Bayesian Network Integration with GIS

Chapter · January 2008

DOI: 10.1007/978-0-387-35973-1_95

2 authors:

Daniel P Ames
Brigham Young University - Provo Main Campus

Allen Anselmo

Balanced Box Decomposition Tree (Spatial Index)

► Nearest Neighbors Problem

Bayesian Estimation

► Indoor Positioning, Bayesian Methods

Bayesian Inference

► Hurricane Wind Fields, Multivariate Modeling

Bayesian Maximum Entropy

► Uncertainty, Modeling with Spatial and Temporal

Bayesian Network Integration with GIS

Daniel P. Ames, Allen Anselmo
Department of Geosciences, Geospatial Software Lab, Idaho State University, Pocatello, ID, USA

Synonyms

Directed acyclic graphs; Probability networks; Influence diagrams; Probabilistic map algebra; Spatial representation of bayesian networks

Definition

A Bayesian Networks (BN) is a graphical-mathematical construct used to probabilistically model processes which include interdependent variables, decisions affecting those variables, and costs associated with the decisions and states of the variables. BNs are inherently system representations and, as such, are often used to model environmental processes. Because of this, there is a natural connection between certain BNs and GIS. BNs are represented as a directed acyclic graph structure with nodes (representing variables, costs, and decisions) and arcs (directed lines representing conditionally probabilistic dependencies between the nodes). A BN can be used for prediction or analysis of real world problems and complex natural systems where statistical correlations can be found between variables or approximated using expert opinion. BNs have a vast array of applications for aiding decision making in areas such as medicine, engineering, natural resources, and decision management. BNs can be used to model geospatially interdependent variables as well as conditional dependencies between geospatial layers. Additionally, BNs have been found to be useful and highly efficient in performing image classification on remotely sensed data.

Historical Background

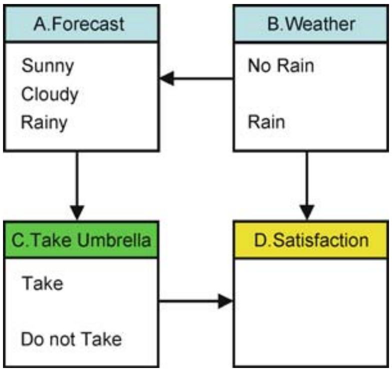
Originally described by Pearl (1988), BNs have been used extensively in medicine and computer science (Heckerman 1997). In recent years, BNs have been applied in spatially explicit environmental management studies. Examples include: the Neuse Estuary Bayesian ecological response network (Borsuk and Reckhow 2000), Baltic salmon management (Varis and Kuikka 1996), climate change impacts on Finnish watersheds (Kuikka and Varis 1997), the Interior Columbia Basin Ecosystem Management Project (Lee and Bradshaw 1998), and waterbody eutrophication (Haas 1998). As illustrated in these studies, a BN graph structures a problem such that it is visually interpretable by stakeholders and decision-makers while, serving as an efficient means for evaluating the probable outcomes of management decisions on selected variables.

Both BNs and GIS can be used to represent spatially explicit, probabilistically connected environmental and other systems, however the integration of the two techniques has only been explored relatively recently. BN integration with GIS typically takes one of four distinct forms: 1) BN-based layer combination (i. e. probabilistic map-algebra) as demonstrated in Taylor (2003); 2) BN-based classification as demonstrated in Stassopoulou et al. (1998) and Stassopoulou and Caelli (2000); 3) Using BNs for intelligent, spatially-oriented data retrieval, as demonstrated in Walker et al. (2004) and Walker et al. (2005); and 4) GIS-based BN decision support system (DSS) frameworks where BN nodes are spatially represented in a GIS framework as presented by Ames (2005).

