number of nodes, and the node number and x-y coordinates for each node in sequence. Other reports, including reports of errors found and a dictionary of corresponding new and old node numbers, are also produced.

However, the main value of the program lies in its ability to produce graphic output and to produce networks capable of graphic display. Figure 2 (1) shows a segment of the Bureau of the Census MMS map of a portion of Trenton, New Jersey. One can see that each street is included as well as block and tract identification, some major open areas (parks and cemeteries), and major rail lines. This map is the basis for the geocoded DIME file for the Trenton area. Figure 3 shows a computer-drawn map of the same approximate area that uses chain files extracted from the Trenton DIME file. A coordinate grid (of the same scale as that used in the DIME file) is overlaid on the map. In Figure 4, the scale has been enlarged four times, and street chain numbers and street names are both included.

These maps were drawn on the face of a Tektronix 4013 cathode ray tube (CRT) terminal by using a preliminary version of the software described above. Standard CALCOMP plotting routines were used for legends and alphanumeric characters. These figures show the great level of detail available in the DIME files.

CONCLUSIONS

Although some work remains to be done in testing and improving the programs, the work reported here has attained two major objectives:

1. The feasibility of using the DIME geocoding system as a basis for transportation network modeling has been established.
2. A methodology for processing the DIME data files has been developed.

Increasingly, transportation planners are being asked to provide more detailed and more specific analysis of transportation systems in urban areas. This requires detailed network models. Use of the software described here and the available Bureau of the Census DIME files will enable the planner to create detailed network models with a minimum of manual intervention. Although many errors and gaps have been found in DIME files, the Bureau of the Census is striving to obtain complete and correct DIME files for all SMSAs. The software described here provides for human intervention, primarily to detect and correct errors. Although the process cannot be completely automated, the amount of effort required to create a network model will diminish as the quality of DIME file data improves. In the future, it is expected that DIME-based networks, which can be readily matched to other demographic census data, will be in widespread use among transportation planning agencies.

ACKNOWLEDGMENTS

I would like to express my gratitude to Granville Paules, Linda Moore, and Robert B. Dial of the Urban Mass Transportation Administration for their support of this work. I would also like to thank Warren B. Powell, William D. Collins, and Brian P. VonHersen for their work in programming the software described above. This research was performed under a university research and training grant and a research and development grant from the Urban Mass Transportation Administration.

REFERENCES

1. Block Statistics Trenton, NJ-PA Urbanized Area. U.S. Bureau of the Census, Metropolitan Map Series, U.S. Government Printing Office, HC(3)-151, Map Sheet P-21, 1971.
2. M. A. Meyer and J. Silver. Present Status and Future Prospects of the Census Bureau's GBF/DIME CUE Program. U.S. Bureau of the Census, presented at Conference on Geographic Base File Systems—A Forward Look, Boston, April 16-17, 1974.

Publication of this paper sponsored by Committee on Computer Graphics and Interactive Graphics.

Computer Geocoding of Travel Surveys

Priscilla Sawyer Cornelio and Joseph M. Manning, Urban Transportation Systems Associates, Inc., Newton, Massachusetts

Computer geocoding of travel survey data, which involves use of a geographic base file and a series of user-oriented computer programs, is a workable and preferable alternative to manual geocoding, which is tedious and time-consuming. The basis for the computer geocoding system is the Dual Independent Map Encoding/Geographic Base File (DIME/GBF) developed by the U.S. Bureau of the Census. The DIME/GBF, which exists for all standard metropolitan statistical areas, contains detailed information on street segments, including street name, direction, range of house numbers, and census tract and block. Two programs developed by the U.S. Bureau of the Census, ZIPSTAN and UNIMATCH, are used to perform the actual geocoding. ZIPSTAN is a preprocessor program that arranges the addresses before being linked by address to the DIME/GBF by use of the UNIMATCH program. Geocoding on-board and travel surveys in the Boston area indicate that 70 to 80 percent of all addresses can be geocoded to a detailed zone level at a processing cost of $2.12/1000 addresses. Addresses not geocoded automatically are generally incomplete or contain invalid information. The basis for this system, the methods and procedures used in the Boston area, the results of the matching operation, and the costs involved in the effort are presented and discussed.

