COMPUTER METHODS FOR GEOGRAPHICAL CODING AND RETRIEVAL OF DATA IN THE DOMINION BUREAU OF STATISTICS, CANADA

I.P. Fellegi and J.I. Weldon
Dominion Bureau of Statistics

A large up-surge in the collection and use of statistics has been experienced in recent years. It is reasonable to expect that the explosion in statistical activities will continue. We shall have to make sure however that future growth will be controlled, well coordinated and that it will be achieved by efficient utilization of the financial and manpower resources.

In view of these considerations and of the recent technological and scientific developments it is important that developmental work should get underway towards the creation of some general tools applicable to several surveys or data files. Such general tools, such automatic survey systems may represent important means to achieve economies to extend our processing and retrieval capabilities, to enable us to deal with massive volumes of data and to build into our data processing systems important elements of standardization. As such, these general survey systems may be the basic technological prerequisities of large-scale national statistical information systems [13].

The present paper will describe briefly the developmental work underway in the Dominion Bureau of Statistics towards the creation of an automatic geographic coding and retrieval system in larger urban areas. Although the system is expected to be of general utility, we shall discuss it in the context of the 1971 Population Census which, we expect, will be the first large-scale application of it.

The major system features which are to be examined in more detail are as follows:

- data retrieval by user specified areas in larger urban municipalities;
- automatic assignment of geographic location identifiers to urban addresses;
- acceptance and recognition of addresses in free format, automatic correction of the spelling and key punching type errors;
- effective data retrieval and tabulation techniques;
- geocoding and geographic retrieval outside of the large cities;
- reliability of data and disclosure.

Conceptually the function of this system is to retrieve and tabulate geographically coded census data for any arbitrarily defined urban area. Geographic coding is achieved by automatic conversion of addresses to unique geographic coordinates. The entire system operation in a nutshell can be described as follows:

- an address conversion file which can convert urban addresses to unique geographic coordinates is to be produced;
- addresses of the enumerated urban households are to be put into a predetermined standard format, verified and corrected;
- the addresses are to be substituted by their respective geographic coordinates (geocoding);
- the geocoded census data is to be stored for future retrieval;
- tabulations by user specified areas are produced on demand, subject to considerations of statistical reliability and confidentiality.

Data Retrieval by User Specified Areas in Larger Urban Municipalities

In the present context a larger urban municipality is tentatively defined as a city or metropolitan area with a population of 50,000 or over. The important consideration is that a municipality must be a certain size, or part of a large urbanized area to be in a position to take advantage of small area information. The user will be able to delineate on a map the area for which he needs statistical tabulations. Such user-specified areas should preferably not cut through block faces and must be sufficiently large to permit the provision of statistical tabulations without violating the principles of confidentiality. Another problem in connection with statistical small area tabulation which will have to be kept in mind relates to sampling and non-sampling errors.

The user specified retrieval areas are conceived as polygons and are described by the coordinate values of the polygon vertices. Data retrieval for the user specified polygon is done by computer. The programme first selects all the block faces (sides of city blocks between neighbouring street intersections) represented by their midpoint coordinates which are within the user specified area, then retrieves and tabulates the census data for the selected block faces. Characteristically the system approximates the arbitrarily specified areas by using block faces as building blocks. The technique enables us to retrieve by streets or street segments as well [1, 3].

Automatic Assignment of Geographic Location Identifiers to Urban Addresses

This operation is commonly referred to as geocoding. The assignment of geographic coordinates to urban addresses enables us to retrieve by the arbitrarily specified areas. The geographic coordinate of an urban address is that of the block face within which the address is located.

Geocoding is performed with the aid of the address conversion file. This file contains street names, address ranges by block faces and the corresponding block face center point coordinates. The geocoding operation is carried out by

computer, which tests address ranges block face by block face until it finds the one which encompasses the submitted household address. After having determined the block face identity the corresponding block face centroid coordinates are added to the household address and merged with the census data. It appears that this method may result in a very efficient computer operation since a full tape reel of census data can, we think, be geocoded in about 10 minutes.

