UNDERSTANDING ERRORS AND THEIR MEASUREMENT IN GEOINFORMATION

[1]Ndehedehe, C., [2]Ekpa, A., [3]Okwuashi, O. and [4]Ogunlade, S.
[1, 2 & 3]Department of Geoinformatics & Surveying,
University of Uyo, Uyo, Nigeria
[4]Department of Surveying and Geoinformatics,
Federal University of Technology, Akure, Nigeria

ABSTRACT

Managing error in GIS datasets is now recognized as a substantial problem that needs to be addressed in the design and use of such systems. Failure to control and manage error can limit severely or invalidate the results of a GIS analysis. The various sources of errors that may affect the quality of a GIS dataset have been highlighted in this paper. One major approach in managing error in a GIS datasets is documenting procedures, products and producing data quality reports. Another is setting of standards and procedures for product.

Key Words: Error, Geoinformation, Spatial data, GIS, Accuracy, Precision

INTRODUCTION

Until recently, people involved in developing and using geoinformation paid little attention to the problems caused by error, inaccuracy, and imprecision in spatial datasets. Certainly there was an awareness that all data suffers from inaccuracy and imprecision, but the effects on geoinformation problems and solutions was not considered in great detail. This situation has changed substantially in recent years. It is now generally recognized that error, inaccuracy, and imprecision can "make or break" many types of geoinformation project. That is, errors left unchecked can make the results of a geographic information systems (GIS) analysis almost worthless. Managing error in GIS datasets is now recognized as a substantial problem that needs to be addressed in the design and use of such systems. Failure to control and manage error can limit severely or invalidate the results of a GIS analysis. The detection and removal of the errors and inconsistency of spatial data in vectors are the main concerns of the Geographic Information Systems (Maras et al 2010). The key point is that even though error can disrupt GIS analyses, there are ways to keep error to a minimum through careful planning and methods for estimating its effects on GIS solutions. Awareness of the problem of error has also had the useful benefit of making IS practitioners more sensitive to potential limitations of GIS to reach impossibly accurate and precise solutions (Foote et al., 1995).

Data quality can be accessed through data accuracy (or error), precision, uncertainty, compatibility, consistency, completeness, accessibility, and timeliness as recorded in the lineage data (Chen and Gong, 1998). Spatial error refers to the difference between the true value and the recorded value of non-spatial and non-temporal data in a database. Error in geoinformation is quite elaborate. This work looks into the general concept of errors, how they can be measured and managed.

CONCEPT OF ERROR PROPAGATION IN GEOINFORMATION DATA

Error propagation is a fundamental issue related to both uncertainty modeling and spatial data quality. According to (Kemp, 2008) *Error propagation* is defined as a process in which error is propagated from the original data set to a resulting data set that has been generated by a spatial operation. The concept of error propagation is illustrated in Figure 1. The data in the original data set(s) or the data set(s) generated through the spatial operation can be spatial data (e.g., the lines representing the road networks), non-spatial data (e.g., the size of a building block), or topological relations (e.g., a building is on the south side of a road). The spatial operation can be, for example, overlay, buffer, line simplification, generating a digital elevation model through a spatial interpolation, or an environmental modeling operation. Within geographic information science, errors can be classified as positional error, attribute error, topological inconsistency error, error on completeness (e.g., omission error or commission error), and temporal error. In the real world, one geographic data set can, and often does, possess more than one type of error simultaneously.

