A Methodological Framework focused on integrating GIS and BBN Data for Probabilistic Map Algebra Analysis

J. D. Morgan[1], M. W. Hutchins[1], J. Fox[1], K. R. Rogers[1]

[1]UNC Asheville's National Environmental Modeling and Analysis Center (NEMAC)
One University Heights Asheville, NC 28804
Email: {jdmorgan;mwhutchi;jfox;krogers}@unca.edu

1. Introduction

The raster cell-by-cell comparison technique commonly referred to as map algebra has become a standard for Geographic Information Systems (GIS) spatial analysis and modelling (Bolstad 2012). Bayesian Belief Networks (BBNs) provide a graphical (and automated) way to display and interact with probabilities of related event information (Pearl 1988). Integrating spatial data with BBNs, at the resolution of pixels, allows for the opportunity to perform probabilistic map algebra (Taylor 2003; Ames & Anselmo 2008). Belief maps (or probability maps) can be created by transferring the results of probabilistic map algebra back into GIS. The methodological framework presented in this paper proposes a set of structured techniques for utilizing GIS and BBNs to inform spatial decision making and analysis.

The methodological framework presented in this paper proposes a set of structured techniques for integrating GIS and BBNs to inform spatial decision making based on belief about causally related datasets. The end result is the ability to produce belief maps based on exploratory analysis of BBNs. Finally, a case study is discussed to put this framework into context.

2. Methodological Framework

Reference to the Oxford English Dictionary (2012) allows us to unpack the terms *methodological* and *framework* which are used in the title of this paper. Methodological is an adjective of the noun methodology which refers to a body of methods, or techniques, used in a particular field of study. Framework is a noun referring to an underlying structure or conceptual scheme or system.

The methodological framework presented in this paper builds on previous work considering GIS/BBN integration by Walker et al. (2004). Unique to this paper is a focus on the raster based technique of map algebra, or more specifically probabilistic map-algebra as described in Taylor (2003) and Ames and Anselmo (2008). We identify four broad methodological steps which move us through the probabilistic map algebra framework. Each step proposes techniques, iterative in nature, which are illustrated in the subsequent case study.

2.1 Identify problem statement and relevant data:

The problem statement is a clear and concise definition of the primary issue that you wish to address in a given project. Along with helping to inform the other steps of the framework, a well defined problem statement guides the data selection process. At this step, a group of decision makers (workgroup) determines what is valued in the study area. For instance, in the case study presented in this paper, the workgroup determined that they valued natural, societal and cultural resources. Geographic datasets that represented these values included land cover, biotic environments and registered historic places.

2.2 Moving data from GIS to BBN:

The cartographic model is a visual way to show the process of a combination of operations performed on spatial datasets (Bolstad 2012). The model presented here (Fig 1.) shows the steps required to prepare both vector and raster datasets for the probabilistic map algebra format. While performing these steps it is helpful to be aware of modifiable areal unit issues introduced during aggregation, disaggregation, and re-sampling of data (Unwin 1996, Openshaw 1984).

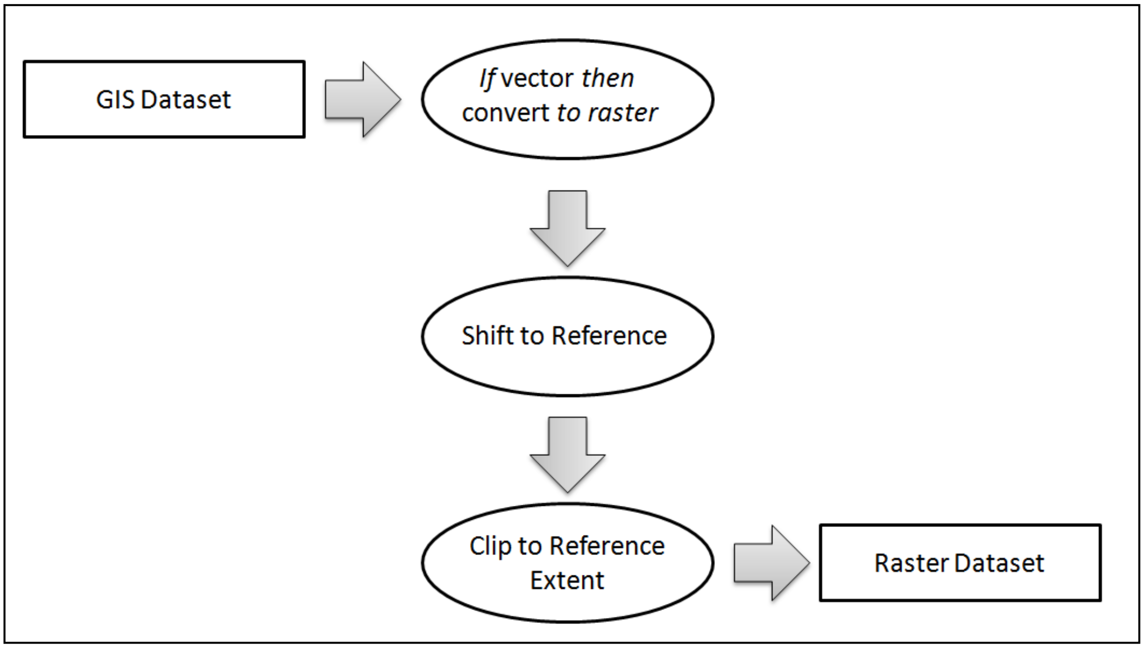
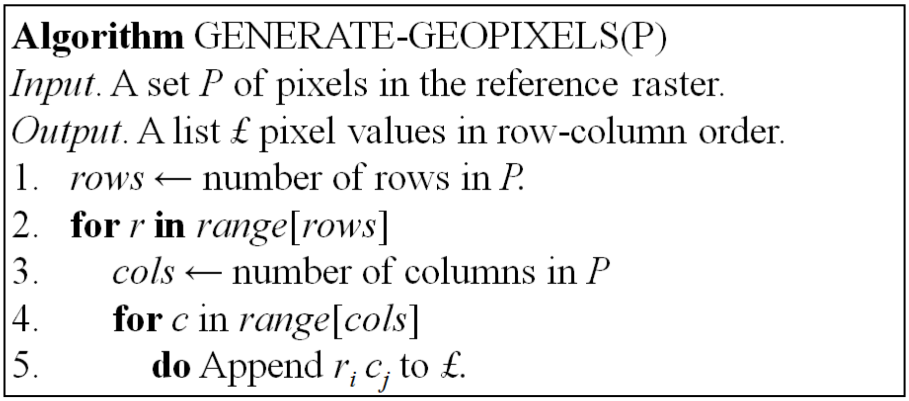