Scientific Fundamentals

As noted above, BNs are used to model reality by representing conditional probabilistic dependencies between interdependent variables, decisions, and outcomes. This section provides an in-depth explanation of BN analysis using an example BN model called the "Umbrella" BN (Fig. 1), an augmented version of the well-known "Weather" influence diagram presented by Shachter and Peot (1992). This simple BN attempts to model the variables and outcomes associated with the decision to take or not take an umbrella on a given outing. This problem is represented in the BN by four nodes. "Weather" and "Forecast" are nature or chance nodes where "Forecast" is conditioned on the state of "Weather" and "Weather" is treated as a random variable with a prior probability distribution based on historical conditions. "Take Umbrella" is a decision variable that, together with the "Weather" variable defines the status of "Satisfaction". The "Satisfaction" node is known as a "utility" or "value" node. This node associates a resultant outcome value (monetary or otherwise) to represent the satisfaction of the individual based on the decision to take the umbrella and whether or not there is rain. Each of these BN nodes contains discrete states where each variable state represents abstract events, conditions, or numeric ranges of each variable.

The Umbrella model can be interpreted as follows: if it is raining, there is a higher probability that the forecast will predict it will rain. In reverse, through the Bayesian network "backward propagation of evidence" if the forecast predicts rain it can be inferred that there is a higher chance that rain will actually occur. The link between "Forecast" and "Take Umbrella" indicates that the "Take Umbrella" decision is based largely on the observed forecast. Finally, the link to the "Satisfaction" utility node from both "Take Umbrella" and "Weather" captures the relative gains in satisfaction derived from every combination of states of the BN variables.



Bayesian Network Integration with GIS, Figure 1 Umbrella Bayesian Decision Network Structure. *A* and *B* nature nodes, *C* a decision node, and *D* a utility node

Bayesian networks are governed by two mathematical techniques: conditional probability and Bayes' theorem.

Conditional probability is defined as the probability of one event given the occurrence of another event and can be calculated as the joint probability of the two events occurring divided by the probability of the second event:

$$P(A|B) = \frac{P(A, B)}{P(B)} . \tag{1}$$

From Eq. 1, the fundamental rule for probability calculus and the downward propagation of evidence in a BN can be derived. Specifically, it is seen that the joint probability of A and B equals the conditional probability of event A given B, multiplied by the probability of event B (Eq. 2).

$$P(A, B) = P(A|B) \cdot P(B) . \tag{2}$$

Equation 2 is used to compute the probability of any state in the Bayesian network given the states of the parent node events. In Eq. 3, the probability of state A_x occurring given parent B is the sum of the probabilities of the state of A_x given state B_i, with i being an index to the states of B, multiplied by the probability of that state of B.

$$P(A_x, B) = \sum_i P(A_x|B_i) \cdot P(B_i) . \tag{3}$$

Similarly, for calculating states with multiple parent nodes, the equation is modified to make the summation of the conditional probability of the state A_x given states B_i and C_j multiplied by the individual probabilities of B_i and C_j.

$$P(A_x, B, C) = \sum_{i,j} P(A_x|B_i, C_j) \cdot P(B_i) \cdot P(C_j) . \tag{4}$$

Finally, though similar in form, utility nodes do not calculate probability, but instead calculate the utility value as a metric or index given the states of its parent or parents as shown in Eqs. 5 and 6.

$$U(A, B) = \sum_i U(A|B_i) \cdot P(B_i) \tag{5}$$

$$U(A, B, C) = \sum_{i,j} U(A|B_i, C_j) \cdot P(B_i) \cdot P(C_j) . \tag{6}$$

The second equation that is critical to BN modeling is Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} . \quad (7)$$

The conditional probability inversion represented here allows for the powerful technique of Bayesian inference, for which BNs are particularly well suited. In the Umbrella model, inferring a higher probability of a rain given a rainy forecast is an example application of Bayes' Theorem.

Connecting each node in the BN is a conditional probability table (CPT). Each nature node (state variable) includes a CPT that stores the probability distribution for the possible states of the variable given every combination of the states of its parent nodes (if any). These probability distributions can be assigned by frequency analysis of the variables, expert opinion based on observation or experience, or they can be set to some "prior" distribution based on observations of equivalent systems.

