QUALITY CONTROL OF COMPUTERIZED GEOGRAPHIC INFORMATION

Kathryn Thomas, William C. Davie, John S. Linebarger
Bureau of the Census

A street map is a graphic representation of the physical features which can be found in our environment. Computerized information files have been developed which contain data and allow the user to associate this data to its specific geographic location in much the same way the use of a map enables us to relate to our physical environment. Combined into one package, these planning tools, maps and computer files become an invaluable data reference source. A geographic base file, then, may be defined as a reference file designed to relate various types of factual data to their geographical location.

In conjunction with the 1970 Census of Population and Housing, the Census Bureau began developing a system whereby a computer could pinpoint the location of a specific address in relation to other addresses, its street location, and so forth. This system -- called the Geographic Base (DIME) File, or "GBF/DIME" -- provided a fast and accurate means of locating addresses to their geography.

The Bureau's Geographic Base File System, currently undergoing expansion to include most major metropolitan areas, was developed jointly by the Bureau of the Census and local area governments and/or planning organizations. The technique for creating the Geographic Base File is a method of translating geographic information such as street names and address ranges from maps and other source materials into a form which can be read by a computer.

The GBF/DIME File is composed of segment records. A segment is defined as a length of a street or other feature between two distinct vertices, or nodes. Other features include imaginary lines defining political or other boundaries; topological features such as rivers and shorelines; other physical map features such as railroad tracks; and any other feature defining a boundary. Nodes are points where features begin, end, intersect, or curve sharply. The Geographic Base File incorporates the DIME (Dual Independent Map Encoding) feature, which gives an identification number at each end of a segment and includes both sides of the street in a single record. Using the node identifications, it is possible to link records together to form a single block. This feature helps in identifying inconsistencies in the information, such as when the computer cannot identify all segments bounding a block. Edits have been formulated to detect such errors. The Bureau provides these edits to local agencies so that the inconsistencies in a file can be eliminated to help improve the file's accuracy.

In order for the file to remain useful over time, the Bureau recommends that local agencies have a regular program for updating and maintenance. The Bureau of the Census has established the CUE program -- a nationwide, standardized approach to the Correction, Update, and Extension of the Geographic Base File/Dual Independent Map Encoding (GBF/DIME) System. The Bureau provides local agencies with clerical procedures, processing methodology, and the computer programs necessary to carry out the CUE operations.

Currently, no formal quality control program exists for the CUE operation, although such a program presently is being tested as a result of recent research discussed in this paper. The Bureau has relied heavily on the local areas to do a high-quality job and on various computer edits to uncover inconsistencies in the file. A formal QC program to evaluate available local reference materials does exist for areas initially setting up a file, but no quality checks are done after the initial reference material evaluation to insure continued updating and high-quality output by every local area.

The purpose of this paper is to discuss the evaluation studies that were undertaken in a few selected areas to determine the adequacy of the creation and correction plans currently in use and to determine whether there was a need for quality control plans for the CUE operation. One site, City A, was chosen for a study to identify phases of the operation where quality control procedures are needed in the "create" portion of the GBF/DIME System program. In order to identify particular phases of the operation where quality control procedures were needed in the CUE program, a separate evaluation was conducted in a different city, City B. Based on these studies, recommendations on quality control requirements were made.

City A Evaluation [1/]

The evaluation of the "create" portion of the GBF/DIME System file was designed with two distinct objectives in mind. One of these was to evaluate the quality of current local reference materials and to verify the blockside error rate estimated from the quality control procedures used in "creating" the file in April of 1974. The procedures employed and analysis of resultant data from this portion of the evaluation study are described below. Further objectives of this geographic evaluation were to compute a segment error rate and an address error rate for the currently existing City A GBF/DIME System file. These results and a description of the methods employed in obtaining them are also described in this document.

1/ It should be noted that evaluation in only one area is inadequate to conclude about the quality of reference materials in all areas. However, this evaluation did produce information that was helpful in planning the QC program. Later evaluations from applications in different areas will give more adequate readings on the quality of reference materials.

Evaluation of Local Reference Materials in City A

The study consisted of systematically selecting (following a random start) a 1-in-40 sample of blocks from a listing of all blocks in City A. This sample was the identical sample of blocks used in evaluating the quality of local reference materials during the creation of that particular GBF/DIME System File in 1974.

