Errors in GIS
Assessing spatial data accuracy

By Paul V. Bolstad and James L. Smith

Effective use of geographic information systems (GIS) requires adjustments in how resource professionals collect and document their spatially referenced data. Prior to GIS, maps were used largely for tactical planning; high spatial accuracy was not required, so errors in spatial data were ignored or considered acceptable.

With GIS, many heretofore costly or time-consuming spatial analyses are now practical. However, this increases the need to measure and document spatial data accuracy (Smith et al. 1991). As models become more sophisticated and rely on site-specific data from many sources, the combined effects of spatial or attribute errors may limit the value of model predictions. For example, errors in stand volume from a GIS-based growth simulator might be due to errors in the growth model, errors in GIS stand area, or erroneous inventory data. We need to know the accuracy of our spatial data before we can effectively perform such spatially based analyses.

This article discusses spatial data accuracy in a resource management setting, beginning with a brief description of spatial data models and an overview of spatial data accuracy (including definition, identification, quantification, and documentation). It concludes with a review of...
accumacies associated with various common data sources and a prescription for improvements in the application of GIS technologies.

Data Model

The data model defines how real-world spatial entities (stands, roads, lakes) are represented in the GIS. Both spatial (location, size, geometry) and tabular (timber type, forage value, site index) characteristics must be represented. Digital "objects" are usually represented by a set of coordinates defining the spatial characteristics and by attribute or tabular data describing the tabular characteristics. Most data models involve several data layers, each corresponding to a different type of data (fig. 1).

Once the data model has been specified, the area is "digitized"—that is, a digital representation is recorded for the area coordinates used to represent each spatial entity, as well as the tabular data associated with the spatial data. For example, the boundaries and tabular data describing all stands, lakes, roads, and property boundaries might be entered. Data entry and update are likely to be the largest share of GIS investments (Congalton and Green 1992). This fact underscores the need to control and document the accuracy of the spatial data.

As described earlier, the real world is represented by both a spatial component (coordinates) and a tabular component (attribute data). Because errors may occur in both, we must discuss two types of accuracy: positional and attribute. Positional accuracies are used to gauge how well the location, size, and shape of real-world features are represented in the database. Attribute accuracies reflect how well the tabular characteristics represent reality.

Positional Accuracy

Positional accuracy can be viewed as a measure of how well the coordinates in the data layer correspond to the "true" coordinates of an entity on the ground. Accuracies, often defined by some measure of positional error, are usually described separately for the horizontal and vertical dimensions.

Positional inaccuracies can be derived from several sources. For example, manual digitizing from paper maps is a common method for entering spatial data (Burrough 1986). Errors may originate in the field surveys, in drafting or map production, in nonuniform distortion of the paper map, in the digitizing equipment itself, or in the activities of the operator. Positional error may also be added in data processing steps. It is possible to measure error associated with each step; but errors may be nonadditive, so it is best to measure the accuracy of the resultant digital data layers.

Ideally, positional errors should be documented for each separate data layer. "True" coordinates would be determined for a number of real-world entities and compared to the coordinates of their corresponding digital objects. "True" coordinates should be accurately determined—at least to the level of accuracy required by all intended geographic analyses. For example, a random sample of Public Land Survey System (PLSS) section corners, paired with the corresponding database coordinates for the digital representations, would allow error calculation for each point, which could then characterize the positional error in the data layer.

While rigorous empirical tests are desirable for all data layers, they are often not performed due to insufficient knowledge, time, funds, or numbers of suitable checkpoints. Many GIS users are not familiar with data accuracy concepts and measurements. In addition, obtaining precise locations for check features requires field surveys.

