[AM-09-106] Error-based Uncertainty

Abstract

The largest contributing factor to spatial data uncertainty is error. Error is defined as the departure of a measure from its true value. Uncertainty results from: (1) a lack of knowledge of the extent and of the expression of errors and (2) their propagation through analyses. Understanding error and its sources is key to addressing error-based uncertainty in geospatial practice. This entry presents a sample of issues related to error and error based uncertainty in spatial data. These consist of (1) types of error in spatial data, (2) the special case of scale and its relationship to error and (3) approaches to quantifying error in spatial data.

Keywords: accuracy, ambiguity, error, resolution, RMSE, scale, Uncertainty, vagueness

Author & citation

Wechsler, S. (2021). Error-based Uncertainty. The Geographic Information Science & Technology Body of Knowledge (3rd Quarter 2021 Edition), John P. Wilson (ed.).
DOI: 10.22224/gistbok/2021.3.3..

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Error-based Uncertainty. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers. (2nd Quarter 2016, first digital)

Explanation

1. Definitions
2. A Re-Introduction to Error
3. Semantic Errors
4. Matching Pattern with Process
5. Representing Error and Error-Based Uncertainty
6. Conclusions

1. Definitions

Error: The departure of a measurement, or spatial representation of data, from its true value (Chrisman, 1991).

Spatial Uncertainty: The result of what is not known about the nature and extent of error in a spatial database (Goodchild, 2007).

Accuracy: The closeness of a measure to its “true” co-located value (Burrough, 1986).

Precision: The exactness of a measurement (Esri, 2021).

Stochastic Model: A modeling approach to address uncertainty that incorporates statistical properties of a spatial dataset including a random component (Esri, 2021).

Probabilistic Model: A modeling approach to address uncertainty that measures the likelihood of a particular outcome (Esri, 2021).

Error of commission: A type of error where data is erroneously included from consideration when it should have been excluded (Lavrakas, 2008).

Error of omission: A type of error where data is erroneously excluded from consideration when it should have been included (Lavrakas, 2008).

Vagueness: Imprecision in language expressed as a lack of unique distinction between objects and classes (Fisher, 1999).

Ambiguity: A component of error that occurs when there are more than two different definitions for a term for example an attribute or map feature (Fisher, 1999).

Semantic Uncertainty: The uncertainty that arises from discrepancies in meanings applied to spatial data, often resulting from ambiguity and vagueness (Fisher, 1999).

2. A Re-Introduction to Error

“In science, the word error does not carry the usual connotations of the term mistake or blunder. Error in scientific measurement means the inevitable uncertainty that attends all measurements...The best you can hope to do is to ensure errors are as small as reasonably possible and to have a reliable estimate of how large they are...” (Taylor, 1997, p. 3)

The term "error" typically has a negative connotation. We are ingrained from a young age to assume that errors are “bad” (the red mark on an exam) and should be avoided. Yet error avoidance is impossible with geospatial data; errors and connected uncertainty are inherent components of spatial data and spatial data analyses. We must adjust our frame of reference regarding our approach to errors in geospatial practice. The first steps in doing so are to recognize sources of errors in spatial data, to identify how these contribute to error-based uncertainty and to use accepted methods to address them.

Here, **error** refers broadly to various factors that impact the quality of a spatial dataset. **Uncertainty** results from our lack of knowledge in how these factors propagate through spatial analyses. Errors in spatial data may result from mistakes, systematic errors, or random errors. Blunders may be associated with the data collection process. Systematic errors are the result of procedures or systems used in the data production or collection and follow fixed patterns that can cause bias. Blunders and systematic errors, when identified can be addressed. Random errors are unpredictable.

Factors that contribute to theses errors are discussed (Sections 2.0-4.0) and examples of methods to address error-based uncertainty are presented (Section 5.0). Understanding and managing error and error based uncertainty is necessary for trust in geospatial

analyses and resulting decisions.

2.1 Scale as an Organizing Framework for Categorizing Error

“Scale is one of the most fundamental aspects of any research, yet is one of the more ambiguous scientific terms...” (Quattrochi and Goodchild, 1997)

Geographic Information Science (GISci) applies standard methods of science to spatial analyses. The first step in the scientific method is observation. Observation in geospatial practice begins with data exploration. Geospatial data are derived from a variety of sources and at varied scales (Dabiri and Blaschke, 2019).

Scale can be categorized as: spatial, measurement, and temporal (Figure 1). Although there is complexity and overlap within and between these types of scales, organizing error as components of these scales can assist users in evaluating error-based uncertainty in geospatial analyses.

