COMPUTER-BASED GEOGRAPHIC CODING FOR THE 1970 CENSUS

William T. Fay and Robert L. Hagan, Bureau of the Census

Almost everyone is aware by this time of the extensive use we plan to make of the mails in the 1970 census. It is possible, however, that many of those who know about it have not considered the amount of planning and preparation necessary to assure its success.

In the time allotted me today I want to tell as much as possible of the part being played by the Census Bureau's Geography Division in this pioneer effort.

I want to make it clear, too, that much of what we are doing now is part of an exploratory phase in which we are mulling over our own ideas and seeking the opinions of others. We want the best possible program and we welcome suggestions from any source.

First of all, there are these reasons for making the change from direct enumeration to mailed questionnaires.

We need a reduction in the time span of the information gathering process to reduce the constant danger of overlooking some people and counting others twice. A speed up in this phase will also contribute to earlier release of data. Publication of most reports on the 1960 census gained 12 to 18 months over the 1950 census but you want even more speed. I need not tell you that the sooner we dispense the information, the more value it has for the Nation.

We want the increased accuracy which will come from making it possible for a family to sit down together to consider the questions, rather than putting the burden of providing the information for the whole family on one person. Quite often the hurried enumerator must question the one member of the family found at home and too often that one person lacks complete information.

We want to improve the quality of census data by diminishing the influence enumerators may have on the answers we receive. Scientific studies have demonstrated that enumerators influence answers to census questions in various ways. While these effects tend to cancel out when the work of a number of enumerators is combined, they can adversely effect the data for individual small areas, each of which is within the territory covered by a single enumerator.

We want to save expense and to reduce the task of recruiting and training the army of enumerators needed in the past. I'm safe in saying that they will never be entirely eliminated but their numbers can be reduced. We plan to use perhaps 100,000 in 1970 as compared with 170,000 in 1960.

A factor which will facilitate a mail census in the increasing drift of our population to urban areas where there are street addresses and where mail is delivered by city carriers. There were 125 million urban dwellers in this country in 1960 and there will be 140 million by 1970.

But if that seems to indicate a simple solution to the need for change in census methods, let's consider some of the complications.

First, there has always been a geographical vagueness about many post office addresses. This is due mostly to the fact that the Giantville post office handles mail for adjacent Dwarftown and even beyond Dwarftown into unincorporated areas.

The Dwarftown citizen usually doesn't mind having a Giantville address as long as he gets his mail promptly. In fact, he sometimes thinks of it as an asset. The folks back home never heard of Dwarftown but they do attach some importance to Giantville and part of that importance tends to rub off on a person who has it as his address.

Sometime the fuzziness about addresses even crosses State lines. For example, the post office at Suitland, Maryland, where the Census Bureau is located, is a branch of the Washington, D. C. post office and we have a Washington address.

In the 1954 Census of Business, the Bureau made an effort to clean up part of that confusion. On the census form a request was printed for a statement of the actual location of the business or plant, regardless of the address. There was nothing complicated about the question but it didn't draw nearly enough accurate answers to solve the problem.

We then decided to see if the post office could help. Our reasoning was that a man trudging the mail route day after day would know the point at which he stepped over the Giantville city limits and entered Dwarftown and that he could make a record of the exact address at which that happened.

He couldn't. Or if he could, he didn't tell us with enough frequency or accuracy to yield us substantial benefit.

Someone had the idea that telephone exchanges were tailored to city boundary lines and that we could identify the locations of business firms from their phone numbers. We found that sometimes this would work but more often it would not. Public utility records didn't help, either.

All of these things foretold some real headaches for anyone attempting a mail census without first clearing up the uncertainties about the relationships between addresses and boundary lines.

We knew it might be possible to blunder along on a mail census by using maps and enumerators in the confused perimeter areas of cities but what we really needed was a master coding file on computer tape which could be used to assign most business locations to States, counties, and cities.

To prepare such a file from scratch would have cost the Bureau far more than we were prepared to pay so we began looking for such a file already in existence.