In the PLSS example cited above, the section corners must be selected and identified in the field, and the coordinates determined. Specialized equipment and skilled personnel are often required to obtain desired accuracies. Multiply this effort by the number of data layers and the costs often become quite large. Finally, many features represented in a digital database are difficult to identify in the field. For example, stand boundaries interpreted from aerial photographs may be difficult to identify on the ground. Despite these difficulties, there are few via-
Error Models
Positional accuracy may be modeled in various ways. Individual points defining geographic features may be considered as a population with both vertical and horizontal error distributions. If the vertical error (the difference between true and database elevation) is assumed to follow a random normal distribution, the mean and variance of this distribution then characterizes the population of vertical point errors. A commonly used metric is the vertical root mean-square error, \( RMSE_v \):

\[
RMSE_v = \left[ \frac{\sum (e^2)}{n} \right]^{1/2}
\]

where \( e \) is the residual error for each measured point (the difference between the true and database elevations) and \( n \) is the number of test points. The validity of the distribution model, and the parameters that define the probability density function, can be determined through sampling. Horizontal error may also be modeled statistically, although more distributions and approaches have been proposed because horizontal error may be considered either univariate or bivariate, with interactions between \( x \) and \( y \) errors. One common metric is the horizontal root-mean square error:

\[
RMSE_h = \left[ \left( \frac{\sum (e_x^2 + e_y^2)}{n} \right) \right]^{1/2}
\]

where \( e_x \) is the error measured in the \( x \) direction and \( e_y \) is the error in the \( y \) direction. Another common measure is the horizontal circular standard error, which can be approximated by \( RMSE_h / 1.4142 \) (Thompson and Rosenfeld 1971).

Describing the positional error for linear or areal features is a bit more complicated. The difference between the “true” and digital data line may vary along the length of the line (fig. 2a). Errors will range from zero at line crossings to high values at the largest separations. One widely applied notion, the epsilon model, defines a distance measured at right angles to the line direction (Perkal 1956, Chrisman 1982). A band around each true line is defined by an epsilon distance (fig. 2b). Various probability distributions have been postulated (Dunn et al. 1990). The proportion or probability of lines within a specified band width can be used to assess line accuracy—for example, 95% of the digital line coordinates may be required to be within an epsilon band of 15 feet for a given data layer.

Error Source and Magnitude
Digital spatial data are derived from a number of sources including maps, aerial photographs, satellite imagery, traditional traverse and leveling surveys, and global positioning system surveys. Each source involves a number of steps and transformations from the original field measurements to the final digital coordinates. The origin and magnitude of common error sources are summarized in Table 1.

Field Measurements. All positional information ultimately relies on field measurements. These may be very precise and painstakingly obtained, such as those that define legal property boundaries, or they may be quite rough, such as

<table>
<thead>
<tr>
<th>Table 1. Reported RMSE ranges for common spatial data sources based on reports of best “commercial” practices.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy component</strong></td>
</tr>
<tr>
<td>Traditional transit surveys</td>
</tr>
<tr>
<td>GPS Carrier phase</td>
</tr>
<tr>
<td>Code-based, stand-alone</td>
</tr>
<tr>
<td>Code-based, differential</td>
</tr>
<tr>
<td>Maps 1/24,000 before digitization</td>
</tr>
<tr>
<td>Manually digitized from 1:24,000 maps</td>
</tr>
<tr>
<td>Digital line graph data</td>
</tr>
<tr>
<td>Aerial photographs</td>
</tr>
<tr>
<td>Uncorrected 9 inch, flat terrain</td>
</tr>
<tr>
<td>Uncorrected 9 inch, steep terrain</td>
</tr>
<tr>
<td>Corrected 9 inch, steep or flat terrain</td>
</tr>
<tr>
<td>Uncorrected 35 or 70 mm</td>
</tr>
<tr>
<td>Corrected 35 or 70 mm</td>
</tr>
<tr>
<td>Satellite data</td>
</tr>
<tr>
<td>Landsat TM</td>
</tr>
<tr>
<td>SPOT HRV, multispectral</td>
</tr>
</tbody>
</table>
the “eyeball” location of an inventory plot on topographic maps. While accuracies for eyeball estimates are in practice unknowable, those determined from the reduction of closed-loop traverse and leveling survey data can be accurately determined (Wolf and Brinker 1989). Traditional surveying practices (theodolite or transit with stadia) commonly achieve accuracies to 1 in 10,000. Since survey points are generally within a few miles of high-accuracy geodetic control monuments, locations surveyed using traditional methods should provide better than one-foot accuracy. Unfortunately, traditional surveying methods are too expensive and time-consuming for most data layers.  