The reference materials were assigned one blockside error whenever one or more address units existing on the blockside could not be located on that blockside via the 1975 reference materials. The results of checking 42 sample blocks containing 175 blocksides yielded a blockside error rate of 2.9 percent with 95 percent confidence limits of 0.3-5.5 percent. No comparison was made to the GBF/DIME System file.

Five blocksides were found to differ when comparing the findings on the ground with the set of 1975 reference materials. Four of the errors were lower and/or higher addresses found on the ground than shown in the reference material. The fifth error was a blockside having odd-even street numbers reversed as to the side of the street on which they were located.

Table A1 below shows the distribution of the sample and error rates for each of the 10 map sheets for City A. Confidence limits in the table were based on the binomial probabilities assuming simple random sampling. When consideration was given to the cluster effect, similar estimates were obtained. The intraclass correlation[2/] was calculated at -0.023, indicating that there was little clustering of errors within a block.

Based on this evaluation, the overall quality of the reference materials was assessed as adequate for use as a source for creating a GBF/DIME System file. However, as indicated in Table A1, the variation on the quality of the materials by map sheet could be large. Therefore, it was concluded that, while the 1-in-40 block sample may be a good means of evaluating overall reference materials, the sample is much too small and spread too thinly to produce good estimates of the quality of the reference materials for a particular map sheet.

Computation of Segment and Addressable Unit Error Rates

A sample was selected from the City A GBF/DIME System file (file after edit dated September 1974); resulting from the CREATE phase of the GBF/DIME System program. The Metropolitan Map Series were used to locate the sample segments in the field. A team of two persons located each segment on the ground verifying the name with street signs, along with the names of intersecting streets. High and low addresses were recorded for both sides of the segment; the total number of addresses contained in the segment (both sides of the street) were noted; odd and even address placement was verified; "out of sequence" addresses and addresses falling outside the file range were documented. Persons in the field had access to the file data, so reconciliation of any differences was conducted at the same time as the original field check. Therefore, all results in this report are after field reconciliation.

After removal of non-street features and inaccessible segments from the sample, there were 280 sample segments. From these it is estimated that 18.6 percent of the segments in the GBF/DIME System file were in error. The 95 percent confidence limits of this estimate are 14.0 and 23.2 percent. Sixty percent of the segment errors (projected at 11.1 percent of the total file) were critical errors; the remaining 40 percent of the segment errors (projected at 7.5 percent of the total file) were non-critical errors. A "critical" error is defined as one that would cause an incoming address to be coded to the wrong geographic area, i.e., block or Census tract. A "non-critical" error segment is defined as one which has an error in the file but which will not result in miscoding of addresses to an incorrect geographic area.

TABLE A1

Sample and Error Distribution by Map

Map No.	No. of Blocks	No. of Blocksides	Blocksides in Error	Blockside Error Rate (%)	95 % Confidence Limits
TOTALS	42	175	5	2.9	0.3, 5.5
1	2	9	1	11.1	0.0,35.5*
2	-	-	-	-	-
3	1	4	-	-	0.0,60.3*
3NW	10	42	2	4.8	1.7, 7.9
3SW	8	32	1	3.1	0.2, 6.0
4	3	12	-	-	0.0,26.5*
4NE	14	60	-	-	0.0, 6.0*
4SE	3	12	-	-	0.0,26.5*
5	1	4	1	25.0	0.0,68.4*
6	-	-	-	-	-

*Constructed by using the binomial distribution for computation of confidence limits. See Hansen, Hurwitz and Madow, "Sample Survey Methods and Theory," Vol. 1, pp. 135-136.

About 23 percent of the critical errors fell in the category "lower and/or higher addresses in segment on the ground than allowed for in the file." Another 26 percent of the critical errors were an "odd-even anomaly;" that means that at least one but not necessarily all of the addresses in the segment were on the wrong side of the segment (i.e., an odd address on the even side of a street or vice versa). It should be noted that the quality control plans that were in existence

2/ Deming, William Edward, "Some Theory of Sampling," p. 203.

for the CREATE phase would not necessarily detect an error of the latter type. A breakdown of all critical errors is contained in Table A2.

