

Validation of Modeling Outcome: Rationalization and Recognition of Realistic Results

Allison Denby, B.A. GIS
Geographic Information Systems Consultant
Integrated Informatics Inc.
1313 7th Street SW
Calgary, AB T2R 1A5
403.829.7575
amdenby@integrated-informatics.com

Jason Humber, B. Eng.
Principal Consultant
Integrated Informatics Inc.
1313 7th Street SW
Calgary, AB T2R 1A5
403.619.8926
jlhumber@integrated-informatics.com

Abstract

Data acquisition, development, and creation technologies are ever evolving and, as a result, are causing a series of revolutions within Geographic Information System (GIS) modeling. GIS modeling of real-world scenarios is finally encroaching upon realism because data sources are reaching levels of granularity only imaginable in the early era of GIS. This, along with other important factors, however, presents its own unique set of issues that are increasing the divide between GIS users and GIS professionals.

Much emphasis is placed on data input quality but, oddly enough, much less is placed on the outcome. Does this make sense considering that business decisions will be made with the output information? Validation of any process or analysis output must be equally important to the input. This paper will outline many of the common issues encountered in the modeling process and offer some cautions and checks for those users striving to understand better spatial modeling.

Introduction

Today GIS is already a daily part of everyone's life, although most people are likely unaware of that fact. Firetrucks, ambulances, and police cars are sent to their destinations via the shortest route available. Health Care uses GIS to study infectious disease and model the affected area and rate of spread. Planning agencies and other organizations are using GIS for decision support. A GIS provides the ability to model an increasingly complex world through integration of different tools and disciplines such as remote sensing, geodesy, cartography, geography, and information technology.

Although the number one use of GIS is primarily no more than digital replacement for large scale maps, the power of GIS is held within spatial modeling. It allows spatial relationships to be identified between multiple disparate datasets. The use of GIS within a business for decision support is on the rise as is the amount of data, quality or not.

The use of GIS as a tool in organizations is leading to a set of problems that could wreak havoc on the industry. This problem stems from the increased "user-friendly" out-of-the-box functionality that is allowing untrained professionals to derive information that is believed to be 'true'. The same amount of emphasis placed on data inputs needs to be placed on validity of the outputs. However, reliance on rigor for adherence to GIS fundamentals is optimistically assumed to be accounted for by "on-the-fly" application abilities.

History of Geographic Information Systems

For more than one hundred years, people have realized the power in analyzing spatial data for decision making. In 1854, John Snow was credited with stopping an epidemic using GIS. In London, England he tracked the source of cholera to a pump that drew water from a contaminated well. His conclusions were drawn from tracking affected peoples movement prior to known infection.

When computer-based GIS began in the 1960's, its main purpose was as a cartographer's tool for digital map making. The correlation between increased power of computers and broadened applications of GIS is very strong. As time progressed from the 60's, computer processing began to increase at a rapid rate. As this happened, the ability to perform new operations in a GIS increased as well. People started to use it as a tool to map interrelationships between spatial datasets.

As with any technology, time has brought much change to geographic information systems, not so much in terms of technology fundamentals but more in terms of technology implementation, for instance, the ease of performing analyses. Today comprehensive modeling is a staple of geographic information systems and modern GIS technology provides an increased amount of out-of-the-box functionality that existed only on the wish lists of pioneering practitioners of GIS.

Nowadays the simple click of a button can produce an output that was at one point in the history of GIS virtually unachievable or at least impractical to achieve. The paradigm shift from mapping to display physical locations to mapping for perspective is leading the industry to spatial reasoning as a core activity of GIS. All that said, on its own, the ease of technology implementation does not guarantee the validity of process or analysis output; the people portion of technology remains vital to validation.

Spatial Modeling

The ability to link geographic and thematic information on spatial objects sets Geographic Information Systems apart from any other information system. Spatial modeling incorporates many datasets to generate useful information for solving complex problems. Spatial models allow users to find and explore relationships among geographic features to understand and address any particular problem. It also provides a framework for understanding and decision making of real world processes. The cost saving of an efficient model is high because many real life instances are too expensive to measure. Models can be extremely simple or very extravagant but in either case the benefits of spatial modeling are very high.

Every model within a GIS is case specific, but a generic framework can be set to outline the basis of spatial modeling. **Figure 1** displays the steps necessary for spatial modeling. By expanding these generic steps to fit a specific model, it will aid in producing the sought after outputs.

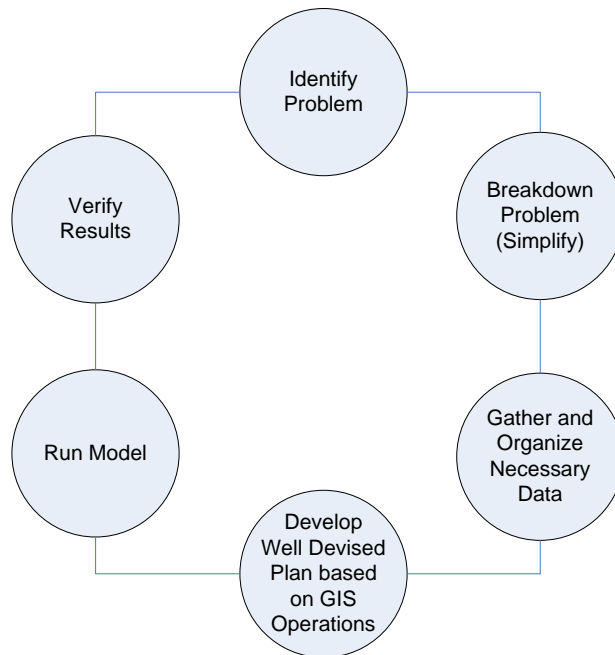


Figure 1: Generic Modeling Process

Data Quality Issues

Due to the case specific nature of spatial modeling, many cautions and concerns need to be discussed in regards to data management and data validation. Clearly, the model's output quality is fully dependant on the input data. If one uses a mixture of good and bad data, the output data will only be as good as the worst dataset input. Error propagation and unawareness of data abilities are the largest cause of incorrect model outputs.