Fig. 1 – Cartographic model of data preparation

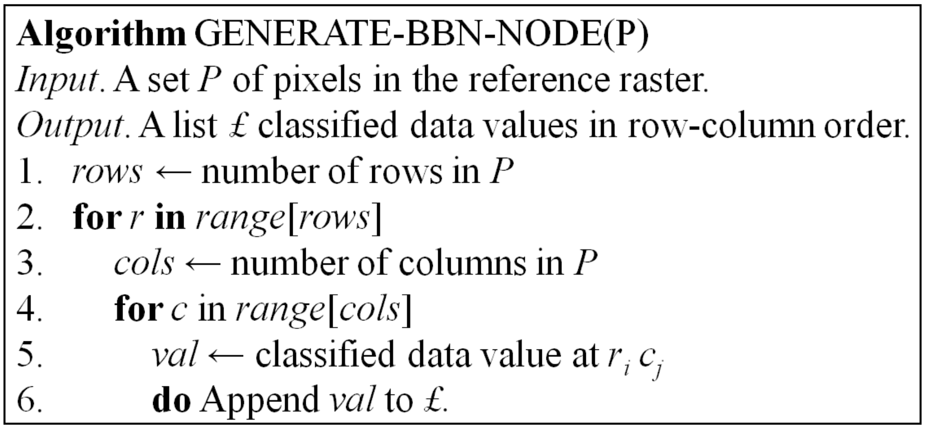
Central to the framework presented in this paper is the concept of the *geopixel* which provides an explicit spatial context to the BBN. The *geopixel* is simply taxonomy for each raster cell of the prepared datasets. This taxonomy can be created by utilizing a scripting language (e.g. python) with access to the spatial and attribute data (e.g. GDAL). Cormen et al. (1990) provide a definition of algorithm as a sequence of computational steps that transform an input set of values to an output set of values. Berg et al. (2008) provide formatting guidelines for algorithms. The following two algorithms illustrate techniques for generating the *geopixel* taxonomy and BBN data content.



```
Algorithm GENERATE-GEOPIXELS(P)
Input. A set P of pixels in the reference raster.
Output. A list £ pixel values in row-column order.
rows ← number of rows in P.
for r in range[rows]
    cols ← number of columns in P
    for c in range[cols]
        do Append $r_i c_j$ to £.
```

Fig. 2 – Algorithm for generating GeoPixels

The output from the algorithm in Fig. 2 is a list of *geopixels*, £, named by row and column number. This list of *geopixels* becomes the spatial hypothesis node in the BBN. Each state of a hypothesis node represents a different hypothesis specific to the relationships defined in the BBN (Krieg 2001). The spatial hypothesis node maintains the ability to make the spatial linkage, via *geopixel*, back to the GIS.



```
Algorithm GENERATE-BBN-NODE(P)
Input. A set P of pixels in the reference raster.
Output. A list £ classified data values in row-column order.
rows ← number of rows in P
for r in range[rows]
    cols ← number of columns in P
    for c in range[cols]
        val ← classified data value at $r_i c_j$
        do Append val to £.
```

Fig. 3 – Algorithm for generating BBN node content from GIS data

The output from the algorithm in Fig. 3 is a list of classified data values, £, which will become the values used to build the BBN node states.