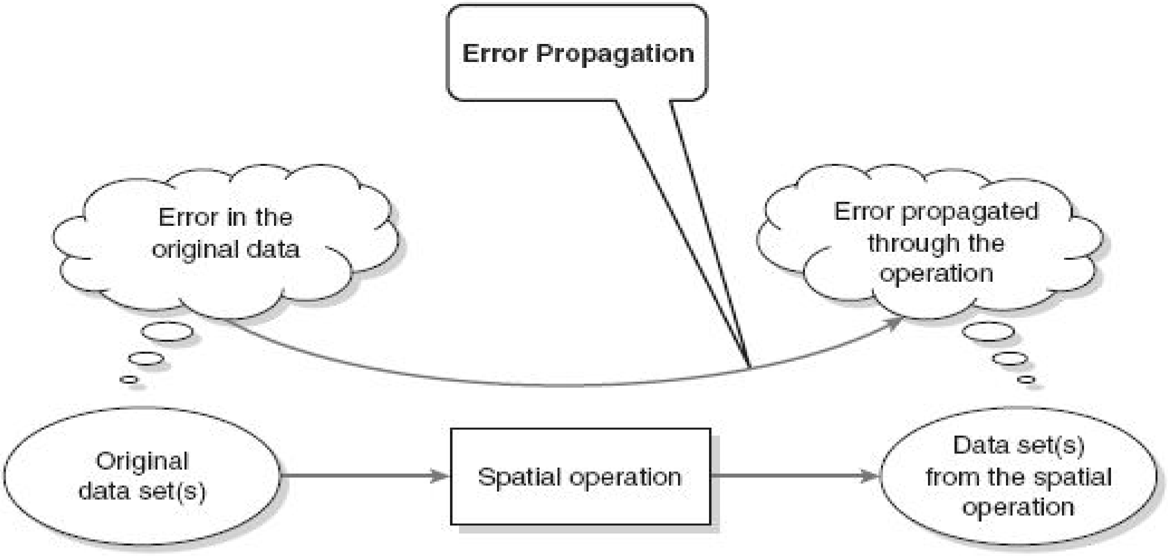


Figure 1 The Concept of Error Propagation

Adapted from Encyclopedia of Geographic Information Science 2008

TYPES OF ERRORS IN SPATIAL DATA

According to (Campbell, 2008) common types of errors in spatial analysis include: cartographic, statistical, thematic, conceptual and measurement. (Foote et al, 1995) listed four major types of errors - Positional error, attribute error, conceptual error and logical error. Positional error is often of great concern in GIS, but attribute error can actually affect many different characteristics of the information stored in a database. There are primarily four types of errors in a GIS database: positional, temporal, attribute, and logical. Logical error refers to the inconsistency of relationship among different features presented in a database. It is usually manifested through other types of errors. Thus, logical relationships of mapped features can be checked for error detection. Positional error has been widely investigated for its determination (Gong *et al.,* 1995; Stanislawski *et al.,* 1996; Kiiveri, 1997; Veregin, 2000), and modelling (Zheng and Gong, 1997; Shi and Liu, 2000). Essentially, positional error is the error contained in the coordinate values of points, lines and volumes. Thus, it is one type of numeric errors. We shall consider (Foote *et al.,* 1995) description briefly.

Positional Errors

This applies to both horizontal and vertical positions. Accuracy and precision are a function of the scale at which a map (paper or digital) was created. For instance The mapping standards employed by the United States Geological Survey specify that: "requirements for meeting horizontal accuracy as 90 per cent of all measurable points must be within 1/30th of an inch for maps at a scale of 1:20,000 or larger, and 1/50th of an inch for maps at scales smaller than 1:20,000" (Foote *et al.,* 1995).

Attribute Errors

The non-spatial data linked to location may also be inaccurate or imprecise. Inaccuracies may result from mistakes of many sorts. Non-spatial data can also vary greatly in precision. Precise attribute information describes phenomena in great detail. For example, a precise description of a person living at a particular address might include gender, age, income, occupation, level of education, and many other characteristics. An imprecise description might include just income, or just gender.

Conceptual Errors

GIS depend upon the abstraction and classification of real-world phenomena. The user determines what amount of information is used and how it is classified into appropriate categories. Sometimes users may use inappropriate categories or misclassify information. For example, classifying cities by voting behaviour would probably be an ineffective way to study fertility patterns. Failing to classify power lines by voltage would limit the effectiveness of a GIS designed to manage an electric utilities infrastructure. Even if the correct categories are employed, data may be misclassified. A study of drainage systems may involve classifying streams and rivers by "order," that is where a particular drainage channel fits within the overall tributary network. Individual channels may be misclassified if tributaries are miscounted. Yet some studies might not require such a precise categorization of stream order at all. All they may need is the location and names of all stream and rivers, regardless of order.