Tables 1 and 2 show CPTs for the Umbrella BN. In Table 1, the probability distribution of rain is represented as 70% chance of no rain, and 30% chance of rain. This CPT can be assumed to be derived from historical observations of the frequency of rain in the given locale. Table 2, represents the probability distribution of the possible weather forecasts ("Sunny", "Cloudy", or "Rainy") conditioned on the actual weather event. For example, when actually rained, the prior forecast called for "Rainy" 60% of the time, "Cloudy" 25% of the time, and "Sunny" 15% of the time. Again, these probabilities can be derived from historical observations of prediction accuracies or from expert judgment.

Bayesian Network Integration with GIS, Table 1 Probability of rain

Weather	
No Rain	**Rain**
70%	30%

Bayesian Network Integration with GIS, Table 2 Forecast probability conditioned on rain

	Forcast		
Weather	**Sunny**	**Cloudy**	**Rainy**
No Rain	70%	20%	10%
Rain	15%	25%	60%

Bayesian Network Integration with GIS, Table 3 Satisfaction utility conditioned on rain and the "Take Umbrella" decision

	Satisfaction	
Weather	**Take Umbrella**	**Satisfaction**
No Rain	Take	20 units
No Rain	Do not Take	100 units
Rain	Take	70 units
Rain	Do not Take	0 units

Table 3 is a utility table defining the relative gains in utility (in terms of generic "units of satisfaction) under all of the possible states of the BN. Here, satisfaction is highest when there is no rain and the umbrella is not taken and lowest when the umbrella is not taken but it does rain. Satisfaction "units" are in this case assigned as arbitrary ratings from 0 to 100, but in more complex systems, utility can be used to represent monetary or other measures.

Following is a brief explanation of the implementation and use of the Umbrella BN. First it is useful to compute $P(\text{Forecast} = \text{Sunny})$ given unknown Weather conditions as follows:

$$\begin{aligned} &P(\text{Forecast} = \text{Sunny}) \\ &\quad = \sum_{i=\{\text{NoRain},\text{Rain}\}} P(\text{Forecast} = \text{Sunny}|\text{Weather}_i) \cdot P(\text{Weather}_i) \\ &\quad = 0.7 \cdot 0.7 + 0.15 \cdot 0.3 = 0.535 = 54\% . \end{aligned}$$

Next $P(\text{Forecast} = \text{Cloudy})$ and $P(\text{Forecast} = \text{Rainy})$ can be computed as:

$$\begin{aligned} &P(\text{Forecast} = \text{Cloudy}, \text{Weather}) \\ &\quad = 0.2 \cdot 0.7 + 0.25 \cdot 0.3 = 0.215 = 22\% \\ &P(\text{Forecast} = \text{Cloudy}, \text{Weather}) \\ &\quad = 0.1 \cdot 0.7 + 0.6 \cdot 0.3 = 0.25 = 25\% . \end{aligned}$$

Finally, evaluate the "Satisfaction" utility under both possible decision scenarios (take or leave the umbrella).