Global positioning system (GPS) technology, another “direct” method, provides lower costs and higher throughput than traditional surveying in many situations. GPS consists of a control segment, a constellation of satellites, and GPS receivers. Positions are determined based on signals broadcast by each satellite. Code-based GPS methods (most commonly applied in a resource setting) provide measurement accuracies of 50-150 feet with one receiver and 5-15 feet with two receivers. Carrier-phase receivers, while more expensive and time-consuming, are capable of providing one-inch RMSEs (Wolf and Brinker 1989).  

Even though GPS technologies have reduced field survey costs dramatically, they are only appropriate for a limited number of data layers. GPS field measurements are best applied where high accuracies are required (e.g., the control data layer) or for sparse linear networks that may be traversed at a high rate of speed such as roads or trails. They are also appropriate when accuracy is required but alternative methods are unavailable—for example, sites far from control points, or remote locations under forest cover. However, GPS is currently not practical where large areas need to be covered in detail.  

Maps. Manual or automated map digitization is currently the most common form of spatial data entry, and as such has the greatest impact on spatial data accuracy. Manual digitizing usually involves putting a paper or mylar map on a digitizer that may be in error.  

Digitization. Positional accuracies during digitization are affected by the equipment and by operator skill or state of mind. Currently digitizers report precision and/or accuracies better than 0.001 inch. One rigorous evaluation (Warner and Carson 1991) identified equipment errors of 0.0035 inch, which at a 1:24,000 scale corresponds to 7 feet. Operator errors vary widely but are generally larger than digitizing table errors. One small study documented manual digitizing precision of approximately 0.0025 inch for well-defined points, approximately twice the observed equipment error (Bolstad et al. 1990). This translates to 4.8 feet on a 1:24,000-scale map and 50 feet on a 1:250,000 map.  

Different errors may be observed for digitized arcs and polygons. Linear features are often depicted with line widths of 0.01–0.04 inch. In addition, generalization of smooth curves introduces additional errors. One study found RMSEs of 22 feet for points from USGS digital line graph data (Vonderhoe and Chrisman 1985). Another study measured approximate average epsilon distances of 0.005 inch, equivalent to 10 feet on 1:24,000 maps (Dunn et al. 1990). Furthermore, when epsilon ranged from 0.005 to 0.04 inch, polygon area error ranged from 1.6% to 16% and was inversely related to polygon size. In another study, polygon area error varied from 5% to 15% of polygon area under realistic assumptions about spatial errors (Priesley et al. 1989).  

Coordinate registration. Registration involves converting from digitizer coordinates to the coordinate system defined by the map projection used for printing the source map. Positional accuracies may suffer at any of the steps involved: identifying control points in both the geographic and digitizer space; obtaining coordinates for the control points in both coordinate systems; choosing a mathematical transformation and estimating coefficients; and applying the transformation to the digitized data, thus producing the output layer.  

While large blunders are easily detected, small blunders or random errors are not. Control point coordinates must be obtained from ground surveys or from controls drafted on the source map. When
field measurements are lacking, control is commonly digitized from geographic coordinate points drafted on the map, e.g., Universal Transverse Mercator graticule intersections drafted on 1:24,000-scale base maps. These controls will contain the positional errors described above. Empirical tests have documented map-derived control errors ranging from 7 to 25 feet (Norberto Fernandez et al., 1991) and 5 to 279 feet (Boistad et al., 1990). GPS technology should facilitate the collection of accurate control data, thus reducing the impact of control uncertainty.

**Imagery.** Imagery is a common source for natural resource spatial data, both for initial database development and for updates. Most current imagery comes from aerial cameras and satellite scanners, although systems based on video cameras and airborne scanners may soon become popular.

