AN EVALUATION OF THE GEOGRAPHIC CODING IN THE 1977 ECONOMIC CENSUSES: AN OVERVIEW

John F. Judge, Jr., U.S. Bureau of the Census

During the 1977 Economic Censuses, the Census Bureau assigned geographic classification codes ("geocodes") to approximately seven million establishments which were within the scope of the economic censuses. To evaluate the geocoding, a random probability sample of addresses was selected for independent research. The evaluation, when completed, will be used (1) to determine the overall accuracy of the geocoding system, and (2) to pinpoint flaws in the geocoding system.

Although not yet completed, the preliminary results show a significant improvement over past economic censuses. These early results indicate that, for those addresses for which geocodes could be determined clerically, the correct geocodes were assigned to approximately 96 percent of the addresses.

There are three elements necessary for geocoding: (1) the addresses, (2) the reference files, and (3) the coding algorithm which ties the first two elements together. Of these three elements, only the reference files and the coding algorithm can be controlled to any great extent. The addresses can be controlled only to the extent that they are processed into a format compatible with the reference files. Too many variables such as spelling errors, keying errors, incorrect ZIP codes, and others, can render some addresses virtually uncodable on the computer.

The 1977 Economic Censuses geocoding system consisted of four major computer programs: (1) the standardizer, (2) the header match, (3) the detail match, and (4) the PAIR program. Together these programs provide the means to link the address files to the reference files, to assign geocodes, and to assess geocode quality. The automated computer geocoding system was further enhanced by a clerical operation to improve the geocode quality of cases of sufficient importance.

When the economic census addresses (house number and street name) were delivered for geocoding, they were freeform. The standardizer program formatted the street addresses to resemble the Address Reference File (ARF) records, analyzed addresses as to the type of address, and substituted standard abbreviations for spelling variations of street type and direction. It also adjusted the post office name, if necessary, for compatibility with the City Reference File (CRF), and it validated the ZIP/State combination. If the ZIP/State combination did not agree, and if the existing State abbreviation was illegal, a new State abbreviation was derived from the Social Security Administration (SSA) or Internal Revenue District (IRD) number, and validation was again attempted. Illegal ZIP/State combinations were flagged. The standardizer also set flags identifying the level of coding required, and clustered header address components (post office name, State abbreviation, and ZIP code) into a separate file for independent coding by the header match process.

In the header match process, the header clusters were matched to the CRF using the ZIP sort code (a device for partitioning the reference files) as the major match key. An exact match on the ZIP sort code and post office name is flagged as a high confidence match (a mismatch on State abbreviation was tolerated).

Clusters that did not match exactly were examined for CRF name similarities within the ZIP key partitions. A character match algorithm which scored name similarities on a scale from 0 to 10 was used to detect possible name misspellings (10 being a "perfect" match). These clusters were also matched to the CRF using the State abbreviation as the major match key. Header cluster addresses that matched only to the State abbreviation or the first three digits of the ZIP key without any post office name similarities were regarded as force coded. The CRF candidate with the highest match confidence (even if force coded) provided the geocode (or geocodes, in case of ties) that was assigned to a header cluster address. These codes were transmitted to the PAIR process.

Individual establishment street addresses were matched to the ARF on ZIP sort key and street name. Establishments that did not match on these keys were transmitted to the PAIR process to be header or force coded. Limited equivocation was tolerated on street type, prefix and suffix directions, ZIP code, and house number for establishments that matched the ARF on ZIP key and street name. The equivocation was based on a scoring system that penalizes combinations of street component mismatches. Geocodes were selected from the ARF record with the highest total point score within tolerated mismatch combinations. Whenever there was a choice of more than one set of geocodes possible for a house number street name combination in the ARF, only two sets of geocodes were transmitted to the PAIR process from the detail match. In case of a tie, the primary geocode was selected on the basis of a previously assigned (1972 Economic Censuses) geocode. Flags were set to indicate ties and to indicate the quality of the ARF match. The detail coded establishment was transmitted to the PAIR process for a final geocode and confidence level assignment.

The geocodes, derived by the header and detail match processes, were aligned and adjudicated by the PAIR process. PAIR was used to perform an analysis of header and detail geocodes, then confidence flags, to select winning geocodes in case of header and/or detail ties. PAIR provided the final confidence level of the selected code based upon the adjudication process used to select the code. It selected a universe for clerical coding based upon the confidence flag assigned and the importance of the establishment. All force coded cases and important establishments that were header coded with low confidence were transmitted to the clerical universe, which was approximately 100,000 cases.

Once the geocodes were assigned, a random sample of addresses was selected from the Bureau's master sample tape of economic census establishments. The sample selection was made on the basis of the unique 11-digit number assigned by the Bureau to identify each economic establishment. It was decided to select the sample by choosing certain of the digits of the ID

equal to randomly chosen values so that if the file were destroyed it could be recreated with a minimum effort. Choosing every n^{th} cases from the master sample could not have accomplished this.

Since certain digits of this number are not randomly distributed, these could not be used. However, since the check digit (MOD-10) portion of the ID number is randomly distributed, a 10 percent sample of the master sample was selected by arbitrarily choosing every establishment with a check digit equal to 5. Even after the sample file had been unduplicated, the sample was too large for our purposes so a 40 percent sample was selected utilizing one of the other randomly distributed digits of the ID. This yielded a desirable sample size of approximately 16,000 cases. Once this was done a subsample was selected, approximately 20 percent, for an "on-the-ground" verification. In the full sample the establishment addresses selected were located in 2,179 of the 3,143 counties or county equivalents of the United States.