TABLE A2

Critical Errors and Non-Critical Errors

In addition to naming the errors, a code for the possible sources of the errors are shown in this table. The sources listed are the best explanation for probable causes of the errors; however, no evidence is available to substantiate or refute these reasons.

Error Type	Number of Segments in Error	Percent of All Critical/ Non-Critical Errors	Percent of File	Possible Source of Error*
Total Errors	52		18.6	
Total Critical Errors	31	100.0	11.1	
1. Odd-Even Sides Reversed	2	6.5	0.7	1, 2, 3
2. Incorrect Block Number	2	6.5	0.7	2, 3
3. Different Address Range on Ground than in File	3	9.7	1.1	1
4. Incorrect or Illegal Place Code Number	3	9.7	1.1	2, 3
5. Lower and/or Higher Address in Segment Than File Allows	7	22.6	2.5	1
6. Incorrect Feature Name	2	6.5	0.7	1, 2
7. Two or More Segments on Ground; One in File	3	9.7	1.1	1
8. Odd-Even Anomaly	8	25.8	2.9	1, 2, 3
9. Improper Placement of Corporate Boundary	1	3.2	0.4	1, 2, 3
Total Non-Critical Errors	21	100.0	7.5	
1. Paper Streets (exist on Map and in File, but not on the Ground)	18	85.7	6.4	1
2. Two or More Segments in File, One on Ground	1	4.8	0.4	1
3. Zero Addresses in File, Numbered Addresses on Ground	1	4.8	0.4	1
4. Duplicate Node Numbers	1	4.8	0.4	1, 2

* (1) Source Materials or Improper Use of Them
(2) Clerical or Transcription Error
(3) Keying/Card Punching Error

Of the non-critical errors, 86 percent were those where the GBF/DIME file indicated that a segment existed, but the segment did not exist on the ground. These are referred to as paper streets and are usually a utility right-of-way, footpath, a street not yet built, or a yard (with no room to put a street). Other non-critical errors are detailed in Table A2. Non-critical errors would not cause an incoming address to be coded to the wrong geographic area.

Speculating on probable causes of the errors, faulty reference materials might first come to mind. However, the first part of this report indicates that the reference materials are of acceptable quality; therefore, wrong interpretation of good source materials is a probable source of error. Also, clerical and transcription errors may have added to the problem.

In each segment the number of individual units were counted and the addressable unit error rate for City A was calculated to be 5.3 percent. The 95 percent confidence limits of this estimate are 2.0 and 8.7 percent. These limits are based on cluster sampling techniques since the selected sampling unit was a segment and addressable units are clustered within the segments.3/ The percentage of addressable units in the sample which would be miscoded to the wrong geographic area was 3.1 percent with 95 percent confidence limits 0.6, 5.6. An additional 2.2 percent (95 percent confidence limits of 0.2, 4.3) of the addresses would not be coded by the file. In an actual Census operation, these would be identified, field-checked and a code would be assigned manually.

The addressable unit error rates by map sheet range from zero (95 percent confidence limits 0.0, 60.2) to 23.5 percent (95 percent confidence limits 0.0, 55.8) -- the confidence intervals are too wide to distinguish among the map sheet error rates. An error does not necessarily cause all addressable units within a segment to be miscoded or no-coded. Some of the addressable units may still be coded to the correct geographic area, even though an error has been made in the segment.

The intraclass correlation coefficient for clustered addressable units within segments was calculated to be 0.69. This is a strong indication that errors are clustered within segments; i.e., if an error affects one addressable unit within a segment, it is likely to affect more of the addressable units within that segment.

The intraclass correlation was calculated (from Hansen, Hurwitz, and Madow4/) using the formula for an estimate of the intraclass correlation coefficient from a sample, made in terms of the variances between and within ultimate clusters. Segments which contained no addressable units were eliminated from the calculations; also, those segments with only one addressable unit were eliminated from this calculation since by definition these cases do not apply.

The City A file was the CREATE file and had not undergone the CUE program. Several cycles through correction and update would be expected to improve the quality of the existing file.