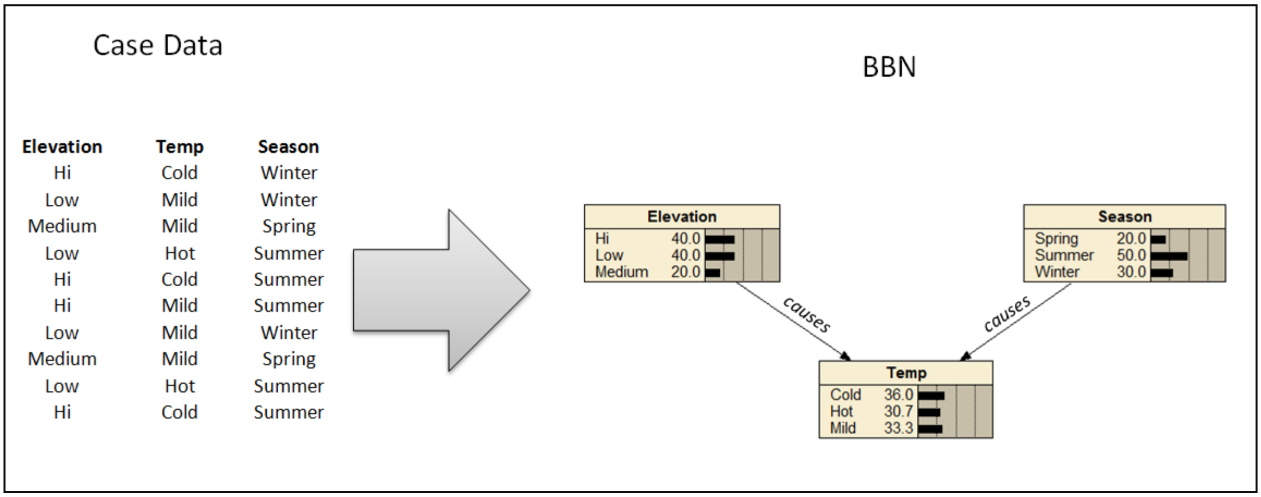
2.3 Exploratory Analysis of data within the BBN:

BBN's are increasingly adopted across a range of applications from medical diagnosis (Lauritzen & Spiegelhalter 1988) to predicting land cover (Aitkenhead & Aalders 2008) and especially relevant to our approach, determining habitat populations (Lee & Rieman 1994). The methodological foundations of BBN's were pioneered in the 18th century by Thomas Bayes.

Bayes' Theorem (Equation 1) rest on the causal connection of events and belief (translated as probability). Bayes' Theorem can be read that the probability of A (hypothesis) given B (evidence) is equal to the probability of B given A multiplied by the probability of A divided by the probability of B.

$$P(A|B) = P(B|A) \times P(A)/ P(B) \qquad (1)$$

What follows is ability to infer beliefs throughout a network. To briefly illustrate BBNs and their usage, consider the following example starting with the belief that temperature is influenced by elevation and annual season (Fig. 4). The initial state(s) of the BBN nodes in Fig. 4 is shown by the belief bar values based on the existing probabilities, a prior, case data (N=10). An important part of understanding the way that BBNs work is in knowing that the belief bar values for a specific node always sum to 100%.

