PRACTICAL PAPER

Quality Control Techniques for a GIS Database Development Project

Jeffrey L. Nugent

Abstract

Basic quality control techniques were developed for the Northern Forest Lands Inventory, a GIS-based data development project. These techniques are both preventative and reactionary. Typical administrative tasks such as procedure sheets and progress tracking serve as preventative quality control tools. Personnel management is also an important preventative factor in a quality control program. The comparison of proof plots against the original source data is perhaps the most effective method for identifying existing database errors. Developing and implementing a quality control program can be a time consuming effort, but many quality control techniques are very effective and require little time to implement. Several such informal techniques are explored.

Introduction

Background

The Northern Forest Lands Inventory (NFLI) is a multi-state, multi-agency effort to develop a geographic information system (GIS) based inventory of natural and economic data in the 26 million acre northern forest of New York, Vermont, New Hampshire, and Maine. One of two data development centers in New York State is the State University of New York (SUNY) College of Environmental Science and Forestry (ESF) in Syracuse, New York. One full time database administrator and several part time graduate student assistants form the data acquisition team. Though basic geographic data analyses are performed on occasion, the primary effort of the NFLI at SUNY-ESF is data development.

Because the NFLI was SUNY-ESF's first major data development project, few previous procedural protocols were available. As a result, most automation procedures, including quality control techniques, had to be developed in-house and implemented without guidance from previous work. To keep within the limitations of a restricted budget, small personnel pool, and completion deadlines, quality control techniques had to be easy to implement and not excessively time consuming.

Defining Quality Control

Quality control, as practiced in the Northern Forest Lands Inventory, comprises a two-stage process: (1) implementing techniques and procedures that attempt to reduce errors and eliminate mistakes; and (2) reviewing all completed work to identify and correct errors before any product is released. (A good overview of the types of error in GIS databases can be found in Thapa and Bossler (1992).) Quality control techniques are implemented throughout the GIS database development, or automation,[1] process. This may contradict peoples' assumption that quality control is merely checking the final product for errors. Quality control in the Northern Forest Lands Inventory (NFLI) attempts to both identify existing errors in the data and prevent errors from occurring. Consequently, the quality control approach is both pro-active and re-active.

The Importance of Quality Control

Developing a quality control program is important for all GIS projects. It is vital, however, in large database production projects for numerous reasons. Large projects naturally deal with large amounts of data, and managing these data becomes a complex problem as more and more are created. Efforts must be taken to insure that all automation steps are executed and that each database file is complete and correct.

GIS production projects employ many people. Most likely, the experience and abilities of these people will vary. Personnel must be trained so that they are well versed in automation procedures. Such familiarity will help reduce the potential for error and ensure that existing errors are recognized. All personnel's work must be checked, especially that created by those recently employed, in order to evaluate their digitizing and automation abilities.

In small GIS projects, the database developers may also be the database users. This use provides an informal, yet effective, check for errors in the database. In contrast, developers of large GIS databases generally do not use the data in any applications, and, consequently, this form of error checking by use is omitted. Other efforts to check the database for errors must be employed.

Components of the NFLI's Quality Control Efforts

General Techniques

A number of quality control tools are employed continuously throughout the automation process. These include the use of procedure sheets and macros, and extensive record keeping procedures. While not typically thought of as such, these are quality control techniques because they help reduce the potential for error.

[1]The term automation is used here to mean the entire process of creating a GIS database, from digitizing through edgematching and final review.

Photogrammetric Engineering & Remote Sensing,
Vol. 61, No. 5, May 1995, pp. 523–527.