Planners and engineers frequently use travel and attitude surveys to determine existing travel patterns and latent demand for new or improved transportation facilities.

These surveys are an important tool to officials who use the data generated to determine the feasibility of improvements.

Trip tables are necessary to analyze origin and destination survey data effectively and are generally constructed to indicate traffic or person movements between analysis zones. These zones can be of any size, ranging in area from a block to a town. To build a trip table, the origins and destinations indicated on each survey form must be assigned to analysis zones, a process known as geocoding.

In many surveys, geocoding is performed manually. This involves locating the address either on a map or in an address coding guide and coding the tract or zone on the survey form for data transcription. The complexity of manual geocoding is proportional to the level of analysis zone used. If a town code is sufficient, geocoding is a simple operation. However, if there are different analysis zones for odd and even house numbers on each street and each block or group of blocks represents a zone, the time involved to code each survey can be substantial.

Obviously, when a large number of addresses must be geocoded, the manual process can be expensive because of the staff time required to code and quality-check the forms. In addition, it is inevitable that results may be inconsistent when the judgment of a coder is necessary—for example, in determining if a building is on the odd- or even-numbered side of the street. A substantial amount of human error also can be introduced during analysis zone coding and data transcription phases when digits can be transposed or incorrectly transcribed.

Computer geocoding is a workable and preferable alternative to manual geocoding. The process uses a geographic base file and a series of user-oriented computer programs. It can result in a better product for less money because the computer can perform sophisticated functions rapidly, economically, and consistently.

DEFINITIONS

Before describing the automatic geocoding process, it is necessary to understand the three major inputs to the process. These are

1. The Dual Independent Map Encoding/Geographic Base File (DIME/GBF),
2. The ZIP Code Standardizer (ZIPSTAN) program, and
3. The Universal Matching (UNIMATCH) program.

DIME/GBF

The DIME/GBF, developed by the U.S. Bureau of the Census, forms the basis of the reference file used for geocoding. The DIME/GBF contains records that describe street segments, legal boundaries, and topographic features including rivers, railroad tracks, and canals (3, 4, 6). Each segment contains a variety of information. Included for street addresses are the street name, street prefix and suffix, highest and lowest house number ranges for both sides of the street, city code, county code, state code, ZIP code, and census tract and block designations. In addition, the file contains X-Y coordinates and to and from node numbers that can be used for mapping.

Figure 1 shows a sample of the map used to construct a DIME/GBF. Each street segment between node points has its own data record, which contains the information given in Table 1 (4).

The U.S. Bureau of the Census, with substantial assistance from local areas, has constructed a DIME/GBF for the urbanized areas of each of the 233 U.S. standard metropolitan statistical areas (SMSAs). The file for any SMSA may be purchased from the bureau for about $80. However, the files for some cities may be more complete than those for others. The bureau encourages planning agencies to correct and update their files to eliminate errors such as incorrect ZIP codes or missing streets. Some DIME/GBFs have been carefully edited, corrected, and updated whereas others have not.

ZIPSTAN

The ZIPSTAN program is a preprocessor that arranges the addresses before being linked by address to the DIME/GBF by using the UNIMATCH program. ZIPSTAN standardizes addresses by ensuring that house numbers and city names are in specific locations on the data records. The program uses equivalency tables that contain possible abbreviations and common misspellings of street prefixes, suffixes, names, and street types and provide the appropriate conversion of each. Because the DIME/GBF uses numeric city, county, and state codes, an alphanumeric table of equivalents is required input for ZIPSTAN. For example, a city table converts the city of Boston to a county-city code of 017005. ZIPSTAN appends to each address record a 100-character matchkey that includes the standardized address (2).

UNIMATCH

The UNIMATCH program is used to match the addresses to be geocoded with the DIME/GBF and to attach zone information to the data record (1). The UNIMATCH program can be set up to attempt an exact match of house number, street name, and city code. If an exact match is not possible, the program applies weights and penalties set by the programmer to misspelled street names and blank or nonmatching street types, directions, and house numbers. Thus, data records are linked with reference file entries that have a high probability of being the same but have some missing elements in their address constructions.