Aerial photographs for resource mapping have been routine for the last 50 years. During this time, correction methods and accuracies of photo-based mapping have been thoroughly documented. Unfortunately, these error correction methods are often ignored. In the past it was not considered a problem because area errors were less than those associated with the technology used to measure map area, e.g., dot grids or planimeters, and the maps were not used in spatial analyses such as multilayer overlays. However, these errors are becoming more obvious and objectionable in a GIS framework.

Tilt and terrain distortion are the two major causes of positional inaccuracies in large-format (9-inch) aerial photographs. Elevation variation causes radial displacement away from the photo perspective center, while camera tilt causes perspective distortion (Wolf, 1983). Tilt distortion may be present even in “vertical” aerial photographs, because vertical photos are commonly defined as those with tilts less than 3 degrees. If not removed, these distortions may result in large positional errors in data layers derived from the photographs (fig. 3). For example, one simulation study of “vertical” photographs reported average positional errors in GIS data layers of 13-52 feet over flat terrain and 125-240 feet over steep terrain (Boistad, 1992). The same study documented area errors of 0%-9% on 9-inch vertical aerial photos.

Terrain effects may be removed by stereoscopic viewing. Stereointerpretation recreates a three-dimensional model, which may then be projected onto a planar surface (Wolf, 1983). Stereomodels may be projected on a base map, as with a stereo zoom transfer scope, in which case the quality of the data can only be as good as the map.

While manual interpretation of stereopairs will remove terrain distortion, it will not remove tilt effects (fig. 3, relief=0). The stereomodel will have at least the minimum of the tilts in the two stereopair photos. Tilt can be removed through a three-dimensional projective transformation (which requires a minimum of four control points), but will not be removed with the two-dimensional transformation commonly used to register digitized spatial data. Commercial software has been marketed with "tilt compensation." However, it does not remove tilt errors; rather, it averages them over the image using a two-dimensional affine transformation. Simultaneous removal of tilt and terrain distortion requires a more rigorous analytical photogrammetric model based on projection geometry.

Terrain- and tilt-corrected digital data may be compiled directly from the stereomodel using microcomputer-based analytical stereoplotter (Warner, 1990). While these devices are more expensive than stereo zoom transfer scopes, they are more accurate and flexible. Instrument errors may be less than two feet for PC-based analytical stereoplotters, so most error depends on control-point accuracy and the operator. These systems, when

![Figure 3. The effects of tilt, terrain, and photo scale on horizontal positional error.](image-url)
Table 2. An error or confusion matrix of photointerpreted points.

<table>
<thead>
<tr>
<th>Photointerpreted information</th>
<th>Ground-checked information</th>
<th>User accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hardwood</td>
<td>Conifer</td>
</tr>
<tr>
<td>Hardwood</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Conifer</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Mixed</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Producer accuracy</td>
<td>10/16 = 62.5%</td>
<td>11/17 = 64.7%</td>
</tr>
</tbody>
</table>

used with GPS for control, are capable of providing highly accurate data over broad areas.

Other factors may affect the accuracy of photo-derived data layers. For example, lens or camera distortion, while usually quite small in mapping cameras, may be large in many small-format (35 or 70 mm) systems. Decentering, radial displacement, or local irregularities may be quite large in lenses or camera systems that are not produced expressly for mapping.

Satellite-based scanners are another source for spatial data, especially when large areas need to be mapped. Automated classification of satellite imagery is an established method of land cover mapping (Lillesand and Kiefer 1987). Classification converts multiband reflectance data to a single-layer land cover map. This land cover map is registered to a geographic coordinate system. Thus the accuracy of class boundary location is a function of classification accuracy, geometry of the image, and quality of the registration.