0099-1112/95/6105-523$3.00/0

College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210.

```
Creating topology
     When all digitizing for a town has been completed, type, in arc
BUILD PARCEL

Topological errors
     Correct any errors that result from intersecting arcs flagged during
BUILD.  Draw the coverage on the screen, turning dangle nodes on.  Locate
and correct any dangling nodes.

Labeling errors
     To check for label errors, type
LABELERRORS PARCEL
The only label error that should occur is that polygon 1 (the background
polygon) has zero label points.  If polygon 1 is not listed at all, there is an
open polygon or misplaced label point.  List the polygon attribute table in
TABLES, and record the label point identification number for polygon 1.
Locate this label point in arcedit, after calling up the coverage and drawing
the arcs and label points.
EF LABEL
SEL $ID =  <id of label point>
SETDRAWSYMBOL 2
DRAWS
The above commands will highlight the label point in red.
     If the LABELERRORS command lists errors in any other polygons,
type
ARCPLOT
DISP 4
MAPE PARCEL
MARKERCOLOR 2
LABELERRORS PARCEL
     Polygons which have no label points are drawn with a star in their
center (since the background polygon has no label, all arcs which border
the background polygon will be displayed, and a star symbol will be drawn
in the center of the coverage).  All polygons with more than one label point
will be drawn with a cross in their center.  To zoom in, type
MAPE *
CLEAR
LABELERRORS PARCEL
Correct any errors that are found.

Adding attributes
     Copy the export file containing the ARLM data into the appropriate
subdirectory.  To import it and add additional items, type, in arc
IMPORT INFO <filename> PRCLINFO
@NEWITEMS
```

Figure 1. Sample procedure sheet for automating tax maps.

Procedure Sheets

Procedure sheets are a very detailed set of instructions to be followed when creating a GIS file, or coverage.[2] In many cases, the instructions include the actual software commands to be used. Use of a procedure sheet insures that the exact same procedure will be followed by all personnel. This consistency is important when an automation step can be performed using two or more techniques or commands which yield slightly different results. (The CLEAN and BUILD commands in Arc/Info are an example of this; both can be used to create topology, but only one can be relied upon not to move arcs or nodes.) Additionally, if personnel follow the procedure sheet in a sequential manner, there is less chance an automation step will be omitted. A portion of a tax map procedure sheet is shown in Figure 1.

For data themes[3] where many coverages will be created, procedure sheets make good managerial sense. In many cases, automation personnel can merely consult the procedure sheet instead of the database administrator when confusion develops. Procedure sheets do have a disadvantage in that they can lure the automation personnel into a false sense of security. The belief that strict adherence to the procedure sheet will result in a problem-free automation sequence, and, therefore, an error-free GIS coverage, can be detrimental. Even though a standardized automation procedure has been developed, problematic situations can still occur, and automation personnel should be attentive to any unusual or non-typical occurrences in the data. To improve their knowledge of the procedure, automation personnel should be individually tutored and supervised at the onset of a data layer automation project. Further, it should be made clear to them that using a procedure sheet *is no substitute* for a thorough understanding of the automation procedure.

Record Keeping

Detailed record keeping throughout the automation process is important. The Northern Forest Lands Inventory at SUNY-ESF uses two forms of record keeping: log sheets and log books. Log sheets are associated with each individual coverage. The sheets allow automation personnel to record the progress and development of each coverage along with any pertinent information concerning each automation step. Each entry in the log sheet includes the date a specific task was performed and who performed it (Figure 2). In most cases, the entries in the log sheet chronologically match the automation steps described in the data layer's procedure sheet.

Whereas log sheets are specific to each coverage, log books are specific to each worker. All automation personnel have a log book in which they record, on a day to day basis, the coverages they worked on, the tasks they performed, and notes or problems associated with their work. Log books are the property of the project and remain in the work place at

digitizing, con't

date	file name and task	RMS	backup	initial	location
______	______________________	____	____	____	____
______	______________________	____	____	____	____
______	______________________	____	____	____	____
______	______________________	____	____	____	____
______	______________________	____	____	____	____
______	______________________	____	____	____	____
______	______________________	____	____	____	____
______	verification______________	____	____	____	____

C. CREATING TOPOLOGY

date		backup	initial	location
______	BUILD______________	____	____	____

D. ATTRIBUTES

completed ______________

date	file name and task	backup	initial	location
______	______________________	____	____	____
______	______________________	____	____	____
______	verification______________	____	____	____

E. ERROR CHECKING AND CORRECTING

completed ______________

date	file name and task	initial
______	______________________	____
______	______________________	____
______	______________________	____
______	______________________	____
______	______________________	____
______	______________________	____
______	______________________	____

Figure 2. Portion of an automation log sheet for regulatory wetlands maps.

[2]The term coverage is used to denote a GIS data file that contains both positional and attribute information. It is, more specifically, a term associated with such files produced using the Arc/Info GIS software package.

[3]A data theme, or data layer, is a set of data representing a singular feature type, such as roads, hydrography, or wetlands.

all times. Notes, clarifications, or additional information on certain procedures can also be recorded in the log book for future reference by the worker.