The 1977 Economic Censuses Geocoding Evaluation is divided into three phases. The first, or control, phase consists of the full sample of 16,087 addresses with the final geocodes to which each address was assigned for 1977 Economic Censuses tabulation and publication purposes. This is the phase which is being evaluated by the other two phases. The second, or clerical, phase consists of the same sample of addresses, each address assigned an appropriate set of geocodes by a qualified geocoding clerk. The third, or field, phase is comprised of the 20 percent subsample of 3,057 addresses verified "on-the-ground."

In Phase II the addresses were clerically researched using whatever geographic reference materials that were available independent of the ARF and the CRF. These references include the Bureau's GBF/DIME-Files, commercial city directories, city atlases, county highway maps, ZIP code directories, local telephone directories, and any other available resources. After each work unit was completed it was dependently verified using a lot acceptance plan (AQL = 2.5 percent) with normal inspection. No provision was made for either tightened or reduced inspection; the verifier reworked the unit if it failed inspection. (N.B.: The verifiers are more highly qualified at geocoding than the clerks, usually having several more years of experience.) At this point every address in the work unit would have a complete set of geocodes, i.e., State code, county code, place code (with a MOD-10 check digit computed on State x county x place), and census tract code (with a MOD-10 check digit computed on the tract code). Since only retail data are tabulated to the census tract level, and then, only if the retail establishment is located inside a locally-defined Central Business District (CBD) census tract, all non-retail and all non-CBD retail cases are consigned to a city-wide census tract number (9999.996).

Some addresses at this stage may still remain partially uncoded (indeterminate) at some level of geographic coding because of insufficient address information or a lack of reference materials. However, during the 1977 Economic Censuses, every address is geographically coded to the "place" level with varying degrees of confidence. Therefore, a special set of geocodes was set up to identify those addresses for which clerical geocodes could not be determined. The verifiers were instructed to research all these cases aside from the quality control plan. Where the verifier could not determine the complete geocodes, an attempt was made to telephone the establishment to ascertain as much information as necessary to assign geocodes (nearest intersections, location on a highway, intervening physical features).

The field phase, or Phase III, is actually a two-step operation. Since the sample selection yielded a random national subsample of 3,057 addresses representing 1,004 of the 3,143 counties or county equivalents in the United States, this was to be a wide-spread operation involving a large number of Field Division interviewers. As most field interviewers are not familiar with the geographic code structure and it is, at best, a difficult subject to teach second-hand in a training memorandum, a simpler method was devised to get the needed address information from the field.

The first step of the field phase, then, was for the interviewer to go to the address given for the establishment and determine if the economic activity was conducted at that address. If it was not, then the interviewer attempted to determine from several sources at what address the establishment was physically located and then went to visit the address. Once the establishment's address had been positively ascertained, the interviewer then drew a sketch map of the location, using a standard sketch map, showing the building in relation to its street and the nearest intersecting (cross) streets. In addition, if the establishment was on or near any political geographic boundaries, these were sketched in relation to the street pattern. The interviewer was also to indicate the name of the State, county and locality where the establishment was located plus any remarks that would assist the geocoding operation.

After all the addresses had been returned from the field, the second step of the field phase went into effect. In the second step, qualified geocoding clerks, selecting the appropriate reference maps, compared the sketch map to the best maps available to determine in which State, county, and place (if any) each address and establishment was located and assigned the appropriate geocodes. If it was a retail firm in a city containing a census-recognized CBD tract(s), the clerk would determine also in which census tract the establishment was located and assign the appropriate tract code.

After all of the addresses from Phases II and III have been coded and verified, the data are to be keyed and transmitted and matched back to the Phase I records for the same establishments. The geographic codes from the three phases (in most cases only two phases) will be compared and, where differences exist, the differences will be adjudicated and a determination of the reasons for coding errors will be made.

When all of the discrepancies have been adjudicated and rectified, the results will be tallied so as to demonstrate coding error rates

in several different categories, such as by State, by Standard Industrial Classification, by header versus detail coding, and several others. For example, one would not expect the same coding error rate for Mississippi as for Massachusetts. The former is more rural with a larger incidence of non-detail codable addresses (P.O. boxes, rural routes, etc.) while the latter is, of course, the converse, being more urban.

As was stated earlier, the results of the evaluation are only preliminary. As of this writing only 14,000 of the 16,000 addresses for Phase II have been geocoded and none of the addresses for Phase III. However, of the 14,000 addresses which have been geocoded, approximately 12,700 were determined to have had the correct geocodes assigned during Phase I while only some 500 were determined to have been incorrectly coded. For another approximately 750 addresses, a complete set of geocodes could not be determined, either because of poor address quality, lack of reference materials, or inability to contact the establishment by telephone. The correct geocodes, then, were assigned to approximately 96 percent of the establishments for which geocodes could be determined. This is a marked improvement over geocoding in the 1972 Economic Censuses where the evaluation revealed that the correct geocodes were assigned to only approximately 90 percent of the addresses for which a geocode could be determined (4,086 correct, 511 incorrect, 202 indeterminate).

This gain in the geocoding rate is in large part due to the improvements made to those two of the geocoding elements that can be controlled: the reference files and the coding algorithm. Because of the importance of accurate geocoding for the use of economic statistics for planning and other purposes, the Bureau is constantly endeavoring to improve the accuracy of its geographic activities. Evaluations of this type enable the Bureau to identify those areas where future improvements can continue to be made.