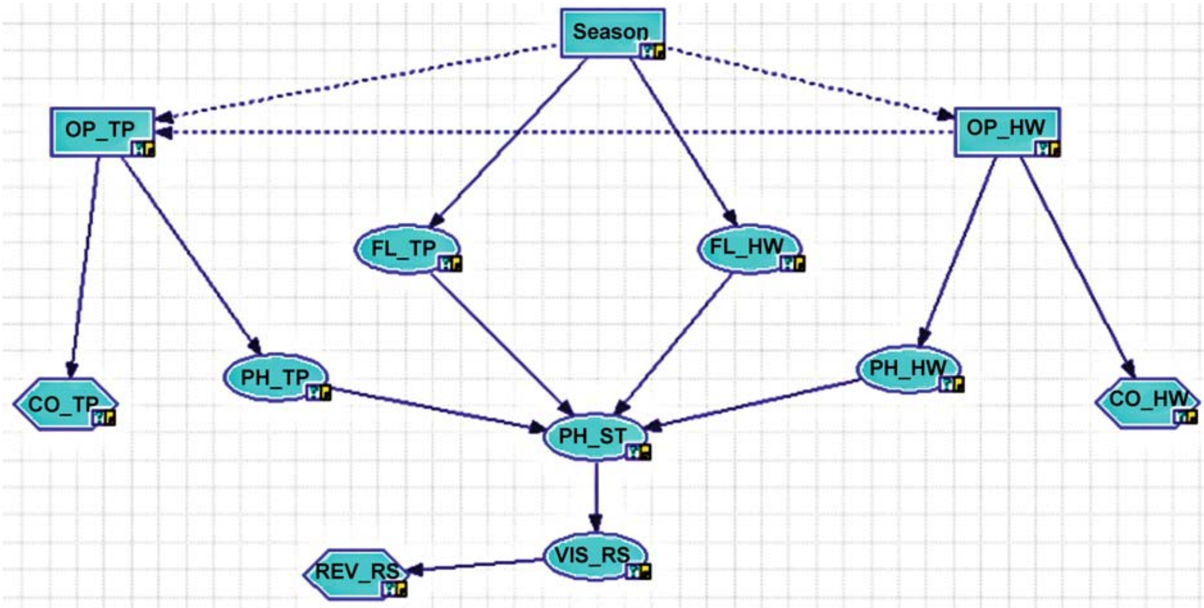
$$\begin{aligned} &U(\text{Satisfaction}|\text{TakeUmbrella} = \text{Take}) \\ &= \sum_{i,j} U(\text{Satisfaction}|\text{TakeUmbrella}, \text{Weather}_j) \\ &\qquad\qquad \cdot P(\text{TakeUmbrealla}_i) \cdot P(\text{Weather}_j) \\ &= 20 * 1.0 * 0.7 + 100 * 0.0 * 0.7 + 70 * 1.0 * 0.3 \\ &\quad + 0 * 0.0 * 0.3 = 35 . \end{aligned}$$

Similarly, the utility of not taking the umbrella is computed as:

$$\begin{aligned} &U(\text{Satisfaction}, \text{TakeUmbrella} = \text{NoTake}, \text{Weather}) \\ &= 20 * 0.0 * 0.7 + 100 * 1.0 * 0.7 + 70 * 0.0 * 0.3 \\ &\qquad + 0 * 1.0 * 0.3 = 70 \end{aligned}$$

Clearly, the higher satisfaction is predicted for leaving the umbrella at home, thereby providing an example of how a simple BN analysis can aid the decision-making process. While the Umbrella BN presented here is quite simple and not particularly spatially explicit, it serves as a generic BN

B



Bayesian Network Integration with GIS, Figure 2 The East Canyon Creek BDN from Ames et al. (2005), as seen in the GeNIe (Decision Systems Laboratory 2006) graphical node editor application

example. Specific application of BNs in GIS is presented in the following section.

Key Applications

As discussed before, integration of GIS and BNs is useful in any BN which has spatial components, whether displaying a spatially-oriented BN, using GIS functionality as input to a BN, or forming a BN from GIS analysis. Given this, the applications of such integration are only limited by that spatial association really. One example mentioned above of such a spatial orientation was showed usefulness of a watershed management BN, but there are other types of BNs which may benefit from this form of integration. For instance, many ecological, sociological, and geological studies which might benefit from a BN also could have strong spatial associations. Another example might be that traffic analyses BNs have very clear spatial associations often. Finally, even BNs trying to characterize the spread of diseases in epidemiology would likely have clear spatial association.