The coverage-based log sheets and the personnel-based log books together serve an important role in the project. Should any problems be identified at any time during or after the automation process, the log sheets can be consulted to identify the person or persons working on the coverage. The problem can then be discussed with that person, or the log book can be checked for a possible explanation of, or insight into, the situation.

Macros

A macro functions similarly to a computer program, but utilizes GIS software commands. Macros are used in a variety of situations, from establishing tolerances and other specifications in a digitizing session to facilitating attribute coding of a coverage. Macros are an effective quality control tool, and automation personnel are motivated to use them because they reduce typing; you can execute numerous software commands by simply typing the macro name. When numerous commands must be issued in sequence, using a macro can insure that all commands will be executed (provided, of course, that the macro itself is executed).

Because many GIS software commands require a complicated set of parameters—some Arc/Info commands may have over one hundred characters in the command line—there is often the potential for typographical or omission errors. A command may still execute despite these errors, and the implications may go unnoticed for some time. Placing complicated command parameters correctly into a macro can help reduce the potential for such errors.

The Digitizing Process

Equipment considerations

In order to minimize (actually, to make consistent) the effects of any equipment errors in the digitizing process, a standardized map set-up procedure is followed each time a map is placed on the digitizing board. Maps are always oriented in the same direction on the digitizing board (for quadrangle sheets, north to the top). During tic registration and digitizing, the digitizing puck, or cursor, is held in the same general orientation. We have found that changing the orientation of the puck while holding it over the same point on a map can produce variable point coordinates. Holding the puck in the same orientation will reduce such variations.

Errors caused by the digitizing board itself have been documented (Giovachino, 1992), sometimes occurring only at isolated locations on the board. Where these errors are significant, i.e., detectable, that portion of the board should be avoided during digitizing, and maps should always be placed in the same general error-free location on the digitizing board. In most cases, however, only very minor, virtually undetectable, errors are present in a digitizing board. Regardless, utilizing the same location on the board will help hold these minor errors constant.

Personnel Considerations

Whenever possible, one person works on any one coverage throughout its automation process. As a result, that person becomes familiar with that coverage and any specific problems associated with it. More importantly, that person is aware of what has and has not been done in terms of the automation procedure. Knowledge of the automation status of a coverage helps reduce the chance of omitting steps in the automation process. In addition, having one person follow a coverage through to completion (i.e., having that person responsible for that coverage) improves the likelihood that any problems will be solved before the coverage is deemed final.

A number of actions can be taken to improve personnel performance. All automation personnel should be familiar with the entire automation procedure for each data theme that they are involved with, even if they are responsible for only a small portion of the procedure. By possessing this knowledge, they will be able to both identify potential problems or errors, and correct them.

The database administrator should thoroughly review each employee's work with him or her soon after employment. The accuracy and quality of their digitizing should be evaluated using both edit plots and large-scale on-screen reviews (edit plots are explained later). Such supervisory feedback is vital to the first-time digitizer. New employees must be able to relate the effort put forth during digitizing to the quality of their product. Moreover, many new employees are unaware of the high level of accuracy required during database automation, as evidenced by the fact that nearly every employee's first digitized product is rejected during the edit plot checking stage because of an excessive number of errors.

Digitizing is a monotonous, fatiguing, and oftentimes boring task. However, it is a task that requires skill, accuracy, and alertness. For this reason, it is important to encourage digitizing personnel to take frequent breaks. Though digitizing utilizes computers and peripherals, it is indeed a physical exercise, and a demanding one at that. The recent awareness of carpal tunnel syndrome problems attest to this. In any event, no person should digitize for more than four hours without a major break from his/her work (Burrough, 1986).

Preliminary Checks Using the Software

When all automation steps for a GIS coverage are complete, a number of preliminary quality control checks are performed by the digitizing personnel using the GIS software. These checks have the advantage of identifying errors before the coverage undergoes final quality checks. Preliminary checks will find errors while the file is still in the automation environment (i.e., at the hands of the digitizing personnel and not in hardcopy format), making error correction quicker, easier, and more likely to take place. Moreover, digitizing personnel are afforded the opportunity to examine the results of their work first-hand. Such self-review can be more effective at insuring a quality automation process than a data administrator's critique. And involving automation personnel in a portion of the quality control program may help foster an added sense of commitment to producing an error free product.