City B Evaluation

Before a QC plan could be devised a decision was required as to what acceptable quality was in terms of a local GBF/DIME file. Cities that used such a file locally would probably need less stringent verification than those that do not use

3/ Cochran, William G., "Sampling Techniques," John Wiley & Sons, Inc., Formula (3.26), p. 65.

4/ Hansen, Morris H., William N. Hurwitz, and William G. Madow, Sample Survey Methods and Theory, Vol. I: Methods and Applications, p. 266.

their files, simply because they would be more likely to discover and correct any errors themselves. City B was selected for evaluation solely because it was further into the CUE program than most other places. City B should not be considered as representative of all cities in the program, but merely a city that was selected as a means of obtaining some insight into the type of problems a city faces in keeping a GBF/DIME file updated and to evaluate the findings with the objective of obtaining a realistic picture of the quality control needs of a CUE program.

In November 1974, a group of 12 field representatives were sent into City B to obtain geographic information for a sample of about 2,000 street segments that were contained in the GBF/DIME System file for City B. These sample segments were each checked on the ground by field personnel and selected address information was obtained independent of the files. These field data were then compared to the current City B GBF/DIME System file as it existed at the time of the study. Subsequently, a field reconciliation was made for all segments where a difference existed between the GBF/DIME System files and the ground.

The evaluation study entailed the matching of two Geographic Base (DIME) System files -- (1) the one prior to the initiation of the CUE program in City B, and (2) the current GBF/DIME System file updated to January 1, 1974. The match was based on a unique six-digit serial number for each segment record. The matched records were then compared for exactness.

The current GBF/DIME System file contained approximately 32,000 street and 6,000[5]/ non-street segments. Each street and non-street record was sorted into one of the following classes, based on their status in the two files cited above: (1) Changes, (2) No Changes, (3) Additions, (4) Deletions. The first two classes were comprised of records which matched on serial number and contained discrepant geographic information (1) or identical geographic information (2). Unmatched records in the January 1974 file comprised class (3), while class (4) consisted of unmatched records in the original file. The first output was a count of the number of records for each sorted class with a subsequent systematic random sample of segments selected from each of the four strata.

A systematic sample of 1,900 segments was selected in order to analyze the data with an adequate degree of confidence. A variable sampling rate was used for each of the four classes to assure adequate representation in each class. The first three classes were selected from the January 1974 file and the fourth class, Deletions, was selected from the first file prior to the initiation of the CUE program.

The total sample was comprised of 1,974 coded segments spread over the four classes as follows:

Class	# of Cases	Sampling Fraction
Changes (same serial number, but some difference in record)	849	1/27
No Changes (no difference in the two files)	357	1/27
Additions (not in the file prior to CUE) and	572	1/10
Deletions (not in the file following Update)	196	1/4
	1,974	

The following table shows the distributions of the updated City B file over the above four classes (first column) and the segment error rate for each portion of the file over the same four classes (second column):

Class	Distribution of Updated City B File (%)	Segment Error Rate (%)
Changes	58.2	20.1
No Changes	25.0	14.0
Additions	14.8	18.9
Deletions (from pre-CUE file)	2.0	7.1

The weighted overall segment error rate for City B was derived by weighting each of the above four classes by the inverse of their sampling fraction. These calculations resulted in a weighted segment error rate for City B of 18.1 percent (see attached Table B1). The 2-sigma limits on this estimate are 16.3 and 20.1 percent.[6]/

One portion of the CUE process is the application of computer edits for checking parity of address ranges, bounding a block, and address-range completeness. An evaluation of the effectiveness of these edits showed that they had caused a reduction in the overall segment error rate from 21 percent to the 18.1 percent level.

More than three-fourths of the file errors were classified as "critical" errors. The remaining errors were classified as "non-critical" errors (See Table B1 attached.) The weighted "critical" error rate is 15.1 percent. The weighted "non-critical" error rate is 3.0 percent. A "critical" error is defined as one that would cause an incoming address to be coded to the wrong geographic area, i.e., block, tract, place or MCD. A "non-critical" error is defined as one that will not

5/ All error rates shown in this report are calculated with a base of 39,288 street and non-street segments. Non-street segments including railroad tracks, etc., were not included in the sample as they would not contribute to the error rate. However, corporate boundaries were field-verified and would contribute to the error rate. Overall, the error rates shown may be underestimates of error rates which are based on a check of all segments.