To illustrate how the UNIMATCH program operates, if the address to be geocoded is missing a street-type suffix (for example, 100 Summer ___), and the reference file has a Summer Street with a house-number range of 1 to 50 and a Summer Road with a house-number range of 75 to 300, the program can be set to select the most probable reference file entry, 100 Summer Road. The programmer also has the option of instructing the computer to assume Street for all addresses with a blank suffix. Thus, if the reference file contained entries for 1 to 200 Summer Street and 75 to 300 Summer Road, the program would select the Summer Street entry.

The UNIMATCH and ZIPSTAN programs have certain computer requirements. They are written in IBM System/360 assembler language and are designed for computers that use the IBM System 370/360 operating system. The programs will operate under the following configurations: (a) primary control program, (b) multiprogramming fixed task, (c) multiprogramming variable task, and (d) VS1 or VS2, the virtual systems.

GEOCODING PROCESS

Initial Steps

To take advantage of automatic geocoding, certain data elements are necessary. The origin and destination addresses must contain a house number or street name including prefix and type. If more than one city or state

Figure 1. DIME/GBF map.

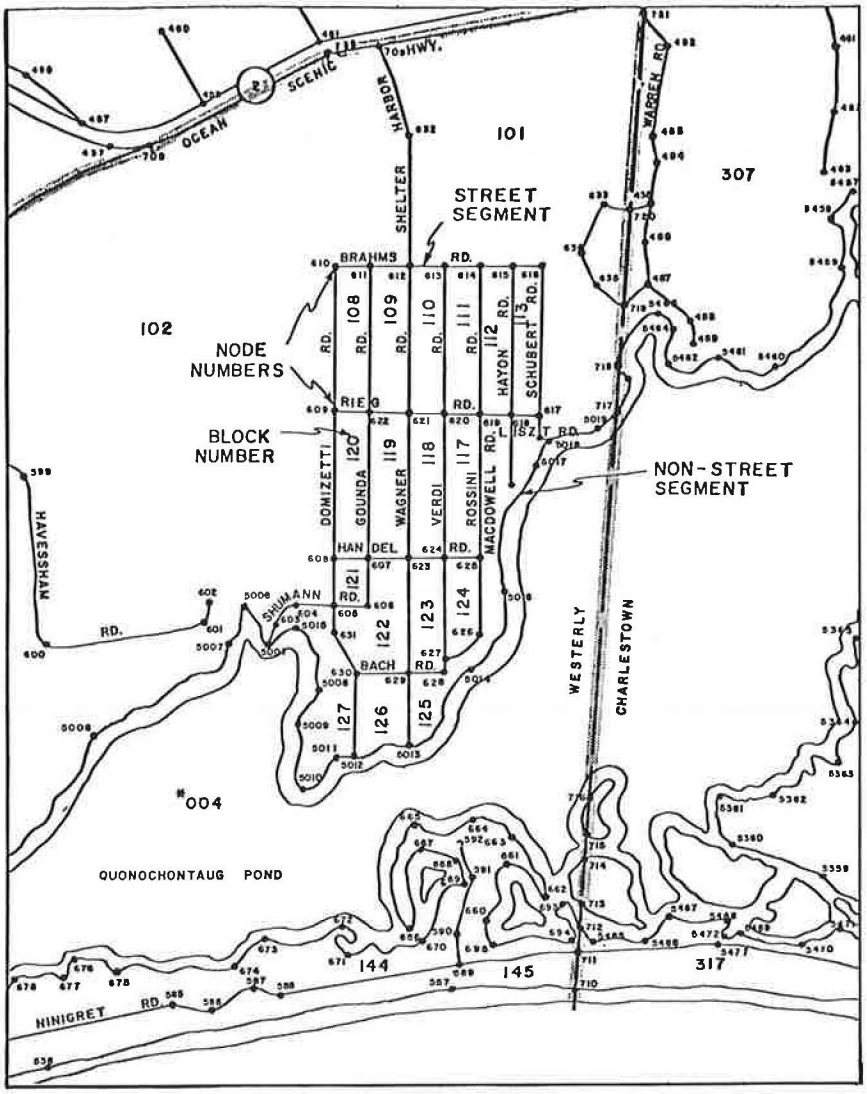


Table 1. Geographic elements contained in DIME/GBF.