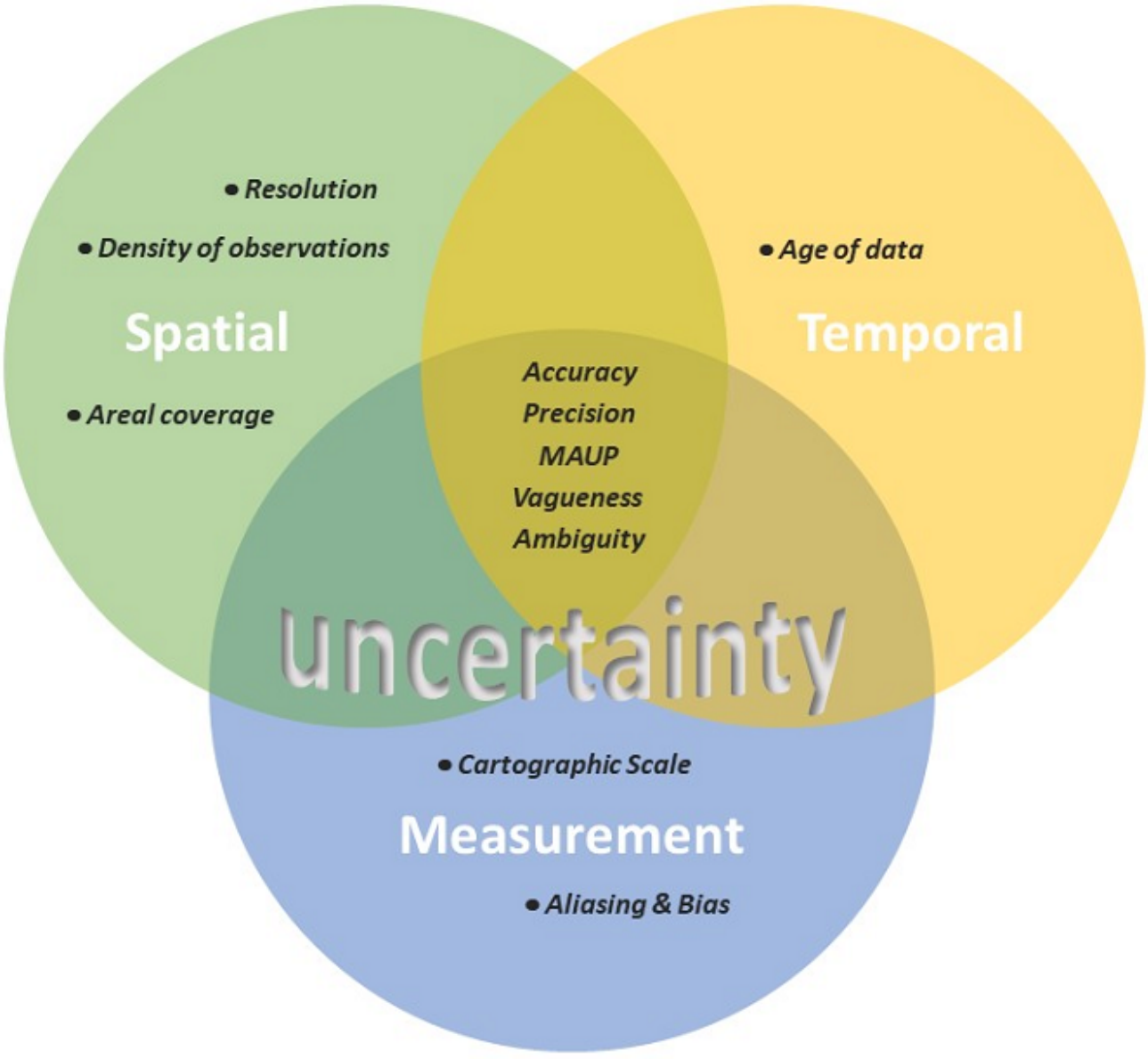


Figure 1. Factors contributing to the reliability of spatial data and error-based uncertainty, categorized under the framework of scale. Source: author.

- Spatial scale refers to the spatial extent of a study area and associated areal coverage of a dataset. Spatial scale determines the representation of data. Errors can be introduced when the source data does not cover a study area consistently. For example, data for a particular study area may not be available from the same source, or at the same spatial or temporal scale. In the raster data structure, grid cell resolution defines the limits of observation. If too coarse, subgrid variabilities are not captured. Errors may arise when observations are not collected at a sampling frequency representative of the feature or process under consideration.
- Measurement scale refers to the precision at which a dataset is generated. Measurement scale is controlled by the map scale imposed on a dataset when it is created. The cartographic scale describes the level of generalization of map features and is typically referenced using the unitless representative fraction. Map accuracy is associated with cartographic scale. Digital datasets are intended for use at the scale of the source data. For example, a spatial dataset derived from data originally produced at a scale of 1:24,000 (e.g. 7.5 minute USGS quadrangles) is not intended to be analyzed with a map of a road network created at a scale of 1:100,000 or a bedrock map created at a scale of 1:250,000. A Geographic Information System (GIS) will however allow us to view and overlay these features concurrently and perform analyses between them. The positional accuracy of the underlying data based on the projection, coordinate system and datum applied, further define the measurement scale.
- Temporal scale refers to the timeframe associated with a dataset. Geospatial data are static representations of features at points in time . Temporal errors may arise from inconsistencies in timeframes between datasets when applied to a particular analysis, for example, use of 2020 census data as a surrogate for present day populations, or the predicting fire hazards using fire perimeters not updated with a current wildfire season.