Logical Error

Information stored in a database can be employed illogically. For example, permission might be given to build a residential subdivision on a floodplain unless the user compares the proposed plat with floodplain maps. Then again, building may be possible on some portions of a floodplain but the user will not know unless variations in flood potential have also been recorded and are used in the comparison. The point is that information stored in a GIS database must be used and compared carefully if it is to yield useful results. GIS systems are typically unable to warn the user if inappropriate comparisons are being made or if data are being used incorrectly. Some rules for use can be incorporated in GIS designed as "expert systems," but developers still need to make sure that the rules employed match the characteristics of the real-world phenomena they are modelling.

Finally, it would be a mistake to believe that highly accurate and highly precision information is needed for every GIS application. The need for accuracy and precision will vary radically depending on the type of information coded and the level of measurement needed for a particular application. The user must determine what will work. Excessive accuracy and precision is not only costly but can cause considerable details.

COMMON SOURCES OF ERRORS IN GEOSPATIAL DATA

There are many sources of errors that may affect the quality of a GIS dataset. Some are quite obvious, but others can be difficult to discern. Few of these will be automatically identified by the GIS itself. For example, smooth changes in boundaries, contour lines, and the stepped changes of chloropleth maps are "elegant misrepresentations" of reality. In fact, these features are often "vague, gradual, or fuzzy" (Burrough, 1986). There is an inherent imprecision in cartography that begins with the projection process and its necessary distortion of some of the data, an imprecision that may continue throughout the GIS process. Recognition of error and importantly what level of error is tolerable and affordable must be acknowledged and accounted for by GIS users. Burrough (1986) divides sources of error into three main categories:

1. Obvious sources of error.
2. Errors resulting from natural variations or from original measurements.
3. Errors arising through processing.

Generally errors of the first two types are easier to detect than those of the third because errors arising through processing can be quite subtle and may be difficult to identify. Burrough (1986) further divided these main groups into several subcategories discussed below:

OBVIOUS SOURCES OF ERROR

Age of Data

Data sources may simply be too old to be useful or relevant to current GIS projects. Past collection standards may be unknown, non-existent, or not currently acceptable. Additionally, much of the information base may have subsequently changed through erosion, deposition, and other geomorphic processes. Despite the power of GIS, reliance on old data may unknowingly skew, bias, or negate results.

Areal Cover

Data on a given area may be completely lacking, or only partial levels of information may be available for use in a GIS project. For example, vegetation or soils maps may be incomplete at borders and transition zones and fail to accurately portray reality. Another example is the lack of remote sensing data in certain parts of the world due to almost continuous cloud cover. Uniform, accurate coverage may not be available and the user must decide what level of generalization is necessary, or whether further collection of data is required.

Map Scale

The ability to show detail in a map is determined by its scale. A map with a scale of 1:1000 can illustrate much finer points of data than a smaller scale map of 1:250000. Scale restricts type, quantity, and quality of data (Star and Estes 1990). One must match the appropriate scale to the level of detail required in the project. Enlarging a small scale map does not increase its level of accuracy or detail.

Density of Observations

The number of observations within an area is a guide to data reliability and should be known by the map user. An insufficient number of observations may not provide the level of resolution required to adequately perform spatial analysis and determine the patterns GIS projects seek to resolve or define. A case in point, if the contour line interval on a map is 40 feet, resolution below this level is not accurately possible. Lines on a map are a generalization based on the interval of recorded data, thus the closer the sampling interval, the more accurate the portrayed data.

