NETWORK DATA, GEOGRAPHIC CODING CAPABILITIES, AND AREA SYSTEMS

George Farnsworth
Southern California Regional Information Study, Los Angeles

I feel strongly that it is past time to discard the notion that the computer is a kind of data factory that merely generates information for the planner to use. On the contrary, the computer must inevitably become a part of the planner's own brain, and he must in turn become part of it. The only possible justification for the tremendous effort now going into the production of census summary tapes and geographic base files is that planners and administrators will use these materials by making them an integral part of their own professional thought processes. This will only happen for those who stop thinking of computers as mysterious and distant oracles tended by a high priesthood of programmers. I stress this point because I am convinced that only those who intimately understand the capabilities and limitations of the machines will be able to make succeesful use of these materials, which are primarily machine oriented.

NETWORK DATA

My house in Los Angeles is high on a hill, and at night, when the smog clears, a panorama of lights spread out below. That is one aspect of network data—a pattern of interconnecting lines. That pattern, however, is the network itself, and the bare net of lines is essentially data free.

The information added to this network in constructing an address coding guide and dual independent map encoding (ACG/DIME) file may be called data for want of a better term, but it is not statistical information. When we speak of census data or transportation data, we are talking of statistical characteristics of groups of people or other units. By network data, on the other hand, I mean specific characteristics of individual elements of the network.

The individual elements of the ACG/DIME network are line segments. The data attached to these segments may be classified into 3 broad interrelated classes: characteristics of the line itself, characteristics of the end points of that line, and information about the areas on each side of the line segment.

Each line segment has a name or other verbal description and a class, such as street, railroad, city limit, or freeway. Other data considered specific to the line itself are its length, ZIP code, and unique sequence number. Additional specific data could be added for transportation planning purposes, such as number of lanes, road surfaces, speed limits, and related information.

Each end of each line segment has an intersection number and a geographic coordinate location and is identified as being on a specific census metropolitan map sheet. All line segments that meet at a particular intersection naturally have one intersection number in common. These numbers are what stitch the line segments together into a network. The 2 ends are further differentiated by order. One is called the from end and the other, the to end. This distinction serves to orient the segment. For streets with addresses, the segments are "aimed" in the direction of increasing house numbers. For other types of segments, the orientation is arbitrary but nonetheless necessary. The end points of segments are a suitable point to attach data such as those on signalization, signs, or separation.

The regions lying to either side of a segment are first denoted left and right by utilizing the orientation established by the order of the end points. Address ranges are recorded for each side (for streets). In general the address range is of even numbers on one side and of odd numbers on the other side, but this is not required; it sometimes happens that both sides are even or odd or that only one side has house numbers. The 2 sides are further identified by geographic codes that range from state to census tract and block. In addition to the standard bureau codes for state, county, city, congressional district, and so on, local codes are provided for traffic zones, school districts, police precincts, and the like.

The ACG elements of the network data are the segment names, address ranges, ZIP codes, and geographic codes. DIME is the process by which the segment is oriented, the intersection numbers added, and the 2 sides brought together. Coordinates are digitized later for each intersection and added to the file. (Coordinates will be recorded in state plane, latitude-longitude, and map miles.)

Without the features provided by DIME, the ACG elements are simply a free-floating set of block sides not connected into a network and not paired off into opposite sides of the same street. The dual part of DIME results from an interesting property of this type of network. Up to this point, I have been describing the DIME network as a set of line segments connected together at intersections. It is possible, however, to consider the same network as composed of a set of block boundaries connected together at block centers. From this point of view, the line "segment" intersections become the "sides" of the boundaries and the block numbers become the "ends."

Figure 1 shows that the solid lines are the streets forming the basic network and the dashed lines are the "dual" of that network. A segment that is usually regarded as connecting point A to point B with block 101 on one side and block 102 on the other can also be viewed as "connecting" blocks 101 and 102 with A and B as the "sides." This principle of duality is what distinguishes the DIME network from other types of node-link networks.

One importance of the dual network is that it provides the capability of identifying which blocks (or other regions) are neighbors. This allows comparison of the data for one block with those of its adjoining neighbors and would thus permit the construction of a large contiguous set of blocks with homogeneous characteristics. Alternatively one might identify those blocks, tracts, or other areas that are different in some respects from their neighbors.