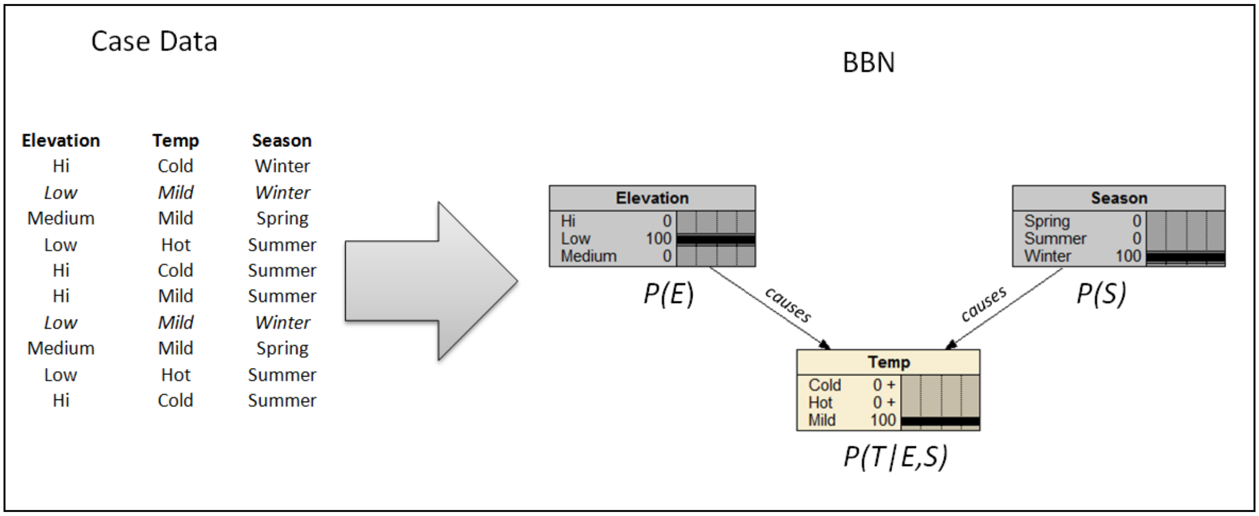


Case Data

Elevation	Temp	Season
Hi	Cold	Winter
Low	Mild	Winter
Medium	Mild	Spring
Low	Hot	Summer
Hi	Cold	Summer
Hi	Mild	Summer
Low	Mild	Winter
Medium	Mild	Spring
Low	Hot	Summer
Hi	Cold	Summer

Fig. 4 – Case Data and BBN Example at Initial State

This nodal relationship defined between the variables in this BBN example can be translated as temperature is dependent on elevation and season as indicated by notation of causal (or influence) linkage arrows. Once compiled, the BBN software (Netica by Norsys Software Corporation) creates conditional probability tables utilizing an algorithm formulated in Lauritzen & Spiegelhalter (1988). When we enter the findings (or evidence), elevation/low and season/winter, we get the resultant beliefs, a posterior, shown in Fig. 5.



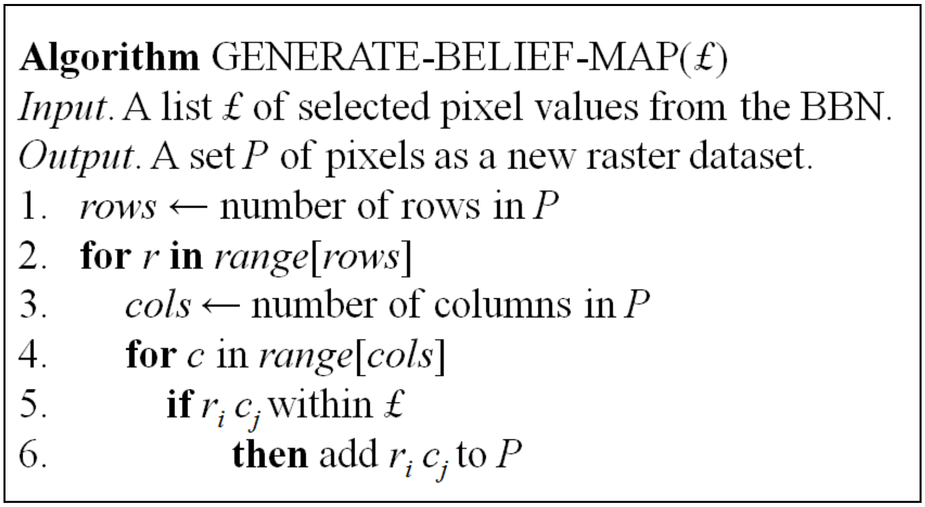
Elevation	Temp	Season
Hi	Cold	Winter
Low	*Mild*	*Winter*
Medium	Mild	Spring
Low	Hot	Summer
Hi	Cold	Summer
Hi	Mild	Summer
Low	*Mild*	*Winter*
Medium	Mild	Spring
Low	Hot	Summer
Hi	Cold	Summer

Fig. 5 – Case Data and BBN Example with Findings Entered

The resultant BBN demonstrates the probability distribution across the temperature node that occurs when evidence of low elevation and a winter season is entered by the user. It is this ability to interact dynamically with the BBN that facilitates the exploratory aspects of this type of analysis. In this sense, the probabilistic map algebra approach supports MacEachren's (1995) presentation of geovisualization where the user is not presented with a single passive static view, but rather an active process where map data, imbued within the BBN, goes beyond simple information communication to that of an enabler of knowledge construction.

2.4 Moving data from BBN to GIS

Fig. 6 shows the algorithm utilized to export the data out of the BBN and back into a format digestible by a GIS. While the iterative process is very similar to the previous two algorithms, a notable difference is the list of pixels (*geopixels*), £, and output of a raster dataset, *P*.



```
Algorithm GENERATE-BELIEF-MAP(£)
Input. A list £ of selected pixel values from the BBN.
Output. A set $P$ of pixels as a new raster dataset.
rows ← number of rows in $P$
for $r$ in range[rows]
    cols ← number of columns in $P$
    for $c$ in range[cols]
        if $r_i\ c_j$ within £
            then add $r_i\ c_j$ to $P$
```

Fig. 6 – Algorithm for moving data from BBN into GIS

From the output of a raster dataset, *P*, we can create what we will call in this paper *belief maps*. These maps are resultant of the probabilistic map algebra approach and BBN exploration. We use the term belief maps because the results are based on interactive state changes (entering evidence) within the BBN which adjust the probability distribution along the *geopixels* node. This resulting distribution can then be used to generate belief maps that illustrate where the probabilistic conditions are met.