Item	Characters	Item	Characters
Street prefix direction	1-2	State code left	139-140
Street or nonstreet feature name	3-22	County code left	141-143
Street type	23-26	Minor civil division code/ census county division code left	144-146
Street suffix direction	27-28		
Nonstreet feature code	29	1970 congressional district left	147-148
1970 enumeration district left (1970 nonmail census areas only)	30-34	1970 area code left	149-151
		Block left (basic)	152-154
Blank (census use only)	35-40	Block left (suffix)	155-156
1970 enumeration district right (1970 nonmail census areas only)	41-45	1960-1970 annexation code left (1970 mail census areas only)	157
From map (basic number)	46-48	State code right	158-159
From map (suffix)	49-50	County code right	160-162
To map (basic number)	51-53	Minor civil division code/ census county division code right	163-165
To map (suffix)	54-55		
Coding limit flag	56	1970 congressional district right	166-167
Left low address	57-62	1970 area code right	168-170
Left high address	63-68	Block right (basic)	171-173
Right low address	69-74	Block right (suffix)	174-175
Right high address	75-80	1960-1970 annexation code right (1970 mail census areas only)	176
File code	81-84		
Record number	85-89	From state plane code	177-178
Check digit	91	To state plane code	179-180
Census tract left (basic)	92-95	From map set mile (X coordinate)	181-186
Census tract left (suffix)	96-97	From map set mile (Y coordinate)	187-192
Census tract right (basic)	98-101	To map set mile (X coordinate)	193-198
Census tract right (suffix)	102-103	To map set mile (Y coordinate)	199-204
ZIP code left	104-108	From latitude (Y coordinate)	205-210
ZIP code right	109-113	From longitude (X coordinate)	211-217
Standard metropolitan statistical area	114-117	To latitude (Y coordinate)	218-223
		To longitude (X coordinate)	224-230
Street code (1970 mail census areas only)	118-122	From state plane (Y coordinate)	231-237
		From state plane (X coordinate)	238-244
From node	123-126	To state plane (Y coordinate)	245-251
To node	127-130	To state plane (X coordinate)	252-258
Place code left	131-134	Blank (census use only)	259-300
Place code right	135-138		

is involved, municipality and state names are required. City names often can be replaced with ZIP codes because ZIP codes generally follow municipal boundaries. Street intersections are not valid input unless an intersection reference file has been created.

Before keypunching is initiated, all questionnaires should be edited to ensure completeness and consistency. The editing procedure consists of checking for blank fields and transferring erroneous field entries to the correct location. For example, a house number that was written in the street-name field should be moved to its proper position on the form. Although many of these deficiencies can be overcome by ZIPSTAN, this editing procedure consistently leads to better matching in subsequent steps.

Computer Geocoding

Figure 2 shows the process involved in a simple geocoding project. In a more complex task, manipulation of the DIME/GBF or additional matches against reference files may be necessary.

Before geocoding, input data record addresses are reformatted to one address per record. If a record contains three addresses (origin, destination, and parking location), the record is reformatted to three separate records, each containing an address and a unique serial number and record position number. These elements permit rejoining the addresses in subsequent processing. Because the ZIPSTAN and UNIMATCH programs reference only one location per record for matching, reformatting the data allows processing of all addresses in one pass.

The steps involved in the geocoding process are the following:

1. The DIME/GBF is arranged in alphabetical order. The records are sorted by city or town and alphabetized by the street name records in each town. They are further sorted within house number, and the file is then ready for accessing. A printout of the file could be used as a guide for manual geocoding.
2. Organization of the data records is initiated, and ZIPSTAN is used to arrange them in a consistent format.
3. After ZIPSTAN has standardized the addresses, they are sorted in the same fashion as was the DIME/GBF. Following that sorting, a coder could manually code most of the data by visually inspecting each list and identifying the matches.
4. The UNIMATCH program is used to link the addresses to be geocoded with the DIME/GBF and to attach the appropriate census tract, block, and traffic zone to the travel data record.