Relevance

Quite often the desired data regarding a site or area may not exist and "surrogate " data may have to be used instead. A valid relationship must exist between the surrogate and the phenomenon it is used to study but, even then, error may creep in because the phenomenon is not being measured directly. An example of surrogate data are electronic signals from remote sensing that are used to estimate vegetation cover, soil types, erosion susceptibility, and many other characteristics. The data is being obtained by an indirect method. Sensors on the satellite do not "see" trees, but only certain digital signatures typical of trees and vegetation. Sometimes these signatures are recorded by satellites even when trees and vegetation are not present (false positives) or not recorded when trees and vegetation are present (false negatives). Due to cost of gathering on site information, surrogate data is often substituted and the user must understand variations may occur and although assumption may be valid, they may not necessarily be accurate.

Format

Methods of formatting digital information for transmission, storage, and processing may introduce error in the data. Conversion of scale, projection, changing from raster to vector format, and resolution size of pixels are examples of possible areas for format error. Multiple conversions from one format to another may create a ratchet effect similar to making copies of copies on a photo copy machine. Additionally, international standards for cartographic data transmission, storage and retrieval are not fully implemented.

Accessibility

Accessibility to data is not equal. What is open and readily available in one country may be restricted, classified, or unobtainable in another. Prior to the break-up of the former Soviet Union, a common highway map that is taken for granted in this country was considered classified information and unobtainable to most people. Military restrictions, inter-agency rivalry, privacy laws, and economic factors may restrict data availability or the level of accuracy in the data.

Cost

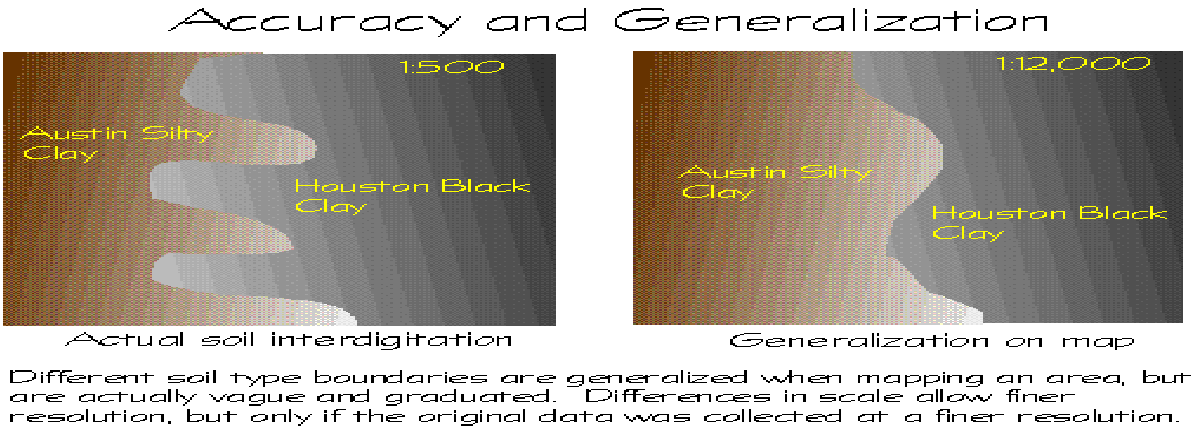
Extensive and reliable data is often quite expensive to obtain or convert. Initiating new collection of data may be too expensive for the benefits gained in a particular GIS project and project managers must balance their desire for accuracy and the cost of the information. True accuracy is expensive and may be unaffordable.

ERRORS RESULTING FROM NATURAL VARIATION OR FROM ORIGINAL MEASUREMENTS

Although these error sources may not be as obvious, careful checking will reveal their influence on the project data.

Positional Accuracy

Positional accuracy is a measurement of the variance of map features and the true position of the attribute (Antenucci *et al.,* 1991). It is dependent on the type of data being used or observed. Map makers can accurately place well-defined objects and features such as roads, buildings, boundary lines, and discrete topographical units on maps and in digital systems, whereas less discrete boundaries such as vegetation or soil type(see fig below) may reflect the estimates of the cartographer. Climate, biomes, relief, soil type, drainage and other features lack sharp boundaries in nature and are subject to interpretation. Faulty or biased field work, map digitizing errors (see fig below) and conversion, and scanning errors can all result in inaccurate maps for GIS projects.



