

UCGIS White Paper 2003
Geospatial Intelligence: Data Integration and Management Issues
Sucharita Gopal
Department of Geography &
Center for Cognitive and Neural Systems
Boston University Boston MA 02215
suchi@bu.edu

Geospatial data and technology is required in every phase of disaster management –prevention, mitigation, response and recovery (Sorenson, 1990). Disaster management requires the full integration of and harnessing of what has been called "the distributed intelligence of the Nation." This research paper highlights two issues related to knowledge management and acquisition using a four faceted approach to knowledge management.

Background: Since 9/11, there is a widespread official concern over the proliferation of biological weapons of mass destruction. One official US estimate notes that the number of countries having or suspected of having, offensive biological weapons programs has increased from 4 in the early 1970's to 10 in 1999 and more in 2001. The following features of biological weapons are worth highlighting: first, biological weapons can have a spatial coverage of impact equivalent to nuclear weapons making them equally effective as weapons of mass destruction. Second, biological weapons have environmental consequences. They can be used in many different ways in a variety of scenarios ranging from contaminating food or water supplies, insects spreading the deadly diseases, deadly toxins to dissemination through the air and inhalation into lungs. Third, a biological attack (using agents such as anthrax, brucellosis, plague and Q fever) can go unnoticed for days, until people start coming down with symptoms (Zoon, 1999). Thus even a small-scale terrorist attack involving chemical or biological substances could cause widespread damage and panic amongst citizens, especially if the state and communities are unprepared.

Four Faceted Knowledge Based Framework: Unlike commercial organizations where needs for data mining and knowledge discovery can be articulated in clear terms and also sufficiently in advance, the set of applications for global security and risk is characterized by the need to respond very quickly to new situations, each of which may require analysis of very large amounts of information. There is an urgent need for approaches oriented to deal very efficiently with such "ad hoc" needs that may require intelligent integration of multiple techniques and data. Gupta (2001) has proposed a four-pronged knowledge-based approach in this context:

- Knowledge Management: This involves access to data from heterogeneous sources and heterogeneous media, including traditional media such as paper and microfiche (Gupta 2001; Carpenter et al., 1997).
- Knowledge Acquisition: This involves integration and reconciliation of material from different sources, domains, and non-text media types (Liu et al. 2001) using fusion, conflation and other techniques.
- Knowledge Discovery: This includes the use of statistical and neural network-based data mining and other sophisticated operations on large sets of data (Reyes et al. 1998). Spatial analysis and spatial data mining (Shekhar 2003) are different from classical data mining and traditional statistics.

- **Knowledge Dissemination:** This involves the creation of advanced prototypes and research tools to apply knowledge in diverse applications. The prototype system should be able to work effectively in wide range of spatial and temporal scales.

Case Study - Geospatial Data and Analysis: Geospatial data is required in every stage of disaster management. The Persian Gulf War study serves to illustrate some practical issues concerning data availability, data fusion, and modeling. Postwar studies point out that veterans deployed in the Persian Gulf region during the Persian Gulf War have a higher chance of reporting health symptoms and medical problems than veterans deployed elsewhere outside of the Gulf region (Proctor et al 1998, Smith et al., 2002). Potential reasons for these unexplained illnesses include exposures to toxic substances during deployment, or other deployment-related experiences. Relevant datasets for the analysis of the Gulf War syndrome include:

- Troop deployment data including location, time and movement;
- Veterans demographic and exposure data (including self-reported environmental exposures and health symptoms);
- Environmental hazards include fumes, smokes from military operations, oil well fires, diesel exhaust, paints, pesticides, sand, depleted uranium, insect repellents, infectious agents, chemoprophylactic agents, immunizations, and possibly chemical or biological weapons;
- Weather conditions such as temperature, prevailing wind, precipitation, and humidity that are input into a *plume model* to estimate the dispersion pattern of chemical agent fallout from the demolition of the Iraqi chemical weapons depot at Khamisiyah in southern Iraq, which was destroyed by U.S. troops in 1991. The DoD (in 2000) estimated that 100,045 veterans had been exposed to low levels of the nerve agent sarin after the demolition.
- Field and satellite data: using remotely sensing and field validation data to map areas of change detection and model pollutant concentrations, plume boundaries, potential oil field emission rates etc.

There were numerous challenges in this study of environmental exposure to chemical agents including differing spatio-temporal resolutions, uncertainty in health symptom and troop deployment data, lack of specific symptom data at the individual level, time lapse between exposure and reporting of symptoms, and downscaling regional satellite weather data to local scales (Proctor et al., 2003). What is required in this context is a flexible approach to handle and analyze large volumes of data and integrate data from heterogeneous sources.