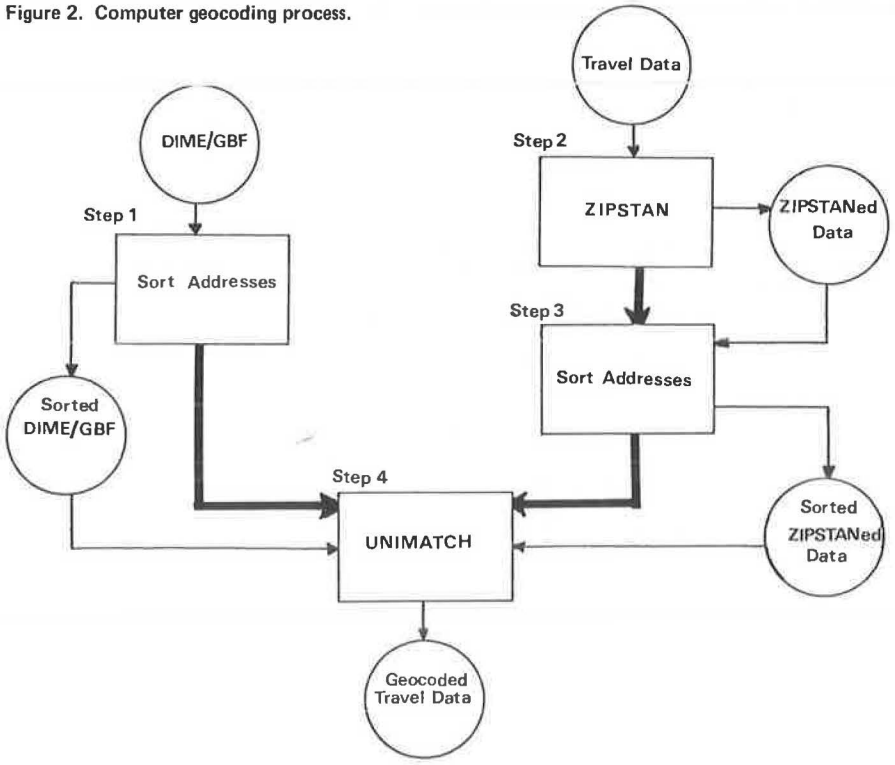
RESULTS

The success rate for computer geocoding varies according to the quality of the respondents' answers, the quality of editing, the accuracy of data transcription, and the allowable weights and penalties specified by the user of the UNIMATCH program.

The editing process is more important in computer geocoding than in manual coding because, to a point, the machine cannot interpret what the respondent meant. For instance, unless specifically informed in the ZIPSTAN preprocessor, the computer does not regard BOS to be the same as Boston. In addition, the ZIPSTAN and UNIMATCH programs do not have the capability of determining that Beaconstreet is actually Beacon Street. Beaconstreet would be treated as a street name, missing a street type. These problems can be solved easily by using the ZIPSTAN equivalency tables; however, it is difficult to anticipate what misspellings and other errors will occur, and to correct such errors requires a second pass.

Inaccurate data transcription can negate the effects of precise editing. A space or letter punched before the beginning of a name, such as Beacon or BBeacon, or a

Figure 2. Computer geocoding process.



reversal of letters, such as Ebacon, precludes a successful match. Experience has shown that the computer time and storage requirements for a match that allows all letters to be misspelled are excessive because almost every reference file entry becomes a possible candidate for a match. A solution to this difficulty is to specify program weights to require an exact match on the first two or three letters in a name, such as Bea___, and allow misspellings for matching the remaining letters. Thus, Beacno Street and Beanco Street can be matched with Beacon Street in the reference file.

The final variables that affect the match rate are the programmer-specified weights and penalties. It is possible to require an exact match in which house number, street name, prefix, suffix, and city must correspond exactly to the reference file entry, but exact matches have a low success rate. If street misspellings and omissions of prefix, suffix, and house numbers are allowed, the match rate increases proportionately.

Experience with on-board surveys conducted for the Massachusetts Bay Transportation Authority (MBTA) and with travel surveys performed as part of the Central Artery Study in Boston has shown that a 70 to 80 percent match rate can be expected with computer geocoding if flexibility is provided within the UNIMATCH program (5). If exact matches had been required in the MBTA and Central Artery Study surveys, the match rates would have been 6 and 4 percent respectively. Note that the UNIMATCH program is, in effect, making the same probabilistic choices that would be made in manual coding. The difference is that the computer is always consistent and unbiased when making these choices whereas a coder is not.