3. Case Study

The case study presented in this paper is taken from a project facilitated by UNC Asheville's National Environmental Modeling and Analysis Center (NEMAC). In this project the Southeast Natural Resource Leaders Group (SENRLG), a consortium of federal agencies, was charged with addressing the issue of climate change impacts and resource sensitivities in the southeast. Several workshops were held early in the project to gain cross-agency consensus through a structured process called Comparative Risk Assessment Framework and Tools (CRAFT). CRAFT was developed by Eastern Forest Environmental Threat Assessment Center (EFETAC) of the US Forest Service and provides a structured approach to identify objectives, develop and compare alternative actions, and display the tradeoffs and risks associated with different decisions and models (Norman et al. 2012). It should be noted that the methodological framework presented in this paper does not present the breadth of the CRAFT approach, but rather focuses on using specific techniques along with probabilistic map algebra for GIS/BBN integration.

An early SENRLG workshop identified the impact of sea level rise (SLR) in the Albemarle-Pamlico area (a 20 county region) in North Carolina as a final candidate for an implementable project. The goals of this project were to prioritize and select at least three locations across the Albemarle-Pamlico that support the natural, cultural and social priorities, and prioritize areas to build adaptive capacity. The goals can be summarized as the following:

- Prioritize areas based on eight essential attributes (or proxies for agency values)
- Identify current resources that represent the agencies and create a map of mutual interest
- Determine prioritized areas that are impacted by SLR.
- From these prioritized areas, identify 3 locations that are good federal investments for increasing adaptive capacity to SLR

The actual implementation of this project occurred during a workshop held on August 2-3, 2012. The first day of workshop brought together representatives from SENRLG constituent agencies and focused on selecting three areas that maximized the mutual interest of the participants. The second day of the workshop brought together the same group, but focused on adaptable activities in the face of SLR.

The 8 essential attributes identified by SENRLG represented common themes across the agencies in an attempt to connect data proxies to existing program values. To guide which datasets to utilize among the essential attributes, agency funding objectives were considered that included but were not limited to wildlife habitat, recreation, ecosystem connectivity, and water quality. Fig. 7 is a roadmap that conceptually associated the identified essential attributes with the contributing factors, threats, and adaptive activities of SLR.

