Volunteered Geographic Information Assessment through Analysis:
Building upon Analogues from Other Spatial Analytical Techniques

Zachary Christman, Department of Geography and Environment, Rowan University

The growing availability of information collected from citizen mapmakers, volunteer contributors, and other informal sources of spatial data has enabled a wealth of new opportunities for analysis and investigation. Of utmost concern from the geospatial analytical community of researchers and practitioners is the quantifiable accuracy and credibility of this volunteered geographic information (VGI), which may differ in format and use from existing conventional spatial data. This paper discusses comparable contexts for the preparation and evaluation of the accuracy of spatial data from conventional sources and new VGI contexts and builds upon the vocabulary and metrics that can be used to quantify and evaluate the contexts in which VGI or other similar data may be used. By incorporating the relevant concepts of vagueness and ambiguity in spatial information (Foody 2004, Leyk etal. 2005) and the quantitative metrics of classification uncertainty (Eastman 2009) or change confusion (Christman 2010), VGI may be effectively evaluated and analyzed using existing techniques that have proven effective in other analytical contexts.

Within existing spatial analytical frameworks, there are a number of techniques and evaluation frameworks that may provide insight and guidance for the evaluation and use of VGI and other informal sources of data:

- From *land use and land cover classification* come challenges of the representation and thematic categorization of continuous phenomena into discrete classes. Additionally the incorporation of ancillary data, in the sense of representations of auxiliary characteristics that may not directly pertain to the topic of interest but may aid in its categorization (Lambin et al. 2011; Miller and Rogan 2007; Herold et al. 2008; Christman 2010; Foody 2010).
- From *species distribution modeling* comes the challenge of lacking "absence" data regarding where a subject of interest may not be found, and instead needing to model only "presence" data regarding the possibilistic range, rather than a normalized probabilistic distribution (Fielding and Bell 1997; Sangermano and Eastman 2007; Machado et al. 2011).
- From *spatial statistical analysis* come quantitative frameworks for the assessment of spatial dependence, spatial connectedness, spatial and statistical interpolation and extrapolation, and Bayesian frameworks of likelihood that may be used to infer conditions beyond or within the range of verifiable cases (Legendre 1993; Miller 2003; Betts et al. 2006; Goodchild et al. 2007)

When using a parametric classification method to delineate discrete categories among the continuous range of variability in spatial data, the intermediate "soft-classified" products may represent the probability or proportion of class membership within the given mapping unit. If prepared as independent measures of class membership, these become possibilistic metrics that can yield information about the contextual knowledge of the overall classification scheme, in addition to the location under investigation (Foody 2004). Using this framework, the discrete category ascribed a particular feature may be defined by rules based on a clear majority or obvious plurality of membership. However, in situations in which there are multiple classes with high membership values, there is *ambiguity* about the best class definition for a feature, which may imply that there is insufficient information to distinguish among the classes of a given legend scheme (Leyk et al. 2005). In situations in which no class in a chosen legend has a high membership value for a particular feature, there is *vagueness*, which may imply that the legend scheme lacks an appropriate thematic class to adequately categorize the feature in question (Fisher 1999). Quantitatively, metrics incorporating these possibilistic memberships can be used to generate indices of *classification uncertainty* (Eastman 2009) to describe the performance of the particular classifier and legend scheme. In the comparison of the resulting map information to independent reference data or another categorical map, the compounded uncertainty can lead to *confusion* in the analysis, which may be quantified to properly contextualize and verify the resulting estimates of error or change (Christman 2010).

The common issues in assessing the validity of spatial information have similar conditions regarding the use of VGI (after Chrisman 1999, Bolstad 2010). The most commonly raised issue pertains to the assessment of *attribute error*, in which a user must assess whether the thematic type ascribed to a feature is correct; the complementary question

ascribed VGI is whether the information is true and trustworthy. Further, the evaluation of *positional error* within conventional accuracy assessments directly relates to whether locations are correctly depicted in the VGI data. Beyond the precision of the feature representations, there is a question of *logical consistency*, in which all features are depicted with the same rules across the data set; within the VGI assessment framework, this relates to whether the magnitudes of values are comparable across space, time, or categories. Similarly, the need for *completeness* in the dataset, in which all qualifying features are depicted raises issues in VGI of where or in what regions aren't features represented in the dataset. Finally, in the more qualitative assessment of the *credibility, aesthetic,* or *conceptual effectiveness* of the spatially represented phenomenon, one might ask of both conventional geospatial data and VGI: is the message of the map clear, and does the information set comprehensively describe the situation to the intended audience?