Finally we investigated a file of punch cards which had been amassed by a directory firm for use in relating census data and sales statistics to dealer areas for purposes of market research and analysis.

That file covered delivery areas of post offices located in cities which had populations of 25,000 or more and it consisted of street names and house number ranges for areas approximating census tracts.

This showed possibilities. It was decided that if the border areas of those cities could be corrected to exact boundaries, the file could be used to assign geographic codes to specific addresses.

In addition, the records could be transferred to computer tapes and programs could be developed which would take an address input, recognize the address elements, and then put them in sequence to facilitate matching against an address reference file.

There were two basic questions to be answered. The first was: "Would a system covering delivery areas of post offices located in cities of 25,000 or more inhabitants provide the base for a solution?"

It was believed that the answer was "Yes," since a major difficulty was the accuracy of coding of establishments located in small cities which had mail delivered from adjacent larger cities. Success, of course, hinged on learning in which community the borderline addresses really belonged.

The other big question concerned the feasibility of developing a computer program which could code a high percentage of addresses accurately.

The file we were studying had been prepared by its owner under a controlled process in which address elements had been standardized to facilitate mechanical coding by use of the card reference file.

Could the Census Bureau use this reference file for coding a large proportion of addresses not prepared under the same controlled conditions?

The answer to this was not so apparent, and so it was decided to put it to a test. Los Angeles County was chosen because it had all of the problems with which we were concerned.

The results were encouraging. About 75 per cent of the addresses from our Census Bureau files were matched to the reference file and coded by computer and the test showed the accuracy level for those addresses to be about 99 per cent.

Following this, a national test was conducted, on a sample basis, with similar results, so the directory firm's entire file for cities of 25,000 or over was obtained as the nucleus of our system. The test procedures were then carried out for the entire United States and the addresses for business firms included in the 1963 Economic Census were coded in this fashion.

The results were satisfactory, that is we did improve accuracy, speed, and cost over past methods. Coverage, however, was not sufficiently extensive, so we are now modifying the basic file to permit coding of business addresses within all city delivery areas in the nation for the 1967 census.

The work described above was a small-scale forerunner of things to come. For the Economic Census the boundaries which had to be recognized in Census publications, and therefore coded by computer or other means, were those for a few thousand urban places, a hundred or so Central Business Districts, a few hundred Major Retail Centers and, of course, the 3,100 counties.

When earnest consideration was given to similar methods for the 1970 Census, the scale changed drastically; then we switched to talk of blocks and block faces with estimated numbers of areas jumping to one and a half and eight million respectively.

Here's where another big problem popped up. Coding of residential addresses to this fine-grained level requires accurate maps with a uniform scale within each urbanized area.

They just didn't exist.

In the 1960 Census, the Bureau needed maps, of course, although not as desperately as now since that was an enumerator census and the periphery problem was not nearly as great. If streets didn't show on maps, or nonexistent ones did, the enumerators were expected to resolve the problems.

The maps used in 1960 were of widely varying sizes and scales. They were barely passable, for the purposes for which we used them, and they resembled an unending series of jigsaw puzzles. In one city, with its suburbs, it was necessary to use 137 different maps for piecing together enumeration districts. There was a constant danger that some areas would be left out or overlapped. That could mean that some people would be overlooked or that some would be approached twice by enumerators of adjoining districts.

For the 1970 Census the problem couldn't be dodged. Something had to be done and the Geography Division decided to tackle it. Here's how it's being done:

We took U. S. Geological Survey 7½ minute quadrangle maps and changed the scale from 1":2,000' to 1":800'. We dropped the topographic detail and updated the street layouts as far as we could with the information we had available.

We knew, of course, that we didn't have enough

information and that errors were certain to creep in. This is where we have to lean heavily on local groups, with a much wider knowledge of their own cities than we have. The cooperation we're getting is excellent.