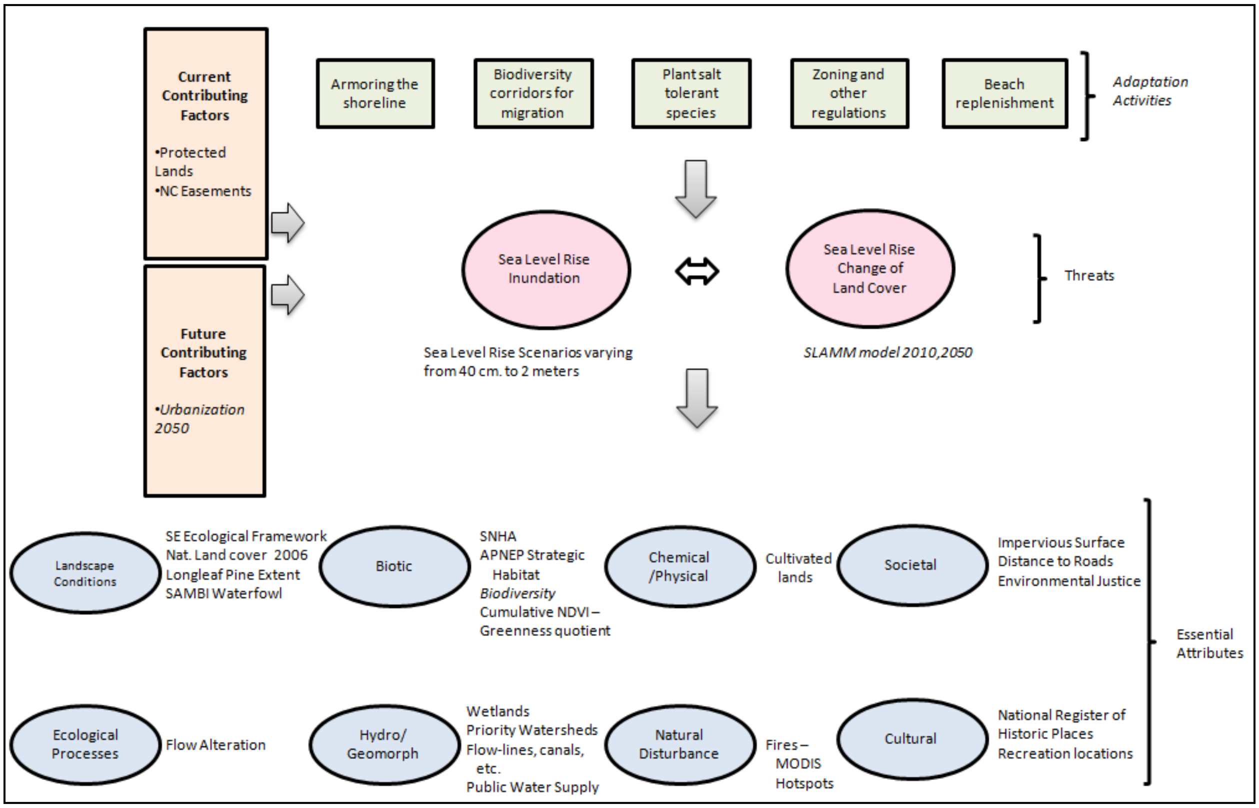


Fig. 7 – Case Study Dataset Roadmap

A contributing factor is defined as something beyond immediate control. Threat refers to the hazardous condition under consideration (SLR in this case). Adaptation activities are considered actions that have the ability to change conditions. The identification of datasets was decided prior to the workshop during a series virtual and face-to-face meetings led by NEMAC and attended by SENRLG stakeholder agency representatives.

At the end of the first day of the workshop the SENRLG group decided on "stories" where specific datasets were selected that fit a theme: drinking water, tourism & recreation, habitat and wilderness. The datasets for these themes relate directly back to the essential attributes (bottom of Fig. 7). Focusing in on results via a thematic story mapping approach proved useful because locating a single *geopixel* that met the condition of evidence of all 30 nodes was practically impossible. Further, the method of story mapping fit well with the exploratory approach of BBNs where a story map is defined by its experiential dimension which is used to better understand places and to mobilize for action (Caquard 2011). Therefore, thematic story maps allowed the group to focus on specific topics and BBN nodes for decision making. Additionally, and just as important, the evidence propagated to the other nodes allowed the group to explore the probabilities of the ancillary datasets which allowed the group to potentially maximize values outside of the selected nodes and identify available federal funding resources to assist on-the-ground conservation and restoration efforts. Further, it was useful to start with a BBN that had *geopixels* aggregated to a county node in order to be able to focus in on specific areas. To illustrate, consider the BBN with evidence entered for the drinking water story (Fig. 8).