Ability to handle and analyze huge volumes of data: The design processes that develop the technology to analyze huge volumes of data are becoming increasingly vital. The development of Knowledge Discovery in Databases (KDD) has been spurred by the exponential increase in digital data. As an example, NASA's EOS (Earth Observing System) includes a coordinated series of polar-orbiting and low-inclination satellites for long-term global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans. ESE is currently managing over 2 petabytes of Earth System Science data and information. Over 2 Terabytes of science data from NASA's Earth observing satellites are added to the archives each day; archived volume of data has doubled each year over the past three years.

When the data are genuinely high dimensional (e.g. some remote sensing data), there is little hope to reliably identify its structure unless sample sizes are extremely high. It is often, however, the case that "most of the structure" is concentrated on a lower-dimensional subspace, which makes the modeling task somewhat more manageable. So, the usual way to cope with the problem of high dimensionality and to simplify estimation and interpretation of results has been to reduce the complexity of the model by imposing structural assumptions. What is needed is a method of

dimensionality reduction, which, on one hand, is free from rigid structural assumptions about the model and, on the other hand, is not as susceptible to the curse of dimensionality of the pointwise estimation of genuinely multivariate functions. Such issues have received very little attention in the geosciences. Some research questions include:

- What are techniques for data reduction and compression ("Data Squashing") that allow for fast real time processing by data-mining techniques?
- Image databases will require some form of automated or semi-automated processing. Automatic generation of image content descriptors in this domain will involve combining image and textual database concepts (lexicon/metadata) into a single system. What are efficient algorithms that can be applied in this context?
- Increasingly multiscale datasets are being generated. How should geospatial features be defined and stored in such databases for subsequent data mining? (Fay et al., 2000; Ross et al., 2000)
- What are relevant geospatial features or concepts in a lexicon relevant in the homeland security context? For e.g. urban regions, roads, water bodies, airports, subway stations, and vehicles. How does one automatically extract such features from images useful for automated feature detection and segmentation? (Waxman et al., 2000).
- What are possible effective visualization and validation tools in dealing with large databases?

Managing and integrating data from heterogeneous sources and heterogeneous media: The challenge is the integration of increasingly large heterogeneous data provided by the Earth Observing System (EOS) of NASA's Earth Science Enterprise (ESE), NOAA, NCAR, USGS, EPA, CDC and others. At present, most geospatial data are archived within numerous meta-data centers such as NOAA's National Climatic Data Center, NASA's Global Change Master Directory, NCAR/UCAR's Community Data Portal, and the International Research Institute/Lamont-Doherty Earth Observatory's Climate Data Library. These centers allow access to complete or partial data collected from numerous Earth system observations, experiments, and simulations. However, Earth system science data is heterogeneous with regard to spatio-temporal scale, phenomenology, and format. The goal is to apply techniques from statistics, Artificial Neural Networks (ANN) and other disciplines to create a very sophisticated repertoire of tools that can model non-stationary and non-linear behavior with high accuracy in *diverse applications* with minimal changes. One of the important hybrid-AI techniques is the combination of neural networks and fuzzy logic, which can potentially offer all the benefits of both techniques and is an emergent and very promising field of research. The so-called neuro-fuzzy systems have been successfully applied to pattern recognition (Carpenter et al., 1999; Abulelgasim et al., 1999).

Hybrid systems offer great potential for managing different pieces of knowledge within an application. Figure 1 shows the uncertainty associated with continental landcover classification using MODIS data. Classification results from decision trees and neural networks were pooled together on a pixel-by-pixel basis using boosting (decision trees) and voting (fuzzy ARTMAP). The results indicate that the hybrid approach provides additional information on uncertainty that could not be obtained using one single classifier alone. A disagreement amongst classifiers (for e.g. around Appalachia, Gulf coast) indicates mixed pixel problem, lack of good training data, and other problems.

The non-precise and context-dependent component of the knowledge is represented using fuzzy logic, while the deep knowledge is represented more efficiently using neural networks. These ensemble or hybrid systems can be designed to be very robust, and perform well even in environments characterized by noisy data. This leads to questions like:

- Geospatial information becomes more useful when data from several sensors is fused. What are general principles that can guide data fusion?
- How should one *combine traditional (statistical) and emerging techniques* (such neural networks) for better results using integrated geospatial databases? How should the forecast space be divided into distinct domains where one method is superior to the other, or should the models be assigned “soft” likelihoods in any domain of interest?
- For many data mining applications, classification and clustering are required. There is a scarcity of labeled data for classification problem. How does one generate a reasonable sized training set?
- How should these tools, or some probabilistic combination thereof, be applied for *extracting knowledge* from large databases about which prior knowledge is limited, and where the data are inherently noisy, nonlinear, and non-stationary?
- How should spatial uncertainty and inherent fuzziness of spatial data be represented in large-scale databases? In addition, how can these concepts be used from the perspective of end user’s decision-making?
- What are innovative approaches for building knowledge repositories (e.g. P2P systems) that support large set of users in this context?
- In the context of homeland security, there is also a need for interactive mining tools including image map interfaces that can provide the analyst easy way to mine the integrated database. How can the user perform real time mining of imagery and other data?