The Bureau can't pay for this help, but we are reciprocating by providing copies of the up-to-date maps being developed, and by creating the capability of providing a vast fund of information which can be blended with local statistics and geared to local governmental units. This is to be facilitated by reserving for local use a five-digit section of the FOSDIC form on which block faces are to be coded. Details of this plan will be given further on in this presentation.

The first step in our map making, the compilation to a scale of 1":800', is done in our Jeffersonville, Indiana, office and it is now more than 40 per cent complete for the urban cores of metropolitan areas. Our goal is to finish the mapping for all of these urban centers by the end of 1968.

Those maps, when finished, will cover 100,000 square miles--about four times the extent covered by the urbanized areas defined in 1960.

When we have developed a map as far as possible in Jeffersonville, we send copies (in 36" x 48" sheets each representing 35 square miles) to a local cooperating agency in the area covered in the map. Usually that agency is a planning group which has agreed to give us the help we vitally need. That group will verify, correct, and update our maps, or they may pass copies on to other local groups in possession of the detailed information necessary.

After the required changes have been noted by local groups, the map is returned to us and we alter our original tracing to conform with the local editing.

Copies are then returned to the local cooperating agency where they are available for local purposes, including the notation of additional changes in streets which should be added to the master maps before the final census deadline.

In the preparation of computerized Address Coding Guides, much the same procedure will be followed.

For this process we'll use FOSDIC worksheets capable of being read electronically. We'll prepare these sheets by printing street names, block face identifications, intersecting streets, and even and odd address range numbers. In doing this, we will use information from a commercial direct mail list and, where available, from directory publishers.

Those partially-completed forms will then be sent to the local cooperating agencies together with copies of our Metropolitan Maps marked with Census area designations. As in the mapping program, the cooperating agencies may themselves verify and complete the worksheets or may farm this process out to other qualified groups within the area.

At this stage, before the forms are returned to us, agencies which are cooperating should determine what use they will make of the five-digit code field provided on the form for local use. We refer to it as the "optional field" because it can be used in various ways, or not used at all if it is not wanted.

You will notice that I say "if it's not wanted." It may be that some communities will not want it; although, if they understand the uses to which it can be put, it seems certain that they will want it. We think the "optional field" is very important, and a little further on I'll speak more of its benefits.

Through the use of address coding guides we will be able, for the first time, to record information for geographic units ranging from one side of a city block to an entire city. The limit to the flexibility of the information available to you after the census will be disclosure rules, computer capacity, and the cost of tabulation. We don't plan to provide this capability for all city delivery areas, but if we can accomplish this for entire urbanized areas and for cities of 25,000 or more inhabitants, the scope of our present plans, far more detailed census data will be available than ever before. A further limitation is the extent of city delivery postal service; beyond these areas we don't plan to code to the block-face level; although, reporting by block is expected to be feasible.

With relatively little added effort a copy of the "Census" Address Coding Guide for an area can be modified locally for broader use by the addition of identification codes for areas such as police precincts, health areas, and so forth. With this accomplished local flexibility is virtually unlimited. Police information, for example, can be matched to the modified coding guide and the police data assigned not only to police precints, but simultaneously to health, school, and other areas, as well as to census tract and block. The same can be done with other local information.

It's hard to imagine better tools for orderly, forward-looking community planning than an accurate address coding guide, but the possibilities are even broader. I'm referring of course to the "optional field" and related methods of securing census data for areas of local interest, rather than being limited to the tabulation areas used by the Bureau.

Let's consider the possibilities in a broad sense and then turn to details.

This tool can be used to secure Census data aggregated to match areas of local interest and then to correlate locally developed statistics with those from the census. As an example, local figures showing a steady growth of juvenile delinquency within a certain police precinct could be related to census population and housing characteristics, coded for the precincts, and from the combined information a clearer picture of the

problem and its solution might emerge.

Another situation might bring the question: "How many children between the ages of 6 and 16 live in the 10th school district and what is the racial proportion there?" A special tabulation would put the answer at the school board's finger tips.