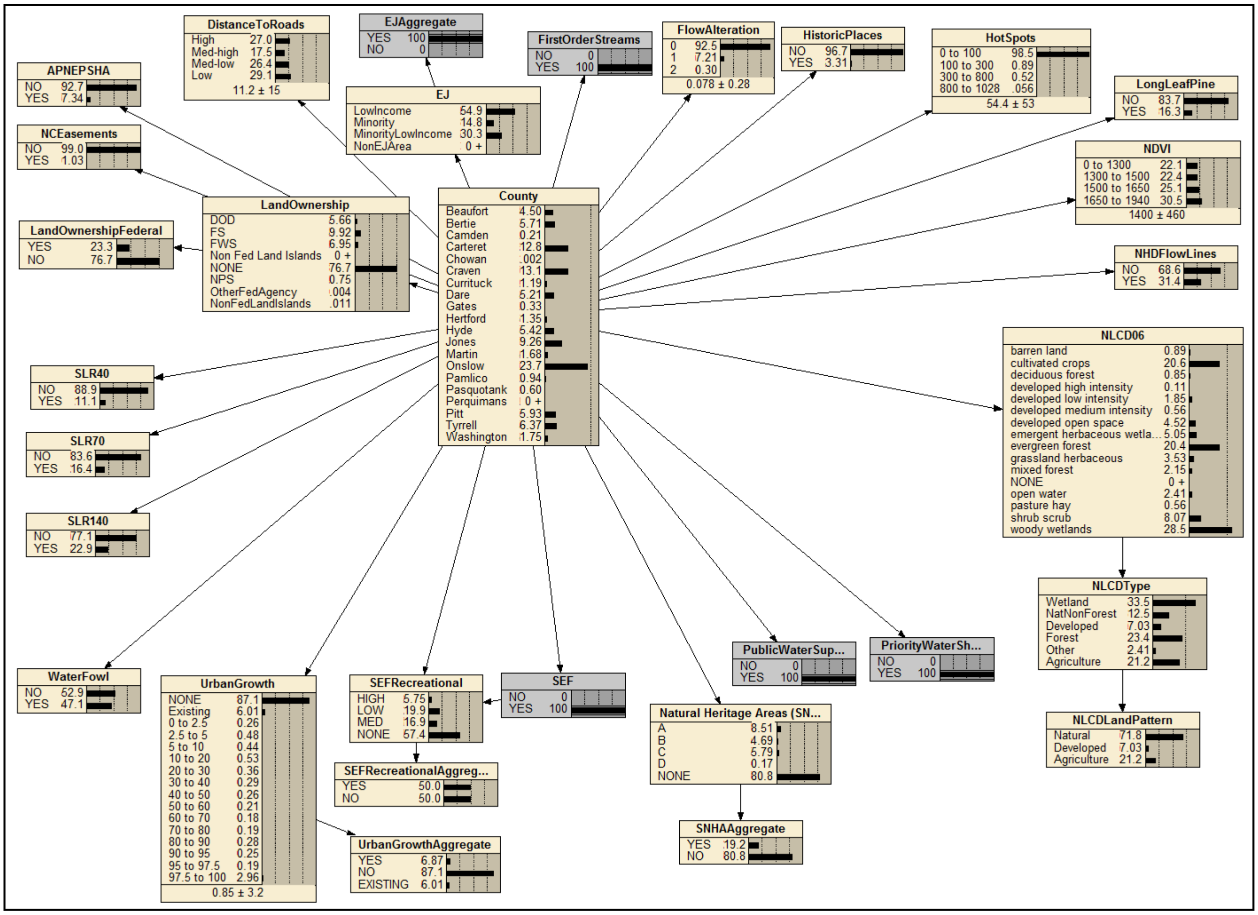


Fig. 8 – Case Study Drinking Water Story (County-wide)

Taking one of these stories as an example the drinking water story included the following datasets:

- Environmental Justice (EJ) - Polygonal delineation of minority, low income, and non-EJ Areas
  - source: Environmental Protection Agency
- Southeast Ecological Framework (SEF) - Raster delineation of natural land and water features - rivers, ridges, estuaries, wetland basins, and upland forests
  - source: GeoPlan Center, University of Florida/EPA
- Public Water Supply (PWS) – Point representation of existing well locations
  - source: North Carolina Division of Environmental Health
- First Order Streams (FOS) – Linear representation of headwater streams features
  - source: North Carolina Division of Environmental Health

Entering the evidence (or a state change) specific to this drinking water story focused the workgroup in on Onslow County because this county had the most *geopixels* that fit the criteria of this particular story (23.7%). And while the software itself (Netica) handled the process of propagating probability distributions throughout the network with each state change, it is useful to understand the specific application of Bayes' Theorem. For instance, on the first state change of the drinking water story evidence of Environmental Justice (EJ=YES) is entered. The resulting probability for Onslow County then is found by the following formula:

$$P(Onslow|EJ) = \frac{P(EJ|Onslow) \times P(Onslow)}{P(EJ)} \qquad (2)$$

With the second state change of the drinking water story evidence of both Environmental Justice and the Southeast Ecological Framework (EJ=YES and SEF=YES) is entered. Now the resulting probability for Onslow County can be found by the following formula:

$$P(Onslow|EJ \cap SEF) = \frac{P(SEF|Onslow \cap EJ) \times P(EJ|Onslow) \times P(Onslow)}{P(SEF|EJ) \times P(EJ)} \qquad (3)$$

With additional state changes it becomes apparent that the primary benefit of the BBN is the ability to perform these calculations on-the-fly and providing a venue to present them in a visually intuitive manner (node and causal linkage arrows). For this reason the use of the BBNs proved useful during preliminary deliberations of the SENRLG workshop when developing maps of mutual interest and prioritized areas.

By interacting and exploring the BBN the SENRLG workgroup was able to discuss and explore the relationship between the datasets specific to the drinking water story. NEMAC staff then exported the *geopixels* that met the certain probability thresholds. In the second day of the workshop the resulting belief map (Fig. 9) from the drinking water story allowed the group to spatially see the results of the probabilistic map algebra approach. More importantly it allowed the workgroup to move the results back into a GIS so that they could layer ancillary datasets on top of the belief maps in order to spatially inform their decision making. For instance 71% of these pixels are on natural land and 6% are predicted to experience urban growth in the future.