2.2 Accuracy and Precision

Accuracy and **precision** are overarching descriptors of data quality. They operate at each of the observation scales and are related to factors contributing to error-based uncertainty. Accuracy, in a geospatial context, refers to the degree to which information on a map matches its co-located real-world values, either positionally, or in the attributes assigned. Precision refers to the exactness of a measurement itself, the tool used to perform the measurement (i.e GPS), or the description of a feature in a spatial database (Burrough, 1986, Foote and Huebner, 2000). GISs often provide a false sense of precision by, for example, assigning excessive decimal places to coordinates and area/length calculations that exceed actual measurement. Precision in language also limits our ability to communicate descriptions of features, for example where one soil type ends and another begins, or the boundary between urban and rural areas. All spatial data are of limited accuracy and precision (Goodchild and Gopal, 1991). Understanding the limits of accuracy and precision can be leveraged to understand the errors they contribute to, and associated

uncertainty.

3. Semantic Errors

Errors can arise due to the lack of precision not only in measurement but also in the language used to describe spatial phenomena. Descriptive information can be imprecise. Words given to describe map features or attributes in a table can have multiple connotations. **Ambiguity** occurs when there are multiple terms used to reference a feature. Vagueness refers to a lack of distinction between boundaries, for example the unclear demarcation between urban and non-urban areas are often represented as discrete crisp boundaries in vector GISs, when the boundary is not so defined. Errors due to ambiguity and vagueness result in semantic uncertainties (Fisher, 1999, Li et al., 2018, Wechsler et al., 2019).

4. Matching Pattern with Process

“…If the process is significantly influenced by detail smaller than the spatial resolution of the data, then the results of the analysis and modeling will clearly be misleading...in addition to the spatial resolution of data it is also important to consider the spatial resolution of the process...” (Goodchild, 2011, p. 6)

A goal of geospatial practice is to match patterns in geospatial data with complex processes that operate at different spatial, measurement and temporal scales. The overlapping of scales and associated errors contribute to uncertainty. Aliasing is a term referring to the consequence of not measuring a process at its natural frequency, resulting in bias.

For example, hydrologic processes can be observed using various scales. Spatial scales of observation can be applied on a hillslope, watershed, basin or larger areas. Measurement scales may vary based on the grid cell resolution of digital elevation models when used to generate the terrain features such as slope, aspect and flow direction through a system. The density of observations as represented by weather stations that measure precipitation inputs and the map scale of stream networks. Temporal scales can include the complexity in representing antecedent moisture conditions in a system that affect how it will respond to a precipitation event, as well as the timeframe of the meteorological measurements. Measuring error associated with each variable enables assessment as it develops and propagates. However, the complexity of overlapping scales and how the errors propagate within and across scales contribute to uncertainty.

Matching the pattern as observed using geospatial data to represent the processes we are trying to understand is a quest in geospatial analysis. An approach offered from the field of signal processing is the Nyquist frequency which states that capturing a process at twice its natural frequency can avoid aliasing and subsequent bias in a sample. The natural frequency of a process is not always known and obtaining spatial data at scales that can accommodate it is difficult. However, attention to appropriate sampling through consideration of the scales of observation is necessary.

Map features, and associated processes, when represented at differing spatial scales, may result in different interpretations. This disconnect is a well-established geographic concept known as the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984; Fotheringham, 1989). The base unit of support in an analysis is controlled by the spatial scale and areal extent. The raster data structure, although continuous by definition, is discrete in construct as each cell represents an observation. Furthermore, where the origin of the grid is placed will influence the classification or representation of the features being mapped.

5. Representing Error and Error-Based Uncertainty

Addressing error and error-based uncertainty requires: identification, quantification and communication. This section provides a review of approaches to addressing error and error-based uncertainty. There is no consensus on visualization of error and resulting uncertainty. Representing uncertainty may cast doubt on the meaningfulness of the results of geospatial analyses thus jeopardizing the credibility of the work. Nevertheless, the scientific method requires reporting the limitations of research that is necessary for further knowledge building. This section provides examples of approaches to address error and error-based uncertainty.