Would recreation officials like to know how many children from 5 to 12 live within a 20-block radius of a proposed playground and the ratio of boys to girls? They can get the answer from the Census Bureau computers.

There are four primary ways that we suggest for local consideration in filling in the optional field and a fifth method that can be used separately or in combination with one of the first three. For convenience let me name the methods and then follow with more detailed explanations. They are:

1. Direct Coding
2. Geographic Unit Coding
3. Local Serial Number Coding
4. Census Bureau Coding
5. No Coding

In the first, the Direct Coding method, the procedure is simple, but the possibilities are quite limited. If a block face lies within the 12th police precinct and data for these areas are wanted, the figure 12 is entered in the optional field. If a school district grouping is wanted and the block face is in the 9th school district, the figure 9 is marked in the field. If the block face is in sanitary district 14, the figure 14 may be put into the optional field.

In this method, as in others, when boundaries cut across blocks, rather than following street lines, the severed block face would be treated as two block faces for the purpose of coding.

The two sections would both be placed on the FOSDIC form as separate block faces, the only distinction in coding of the two being in the range of addresses within each section and the code in the optional field. Thus, house numbers 1 to 19 of the block face might be coded as belonging to the 11th police precinct while numbers 21 to 49 might be coded as being part of the 12th police precinct. (The Census Bureau will limit such block face splits to 10 per cent of the total block faces listed.)

The direct coding system has limitations, the three administrative areas cited above and the total of five digits needed to code them into the optional field (12-9-14) exhaust the limits of the field and prohibit coding of the block face into other civic classification on the FOSDIC forms.

In theory, five such area codings could be made for each block face, but in practice such a possibility is rare since it depends on the unlikely supposition that there would be no more than nine (1, 2, 3, 4, 5, 6, 7, 8, 9) each of such administrative areas as districts and precincts and, hence, each would require only one of the five digits in the optional field. A more realistic number of codings would be one or two-three at most, thus drastically limiting the amount of local statistical groupings possible.

Another disadvantage of this system is that once a block face is coded into the optional field as being a part of a local administrative area, it goes into the Census records in that form. If the block face is later shifted, as from one police precinct to another, Census Bureau reference tapes would have to be altered to conform to the change. Otherwise, the value of that particular optional field coding would be lost.

The second method, Geographic Unit Coding, requires preparation of a map on which are shown the boundaries of all the local administrative units for which data will be desired.

The map will look like a hodge-podge of lines, lines which serve to cut the map into the "Geographic Units" referred to in the title of this method. The map will show clusters of blocks, ordinarily, bounded by lines drawn on the map. Each such cluster, or geographic unit, can be described as being entirely within one police precinct, one school attendance area, one traffic zone, and so forth.

On the map, numbers are assigned to each geographic unit and these numbers, or codes, are then inserted in the optional field. One more local action is necessary; a master list prepared to show the combination of codes required to provide data for each area.

The method creates a large number of sub-areas, smaller than tracts but larger than blocks, which can be combined to produce the information wanted for various areas.

The third, or Local Serial Number method, would permit almost unlimited use of the optional field. The flexibility, however, is at least partially offset by its complexity and somewhat greater cost.

In operation, it would require that a serial number be inserted in the optional field for each block face coded on the FOSDIC worksheets. That number would serve as a code link between local records and those of the Census Bureau and would enable the Bureau to extract information from its files to match local areas for which data are desired. One way in which local officials might handle the bookkeeping in this system is shown below.

Giantville	Local Serial No. of Block Face	Police Precinct	Transportation Zone	Sanitary District	School District	Improvement District
Tract						
1	1	8	12	7	18	6
1	2	8	12	7	18	6
1	3	9	11	6	17	7
1	4	9	11	6	17	7
1	5	9	10	5	16	8

As can be seen, the block face in Tract 1 which has been identified as No. 1 in the optional field is a part of police precinct 8, transportation zone 12, sanitary district 7, school district 18, and improvement zone 6. The same type of record is shown for block faces 2, 3, 4, and 5 and can be applied to all block faces within a tract or other given area through use of maps and civic records.