The work required for establishing an address conversion file represents a major effort at the present time. It is estimated that the preparation of the conversion file for a city of 1,000,000 people currently would take three clerical man years work. It is very likely, however, that this time will be halved by improved system design, methodology and by gaining on the job experience. The creation of the conversion file requires the selection of an accurate map of the municipality; the updating of it; preparation and key punching of a street index; the digitization of strategic points along all streets representing beginnings, ends, intersections and changes in direction; and the preparation, coding and key punching of address ranges by block face. The input data are edited, verified and processed by computer; block face center points are calculated and the address conversion file is produced. A by-product of the operation is a plotted street map for the municipality. Having produced the address conversion file its periodic updating will require a few days clerical work at a time. The real problem in updating the address conversion file is to obtain street data update information. We hope to get this directly from the respective municipalities, since they might be important beneficiaries of the system. We hope to make the address conversion capability available to interested municipalities to permit them to geocode their locally collected data. The address conversion file can also be used for geocoding any survey data containing addresses. We are currently completing the preparation of an address conversion file for the City of London, Ontario. This work was performed in conjunction with our 1967 Census Test for London, Ontario.

Acceptance and Recognition of Addresses in Free Format, Automatic Correction of the Spelling and Key Punching Type Errors

In free-form address neither the fields of the address components (such as house number, street name, street type, city name, etc.) nor their position sequence have to be specified. The identification of address components is performed by separating the words of the address into numeric and alpha fields and by relating the positions of the words of the address to recognizable key words (e.g. "Street", "Ave.", "Apt.", "County", "Rural Route", etc.). The resulting pattern of numeric fields, alpha fields and key words is unique enough to identify the address components in a large proportion of cases (present limited experience indicates that this proportion is well over 90%). Upon recognition of the address components it is necessary to verify at least the street and city names by comparing them with a file of "correct" names.

The census operation will have to deal with some three million urban addresses. These addresses may be obtained from existing lists in machine readable form or they may be key punched from field listings prepared by the Bureau, or both. These addresses will contain errors and they may be in different formats as well. The rewriting of these addresses on coding sheets in fixed format would require hundreds of clerks for many months. This operation, besides being error-prone, is impractical because of manpower, space, equipment and other limitations. To overcome these difficulties, we have developed a computer operation to accept and to recognize addresses in free-form. It is estimated that there are some three million addresses in Canada, which may require 100 magnetic tape reels for recording. It is reasonable to expect that 10% of the addresses will contain spelling or key punching type errors, especially if the addresses are produced without key punch verification. These errors would amount to some 300,000 address rejects requiring further manual intervention in the form of correction, key punching and reintroduction.

This major clerical operation might be substantially reduced by developing a computer programme for automatic error correction of the key punching and spelling type errors. A good proportion of these types of errors is due to a few different, missing or surplus characters. The error correction logic of the system is based on checking street or city names of similar (but not necessarily identical) lengths and on finding the name which produces the smallest number of discrepancies. The maximum allowable number of discrepancies is some variable function of the name length. The recognition and error correction of three million addresses would probably require about three to five days of continuous processing on a large scale computer. Judging from a performance of a similar system, it is to be expected that 1 1/2 or 2 1/2% of the addresses would still be rejected, amounting to 45 to 75,000 rejects. These address rejects would then have to be processed manually. We have an operational computer program now to decode addresses in free-form with certain restrictions. The complete system described above should be operational by summer of 1968. The concepts described in the following sections are still in the planning stage at the present time.

Effective Data Retrieval and Tabulation Techniques

The proposed census file would contain enumeration data with urban addresses organized in block face sequence. We expect to produce from this file the traditional tabulations by census tracts and enumeration areas, as well as tabulations by any combination of characteristics by ad hoc user specified areas. The difficulty in

providing tabulations by user specified areas is, of course, that the requirements cannot be known in advance, yet the Statistical Bureau has to satisfy these demands without much delay and at a reasonable cost. These restrictions will quite possibly necessitate that census data be organized in random access storage. We also hope to be able to satisfy at least the simpler types of special tabulations by using a generalized, efficient retrieval and tabulation program.

Random Access Storage of the Census Data

The random access file organization appears to hold out several promises for storing and retrieving census data on users' requests which we intend to carefully investigate. If we shall use randomly accessible storage devices we may be able to compress the data to the extent where little storage will be wasted; and we may be able to increase the efficiency of retrieval to the extent that only data required for retrieval would be accessed. This type of file might consist of two modules, which are the data file and the index file. The organization of the records in the data file would be by cities or metropolitan areas and by block faces within them. A record in the census file today typically contains all the characteristics relating to one person. The records of the proposed random access file would be organized by characteristics in a string form, each string containing one of the characteristics for all the enumerated persons. This means that if in a metropolitan area there are one million persons enumerated and there are, say, 50 characteristics reported per person, the proposed file will consist of 50 strings, each one of them one million characters, digits or bits long depending on the data content.