In the MBTA survey, a major generator file, later expanded for the Central Artery Study, was constructed. This file contained the names of buildings such as City Hall and the State House and places such as Harvard Square and Copley Plaza, which were not in the DIME/GBF. The locations of these generators were manually coded, transcribed on tape, and processed against addresses that did not match with the DIME/GBF. In smaller surveys, this additional step is often unnecessary because the addresses of certain major generators can be entered on the survey form during the editing step. However, in large surveys, this secondary match is required.

The table below gives a summary of the results of geocoding in the Central Artery Study survey:

Condition	Number of Addresses	Match Rate (%)
Exact match	5 026	4
Probable match	84 375	69
Generator file match	22 939	19
Nonmatch	10 524	8
Total	122 864	100

The supplemental match with the generator file increased the match rate from 73 to 92 percent. Altogether, 112 340 records were geocoded to a detailed zone level. The remaining 10 524 addresses were subsequently coded

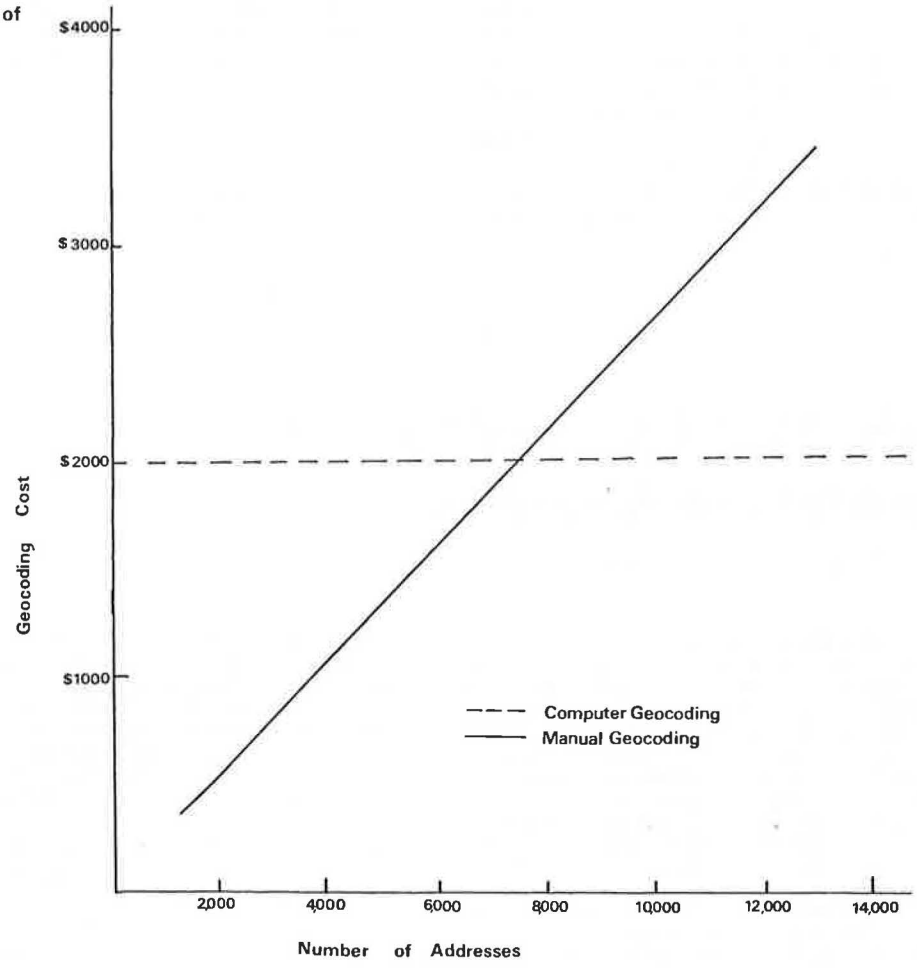


Figure 3. Cost comparison of computer and manual geocoding.

to a town level. One hour and 17 min of computer processing unit (CPU) time and 576 K of virtual storage on an IBM 370 computer were required to geocode 89 401 addresses by using the DIME/GBF exact and probable match capabilities of the UNIMATCH program. The match with the generator file, which had 1900 entries, took about 13.5 min and used 64 K of virtual storage. As these figures indicate, the CPU time and storage requirements are a function of the number of fields being matched (city, street name, prefix, suffix, and house number versus city and building name) and the number of reference file and data file entries.