5.1 Accuracy Statistics

We rely on accuracy statistics to quantify the quality of a dataset and thus the measure of error. Such validation of a dataset often requires a measures of ground truth or higher accuracy data for use in a statistical description of accuracy. One of the more common accuracy statistics is the Root Mean Square Error (RMSE) which compares a sample of the data with a validation dataset of higher quality (equation 1). The RMSE squares the difference between predicted and actual (higher accuracy/measured) values. Squaring differences removes negative values in the equation, yet places more emphasis on values that are farther apart. The Mean Absolute Difference (MAD) accuracy statistic (equation 2) applies the absolute value of the predicted minus the actual (higher accuracy/measured) values. The standard deviation of difference (Equation 3) measures the dispersion of this shift. Accuracy can therefore be further reported as the mean absolute difference plus or minus the standard deviation of difference (Li, 1992).

Equation 1.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{(N-1)}}$$

Equation 2.

$$MAD = \frac{\sum_{i=1}^{N}|(Predicted_i - Actual_i)|}{(N-1)}$$

Equation 3.

$$\text{Standard Deviation of Difference} = \sqrt{\frac{\sum_{i=1}^{N}(|Predicted_i - Actual_i| - MAD)^2}{(N-1)}}$$

Accuracy statistics can be calculated, for example, to quantify the quality of a digital elevation model (DEM) by collecting a sample of higher accuracy elevation measurements in the field, or by extracting a set of validation points from a LiDAR point cloud, and comparing these to the derived DEM.

5.2 Applying Map Accuracy Standards

Map features are generalized in order to be represented. Map accuracy standards, initially released in 1947 for paper maps, quantify this generalization and provide levels of expected confidence in map data (USGS, 1999; US Bureau of the Budget, 1947). According to these original map accuracy standards, no more than 10% of the points tested for maps at scales of 1:20,000 or larger may have an error of 1/30th of an inch, and 1/50th of an inch for maps smaller than 1:20,000. This means that features on a 1:24,000 quadrangle map may have a horizontal error of ±40 feet and ±166.7 feet for features on maps at scales of 1:100,000. Because many digital maps in use were generated from paper maps, these map accuracy standards hold true for the derived digital representations. Horizontal accuracy specifications should be provided with a dataset's metadata. For example, the allowable horizontal positional accuracy of the Federal Emergency Mapping Agency (FEMA) flood hazard maps is reported to be ±38 feet (FEMA, 2021, p. 32).

In map overlay operations two or more features are overlain producing output based on a topological relationship, such as their intersection. Typically, such results are presented using discrete boundaries, yet boundaries may vary based on scale. Map accuracy

standards can be applied to visualize the fuzziness of boundaries around a map feature thus quantifying potential errors of omission and commission.

Figure 2 depicts an approach where uncertainties in the boundaries of a FEMA flood zone layer are visualized and quantified based on the specified horizontal accuracy (e.g. ±38 ft). Areas beyond the given boundary may potentially be in a flood zone (error of omission) while areas within the given boundary may be included unnecessarily (error of commission) (Figure 2).