6/ Cochran, William G., "Sampling Techniques," John Wiley & Sons, Inc., Formula (5.43), p. 106.

cause miscoding of addresses to a geographic area. The major category accounting for about 50 percent of the "critical" errors was defined as a smaller and/or larger address found existing on the ground than allowed for in the files. The major category of "non-critical" errors was defined as a segment with no addresses in the file and numbered addresses on the ground, accounting for about one-third of the total "non-critical" errors. This type of error was classified as non-critical since it would not cause an incoming address to be coded to the wrong geographic area. This error type would initially result in no coding. However, in a census, the item would be identified by the computer as a problem, followed up in the field, and a code eventually assigned.

The range of error rates by map sheet for City B areas is a low of 9.7 percent in the northeast to a high of 30.2 percent in the southwest. The error rates for the central part of City B (Maps 7, 7NE, 7NW) are somewhat lower than the areas surrounding the central city. The highest error rates appear in the west and southwest sections of the city (Maps 5, 6, 10 and 11). This distribution pattern of error rates was expected since the central city areas are usually more stable and more systematically arranged than the suburban areas.

The corrective process for CUE has an error associated with it. The "Change" category makes up the largest part of the City B file (about 3/5 of the file) and had the largest error rate (both number of errors and percent of total errors) among the four strata, Changes, No Changes, Additions and Deletions. Because the "Change" category is a significant part of the total City B file, further analysis was done to determine the extent to which a change occurred and either corrected the error or did not correct the error.

Table B3 below shows the source of errors (combined critical and non-critical) for the "change" category as an aid to determining whether an existing error was corrected (77.0 percent of 849) or was left uncorrected by the change action (11.3 percent of 849), whether the change replaced one error with another (9.3 percent of 849) or whether the change caused an error to a previously correct record (2.4 percent of 849).

In column 4, a count is given by map of those segments for which a change was made to a portion of the segment but did not affect the error; that is, the error was present in the original file (dated 1971) and remains after the change (49.2 percent of the 195 errors).

In column 5, a count is given of those segments for which changes were made to the error portion of the record. However, an error remains following the change (40.5 percent of 195 errors).

Column 6 gives the count of those segments for which the change created an error. The portion of the segment which is now in error was correct in the original file (10.3 percent of 195 errors).

TABLE B3

Status of Error Changes Made to the "Changes"

Map No. (1)	Total Change* (2)	Change Corrected Error (3)	Error Exists After Change: Original Error Not Corrected by Change (4)	Error Exists After Change: Change Replaced One Error with Another (5)	Error Exists After Change: Change Introduced Error (6)
Total	849	654	96	79	20
1	28	25	2	0	1
2	113	81	23	8	1
3	31	25	3	1	2
4	150	119	9	18	4
5	52	33	7	12	0
6	81	49	14	11	7
7	44	36	3	5	0
7NW	137	116	12	8	1
7NE	80	71	4	4	1
8	66	50	8	6	2
9	13	9	2	2	0
10	31	24	4	2	1
11	23	16	5	2	0

*There was a total of 849 segment records in the "Change" category, of which 195 were in error prior to the review of Topological and Address Range Edits.

An attempt was made to estimate the address error rate for the City B GBF/DIME System file. All segments that contained "critical" errors were reviewed and an estimate of the number of error addresses was derived. For those segments where a larger and/or smaller address existed on the ground in the sample segment than allowed for by the file, one address was assumed as the minimum number of address errors and a maximum was derived by taking the difference between the actual and the allowable (file) addresses and dividing by two. (Example -- if the first address of one sample segment on the ground was 93, and the low address in the file for the segment was 99, the difference of six was divided by two and the resulting three addresses were considered to be in error -- 93, 95, and 97).

When an entire segment would have been miscoded -- a wrong block number or wrong MCD/Place was in the file -- eight addresses were assumed to be in error (eight addresses as the average per segment was derived from the estimated total number of addresses in City B divided by the number of street segments in the City B file). Using these methods, the range for the address error rate for the City B GBF/DIME System is estimated to be 8 percent to 21 percent.