In the context of VGI, these frameworks can yield important insights regarding the data collection scheme and cases of conflicting or incomplete knowledge about the conditions at a particular location (c.f. Goodchild and Glennon 2010). For example, in the use of VGI in the aftermath of a natural disturbance, as is often the case with the implementations like the crowdmapping platform Ushahidi (http://ushahidi.com) by members of both the public and geospatial community to map "help needed" or "help offered" scenarios, several cautionary lessons could be crafted using this framework. First, in the context of many mapped VGI feature entries, in which there is a majority of entries of one kind or the other, the resulting message is clear: the situation is either dire or stable at a given location. By contrast, in regions in which few VGI feature entries have been recorded, there may be insufficient information to draw a conclusion; importantly, one must maintain caution in interpreting a lack of negative responses as positive (perhaps responses are not possible). Further, one must also exercise caution in relating population density to the density of VGI feature entries, as the situations of vagueness in either case may yield important information, such as isolated rural residents. Finally, in situations of ambiguity, regardless of the density of feature entries, additional information may be critical to correctly categorize the situation.

Volunteered Geographic Information has enabled a broad population to contribute to new datasets previously unavailable or requiring intensive preparation. While the sources and frameworks for collecting and managing these data are rapidly evolving, many methods for evaluation and interpretation of these data exist within the range of analogous geospatial analytical techniques. In this context, frameworks that enable the consideration of the credibility of both the data and the categorization scheme can be an important tool in generating timely insights from these new sources.

References

Betts et al. 2006. The importance of spatial autocorrelation, extent, and resolution in predicting forest bird occurrence. *Ecological Modeling* 191: 197-224.

Bolstad, P. 2012. *GIS Fundamentals. 4th Edition.* St. Paul: Eider Press

Chrisman, N. 1999. "The Error Component in Spatial Data"Chapter 12 in *Geographical Information Systems,* Ed. McGuire, D. J., Goodchild, M. F., and Rhine, D. W. New York: John Wiley & Sons, Ltd.

Christman, Z. 2010. *Land Change in Central Mexico: Landscape Heterogeneity, Natural Variability, and Classification Uncertainty.* Ph.D dissertation, Clark Unniversity Graduate School of Geography, Worcester, Massachusetts.

Eastman, J. R. 2009. *Idrisi Taiga Guide to GIS and Image Processing.* Worcester, MA: Clark Labs.

Fielding, A. and Bell, J. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24(1): 38-49.

Fisher, P. 1999. "Models of uncertainty in spatial data," Chapter 13 in *Geographical Information Systems.* Ed. McGuire,D. J., Goodchild, M. F., and Rhine, D. W. New York: John Wiley & Sons, Ltd.

Foody, G. 2004. "Sub-pixel Methods in Remote Sensing" in *Remote Sensing Image Analysis: Including the Spatial Domain,* ed. S. de Jong and F. van der Meer. Dordrecht: Kluwer Academic Publishers.

Foody, G. 2010. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment* 114: 2271-2285.

Goodchild, M., et al. 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science* 21(3) 239-260.

Goodchild, M., and Glennon, J. A. 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth* 3(3): 231-241.

Herold, M, et al. 2008. Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. *Remote Sensing of Environment* 112(5): 2538-2556.

Lambin, E., et al. 2011. Global land use change, economic globalization, and the looming land scarcity. *Proceedings of the National Academy of Sciences* 108(9):3465-3472.

Leyk, S. et al. 2005. A Conceptual Framework for Uncertainty Investigation in Map-based Land Cover Change Modeling. *Transactions in GIS* 9(3):291-322

Legendre, P. 1993. Spatial Autocorrelation: Trouble or New Paradigm? *Ecology* 74(6): 1659-1673.

Machado, E. 2011. Empirical mapping of suitability to dengue fever in Mexico using species distribution modeling. *Applied Geography* 33: 82-93.

Miller, J. et al. 2007. Incorporating spatial dependence in predictive vegetation models. *Ecological Modeling* 202: 225-242.

Miller, J. and Rogan, J. 2007. "Using GIS and remote sensing for ecological mapping and monitoring" in *Integration of GIS and Remote Sensing,* ed. V. Mesev. New York: John Wiley & Sons, Ltd.

Sangermano, F. and Eastman, J. R. 2007. "Linking GIS and Ecology—The Use of Mahalanobis Typicalities to Model Species Distribution" in *Memorias XI Conferencia Iberoamerica de Sistemas de Información Geográfica,* G. D Buzai (Ed.), pp. 1-13. Luján: Universidad Nacional de Luján.