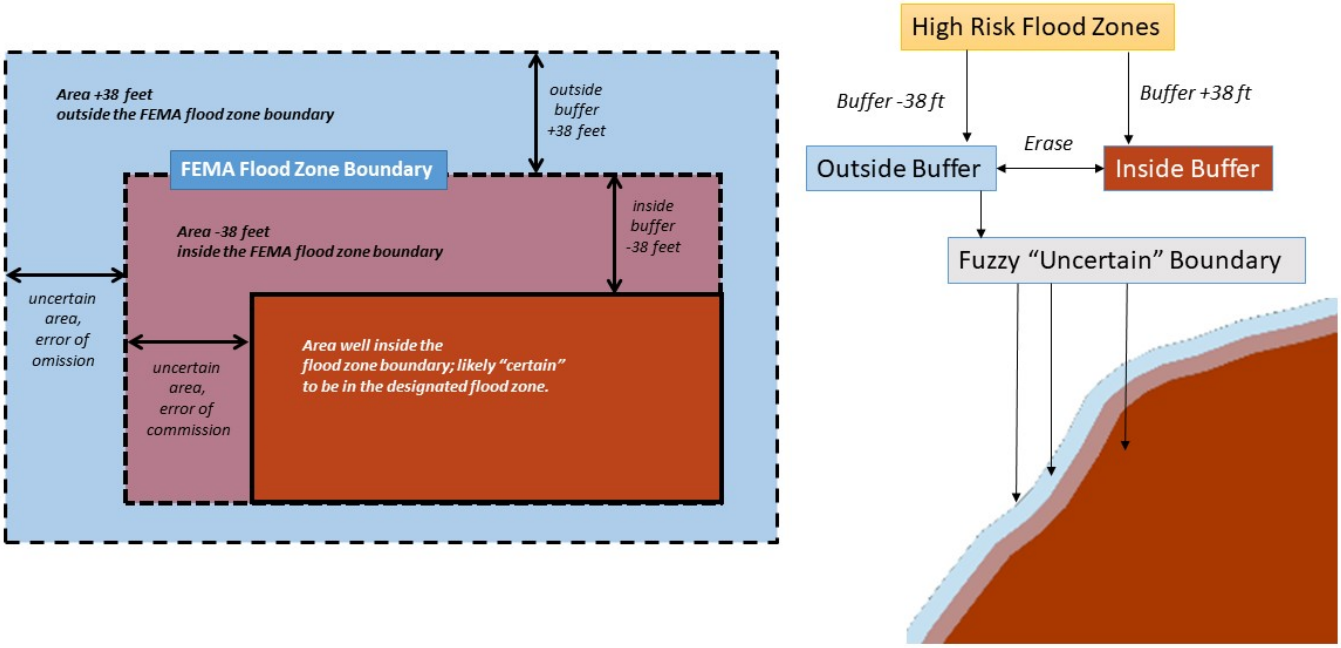


Figure 2. Application of Map Accuracy Standards to an Overlay Operation. Source: author.

5.3 Polygon Mismatch

Boundaries are dependent on the scales at which they are mapped. Differences may arise between polygons that represent the same features, for example: (a) comparison of vegetation species calculated using different methods, (b) land cover represented by different years (c) comparing field derived ground truth data with computer-derived classifications, or (d) comparing watershed boundaries derived using different grid cell resolutions. These area differences can be quantified using geoprocessing techniques. The percent overlap between the polygons can be calculated by dividing the total area of intersect by the total area of their union. Errors of omission (not including areas that should have been included in the polygon) and errors of commission (including areas that should not be included) can be represented using a geoprocessing technique (referred to as the symmetrical difference in Esri's ArcGIS), where areas in common are erased and areas that are not coincident can be measured. These are straightforward approaches to quantify changes over time, or accuracy of certain classification methods.

5.4 Stochastic Modeling

Stochastic modeling applies a statistical technique to address error-based uncertainty. One

example of the application of stochastic methods is in the representation of error propagation in digital elevation models (DEMs) and derived terrain parameters. This technique adds N random statistically uncorrelated or spatially autocorrelated error fields, such as those that incorporate the DEM's RMSE or the spatial dependence within a DEM that is based on a semivariogram analysis (Wechsler and Kroll, 2006, Wechsler, 2007). When added to a surface these outcomes represent realizations of the DEMs under uncertain conditions. Derived terrain parameters obtained from each realization are compared with those calculated from the original dataset yielding residual surfaces. Analyses on a per-cell basis can be summarized using various statistical measures to approximate the range of uncertainty due to DEM error. This modeling approach is referred to as Monte Carlo simulation.