The GBF/DIME System file for the City B SMSA had an overall segment error rate of 18 percent at the completion of the first update cycle. Further

updates, using reference material of current quality and with continued acceptable quality on manual coding operations, should result in a segment critical error rate of about 10 percent. This estimate is derived, based on the assumption that the computer edits would effectively locate and identify errors for clerical correction. Reference material quality and the quality of the resultant GBF/DIME System is not inconsistent with quality levels achieved in the 1970 Census of Population for similar processes.

On-site investigation and review of sample information indicate that manual coding operations are of acceptable quality and that the designed edit operations of the CUE program are well-conceived and adequate for purposes of detecting file inconsistencies. An area which could be improved through careful quality control processes is the reference or source material used in the CUE operations. Overall, based on a small sample, the reference material for the City B area is estimated to have an 8.5 percent segment error rate, and it is felt that a feasible goal would be to achieve an average outgoing quality limit of 5 percent for references and sources. A quality control plan for this purpose, and incorporating further checks on manual processes, is now under development.

CONCLUSIONS

1. The reference material used in the creation of the City A file is of acceptable quality based on the quality control evaluation and verified by this study using a similar set of reference materials updated to 1975.

2. The 1-in-40 block quality control sample is sufficient for an overall estimate of the quality of reference material for a large area. However, it is too small a sample and spread too thinly to provide error rates on a map sheet or smaller area basis.

3. Based on the estimated 18.1 percent segment error rate in City B, a quality control program is needed for the GBF/DIME System CUE operation.

RECOMMENDATIONS

Based on the findings of the evaluation studies, the following recommendations were made:

1. While the 1-in-40 block sample is adequate for overall error rate estimates, it is recommended that a larger sample, such as that used in the procedures recommended for CUE, be selected to provide detailed error rates by map sheet. Based on the City B experience, the variability of error rates between map sheets may be great. Thus, an area might be acceptable from an overall determination of the error rate, but one or more map sheets or small areas might not be acceptable when individual map sheet or small area error rates are known.

2. When possible, use the same type of quality control plans for CREATE as for CUE.

3. More attention must be given to the clerical and keying operations, since even with good reference materials, a high ultimate error rate can occur in GBF/DIME. It is suspected that the 1-in-40 block sample is too small to quality check on individual coders -- similar to "by map" problem.

As a result of the recommendations based on these evaluation studies, a set of quality control plans for the creation, correction, update and extension of the GBF/DIME System file has been proposed and is currently being tested in several areas. Results of these tests will be documented, and based on these results, quality control plans will be designed for use in all areas in the CUE program.

REFERENCES

"Geographic Base Files - Plans, Progress, and Prospects," Conference Proceedings from Jacksonville, Florida, April 1-2, 1971.

"CUE - Correction-Update-Extension," Procedural Manual for the Review and Correction of the Address Range Edit (ADDEDIT) Listing, January 1973.

"CREATE Coder's Manual," CUE, Correction-Update-Extension, January 1, 1974, Revised July 1, 1974.

CUE Evaluation Study--Field Procedures, October 1974.

"Final Quality Control Specifications for CUE Operation" from C. Jones, SMD to J. Silver, Geography Division, February 23, 1976.

"An Outline on the Geographic Base File and CUE Maintenance Program to the Ohio Department of Transportation," by Kenneth Graska, May 31, 1975.

Table B1

City B CUE Evaluation

Weighted Breakdown of Source of Errors by Type of Error

Error Type	Estimated Number of Segments with Errors	Percent of Total Errors	Percent of Total File
(1)	(2)	(3)	(4)
All Errors	7103	87.3	18.1
Non-Critical Errors	1188	14.6	3.0
Critical Errors	5915	72.7	15.1
Source Materials	3978	48.8	10.2
1. Smaller and/or larger addresses in segment than file allows	3658	44.9	9.3
2. Wrong Address Range	101	1.2	0.3
3. 2 segments on ground, 1 segment in file	219	2.7	0.6
Clerical			
(Transcription, Keying, Card Punching)	1475	18.1	3.8
1. Wrong block No.	1058	13.0	2.7
2. Wrong MCD/Place	336	4.1	0.9
3. Wrong Tract No.	81	1.0	0.2
Combination of Source Materials and Clerical/Computer			
1. Odd-even sides reversed	462	5.7	1.2