The information collected on the tabular forms can be put into machine-readable form to show all of the administrative districts in which each block face lies.

When census information is required for any area for which block faces have been coded, local officials can prepare and send the Bureau a computer tape, or punched cards which include, for each block face, the census tract numbers, the local serial numbers, and the identification of the area for which a tabulation is needed, for example, the school district number. We can then match the tract and serial numbers to our own records and "instruct" our computers to prepare a tabulation for the school districts. The serial number, then, is a common identification, in your records and ours, to permit ready identification of the units we must aggregate to prepare tabulations that you need for local use.

Unlike the direct coding system, block faces can be shifted from one local area to another with no complications other than an adjustment of local records.

Numbers may be assigned in two ways in the Local Serial Number method. In either case they are chosen by the local group for entry in the optional field of the worksheets. The numbers may bear some systematic relationship to a set of locally defined areas or, more frequently, they will be arbitrarily assigned, that is 1, 2, 3, and so forth, as the name suggests. Alternatively numbers will be assigned, upon request, by the Bureau to each block face record as described below.

Marking serial numbers on FOSDIC worksheets can be a tedious and time consuming task with many chances for errors. A computer, however, can do this work rapidly and accurately once appropriate instructions have been written. In recognition of the savings that will result, we are prepared to assign unique identification numbers to each block face, within census tract, if asked to do so.

As FOSDIC worksheets are transferred to computer tape the computer will be programmed to identify all block faces for each block and to assign a two-digit "block face number" to each. These numbers may be changed from time to time. However, we can reproduce these numbers in the optional field on the computer tape and that identification will not be altered even though our "label" changes. For example, block face 12 of block 307 might become block face 03 of block 309 in our part of the record. Nonetheless, if we had entered 30712 in the optional field that identification would remain fixed despite other manipulations, just as if it were a number coded in the optional field initially, and the local participants would have a positive identification of that segment of the address coding guide, once they had received a copy of the product.

Note, further, that this would provide a "structural" code in that the first three digits would be a block number, the last two a block face number. This serves to simplify the preparation of a cross-reference table or dictionary such as the one illustrated above. Specifically, for each block that is not cut by a "local" boundary, and this should include the vast majority of blocks, only one record need be prepared for the block, rather than one for each block face. That record would have a "00," "99," "xx," or other distinctive symbol, yet to be specified, as the last two digits of the equivalent of "Local Serial Number" with a block number as the first three digits. Later we would instruct our computers to recognize that this symbol means "assemble all block faces (within tract) having the specified first three digits."

The fifth method, "No Coding," is included to indicate that "all is not lost" if the optional field is not marked, or if it turns out that a need was overlooked in using this field. In fact most of the benefits of this field can be realized even though it is left blank.

As noted above, we will assign a number to each block face within a block, a number that cannot be considered permanent for the 1970 Census until perhaps four months after the census enumeration. At that time, however, anyone who wishes can secure a copy of the appropriate maps and the address coding guide and can then record the Bureau's identifications of blocks and block faces that correspond to any area of interest. He can then list our identifications to match his areas and we will be able to prepare the desired tabulation.

The disadvantages of the "No Coding" method are two. First, the splitting of a single block face, cut by a local administrative boundary, into two block faces is not feasible. Second, the preparation of local records to relate local areas to Census area identifications cannot be carried to completion until the Bureau's identification of areas is stabilized for the 1970 Census. If, for example, we identify a block face as number 2 within a specific block we must be free to change it to, say, 7 if we wish. If block face 2 is changed to 7, any coding in the optional field for face 2 will be carried over to the new 7 and that link between your records and ours will be just as valid as before. However, any record you prepared based on the fact that we had assigned "2" to a specific block face would be of doubtful value if you could not be sure, and you wouldn't be, that we were not changing identifications to satisfy our internal requirements.

We believe this method will be especially useful as a supplementary aid to those who use the direct coding method and will influence many to use it.