The other file module mentioned was the index file. The index file might be organized in a hierarachical fashion in list mode. The first level of this hierarchy might contain the list of province names and address pointers which are directing to the list of city names within the respective provinces. The second level of the hierarchy might contain the list of city names by provinces and address pointers which are directing to the list of block faces within the respective cities. The third level of the hierarchy contains the list of block faces for each of the cities in the form of block face centroid coordinates and address pointers which are directing to the first sequential appearances of block faces in the various census data characteristic strings.

Retrieval by arbitrary areas can be achieved by listing the coordinate points of the retrieval polygon vertices, accessing the block face centroid list for the requested municipality by descending through the hierarchy of the index file, determining the block face centroids which are contained within the arbitrarily specified retrieval polygon, retrieving the desired characteristic data string portions for the selected block face groups from the census data file, and performing the requested tabulation. The entire operation is an integrated computer process.

Generalized Retrieval Programme

An important aspect in providing fast turn around time at a low cost to users is the availability of a generalized retrieval programme. Input to such a retrieval programme requires the designation of the province, municipality, the listing of the desired characteristics and retrieval conditions for tabulation or cross-tabulation, and the coordinate points of the vertices for the requested retrieval polygon. The significance of such a generalized programme would be that at least simpler types of special tabulations could be specified through the use of the programme without extensive training in programming. The data file organization by characteristic strings and the index file organization in hierarchical structure would greatly facilitate the utilization of such a generalized programme. The most significant advantage of such high level retrieval languages is that they permit the description of the retrieval and tabulation requests in some restricted English language form, which then can be used as an input to a computer programme to perform the designated operations.

The system must also be designed to facilitate an inverse retrieval function. This refers to the type of request which seeks the delineation of an area (or areas) which satisfy some stated conditions. After having determined the desired area its boundary could be mapped by means of computer graphics.

The Problem of Geocoding and Geographic Retrieval Outside the Larger Cities

Automatic geocoding assigns to postal addresses, with a minimum of manual intervention, the location-specific coordinates of the center point of the block face in which the address is located. In this fashion the traditional coding is carried out in that the address is identified as belonging to a particular pre-designated standard area, the block face in the present case. Automatic geocoding differs, however, from the traditional geographic coding in three important ways. First, it carries the coding to much smaller areas (block faces) than would be conceivable using manual methods, hence it provides very small building blocks for future aggregations. Second, it provides a reasonably error-free general tool that can be applied to data files, whatever their origin, as long as the data fields are identified

by postal addresses*. Third, the codes which identify the "building blocks" of the coding system directly identify their location as well, hence future aggregations of contiguous "building blocks" into larger areas are greatly facilitated, no matter how the larger areas are specified.

The system, as outlined above, is designed to handle addresses in the larger cities or metropolitan areas. In smaller urban areas it would not be economically feasible. In rural areas it is not even conceptually feasible since rural addresses often are not specific enough to determine their location (e.g. the address John Smith, 29 Bank Street, Ottawa is location specific even without the name of the occupant; John Smith, R.R. 2, Cornwall, Ontario is not location specific). It is unlikely, therefore, that we shall have in these areas an automatic and general geocoding system capable of coding addresses to sufficiently small areas**. Present plans for the 1971 Census in these areas [6] indicate that enumerators will be canvassing the addresses and will, at the same time, code them in the traditional fashion to Enumeration Areas (similar to the Enumeration Districts in the U.S.). In a separate manual operation the coordinates of the center points of these Enumeration Areas will be determined. Each census record in an Enumeration Area (about 150 households) will carry the coordinates of this center point. The number of affected Enumeration Areas is expected to be less than 20,000. Of the three benefits of automatic geocoding two will be lost using this rather primitive method: it will not carry the coding process to building blocks smaller than the traditional ones and it will not provide a general tool applicable to data files other than the census. It will, however, retain the third important advantage, namely it will facilitate the aggregation of contiguous Enumeration Areas into larger areas, no matter how the larger areas are specified provided they do not cut across Enumeration Areas.