Accuracy and Generalization

Actual soil interdigitation

Generalization on map

Different soil type boundaries are generalized when mapping an area, but are actually vague and graduated. Differences in scale allow finer resolution, but only if the original data was collected at a finer resolution.

Fig. 3.1: Overlay of digitised vector from Google earth imagery on another georeferenced Satellite imagery of the same location.

Accuracy of Content

Maps must be correct and free from bias. Qualitative accuracy refers to the correct labelling and presence of specific features. For example, a pine forest may be incorrectly labelled as a spruce forest, thereby introducing error that may not be known or noticeable to the map or data user. Certain features may be omitted from the map or spatial database through oversight, or by design.

Other errors in quantitative accuracy may occur from faulty instrument calibration used to measure specific features such as altitude, soil or water pH, or atmospheric gases. Mistakes made in the field or laboratory may be undetectable in the GIS project unless the user has conflicting or corroborating information available.

Variation in Data

Variations in data may be due to measurement error introduced by faulty observation, biased observers, or by mis-calibrated or inappropriate equipment. For example, one cannot expect sub-meter accuracy with a hand-held, non-differential GPS receiver. If one is not aware of this natural variation, incorrect assumptions and decisions could be made, and significant error introduced into the GIS project.

ERRORS FROM DATA PROCESSING

Processing errors are the most difficult to detect by GIS users and must be specifically looked for and require knowledge of the information and the systems used to process it. These are subtle errors that occur in several ways, and are therefore potentially more insidious, particularly because they can occur in multiple sets of data being manipulated in a GIS project. These errors include:

Numerical Errors

Different computers may not have the same capability to perform complex mathematical operations and may produce significantly different results for the same problem. Burrough (1990) cites an example in number squaring that produced 1200% difference. Computer processing errors occur in rounding off operations and are subject to the inherent limits of number manipulation by the processor. Another source of error may be from faulty processors, such as the recent mathematical problem identified in Intel's Pentium(tm) chip. In certain calculations, the chip would yield the wrong answer.

A major challenge is the accurate conversion of existing analogue maps to digital form. Because computers must manipulate data in a digital format, numerical errors in processing can lead to inaccurate results. In any case numerical processing errors are extremely difficult to detect, and perhaps assume sophistication not present in most GIS workers or project managers.

Errors in Topological Analysis

Logic errors may cause incorrect manipulation of data and topological analyses (Star and Estes 1990). One must recognize that data is not uniform and is subject to variation. Overlaying multiple layers of maps can result in problems such as Slivers , Overshoots , and Dangles . Variation in accuracy between different map layers may be obscured during processing leading to the creation of "virtual data which may be difficult to detect from real data" (Sample 1994). Figure 3.2 shows a demonstration of some examples of topological errors