As outlined above, GIS-based BN analysis typically takes one of four distinct forms including:

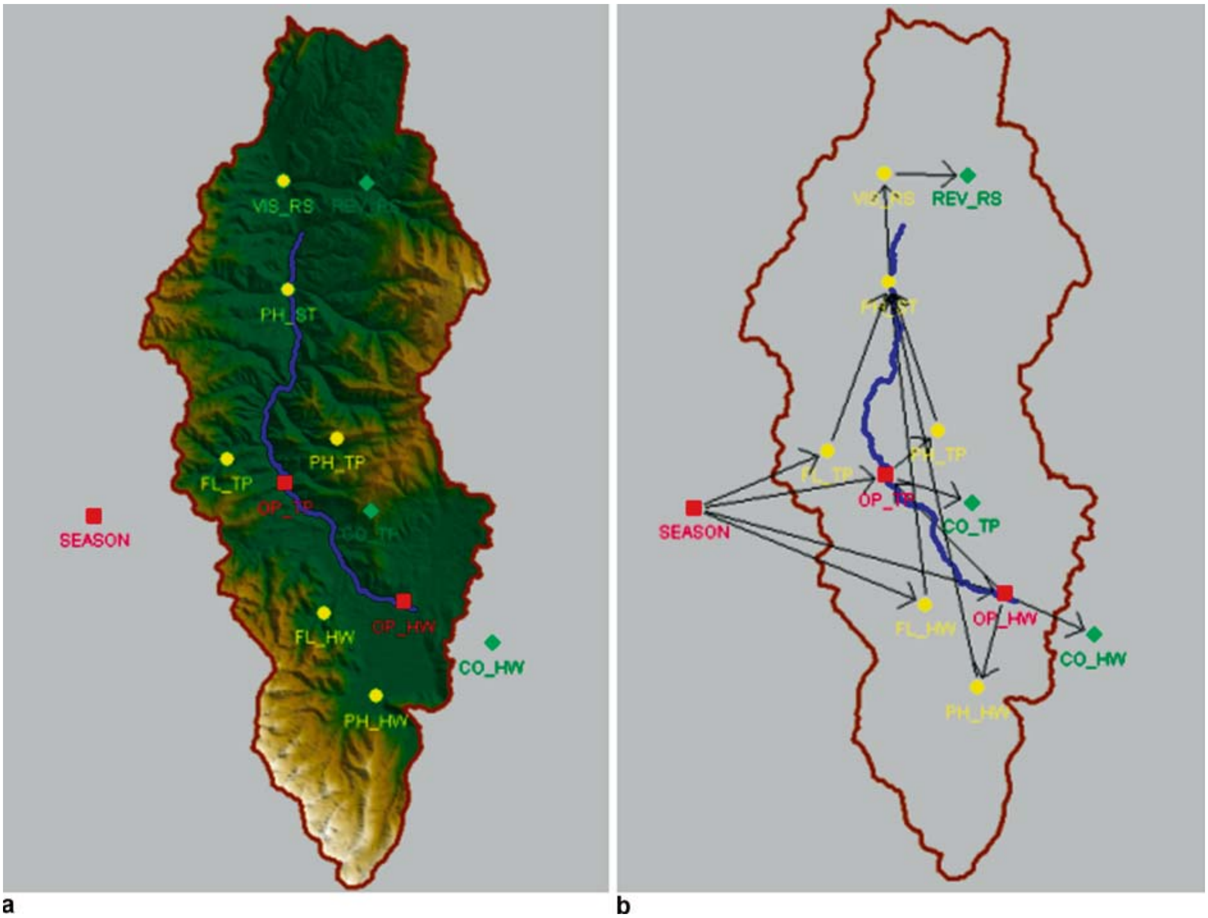
- Probabilistic map-algebra
- Image classification
- Automated data query and retrieval
- Spatial representation of BN nodes

A brief explanation of the scientific fundamentals of each of these uses is presented here.