Geospatial data modeling in the context of homeland security requires the creation and use of databases of increasingly heterogeneous sources and heterogeneous media. To mine and analyze these complex databases effectively, demands the continual development and application of innovative methodologies that exploit advances in new technologies.

References:

Abuelgasim, A., Ross, W. D., Gopal, S., and Woodcock, C. E. (1999). Change detection using adaptive neural networks: Environmental damage assessment after the Gulf War, *Remote Sensing of the Environment*, 70 (2), 208-223.

Carpenter, G., Gjaja, M., Gopal, S., and Woodcock, C. (1997). ART networks in Remote Sensing, *IEEE Transactions on Geoscience and Remote Sensing*, 35(2), 308-325.

Carpenter, G., Gopal, S., Martens, S., and Woodcock, C. (1999). A Neural Network Method for Mixture Estimation for Vegetation Mapping, *Remote sensing of the Environment*, 70 (2), 138-152.

Fay, D.A., Waxman, A.M., Aguilar, M., Ireland, D.B., Racamato, J.P., Ross, W.D., Streilein, W.W., Braun, M.I. (2000). Fusion of Multi-Sensor Imagery for Night Vision: Color Visualization, Target Learning and Search. In *Proceedings of 3rd International Conference on Information Fusion*, Paris, France.

Gupta, A. (2001). A Four-Faceted Knowledge Based Approach to Surmounting National and Other Borders. *The Journal of Knowledge Management*, Volume 5, No. 4, 2001.

Liu, W., Gopal, S., and Woodcock, C. (2001). Spatial data mining for classification, visualization and interpretation with ARTMAP neural networks, in Robert L. Grossman (Editor), et al *Data mining for scientific and engineering applications*, Kluwer Academic Publishers, Netherlands.

Proctor S P, T Heeren, R F White, J Wolfe, M S Borgos, J D Davis, L Pepper, R Clapp, P B Sutker, J J Vasterling, and D Ozonoff (1998). Health status of Persian Gulf War veterans: self-reported symptoms environmental exposures and the effect of stress. *International Epidemiological Association* 27: 1000-1010

Proctor S.P, Gopal, S., Imai, A., White, R., Wolfe, J., and Ozonoff, D. (2003). Spatial analysis of Gulf War troop location data in relationship with symptom reports using GIS techniques. Manuscript submitted to *Transactions in GIS*.

Reyes, C., Ganguly, A., Lemus, G., and Gupta, A. (1998). A Hybrid Model Based on Dynamic Programming, Neural Networks, and Surrogate Value for Inventory Optimization Applications, *Journal of the Operational Research Society*, Vol. 49, 1998, pp. 1-10.

Ross, W.D., Waxman, A.M., Streilein, W.W., Aguilar, M., Verly, J., Liu, F., Braun, M.I., Harmon, P., Rak, S. (2000). Multi-Sensor 3D Image Fusion and Interactive Search. In *Proceedings of 3rd International Conference on Information Fusion*, Paris, France..

Shekhar, S. (2003). Spatial data mining.
<http://www.ahpcrc.org/publications/archives/v12n2/Story4/>

Sorensen, J. (1990). 'Society and emergency preparedness' in Kirby, A. (Ed.) *Nothing to Fear: Risks and Hazards in American Society*, pp. 241-260. Tucson, Ariz.: University of Arizona Press.

Smith T.C., Heller J.M., Hooper T.I., Gackstetter G.D., Gray G. (2002). Are Gulf War veterans experiencing illness due to exposure to smoke from Kuwaiti oil well fires? Examination of Department of Defense hospitalization data. *American Journal of Epidemiology*, 155(10):908-917. May 2002.

Streilein and Waxman, A. (2000). Fused multi-sensor image mining for feature foundation data. *Proceedings of the Third International Conference on Information Fusion (FUSION 2000)*, Volume: 1, 2000.

Zoon, K. (1999). Vaccines, pharmaceutical products and bioterrorism: Challenges for the USFDA, *Emerging Infectious Diseases*, 5(4), <http://www.cdc.gov/ncidod/EID/vol5no4/zoon.htm>.

Source: Liu, W., Gopal, S., and Woodcock (2003). Hybrid approaches to landcover mapping. Manuscript submitted to PE & RS.