One preliminary check performed by the digitizing personnel is for errors of commission and omission. Errors of commission involve a feature being digitized more than once, while errors of omission involve features that were overlooked during digitizing (commission error may also be attributed to a sliver polygon). To check for such errors, the GIS software is used to generate a report on the number of map features (in most cases, points or polygons) in a coverage. The digitizing personnel tally the number of features on the corresponding map or maps, and compares the map tally with the software report. If the number of features in the coverage exceeds those on the map, commission errors have

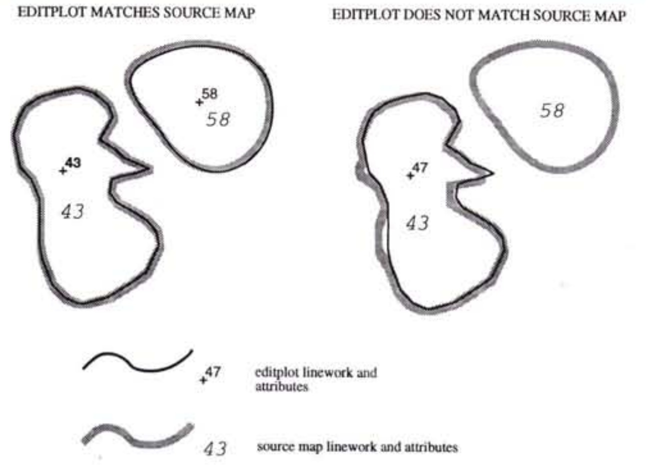

Figure 3. Hypothetical comparison of an edit plot to the source map. The diagram on the left shows digitized linework within an acceptable accuracy tolerance, attributes for polygons coded correctly, and no omission errors. The diagram on the right shows inaccurate digitizing and attribute coding, as well as an omission error.

occurred; if the number of features in the coverage is less than the number on the map, map features were omitted during digitizing. In a similar sense, attribute files can be reviewed and the number of records in the files compared with the number of features that should have been coded. Attribute coding omissions can also be identified by empty or zero-value fields.

Such omission/commission checks are not exhaustive. They would not, for instance, be effective when errors of commission balance errors of omission. These checks do, however, serve as an effective preliminary check and have the benefit of being easy and quick to perform.

Another way to identify digitizing errors (in vector format) is to check the coverage for logical consistency (NCDCDS, 1988), also termed topological integrity. Logical consistency indicates features are defined with a correct data structure. In other words, all lines that are supposed to intersect do indeed intersect, all polygons close, and, if attributes are to be included, all features are labeled. These topological checks are easy to perform. The NFLI's GIS software (pcArc/Info) flags non-connected lines or open polygons on-screen with a red box. Such an error would indicate either sloppy digitizing or omitted lines. Unlabeled polygon errors can be identified by comparing the number of attribute labels with the number of digitized polygons; the values should be the same.

Edit Plots

The final step in the quality control program is checking an edit plot. An edit plot is a hardcopy output (usually produced by a pen plotter) of the coverage. The edit plot is at the same scale as the source map, and contains all features (points, lines, or polygons) and their attribute codes. The source map is placed on a light table, and the edit plot is laid over the source map (Campbell and Morteson, 1989). Every feature on the source map is checked against every feature on the edit plot for three conditions (Figure 3):

- no errors of commission or omission—all features that should have been digitized were indeed digitized (and digitized only once);
- no positional errors beyond a specified tolerance—the edit plot linework overlies the source map linework; and
- no attribute coding errors—the attributes listed on the edit plot match those on the source map.

For most GIS applications, including the NFLI, errors of commission, omission, or attribute coding are not acceptable. In contrast, it is impossible, or nearly so, to achieve or measure the exact positional accuracy of hand digitized features. Typically, some sort of tolerance standard is set. Such a standard specifies the acceptable distance the digitized feature may deviate from its location on the source map. In many cases, this distance is the width of one pen line on the edit plot.

The edit plot checking procedure is essentially a 100 percent sample of the data. It is very costly (in terms of time). However, it is an extremely effective quality control technique and perhaps the most effective method available to identify errors of omission and attribute coding or assess the positional accuracy of digitized features. There is no substitute for viewing the coverage in hardcopy format; examining the file on-screen is simply not as effective.

Conclusions

While a quality control program is a necessary part of any GIS effort, implementing such a program can be expensive and time-consuming. There are, however, a number of quality control procedures, including some described above, that are easy to implement and quick to execute. With this in mind, it is fair to say that no GIS effort should be without some form of quality control.