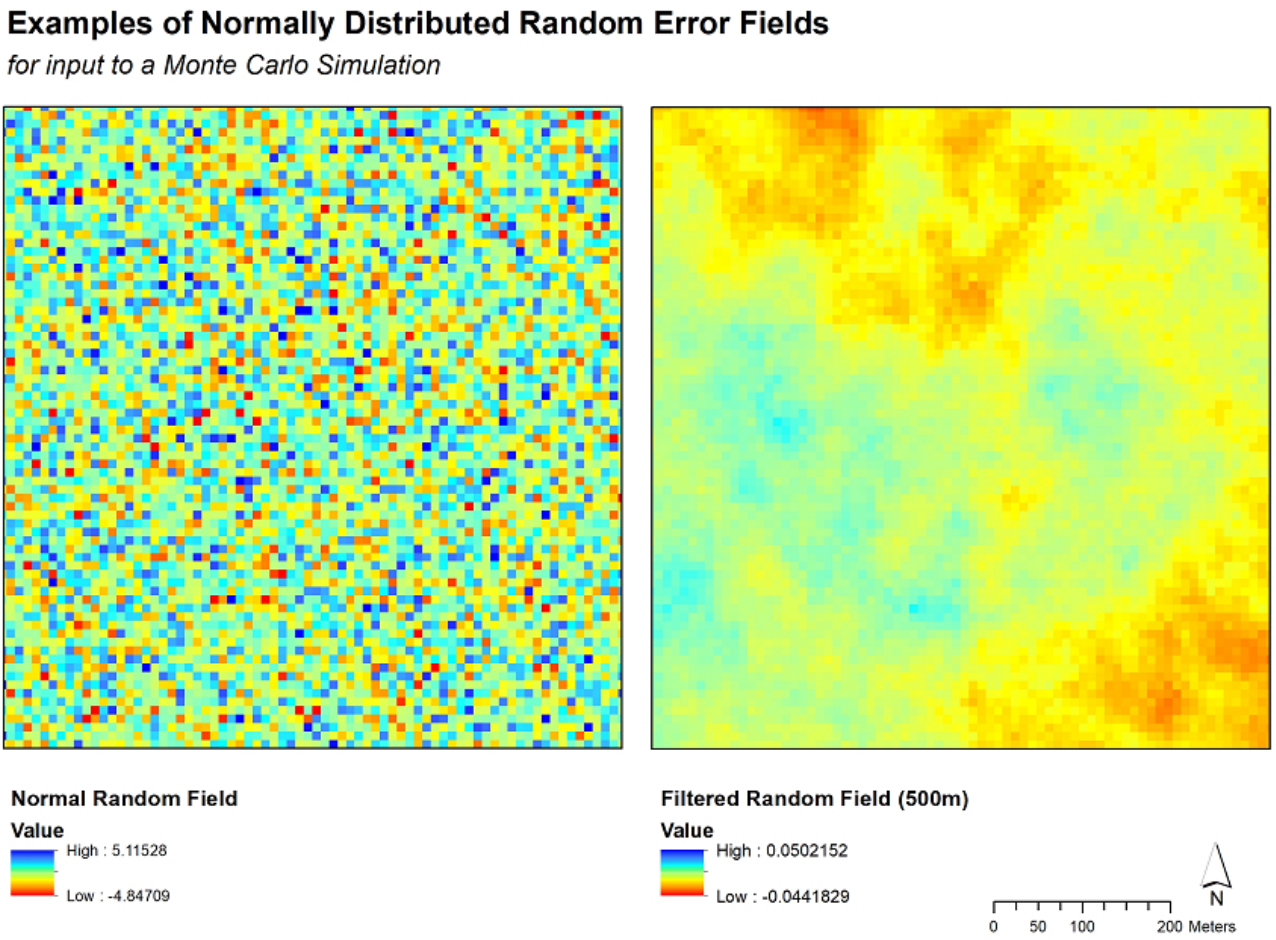


Figure 3. Random fields for use in Monte Carlo simulations. Source: author.

5.5 Probability Modeling

Probability modeling is also referred to as multicriteria analysis. Often a geospatial user is not a subject matter expert. Even experts may not agree on appropriate criteria for a spatial analysis. Probability modeling enables exploration of a variety of input assumptions. Rather than reporting one outcome of a spatial analysis, input assumptions can be varied, and the analysis run N times. The range of outputs when integrated, represent the

likelihood that a result meets a set of criteria. Figure 4 represents the integration of six equiprobable outcomes of a weighted overlay analysis based on varying weights applied to vegetation, soils and slopes. This approach can be applied to other analyses such as a cost path analysis. A variety of inputs can be varied, generating equi-probable total cost surfaces from which pathways are derived. The overlap between these pathways can be represented to explore the most likely scenario.

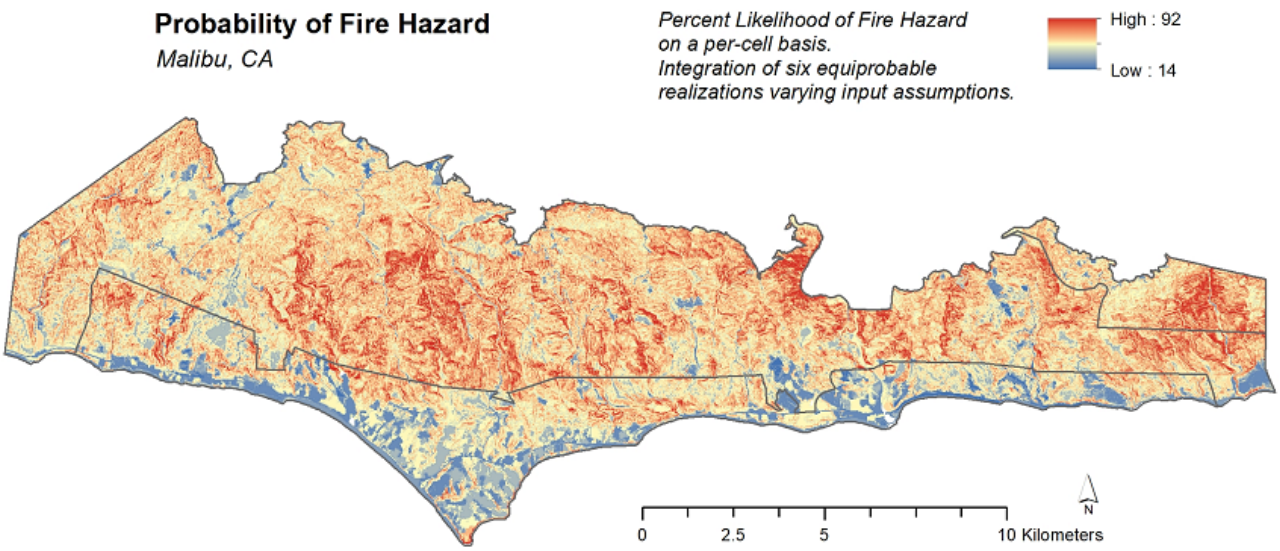


Figure 4. Multicriteria analysis representing likelihood of areas to burn on a per-cell basis from varied input assumptions. Source: author.