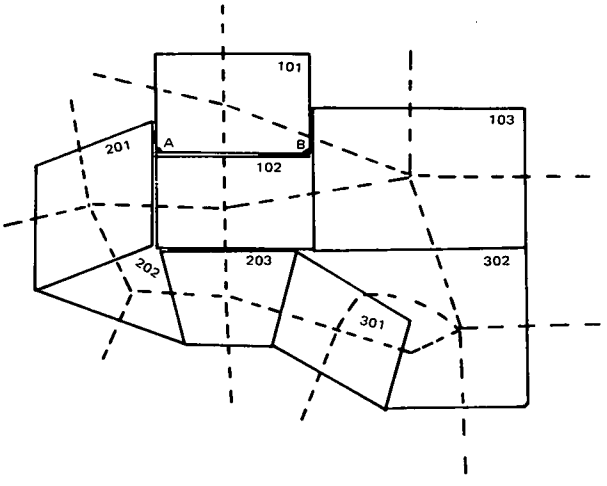

Figure 1.

In addition, of course, duality provides an important and powerful editing feature by ensuring that the clerically created primary network is consistent with its "independently" encoded dual.

To get back temporarily to my starting point, it is now possible for one with real appreciation of the capabilities of the computer and with solid knowledge of his objectives, to manipulate the ACG/DIME network in the same way that anyone can manipulate a map. The ACG/DIME file is in principle a computer-usable map but one that contains much more information than is usually found on maps.

The steam engine did much more than replace muscle power, however; and the computer makes it possible to go far beyond the manipulation ordinarily possible for clerks using maps and statistical tables. This is a perfect example of the "medium" becoming the "message."

It is one thing to imagine asking a clerk to identify a set of contiguous blocks that are homogeneous with respect to some specific demographic characteristic. With a machine, however, it is possible to go far beyond this and ask for a definition of homogeneity that depends on a set of characteristics, giving more weight to some than others and including numerous other criteria.

A draftsman can easily draw a 2-mile radius around a planned facility, but with the geographic base file you and the computer can draw a line indicating the 5-minute driving distance radius and include adjustments for varying speeds, traffic conditions, and other factors. It can be done for a hundred planned facilities at once as easily as for one, and relevant statistical data for the areas can be generated as a by-product.

These and many other applications are possible by using the network data on the census summary tapes and from other data sources. The Southern California Regional Information Study (SCRIS) and many others are working to develop generalized software to aid such manipulations. None of this work, however, will really benefit those who cannot appreciate the power of computing machinery or those who are not capable of seeing how their problems can be aided by such solutions.

The geographic base file, however, is somewhat limited by the fact that it relates to a specific moment in time. SCRIS, therefore, is devoting considerable effort to the development of a continuous maintenance system so that local areas can keep their files current. The system now envisioned is built around the census metropolitan map with intersection numbers. The plan is that all references to existing or new line segments will be on the basis of information available from the map alone, thus avoiding the problem of having to refer to listings for serial numbers or other arbitrary references.

In general, the system will allow changes to all characteristics of existing segments as well as deletions and additions. All added segments and certain changes to existing segments will carry dates to allow historical analysis. A provision for splitting census blocks in a way analogous to that already used for census tracts is expected to allow creation of new blocks while preserving comparability with 1970 block statistics.

SCRIS is now preparing a publication describing the proposed characteristics of the system and outlining methods for using it.

GEOGRAPHIC CODING

Geographic coding (or geocoding) is the process of locating things to particular geographic areas. Normally the things are places or events that enter the process with some very detailed geographic code already present, such as street address, and the geocoding process then consists of assigning various higher level geographic codes such as census block, census tract, and city. Other original location information may, however, be present instead, particularly if the things are not point related, such as school statistics, census data, or high statistics.

To cover such situations, the concept of geographic coding may be expressed as follows: "Given an exhaustive and mutually exclusive set of subsets and an element that fits in only one of the subsets, determine within which subset the element belongs." This excludes problems such as "Which state is US-66 in?" because US-66 is in many states. As long as one of the subsets is defined to be "everything else," it

is possible to handle problems such as "Which street is parcel 17B, lot A on?" even if the particular parcel is not located on any street. This definition also includes problems such as "Is this address on the list of all restaurants?"

In practical applications, 2 particular situations are normally of interest. The most common of these is geographic coding of street addresses. This is discussed further later. Another common situation is geographic coding of points identified with coordinates. This situation arises with points determined by digitizing or surveying and is normally handled by defining the subset geographic areas as polygons and performing point-in-polygon processing.

SCRIS has also done research on this aspect of geographic coding and has developed techniques for using the geographic base file to define census blocks, tracts, or other areas as polygons. A point-in-polygon routine is also available, but we have not yet developed a complete system to the level of ADMATCH, which is used for geocoding of addresses.