Probabilistic Map Algebra

Probabilistic map algebra involves the use of a BN as the combinatorial function used on a cell by cell basis when combining raster layers. For example, consider the ecological habitat models described by Taylor (2003). Here, several geospatial raster data sets are derived representing proximity zones for human caused landscape disturbances associated with development of roads, wells, and pipelines. Additional data layers representing known habitat for each of several threatened and endangered species are also developed and overlaid on the disturbance layers. Next a BN was constructed representing the probability of habitat risk conditioned on the both human disturbance and habitat locations. CPTs in this BN were derived from interviews with acknowledged ecological experts in the region. Finally, this BN was applied on a cell by cell basis throughout the study area, resulting in a risk probability map for the region for each species of interest.

Use of BNs in this kind of probabilistic map algebra is currently hindered only by the lack of specialized tools to support the analysis. However, the concept holds significant promise as an alternative to the more traditional GIS based “indicator analysis” where each layer is reclassified to represent an arbitrary index and then summed to give a final metric (often on a 1 to 100 scale of either suitability or unsuitability). Indeed the BN approach results in a more interpretable probability map. For example, such an analysis could be used to generate a map of the probabil-



Bayesian Network Integration with GIS, Figure 3 **a** East Canyon displayed with the East Canyon BN overlain on it. **b** Same, but with the DEM layer turned off and the BN network lines displayed

ity of landslide conditioned on slope, wetness, vegetation etc. Certainly a map that indicates percent chance of landslide could be more informative for decision makers, than an indicator model that simply displays the sum of some number of reclassified indicators.

Image Classification

In the previous examples, BN CPTs are derived from historical data or information from experts. However, many BN applications make use of the concept of Bayesian learning as a means of automatically estimating probabilities from existing data. BN learning involves a formal automated process of "creating" and "pruning" the BN node-arc structure based on rules intended to maximize the amount of unique information represented by the BN CPTs. In a GIS context, BN learning algorithms have been extensively applied to image classification problems. Image classification using a BN requires the identification of a set of input layers (typically multispectral or hyperspectral bands) from which a known set of objects or classifications are to be identified.

Learning datasets include both input and output layers where output layers clearly indicate features of the required classes (e. g. polygons indicating known land cover types). A BN learning algorithm applied to such a data set will produce a optimal (in BN terms) model for predicting land cover or other classification scheme at a given raster cell based on the input layers. The application of the final BN model to predict land cover or other classification at an unknown point is similar to the probabilistic map algebra described previously.

Automated Data Query and Retrieval

In the case of application of BNs to automated query and retrieval of geospatial data sets, the goal is typically to use expert knowledge to define the CPTs that govern which data layers are loaded for visualization and analysis. Using this approach in a dynamic web-based mapping system, one could develop a BN for the display of layers using a CPT that indicates the probability that the layer is important, given the presence or absence of other layers or features within layers at the current view extents. Such a tool would supplant the typical approach which is to activate or deactivate layers based strictly on "zoom level". For example, consider a military GIS mapping system used to identify proposed targets. A BN-based data retrieval system could significantly optimize data transfer and bandwidth usage by only showing specific high resolution imagery when the probability of needing that data is raised due to the presence of other features which indicate a higher likelihood of the presence of the specific target.

BN-based data query and retrieval systems can also benefit from Bayesian learning capabilities by updating CPTs with new information or evidence observed during the use of the BN. For example, if a user continually views several datasets simultaneously at a particular zoom level or in a specific zone, this increases the probability that those data sets are interrelated and will should result in modified CPTs representing those conditional relationships.

Spatial Representation of BN Nodes

Many BN problems and analyses though not completely based on geospatial data have a clear geospatial component and as such can be mapped on the landscape. This combined BN-GIS methodology is relatively new but has significant potential for helping improve the use and understanding of a BN. For example, consider the East Canyon Creek BN (Ames et al. 2005) represented in Fig. 2. This BN is a model of streamflow (FL_TP and FL_HW) at both a wastewater treatment plant and in the stream headwaters, conditional on the current season (SEASON). Also the model includes estimates of phosphorus concentrations at the treatment plant and in the headwaters (PH_TP and PH_HW) conditional on season and also on operations at both the treatment plant (OP_TP) and in the headwaters (OP_HW). Each of these variables affect phosphorus concentrations in the stream (PH_ST) and ultimately reservoir visitation (VIS_RS). Costs of operations (CO_TP and CO_HW) as well as revenue at the reservoir (REV_RS) are represented as utility nodes in the BN.