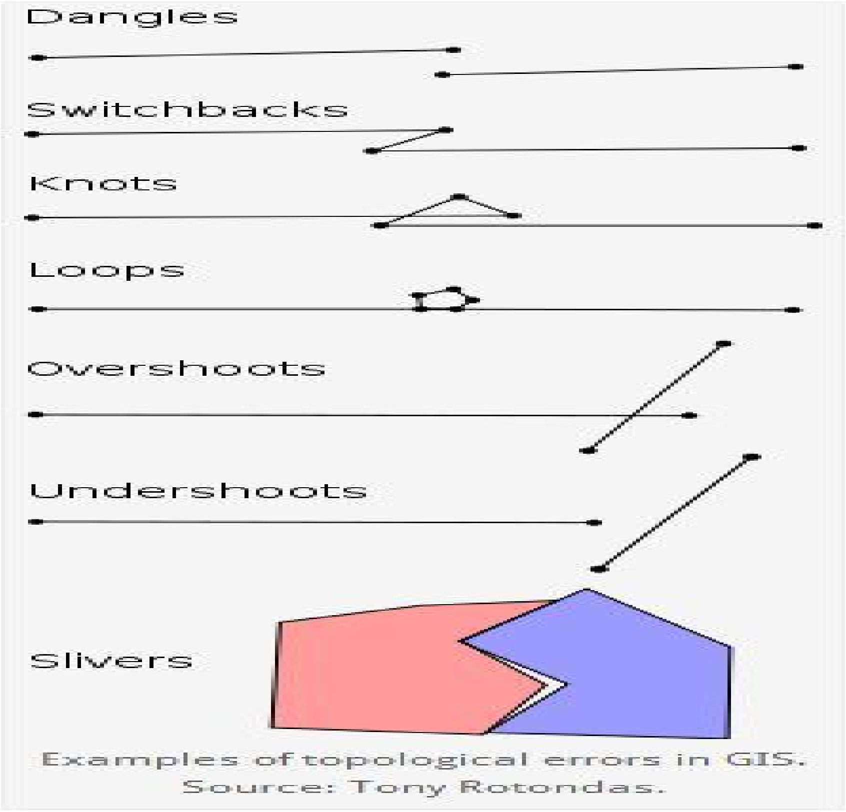


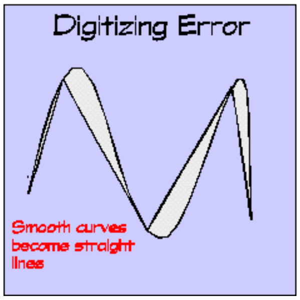
Fig 3.2: Examples of Topological Errors in GIS (Adapted from Tony Rotondas)
In (Maras et al., 2010) most common topological error types in spatial vector data are: Floating or short lines, Overlapping lines, Overshoots and undershoots, Unclosed and weird polygons. On the correction of topological errors in GIS see again (Maras et al., 2010).

Classification and Generalization Problems

For the human mind to comprehend vast amounts of data it must be classified, and in some cases generalized, to be understandable. According to (Burrough, 1986) about seven divisions of data is ideal and may be retained in human short term memory. Defining class intervals is another problem area. For instance, defining a cause of death in males between 18-25 years old would probably be significantly different in a class interval of 18-40 years old. Data is most accurately displayed and manipulated in small multiples. Defining a reasonable multiple and asking the question "compared to what" is critical (Tufte, 1990). Classification and generalization of attributes used in GIS are subject to interpolation error and may introduce irregularities in the data that is hard to detect.

Digitizing and Geocoding Errors

Processing errors occur during other phases of data manipulation such as digitizing and geocoding, overlay and boundary intersections, and errors from rasterizing a vector map. Physiological errors of the operator by involuntary muscle contractions may result in spikes, switchbacks, polygonal knots, and loops. Errors associated with damaged source maps, operator error while digitizing , and bias can be checked by comparing original maps with digitized versions. Other errors are more elusive.



BASIC SOURCES OF ERROR ENCOUNTERED IN USING A GIS

There is error associated with every stage in the GIS process. It is necessary to understand where the error originates in order to ensure a quality final product suitable for its intended use.

From (Aronoff, 1995) Table 5.2 shows examples of errors encountered in GIS.

Data Collection	Errors in field data collection
	Errors in existing maps used as source data
	Errors in the analysis of remotely sensed data
Data input	Inaccuracies in digitizing caused by operator and equipment
	Inaccuracies inherent in the geographic feature (e.g. edged, such as forest edges, that do not occur as sharp boundaries
Data storage	Insufficient numerical precision
	Insufficient spatial precision
Data manipulation	Inappropriate class intervals
	Boundary errors
	Error propagation as multiple overlays are combined

	Slivers caused by problems in polygon overlay procedures
Data output	Scaling inaccuracies
	Error caused by inaccuracy of the output device
	Error caused by instability of the medium
Use of results	The information maybe incorrectly understood
	The information may be inappropriately used

The T-junction is not intersecting accurately.