5.6 Fuzzy Analysis

Often results of geospatial analyses are reported using discrete boundaries confining representation of output as ‘in-or-out’. The “fuzziness” of boundaries is typically not quantified nor visualized using the vector data structure, however they can be approximated in the raster data structure which permits continuous representation of variables. Fuzzy set theory was developed to address vagueness and ambiguity in representation of boundary features. For example – where do urban areas end and suburban areas begin, or where does the deciduous forest transition to coniferous? Fuzzy logic converts semantic descriptions into a spatial representation using tools that reclassify variables using fuzzy membership classes rather than Boolean classifications. For example, when identifying potential fire hazard areas, input variables are based on membership curves that describe the likelihood of membership and reclassifies a cell in ranges of 0 to 1 accordingly. Reclassifying inputs based on membership classes enables representation of complex phenomenon that are not adequately modeled using discrete variables. (Figure 5).

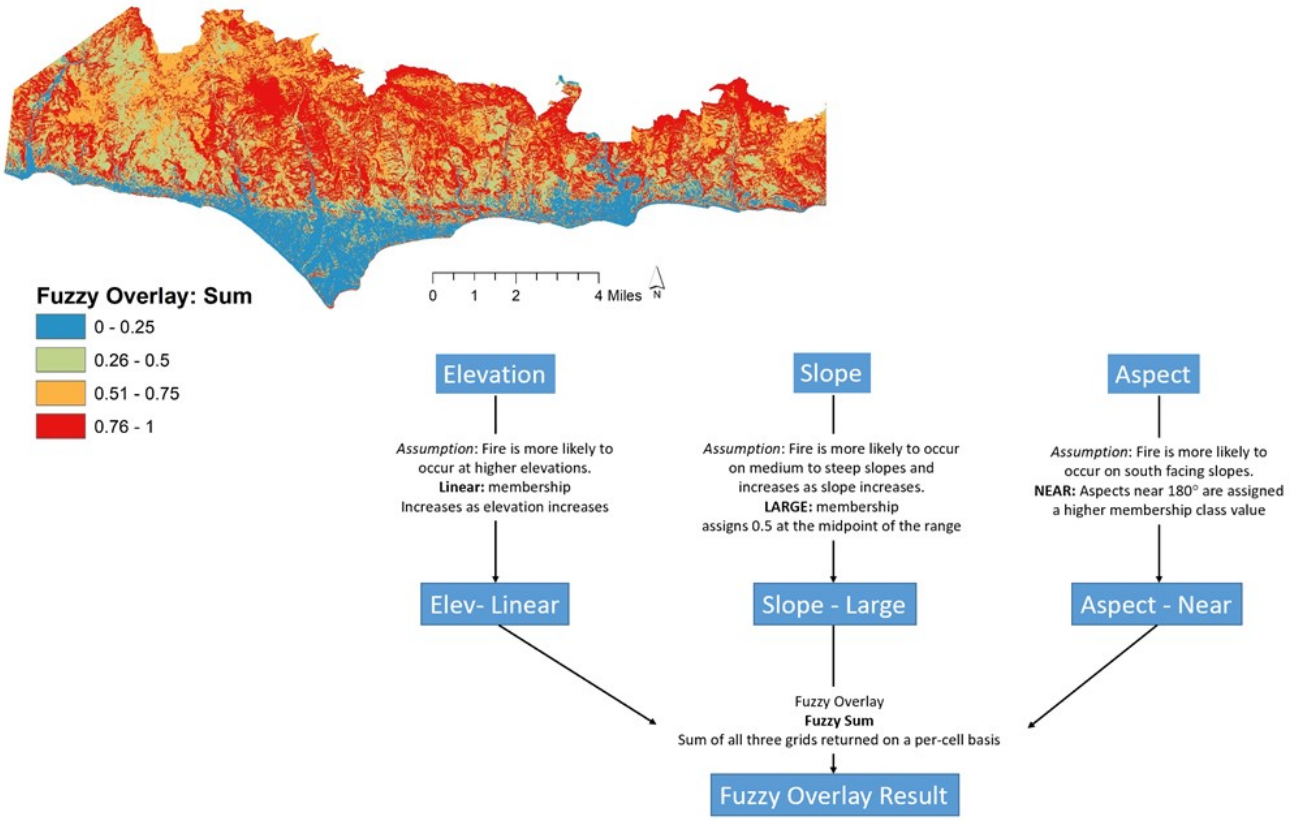


Figure 5. Spatial model of a fuzzy overlay analysis. Source: author.