Most of the nodes in this BN (except for SEASON) have an explicit spatial location (i. e. they represent conditions at a specific place). Because of this intrinsic spatiality, the East Canyon BN can be represented in a GIS with points indicating nodes and arrows indicating the BN arcs (i. e. Fig. 3). Such a representation of a BN within a GIS can give the end users a greater understanding of the context and meaning of the BN nodes. Additionally, in many cases, it may be that the BN nodes correspond to specific geospatial features (e. g. a particular weather station) in which case spatial representation of the BN nodes in a GIS can be particularly meaningful.

Future Directions

It is expected that research and development of tools for the combined integration of GIS and BNs will continue in both academia and commercial entities. New advancements in each of the application areas described are occurring on a regular basis, and represent and active and interesting study area for many GIS analysts and users.

Recommended Reading

1. Ames, D.P.: Bayesian Decision Networks for Watershed Management. Utah State University, Logan, UT (2002)
2. Ames, D.P., Neilson, B.T., Stevens, D.K., Lall, U.: Using Bayesian networks to model watershed management decisions: an East Canyon Creek case study. IWA Publishing, J. Hydroinform. (2005)
3. Borsuk, M.E., Reckhow, K.H.: Summary Description of the Neuse Estuary Bayesian Ecological Response Network (Neu-BERN). (2000) http://www2.ncsu.edu/ncsu/CIL/WRRI/neuseltm.html (Dec. 26, 2001)
4. Haas, T.C.: Modeling waterbody eutrophication with a Bayesian belief network. School of Business Administration, Working Paper. University of Wisconsin, Milwaukee, Wisconsin (1998)
5. Heckerman, D.: Bayesian networks for data mining. Data Mining and Knowledge Discovery, **1**, 79–119. MapWindow Open Source Team (2007). MapWindow GIS 4.3 Open Source Software, Accessed February 06, 2007 at the MapWindow Website: http://www.mapwindow.org/ (1997)
6. Kuikka, S., Varis, O.: Uncertainties of climate change impacts in Finnish watersheds: a Bayesian network analysis of expert knowledge. Boreal Environ. Res. **2**, 109–128 (1997)
7. Lee, D.C., Bradshaw, G.A.: Making Monitoring Work for Managers: Thoughts on a conceptual framework for improved monitoring within broad-scale ecosystem management. (1998) http://icebmp.gov/spatial/lee_monitor/preface.html (Dec. 26, 2001)
8. Norsys Software Corp.: Netica Bayesian Belief Network Software. Acquired from http://www.norsys.com/ (2006)
9. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco (1988)
10. Stassopoulou, A., Petrou, M., Kittler, J.: Application of a Bayesian network in a GIS based decision making system. Int. J. Geograph. Info. Sci. **12**(1) 23–45 (1998)
11. Stassopoulou, A., Caelli, T.: Building detection using Bayesian networks. Int. J. Pattern Recognit. Artif. Intell. **14**(6), 715–733 (2000)
12. Taylor, K.J.: Bayesian Belief Networks: A Conceptual Approach to Assessing Risk to Habitat, Utah State University, Logan, UT (2003)
13. Varis, O., Kuikka, S.: An influence diagram approach to Baltic salmon management. Proc. Conference on Decision Analysis for Public Policy in Europe, INFORMS Decision Analysis Society, Atlanta, GA (1996)
14. Walker, A., Pham, B., Maeder, A.: A Bayesian framework for automated dataset retrieval. In: Geographic Information Systems,

mmm, p. 138. 10th International Multimedia Modelling Conference (2004)
15. Walker, A., Pham, B., Moody, M.: Spatial Bayesian learning algorithms for geographic information retrieval. In: Proceedings 13th annual ACM international workshop on Geographic information systems, pp. 105–114. Bremen, Germany (2005)
16. Shachter, R., Peot, M.: Decision making using probabilistic inference methods. In: Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence, pp. 275–283 (1992)