As mentioned earlier, the most common geographic coding problem involves locating street addresses within census blocks, census tracts, or other areas. The census use study ADMATCH system, in conjunction with the geographic base file, has proved a very good solution for such situations. ADMATCH has been used by the use study and SCRIS in coding scores of different local address files to ACG/DIME files and census tract street index files. The kinds of files processed have included customer billing lists, survey cases, administrative records, and tax assessment files.

All files processed by ADMATCH have street address as a common characteristic, but in other ways they have been more different than alike. Some have been as short as 80-character card images, while others have been as long as 150 characters. ZIP code, post office name, city code, and no code have all been used as the major match key. Codes assigned to the successful matches have ranged from ZIP code down to block face and coordinate location.

The 2 essential elements of ADMATCH are its ability to unscramble ordinary mailing addresses and its ability to match by means of a weighted score. These together give the ADMATCH system considerable flexibility and a relatively high rate of match.

Address unscrambling consists of finding the essential elements of the address, standardizing them, and inserting them into a fixed format "match key." The unscrambling is accomplished through a scan for all possible abbreviations such as of street directions and street types. Following the scan, a pattern recognition table is used to determine the actual standardized address. In this way an address such as 918 Way St. can be processed and "Way" can be established as the street name rather than a street type. A large number of other examples are illustrated in the ADMATCH Users Manual that accompanies the program. The essential point is that ADMATCH does not require elaborate clerical standardization of either formats or abbreviations.

The ADMATCH programs are written in IBM-360 Assembly Language and will run on a machine as small as a Model 25 under DOS with 32 K bytes of memory. This version is currently being distributed by the Bureau of the Census at the cost of reproduction ($60). A larger and faster version of the system is currently being prepared for use by those with large 360 machines under OS. A version for use on RCA SPECTRA-70 machines is also in the works. On a 360/30 under DOS the system processes some 200,000 addresses per hour, and the OS version on a 360/65 can pass about 1.5 million addresses per hour. Each address must be processed twice; once to unscramble it and once to match it. An intervening sort is also required.

AREA SYSTEMS

I use the phrase "area systems" as a short way to refer to a very large and complex situation. Many rural areas and all urbanized areas are crisscrossed by a very complicated set of boundary lines for a large variety of districts ranging from state boundaries down to individual parcel lines. These districts are used for legal, statistical, administrative, or other purposes; many districts have multiple uses. In Los Angeles County SCRIS staff members have counted more than 200 different types of districts. Almost all of these districts interpenetrate and overlap, and nearly all of them change frequently.

In most cases it is convenient and necessary to establish at least basic census data for the districts, and it is frequently useful to use local data from one type of district for others. Traffic zones, for example, usually have considerable data available from surveys and other sources, but it might often be convenient to tabulate accident data, employment statistics, or tax assessment information for traffic zones. If the accidents are currently totalled by police precincts, the employment data by state tax board districts, and the assessor information by parcel, it may become quite difficult to reconcile these statistics. The central problem of area systems, then, is to reconcile these conflicting and overlapping districts and to establish correspondence between them.

The SCRIS approach to the problem of area systems has centered around use of the ACG/DIME segments as the lowest common denominator on which to build the necessary correspondences. We have devised systems for defining districts either in terms of census tracts and blocks or in terms of their boundaries. A computer program then takes over and assigns the proper district code to each segment within it. Editing tests are performed to check for overlapping of the same type of district, and if the district is specified as exhaustive a check is made for any unassigned areas.

These systems are intended for general use to increase the power of the ACG/DIME materials and to improve the utility of the census data. The manual effort is very moderate and considerable accuracy is ensured through the machine editing. A revision in the ACG/DIME format is planned to allow room for about 20 additional area codes for each side of each segment. When addresses are coded with ADMATCH, then, it would be possible to extract not only the census tract, block, and coordinates but also the traffic zone, school district, police precinct, or other areas.

A fairly simple computer program is also planned to create correspondence tables among the various districts. These tables may also show, for the various correspondences, the percentage of total road distance or total land area included. The table would then indicate, for example, that 20 percent of census tract 1466 is in school district 3 or 100 percent of it is in traffic zone 9738.

SUMMARY

The ACG/DIME files with census and local data constitute an invaluable resource for transportation planning, city planning, and administration. These files provide computer-usable data tied directly to the highway and road network and to all significant census and local geographic codes.

The ACG/DIME files and the morass of census statistics, however, are relatively worthless to the unaided human brain—even when they are printed out or mapped out by a computer. The same information is, of course, equally worthless to the unaided computer. Only when the two work together can real benefits be obtained from these network data files, geographic coding capabilities, and area systems.