6. Conclusions

"…the importance of an uncertainty depends on how much it could affect the decision, not simply the outcomes." (Morgan and Henrion, 1990, p. 197).

Error and resulting error-based uncertainty have long been acknowledged in the geospatial community, yet are not consistently addressed. According to Oreskes (2021), consensus is the key to scientific acceptance. There is growing consensus regarding the presence and sources of error in geospatial data, but no consensus regarding how to utilize such information in reporting or analytic results and to inform decision making.

Addressing uncertainty requires a recalibration of our perspective related to error, changing the focus from avoidance to utilization. To manage uncertainty requires knowledge about how errors are introduced into spatial data and how errors are measured. Error should be minimized to the best of our ability and then quantified to the extent possible. Uncertainty measures should be developed and applied routinely as a part of spatial analyses and communication of results. These tools are essential to capture the complexity in geospatial analyses and in turn used for more trustworthy applications and robust knowledge building.

Identifying sources of error and predicting or judging the potential impact of observed error and associated error-based uncertainty should become a basic component of the geographic approach to spatial inquiry and analyses. Addressing uncertainty associated with spatial analyses will lead to better decisions, responsible practice, and greater trust in

the results.

References

Burrough, P. A. (1986). Principles of Geographical Information Systems for Land Resources Assessment, First Edition. Oxford University Press.

Chrisman, N. R. (1991). The error component in spatial data. Geographical information systems, 1(12), 165-174.

Dabiri, Z., & Blaschke, T. (2019). Scale matters: a survey of the concepts of scale used in spatial disciplines. European Journal of Remote Sensing, 52(1), 419-434.

Esri. (2021). GIS Dictionary. Last Accessed August 2021.

Federal Emergency Management Agency (FEMA). (2021). Flood Risk Analysis and Mapping (Rev. 11) FEMA Policy #204-078-1. Last accessed June 2021.

Fisher, P. F. (1999). Models of uncertainty in spatial data. Geographical Information Systems, 1, 191-205.

Foote, K. E., and Huebner, D. J. (2000). Error, accuracy and precision—the geographer's craft project. Department of Geography, University of Texas, Austin (2000). Reprinted at PSU with permission.

Goodchild, M. F. (2007). Imprecision and Spatial Uncertainty. In: Shekhar, S., & Xiong, H. (Eds.) Encyclopedia of GIS. Springer Science & Business Media, p. 480-483.

Goodchild, M. F. (2011). Scale in GIS: An overview. Geomorphology, 130(1-2), 5-9.

Goodchild, M. F. & Gopal, S. (Eds.). (1989). The Accuracy of Spatial Databases. CRC Press.

Lavrakas, P. J. (editor) (2008). Encyclopedia of Survey Research Methods. Sage Publications.

Li, Z. (1992). Variation of the accuracy of digital terrain models with sampling interval. The Photogrammetric Record, 14(79), 113-128.

Morgan, M. G., Henrion, M., & Small, M. (1990). Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press.

Oreskes, N. (2021). Why Trust Science? Princeton University Press.

Quattrochi, D. A., & Goodchild, M. F. (Eds.). (1997). Scale in Remote Sensing and GIS. 1st Edition. CRC Press.

Taylor, J. (1997). Introduction to error analysis, the study of uncertainties in physical measurements. University Science Books, Sausalito, CA, 327 p.

United States Bureau of the Budget. (1947). "United States National Map Accuracy Standards."

United States Geological Survey (USGS). (1997). Standards For Digital Elevation Models. Part 1: General, Part 2: Specifications, and Part 3: Quality Control. Department of the Interior, Washington, DC.

United States Geological Survey (USGS). (1999). Map Accuracy Standards, USGS Fact Sheet 171-99, November 1999. Accessed June 2021.

Wechsler, S. P. (2007). Uncertainties associated with digital elevation models for hydrologic applications: a review. Hydrology and Earth System Sciences, 11(4), 1481-1500

Wechsler, S. P., & Kroll, C. N. (2006). Quantifying DEM Uncertainty and its Effect on Topographic Parameters. Photogrammetric Engineering & Remote Sensing, 72(9), 1081-1090.

Wechsler, S. P., Ban, H., and Li, L. (2019). The Pervasive Challenge of Error and Uncertainty in Geospatial Data. In: Koutsopoulos, K., de Miguel González, R., Donert, K. (eds) Geospatial Challenges in the 21st Century. Key Challenges in Geography. Springer, Cham.