Resolution (or Scale)

Raster data is based on finite resolution (pixel size); therefore as long as the pixel size is greater than the positional accuracy of the data, we have no risk of introducing error. For example, if you have a ten meter resolution dataset and a 30 meter resolution dataset that you are using in a model. The output resolution will be 30 meters. In the case of vectors, they often have a precision that is different than the accuracy. The problem for vectors occurs in digitizing and in spatial overlays. [2]

Logical Consistency

Logical consistency is not error created from the input datasets, but can be created from the outputs. It refers to how the data was created. Certain rules must be followed depending on the type of dataset being created. If development rules are not followed, errors can exist in data. For example, if you are geocoding addresses to land parcels and misrepresentation exists in one of the source datasets, the address may end up attached to an incorrect parcel.

Data Completeness

A dataset that is designed to represent a particular feature may be incomplete if some of the features are missing. A dataset that contains the water bodies throughout a city is incomplete if a river is missing.

Interpolation

Interpolation is a method of creating data by predicting values based on a sample dataset. Some continuous data can be accurately derived using this method, however one must take caution in using this method. Resources and money is not unlimited, therefore interpolation aids in reducing cost of data creation and

acquisition. This method is appropriate for data sets such as rainfall, snowfall, elevation, noise levels, and chemical concentrations.

Spatial Error

Spatial errors or positional errors occur when geographic features are not located in the correct place on the earth's surface. Remember that scale is not a spatial error, but rather the extensibility of a dataset. Positional errors are shown when one dataset is overlaid on another. Depending on the operation and necessary correctness of the dataset, compare the two and use the dataset that is more correct.

Attribute Error

Sometimes attribute errors will exist. An example of an attribute error is when the street dataset has streets not named or incorrectly named. These errors often link to temporal, logical consistency, or data completeness errors.

Validation Approaches

In addition to facilitating the amalgamation of input data, GIS techniques are useful in evaluating and analyzing the model's output data. Depending on the organization and the model being used, methodologies may have been developed for analysts to validate model results. Validation is very important for proof that the model has produced the correct (or near correct) results and to achieve buy-in within an organization. It will provide the supporting evidence necessary to prove the results display factual information.

Validation is very case specific and data dependant. An obvious way to validate data is through ground truthing. The cost of it is extreme and not always an option. There are other reliable methods that will produce the results necessary. Validation could be performed using visualization techniques. For example, derive a hillshade from a DEM to help identify artifacts in its source data. Another way to test for consistent results is by re-running the model using a subset of the original data. Many models have rules and methods built into them; this allows users to apply them also as a method for validation. Once the outputs are derived, simply check to see if the constraints hold true.

When validating results, one must remember the implications of a model. All models are abstractions of reality so no model is perfect. A basis of validation should be knowing whether your model outputs predict conservatively or non-conservatively, how much does it over or under predict? By addressing error propagation, the model is likely more realistic because one can account for the inherent errors of spatial modeling.

Conclusion

The lack of documentation and set procedures for many processes within a GIS is reason enough to demand each individual set and maintain quality assurance and rigorous processes. In the book, *The Unreality Industry*, Mitroff and Bennis (1989) aim their critiques at the TV news because they feel television has turned current events into a mere form of entertainment. There are "some strong parallels between the representation of current events on TV, as viewed by Mitroff and Bennis, and the representation of the geographic world in GIS. While the goals of the two enterprises are clearly very different (luring audiences through entertainment, versus winning professional markets through increased efficiency and productivity), both the means (electronic manipulation of facts and images) and the results (production of unreality of both the first and second kind) have strong similarities." [1] Televisions have become ubiquitous, has GIS also? To ensure the science and professional industry remains credible, there are procedures and standards that need to followed. This will ensure spatial modeling integrity continues to increase in the professional world.

References

- [1] Couclelis, Helen, "Geographic Illusion Systems: Towards a (Very Partial) Research Agenda for GIS in the Information AgeGIS Fundamentals", 1998, NCGIS
- [2] Bolstad, Paul, "GIS Fundamentals", May 2001, Eider Press: White Bear Lake, MN

Biography: Allison Denby

Allison Denby joined Integrated Informatics in May 2004 as the lead Geographic Information Systems Consultant. Her activities within the company encompass many aspects of spatial data management including database development, analysis techniques, research and development for raster based modeling, and GIS implementation.

Allison has successfully completed her Bachelors of Applied Geographic Information Systems degree. Her work background is varied and has involved projects ranging from promotion of geomatics and GIS education awareness across Canada to process improvement for pipeline integrity data integration.

Biography: Jason Humber

Jason Humber founded Integrated Informatics Inc. in October of 2002 to provide data management and system design consulting services to the pipeline industry for new construction, operations, and pipeline integrity. As a Principal Consultant, Jason is responsible for corporate level development and delivery of Integrated Informatics unique suite of services.

In 1999, Jason began his career with the Natural Gas Business Unit of BP Canada Energy Company in Calgary, Alberta. His primary focus within BP was development of a business unit wide data management system that supported the analytic and integration needs of pipeline integrity. While working with BP, Jason also took on a pivotal role on the project management team within the Alaska Gas Producers Pipeline Team and helped to establish the processes required for Project Data Management. More recently, Jason has completed a similar advisory role with the Mackenzie Gas Project, and has broadened and implemented these data management approaches to encompass the needs of developing oil sands projects.