Fig. 3.3: Errors from clipped Google Earth Imagery using Elshayal Smart GIS

Fig. 3.4: GIS error resulting from variations in original data sets (Black-digitised from Google Earth Imagery and the Red-digitised from Orthophoto)

STRATEGIES FOR ERROR MANAGEMENT IN GEOINFORMATION DATA

Methods for controlling, measuring, and managing error are relatively very important in spatial data analysis. A Few Simple Strategies for Error Management according to (Campbell, 2008) include:

Use of decision trees–heuristics, Know your assumptions/double-check your logic, Know your relevant statistics, geostatistical analysis of model error can identify spatial patterns and find error sources, Specify a minimum standard of accuracy, Consider merging layers for more detail.

PRINCIPLES OF MANAGING ERROR

Managing error in GIS datasets is now recognized as a substantial problem that needs to be addressed in the design and use of such systems. Failure to control and manage error can limit severely or invalidate the results of a GIS analysis. The approaches in managing error in a GIS datasets as suggested by Kenneth E. Foote and Donald J. Huebner, Department of Geography, University of Texas at Austin, in 1995 are highlighted below:

Setting Standards for Procedures and Products

No matter what the project, standards should be set from the start. Standards should be established for both spatial and non-spatial data to be added to the dataset. Issues to be resolved include the accuracy and precision to be invoked as information is placed in the dataset, conventions for naming geographic features, criteria for classifying data, and so forth. Such standards should be set both for the **procedures** used to create the dataset and for the final **products**. Setting standards involves three steps.

Establishing Criteria That Meet the Specific Demands of a Project

Standards are not arbitrary; they should suit the demands of accuracy, precision, and completeness determined to meet the demands of a project. The Federal and many state governments have established standards to meet the needs of a wide range of mapping and GIS projects in their domain. Other users may follow these standards if they apply, but often the designer must carefully establish standards for particular projects.

Training People Involved To Meet Standards and Practice

The people who will be compiling and entering data must learn how to apply the standards to their work. This includes practice with the standards so that they learn to apply them as a natural part of their work. People working on the project should be given a clear idea of why the standards are being employed. If standards are enforced as a set of laws or rules without explanation, they may be resisted or subverted. If the people working on a project know why the standards have been set, they are often more willing to follow them and to suggest procedures that will improve data quality.

Ensuring That Standards are being Employed throughout the Project Cycle

Regular checks and tests should be employed through a project to make sure that standards are being followed. This may include the regular testing of all data added to the dataset or may involve spot checks of the materials. This allows the designer to pinpoint difficulties at an early stage and correct them.

Documenting Procedures and Products: Data Quality Reports

Standards for procedures and products should always be documented in writing or in the dataset itself. Data documentation should include information about how data was collected and from what sources, how it was pre-processed and geocoded, how it was entered in the dataset, and how it is classified and encoded. On larger projects, one person or a team should be assigned responsibility for data documentation. Documentation is vitally important to the value and future use of a dataset. The saying is that *an undocumented dataset is a worthless dataset*. By and large, this is true. Without clear documentation a dataset cannot be expanded and cannot be used by other people or organizations now or in the future. The following questions on undocumented data may arise; what is the age of the data? Where did it come from? In what medium was it originally produced? How accurate are positional and attribute features? What projection, coordinate system, and datum were used in maps? To what map scale was the data digitized? Etc.

Documentation is of critical importance in large GIS projects because the dataset will almost certainly outlive the people who created it.The staff who enters the data may have long retired when a question arises about the characteristics of their work. Written documentation is essential. Some projects actually place information about data quality and quality control directly in a GIS dataset as independent layers.

Measuring and Testing Products

GIS datasets should be checked regularly against reality. For spatial data, this involves checking maps and positions in the field or, at least, against sources of high quality. A sample of positions can be resurveyed to check their accuracy and precision. The USGS employs a testing procedure to check on the quality of its digital and paper maps, as does the Ordnance Survey. Indeed, the Ordnance Survey continues periodically to test maps and digital datasets long after they have first been compiled. If too many errors crop up, or if the mapped area has changed greatly, the work is updated and corrected. Non-spatial attribute data should also be checked either against reality or a source of equal or greater quality. The particular tests employed will, of course, vary with the type of data used and its level of measurement.