As stated previously, checking for commission and omission errors is an easy process, and, though not exhaustive, is quite effective. Due to its simplicity, it is a good check of completeness for a coverage, and a check that identifies errors early in the automation process.

Should the 100 percent sample during an edit plot check be infeasible, spot checking certain features is a good substitute. It is recommended that efforts be concentrated on those features more prone to errors. For example, in a roads database, digitizing very curvy or sinuous roads entails more effort and concentration, and therefore a greater opportunity for error, than digitizing straight roads. If time is a concern when checking this coverage for positional accuracy, effort would be better spent checking the curvy roads instead of the straighter roads. When digitizing a map or coding a coverage for attributes, small or isolated features are more likely to be overlooked than larger ones. It is these features that should receive priority when checking the coverage for omission errors.

One technique that can be used to check for feature omission and attribute coding errors involves comparison of spatially adjacent GIS coverages. Such a check can be performed on-screen or by using hardcopy plots. When examining the common border between two adjacent coverages, each feature on one side of the border should have a corresponding feature on the other side (Figure 4). The attribute coding of these corresponding features should also be the same. Be aware that original source maps often contain omission and coding errors along their borders. If the maps are

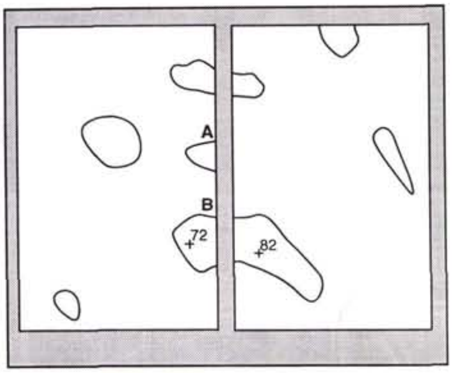

Figure 4. Errors on two spatially adjacent coverages are obvious when examining the coverages side by side. An omission error is present on the right-hand coverage at "**A**", while an attribute coding error exists at "**B**".

digitized faithfully, these inconsistencies will appear in the coverages as well. However, they do not represent automation errors. Regardless, this technique is a useful one, and is especially appropriate when performed prior to edgematching.

Finally, there is an unconventional and frequently overlooked technique (perhaps better described as a skill) useful for identifying significant errors in a GIS coverage. Simply put, it is helpful to have a first-hand knowledge of the study area. Such knowledge allows one to develop a cognitive map of the area which can be compared with the database. Situations such as missing roads or lakes become obvious as omission errors, or at least features in need of verification against the source map. In a similar sense, automation personnel must be encouraged to be fully aware of the data with which they are working, and be inquisitive about situations that simply "don't make sense." Peculiarities such as an interstate highway coded as a gravel road or a wetland located on the side of a ridge should invite attention. Sometimes these situations are real life anomalies, but more often than not, they are the results of automation errors.

Acknowledgments

I wish to acknowledge Lee P. Herrington for his comments on this paper, and the New York State Department of Environmental Conservation's Habitat Inventory Unit in Latham, New York, whose automation techniques provided the stimulus for several of the procedures outlined in this paper. The Northern Forest Lands Inventory is funded by Congress as part of the Northern Forest Lands Council's efforts.

References

Burrough, P.A., 1986. *Principles of Geographical Information Systems for Land Resources Assessment*, Clarendon Press, Oxford, 194 p.

Campbell, W.G., and D.C. Morteson, 1989. Ensuring the Quality of Geographic Information System Data: A Practical Application of Quality Control, *Photogrammetric Engineering & Remote Sensing*, 55:1613-1618.

Giovachino, D., 1992. Graphic Digitizer Accuracy, *Technical Papers of the ASPRS/ACSM/RT 92 Convention*, 3:380-389.

NCDCDS, 1988. *The American Cartographer*, 15(1):11-140.

Thapa, K., and J. Bossler, 1992. Accuracy of Spatial Data Used in Geographic Information Systems, *Photogrammetric Engineering & Remote Sensing*, 58:835-841.

(Received 10 May 1993; accepted 9 September 1993)

Jeffrey L. Nugent

Jeffrey L. Nugent received his MS in Environmental and Resource Engineering from the State University of New York College of Environmental Science and Forestry (SUNY-ESF) in Syracuse. He is the technical director of the Northern Forest Lands Inventory project at SUNY-ESF, and is actively involved in GIS data development and exchange in Northern New York State. His interests include cartography and development of data documentation standards.