The MBTA and Central Artery Study addresses were geocoded on state computer facilities. On other geocoding projects, the computer costs associated with ZIPSTAN and UNIMATCH processing on a private computer facility average approximately $2.12/1000 addresses. Depending on the complexity of the DIME/GBF, the entire geocoding process, including development of ZIPSTAN tables and UNIMATCH specifications, can be developed, set up, and pretested at a cost of from $1000 to $4000. The costs for a major DIME/GBF such as Boston's, which has over 90 000 records, is at the higher end of the scale, and the costs for smaller DIME/GBFs are correspondingly lower.

As previously mentioned, some manual coding was necessary to construct the major generator file. Of the 1891 entries in the file, about 700 were unique locations; the remainder were variations of a place or building name. The 700 unique locations required about 48 person hours, or about 4 min/address, to geocode. At a cost of $4/h, excluding overhead and supervision of coders, this results in a cost of $0.267/address or $267/1000 addresses.

Figure 3 shows a graph that illustrates the cost relation between manual and computer geocoding. In constructing this figure, the setup cost for the computer geocoding was assumed to be $2000 and the processing cost $2.12/1000 addresses. The manual cost is $267/1000 addresses. As Figure 3 indicates, at these costs it is more economical to geocode by computer when there are about 7500 addresses or more. However, if smaller surveys, such as on-board bus surveys, are to be conducted on a continuing basis, it is more cost-effective to develop the computer geocoding capability.

Once they are set up, the DIME/GBF, major generator file, and ZIPSTAN and UNIMATCH programs can be used almost interchangeably on any on-board or other type of travel survey an agency may wish to conduct. Although a somewhat substantial start-up cost is required to implement a computer geocoding system, it can be more than recovered through the multiple uses of the system.

ACKNOWLEDGMENTS

The computer geocoding system described in this paper was prepared for the Community Affairs Division of the Massachusetts Bay Transportation Authority, the Central Artery Study of the Massachusetts Department of Public Works, and the Federal Highway Administration, U.S. Department of Transportation.

REFERENCES

1. Census Use Study: UNIMATCH 1 Users Manual. U.S. Bureau of the Census, 1974.
2. Census Use Study: ZIPSTAN Address Standardizer. U.S. Bureau of the Census, 1974.
3. CREATE: Coders' Manual. U.S. Bureau of the Census, 1976.
4. Documentation for the GBF/DIME-File. U.S. Bureau of the Census, 1976.
5. Documentation of On-Board Surveys—Volume 1. Urban Transportation Systems Associates and Massachusetts Bay Transportation Authority, 1976.
6. GBF/DIME System: A Geographic Dimension for Decision-Making. U.S. Bureau of the Census, Conference Proceedings, April 1976.

Publication of this paper sponsored by Committee on Computer Graphics and Interactive Graphics.

Network Base File System for Transportation Analysis

Claus D. Gehner, University of Washington

Development of the Network Base File System (NETBASIS), which has been under way at the University of Washington since 1974, is described. NETBASIS, which is implemented in an interactive graphic computer environment, explicitly builds on geographic base file data to provide a general-purpose transportation network data base together with the required data manipulation and display software. In addition to the standard capabilities of data base input, editing, and retrieval provided by the host data base management system, software has been implemented to allow the user to extract geographic or functional subsets of the network (e.g., arterials only). This extracted network can then be abstracted to remove nonintersection nodes and thus produce an efficient network model that can be used within most existing transportation planning tools. An interactive graphic network editor is also provided that allows the user to modify the extracted and abstracted network to reflect planning options to be analyzed. Segment-specific transportation data, such as transit lines, speed, and capacity, are added to the data base by a FROM-ON-TO coding scheme, which significantly reduces costly and error-prone intermediate data coding. Use of a commercial data base management system with a FORTRAN interface has resulted in significant cost savings in the development of the data input, editing, and retrieval software required to allow users easy access to the network data and to provide maximum flexibility of base file content and structure for adaptation to changing user requirements.

The existence of U.S. Bureau of the Census geographic base files (GBFs) for most larger urban areas offers a significant resource for the network models required as part of many transportation studies. The thrust of the development of the Network Base File System