CONCLUSION

It is now recognized that error, inaccuracy, and imprecision can "make or break" many types of geoinformation project. That is, errors left unchecked can make the results of a GIS analysis almost worthless. Common types of errors in spatial analysis include: cartographic, statistical, thematic, conceptual and measurement. Four major types of errors have been considered - Positional error, attribute error, conceptual error and logical error. The various sources of errors that may affect the quality of a GIS dataset have been highlighted. One major approach in managing error in a GIS datasets is documenting procedures, products and producing data quality reports. Another is setting of standards and procedures for product. At some point GIS datasets should be checked regularly against reality. For spatial data, this might involve checking maps and positions in the field or, at least, against sources of high quality. Also a National Geospatial Data Infrastructure (NGDI) should be established to reduce data duplication and error propagation from poorly acquired and processed GIS data sets.

REFERENCES

Antenucci J.C, Brown K., Croswell L.P, Kevany J.M, Archer H. (1991). Geographic Information Systems. Van Nostrand Reinhold. New York.

Aronoff, Stan. 1995. Geographic Information Systems: A Management Perspective. WDL Publications. Ottawa, Canada.

Brook Campbell 2008: Error in Spatial Analysis. GEOG 570 Presentation 2008.

Burrough, P.A. 1990. *Principles of Geographical Information Systems for Land Resource Assessment*. Clarendon Press. Oxford

Campbel, Brooke 2008. Identifying Common Types of Error in Spatial Analysis. GEOG 570 Presentation March 25th, 2008.

Chen, J. and P. Gong, 1998. Practical GIS: Building and Maintaining a Successful GIS, Science Press, Beijing, p.186.

Dutton, Geoffrey, Improving Locational Specificity of Map Data - a Multi-resolution, Metadata-driven Approach and Notation. International Journal of Geographic Information Systems. Vol. 10 No 3 pg 253-268: 1996.

Foot, Kenneth *and Donald Huebner* 1995: *The Geographer's Craft Project, Department of Geography,* The University of Colorado, Boulder Land Information Systems. Vol. 53 No. 1 pg 29-40 : 1993

Gong, P., X. Zheng, J. Chen, 1995. Boundary Uncertainties in Digital Maps: An Experiment on Digitization Errors, Geographic Information Sciences, 1(2):65-72.

Kemp, K. (2008): *Encyclopaedia of Geographic Information Science,* page 129-130 SAGE Publications Ltd. United Kingdom.

Kiiveri, H.T., 1997. Assessing, Representing and Transmitting Positional Uncertainty in Maps, International Journal of Geographical Information Science, 11(1):33-52

Shi, W., W.B. Liu, 2000. A stochastic Process-Based Model for the Positional Error of Line Segments in GIS, International Journal of Geographical Information Science, 14(1):51-66.

Stanislawski L.V., B.A. Dewitt, R.L. Shretha, 1996. Estimating Positional Accuracy of Data Layers within a GIS through Error Propagation. Photogrammetric Engineering and Remote Sensing, 62(4):429-433.

Star JL, Estes JE (1990). Geographic Information Systems: An Introduction, pp. 2-6, Prentice-Hall Inc. Englewood Cliffs. New Jersey.

Süleyman Sırrı Maras, Hakan Hadi Maras, Bahadır Aktu2, Erdem Emin Maras and Ferruh Yildiz 2010. Topological Error Correction of GIS Vector. International Journal of the Physical Sciences Vol. 5(5), pp. 476-483.
www.academicjournals.org/IJPS ISSN 1992 - 1950 © 2010 Academic Journals

Veregin, H., 2000. Quantifying positional error induced by line simplification. International Journal of Geographical Information Science, 14(2):113-130.

Zheng, X., P. Gong, 1997. Linear Feature Modelling with Curve Fitting: Parametric Polynomial Techniques, Geographic Information Sciences, 3(1):7-19.

Reference to this paper should be made as follows: Ndehedehe C., *et al.,* (2013), Understanding Errors and their Measurement in Geoinformation, *J. of Environmental Sciences and Resource Management, Vol.5, No.1, Pp.74-87.*