The Problems Relating to Reliability of Data and Disclosure

References have been made to many remaining unsolved problems in system design and programming, as well as to others on which developmental work is well under way towards a satisfactory solution. In conclusion two different and very important problems should be at least briefly mentioned.

The first problem relates to the reliability of data. It is well known to this audience that census data, whether they are based on a full count or on sampling, are subject to potentially large measurement errors and in the latter case also to sampling errors. It is important to understand that although in the larger urban areas we will code to the block face level this will not be a level at which data can be made available (except possibly some very simple counts). The purpose of coding to the block face level is to achieve a degree of flexibility in aggregating to larger areas which was not open to us before. Just the same, the temptation will be substantial to ask for data for areas smaller then, say, census tracts. Also undoubtedly there will be a greater number of people using census data than before. This would make it very important to be able to associate with census tabulations, particularly with those referring to smaller areas, a measure of the reliability of the data, including the contributions of measurement errors as well as sampling errors where applicable. This will pose several serious problems. It is well known that the measurement of such errors is notoriously difficult [2, 4-12, 14-16]. It would be even more difficult to measure them early enough to be available when the main census publications are prepared. And even if they were available for some of the counts in time, there would be some difficulty in imputing them for the remaining counts (since direct calculation for all items would be inconceivable) and presenting them in a readable and useful form. A series of special problems arise related to sample estimation but these will not be discussed here.

The second problem area relates to disclosure of individual information. The flexibility of retrieval of information for areas which are not the standard, pre-coded ones greatly increases the danger of the so-called residual disclosure. The monitoring of all the previously produced small area tabulations appear to be the best solution to protect against residual disclosure, at this time. Whether or not this monitoring function can be performed automatically by computer, we are going to ensure the confidentiality of census data.

* Such a general tool has an important unifying influence in that it facilitates the production of comparable and compatible geographic tabulations from different data files.

** A solution might be found for the problem if an appropriate ZIP-code type system was adopted by the Canadian Post Office.

REFERENCES

[1] Calkins, H. W., Research report No. 2, Operations Manual for street address conversion system.

[2] Cochran, W. G., Sampling Techniques, Second Edition. New York: John Wiley and Sons, 1963. Pp. 141-2.

[3] Dial, R. B., Research report No. 1, street address conversion system.

[4] Fellegi, I. P., "An analysis of response variance", Bulletin of the International Statistical Institute, 34th Session, 40 (1963), 758-9.

[5] Fellegi, I. P., "Response and its estimation", Journal of the American Statistical Association, 59 (1964), 1016-1041.

[6] Fellegi, I. P., and Krotki, K. J., "The Testing Programme for the 1971 Census in Canada", Proceedings of the Social Statistics Section, American Statistical Association (1967).

[7] Hansen, M. H., Hurwitz, W. N., and Bershad, M. A., "Measurement errors in censuses and surveys", Bulletin of the International Statistical Institute, 32nd Session 38 (1959), 359-74.

[8] Hansen, M. H., Hurwitz, W. N., and Pritzker, L., "The estimation and interpretation of gross differences and the simple response variance", Unpublished report, Bureau of the Census, U.S.A. (1963),(honouring Professor P. C. Mahalanobis).

[9] Hanson, R. H., and Marks, E. S., "Influence of the interviewer on the accuracy of survey results", Journal of the American Statistical Association, 53 (1958), 635-55.

[10] Kish, L., and Lansing, J. B., "Response errors in estimating the value of homes", Journal of the American Statistical Association, 49 (1954), 520-38.

[11] Kish, L., "Studies of interviewer variance for attitudinal variables", Journal of the American Statistical Association, 57 (1962), 92-115.

[12] Mahalanobis, P. C., "Recent experiments in statistical sampling in the Indian Statistical Institute", Journal of the Royal Statistical Society, 109 (1946), 325-70.

[13] Nordbotten, S., "Automatic files in statistical systems".

[14] Pritzker, L., and Hanson, R. H., "Measurement errors in the 1960 Census of Population", Proceedings of the Social Statistics Section, American Statistical Association (1962), 80-90.

[15] Sukhatme, P. V., Sampling Theory of Surveys with Applications, Ames, Iowa: Iowa State University Press and New Delhi, India: Indian Society of Agricultural Statistics, 1954. Pp. 445-6.

[16] Sukhatme, P. V., and Seth, G. R., "Non-sampling errors in surveys", Journal of the Indian Society of Agricultural Statistics, 4 (1952), 5-41.