Direct coding is inexpensive, locally, and permits retrieval of census information at minimum cost. The procedures noted above make it feasible, despite the limitations of direct coding, to retrieve other information, when it proves to be desirable to do so. Direct coding, we believe, is the best method for smaller metropolitan areas, especially those that do not now have an ongoing program of computerized data processing.

Geographic Unit Coding is an excellent method, with the exception of the problems that may arise when local area units have boundary changes. It opens up to users the possibility of obtaining census summary data tapes that they may use to regroup our data in various ways of local interest. The problem of preserving confidentiality and the problem of excessive sampling variability rule against the release of detailed census data for individual blocks and block faces and even for block clusters that are small in population. For clusters as large as census tracts, we have in the past provided quite voluminious summary data. We are working hard on the problem of how to present data for various types of small areas without disclosing the characterisitics of any individual or tabulating meaningless numbers. For clusters that contain 1,000 or more persons, I now believe we will be able to provide quite useful summary data tapes.

Local Serial Number coding involves the preparation of an extensive cross-reference table, but is an excellent method of great flexibility and presents no serious problems when local boundaries change, except where a new boundary cuts a block face. That problem, I suggest, is not a serious one.

Census Bureau Coding, in our view, provides all or nearly all of the benefits of Local Serial Number Coding and markedly reduces local effort and costs in "keying in" to census records. If you lean toward the serial number method, we suggest that this alternative is almost certainly preferable.

"No Coding" is an "escape valve" if local groups cannot agree on the use of the optional field, or if an unanticipated need arises.

Further variations are possible through combination of methods shown above. Direct Coding and Small Area Coding may be combined with each other or with Serial Number coding within the five-digit limit.

As an example, the local group may wish to use the first one or two columns in the optional field for small area identification and the last three or four columns for insertion of an arbitrarily chosen code number. Such a combination would result in ready tabulations for the directly-coded areas while at the same time maintaining in census data files the greatest capability for data tabulation for other areas.

I have no doubt that much of this sounds complicated to almost everyone except computer people. With that in mind I assure you that when interested groups with specific problems need further explanation or advice, we will do all we can to help.

Another point of interest; we plan to identify the locations of blocks or block faces by grid coordinates. While the word is "plan," not "promise," I believe we'll carry out this proposal.

Within the areas covered by address coding guides, we expect to have coordinates for block faces; for other parts of urbanized areas, coordinates for blocks. Our coordinates will be recorded in latitude and longitude to four decimal places, that is to 36 feet at most, but those who wish State plane coordinates, rather than latitudes and longitudes, will be able to secure them. The coordinates will be available, at cost of copying, to those who wish them and can be used, within the Bureau, for special tabulations you may desire.

That's all on that topic. More would bore many of you and those who wouldn't be bored can probably imagine potential uses far better than I can describe them.

Somehow we always get around to talking about money; in this case expense involved in local cooperation with the Bureau of the Census. While we believe that local financing of this effort would be successful in many areas, we are not so optimistic as to believe that a substantial proportion of metropolitan area officials will be both willing and able to provide the required funds on short notice. However, the Department of Housing and Urban Development shares our enthusiasm for the new methods and the potential benefits, and will, in about one month, formally announce a program through which eligible agencies can secure "701" grants to cover two-thirds of the costs involved in editing our maps and completing our FOSDIC worksheets.

Further, we are encouraged by the reaction of many top level planners that no metropolitan planning group that really wants a good information base for its work can fail to take advantage of this program.

This paper is a continuation of our efforts to tell local groups of our plans, efforts that will be continuing in the months ahead. We plan to bring our program to the attention of metropolitan area groups and officials of cities with 25,000 or more inhabitants. In this way we hope to secure the local aid we need to change "potential" to "reality" in speaking of the improvements in the 1970 Census.

1970 is still almost four years away, but it is now time to begin planning ways in which this new and different census can best be made to serve the nation at all levels. We'll be pleased to hear your suggestions and comments and, especially, your offers of assistance in our endeavors.