Empirical studies indicate that systematic geometric errors can be removed from current high-resolution satellite data. Linear registration of Landsat Thematic Mapper data resulted in RMSEs of approximately 23–46 feet over flat terrain and 45–90 feet in areas of approximately 3,000 feet of relief (Welch et al. 1984). Positional accuracies for data from the SPOT satellite system are also in the 16- to 83-foot range. Much of this error is due to control-point identification error, since it is difficult to locate subpixel points consistently in a digital image. Positional accuracies in high-relief areas can be improved through digital stereopair analysis or through geocoding based on digital elevation models (Labovitz and Marvin 1986, Labovitz and Wolf 1988).

Attribute Accuracy

Where does tabular (attribute) information come from? Generally from either remote sensing (cameras, scanners, video) or field inventory. Forest stand boundary lines are drawn on aerial photographs, and the cover type is identified by the photointerpreter. Soil type characteristics are based on field information, but soil type boundaries are interpreted on aerial photographs. Satellite image classifications result in identified categories, such as homogeneous regions of land use or land cover. Road and stream attributes are typically collected in the field or from local experience. For decades, forest inventory was the accepted method for gathering mensurational information. None of these methods produces information that exactly represents actual feature attributes.

Photointerpretation. Photointerpretation quality results from complex interaction among the skill and background of the person performing the task, the methods used, the classes or categories identified, and the characteristics of the aerial photography (Avery and Berlin 1985). With all these factors, it is impossible to present general conclusions about the quality of photointerpretation.

Few studies have assessed the accuracy of photointerpretation in operational forestry settings (Biging et al. 1991), although methods have been developed (Congalton and Mead 1983) wherein a number of photointerpreted sites are checked on the ground and the results are cross-tabulated into an error or confusion matrix. In the example presented in table 2, 11 sites photointerpreted as being conifer actually were hardwood and four were mixed. The overall photointerpretation accuracy (67.2%), accuracy for individual categories from both the map reader's (user) and map maker's (producer) perspective, and the types of photointerpretation errors committed can be determined. The user must then evaluate the quality of photointerpretation under the specific conditions encountered, and remember that this quality is translated into the tabular component of the GIS database.

Digital imagery. Since the use of digital images in natural resource management is rising, the accuracy of attributes derived from these sources must be considered. Factors such as sensor type, classification method, season of acquisition, and categories to be identified affect the quality of information developed from digital images. The practicality of satellite image data has been debated for nearly 20 years, even as methods are developed to ease and improve their use; thus confusion matrices are routinely applied to satellite-sourced digital maps. Many accuracy studies have used photointerpretation as "ground-truth." This misnomer has clearly had a negative impact on the results of satellite-image classification accuracy. Methods are available for incorporating these accuracy results into particular applications, including GIS (Pisley and Smith 1987; Czaplewski 1992). The need to evaluate the quality of nonphotographic, remote sensing information has been recognized by researchers; a GIS practitioner who uses digital imagery as a source of attribute information must do the same.

Field-collected information. The quality of field-collected attribute information used in a GIS must also be considered. Resource managers generally realize that since most inventory information is collected using a sampling scheme, there is inherent variation in the resulting data. This variation is reported in the form of a sample variance. For instance, the results of a forest cruise might be reported as 100,000 cubic feet plus or minus 8,000 cubic feet. This is entirely appropriate, but the standard deviation figure almost never becomes part of an at-
tribute database. In addition, field information is often averaged over a large region and then applied to smaller areas.

Road attributes are rarely inventoried—they are usually determined either from existing maps or from local experience. This kind of information is often out of date, which introduces the attribute quality of timeliness. Clearly, even the quality of field-collected attributes must be evaluated when using a GIS database.

Conclusions

The foundation of a geographic information system is spatial and tabular data. Most of the time and money in a GIS are spent on basic data. Data quality affects virtually every use of the GIS to either a small or large extent. The various components or aspects of GIS database quality discussed in this article need to be recognized so that system developers, managers, and users can apply the database appropriately and effectively to natural resource management problems.

Literature Cited


Paul V. Bolstad is assistant professor and James L. Smith is associate professor, Department of Forestry, Virginia Polytechnic Institute and State University, Blacksburg.