Bayesian Spatial Regression

► Bayesian Spatial Regression for Multi-source Predictive Mapping

Bayesian Spatial Regression for Multi-source Predictive Mapping

ANDREW O. FINLEY[1], SUDIPTO BANERJEE[2]
[1] Department of Forestry and Department of Geography, Michigan State University, East Lansing, MI, USA
[2] Division of Biostatistics, School of Public Health, University of Minnesota, Mineapolis, MN, USA

Synonyms

Pixel-based prediction; Bayesian spatial regression; Spatial regression

Definition

Georeferenced ground measurements for attributes of interest and a host of remotely sensed variables are coupled within a Bayesian spatial regression model to provide predictions across the domain of interest. As the name suggests, multi-source refers to multiple sources of data which share a common coordinate system and can be linked to form sets of regressands or response variables, $y(\mathbf{s})$, and regressors or covariates, $\mathbf{x}(\mathbf{s})$, where the $\mathbf{s}$ denotes a known location in $\mathbb{R}^2$ (e. g., easting–northing or latitude–longitude). Interest here is in producing spatially explicit predictions of the response variables using the set of covariates. Typically, the covariates can be measured at any location across the domain of interest and help explain the variation in the set of response variables. Within a multi-source setting, covariates commonly include multitemporal spectral components from remotely sensed images, topographic variables (e. g., elevation, slope, aspect) from a digital elevation model (DEM), and variables derived from vector or raster maps (e. g., current or historic land use, distance to stream or road, soil type, etc.). Numerous methods have been used to map the set of response variables. The focus here is linking the $y(\mathbf{s})$ and $\mathbf{x}(\mathbf{s})$ through Bayesian spatial regression models. These models provide unmatched flexibility for partitioning sources of variability (e. g., spatial, temporal, random), simultaneously predicting multiple response variables (i. e., multivariate or vector spatial regression), and providing access to the full posterior predictive distribution of any base map unit (e. g., pixel, mulitpixel, or polygon). This chapter offers a brief overview of remotely sensed data which is followed by a more in-depth presentation of Bayesian spatial modeling for multi-source predictive mapping. Multi-source forest inventory data is used to illustrate aspects of the modeling process.

Historical Background

In 1970, the National Academy of Sciences recognized remote sensing as "the joint effects of employing modern sensors, data-processing equipment, information theory and processing methodology, communications theory and devices, space and airborne vehicles, and large-systems theory and practices for the purpose of carrying out aerial or space surveys of the earth's surface" [26] p1. In the nearly four decades since this definition was offered, every topic noted has enjoyed productive research and development. As a result, a diverse set of disciplines routinely use remotely sensed data including: natural resource management; hazard assessment; environmental assessment; precision farming and agricultural yield assessment; coastal and oceanic monitoring; fresh water quality assessment; and public health. Several key publications document these advancements in remote sensing research and application including *Remote Sensing of Environment*, *Photogrammetric Engineering and Remote Sensing*, *International Journal of Remote Sensing*, and *IEEE Transactions on Geoscience and Remote Sensing*.

With the emergence of highly efficient Geographical Information Systems (GIS) databases and associated software, the modeling and analysis of spatially referenced data sets have also received much attention over the last decade. In parallel with the use of remotely sensed data, spatially-referenced data sets and their analysis using GIS is often an integral part of scientific and engineering investigations; see, for example, texts in geological and environmental sciences [38], ecological systems [33], digital terrain cartography [21], computer experiments [31], and public health [11]. The last decade has also seen significant development in statistical modeling of complex spatial data; see, for example, the texts by [9,10,24,32,36] for a variety of methods and applications.

A new approach that has recently garnered popularity in spatial modeling follows the Bayesian inferential paradigm. Here, one constructs hierarchical (or multi-