systematically resolved. The methods of resolution must respect the basic principle of probability sampling: that every person in the target population must have a calculable non-zero probability of selection. In order to ensure this, the procedures for problem detection must be exhaustive and the rules for problem resolution must be replicable. Replicability here means:

i. that any of the claimant EA's, if selected, would have lead to the identification of the problem;
ii. that any cartographic team would have identified and resolved the problem in the same way.

The notion of replicability extends, it should be noted, to cases in which remapping of one or more whole census districts, followed by reselection of one or more EA's within the remapped area, is the only viable solution. This means, at least in principle, that regardless of which EA(s) had been selected the same decision would have been reached.

Survey sampling and process designers will need to think carefully on the appropriate level of responsibility for decision making in problem resolution, in their own organizational context. It may be that they will need two levels of rules: the first level, for application by the mapping and enumeration teams themselves, would be simple "closeness" or "proximity" rules of the type given for our rural example; the second level, to be used when the first level rules fail, would be applied by the next level of supervision. All problem identification and resolution would be subject to further review at a still higher level of supervision.

6. References

Kiregyera, B. (1987): Types and Causes

of Nonsampling Errors in Household Surveys in Africa. Journal of Official Statistics, Vol. 3. No. 4. pp. 349–358.

Olsson, U. (1984): A Master Sample for Tanzania: Evaluation, Discussion and Proposals. Central Bureau of Statistics, Dar es Salaam and Statistics Sweden, Stockholm.

Sunter, A. (1981): Nigeria Fertility Survey – Cartographer's Manual WFS/TECH, 1581. International Statistical Institute.

U.S. Bureau of the Census (1978): Mapping for Censuses and Surveys. Statistical Training Document ISP-TR-3, Washington, D.C.

U.S. Bureau of the Census (1979): Popstan: A Case Study for the 1980 Censuses of Population and Housing. Part B: Planning and Preparation for Popstan Census. Statistical Training Document ISP-TR-4B, Washington, D.C.

United Nations Economic Commission for Africa (UNECA) (1984): Classifications, Definitions and Concepts of Locality in Africa. Statistical Information Bulletin for Africa, Addis Ababa, Ethiopia.

Received April 1987
Revised June 1989