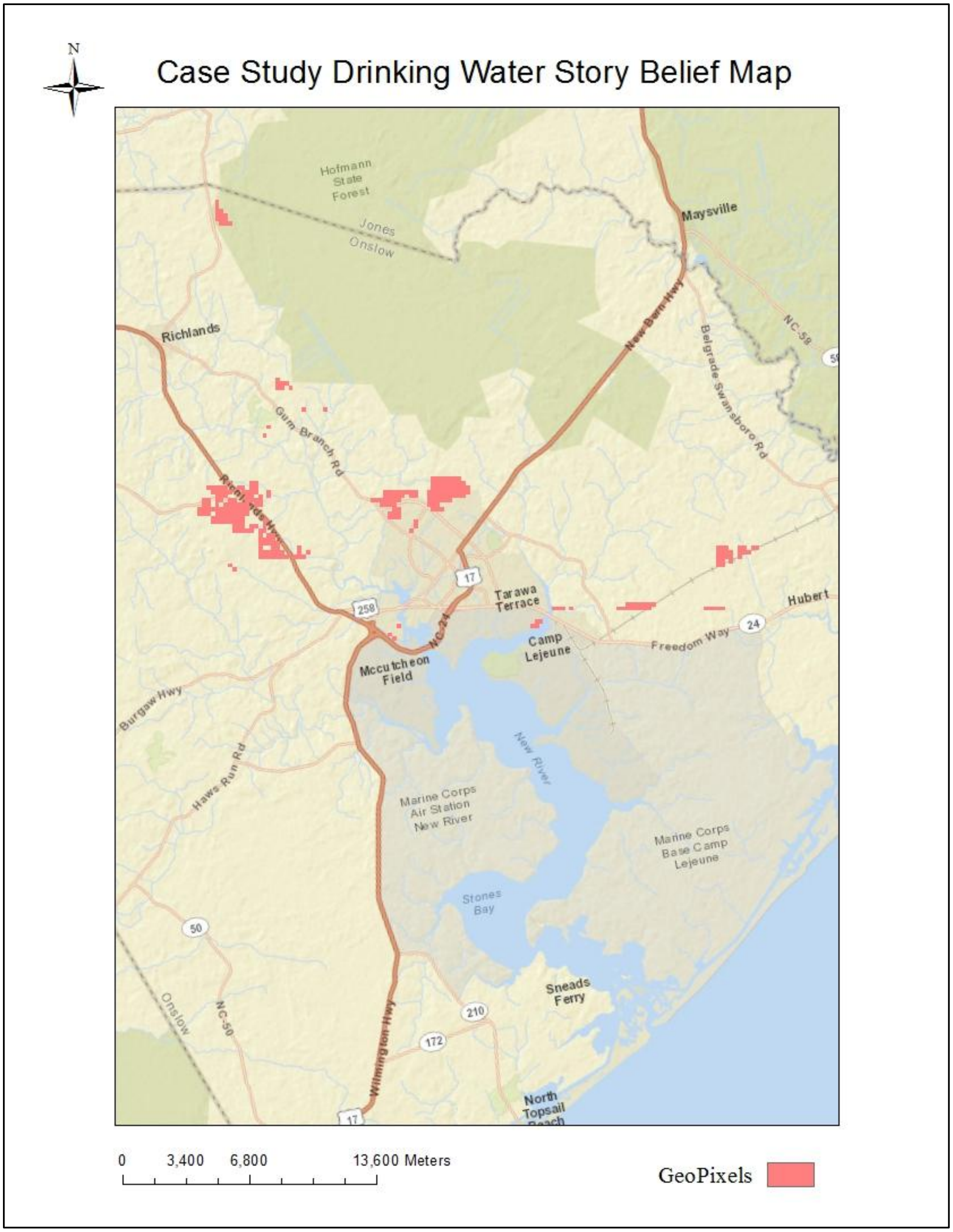


Fig. 9 – Case Study: Drinking Water Story Belief Map

4. Conclusion

While the ability to integrate BBNs (probabilistic map algebra) with spatial analysis and geovisualization proved useful during the SENRLG decision process, the methodology was not without its limitations. A notable limitation that manifested during the SENRLG workshop was that the usage of spatial data within a BBN is hindered by the simple fact that it is not a map in the cartographic sense. While informative from an analytical standpoint, during the SENRLG workshop the BBN exploration played a smaller than expected role in the group discussion and decision making process. This likely had to do with the inability of the group to see the topological relationships within a BBN. The final decision making occurred with the use of a GIS map interface, and in part with the ability to overlay related ancillary datasets on top of the belief maps.

Another limitation of the probabilistic map algebra approach is in dealing with datasets where the spatial extent is not indicative of the true value of the resource it represents. Consider the case of national registered historic places, which were provided in the vector format of points (NPS 2012). This dataset clearly carries with it a cultural significance and value that exists at specific locations. However, the challenge was in representing the value of this resource across the *geopixels* so that it carried influence in the probability distribution of the BBN. To circumvent this problem you could provide a weighted buffer, as we did during the SENRLG project. However, taking a cue from Tuan (1975) we can conclude that there are certain datasets which refer us to a *place* constructed by meaning based on experience. And these stand in contrast to *space* which is a more abstract and hence easier to fit into the bounds of representation by way of pixel.

Even with these limitations the ability to utilize GIS data to do cell-by-cell raster-based map algebra operations within the context of probability holds great potential. As noted by Ames and Anselmo (2008) the use of BBNs in this kind of probabilistic map algebra is currently hindered only by the lack of specialized tools to support the analysis. This paper adds to the breadth of both the BBN and map algebra literature in two ways: 1) a generalized framework for moving from GIS to BBN and back to GIS again; and 2) a set of data integration techniques for moving through the proposed framework that can be applied in a variety spatial decision making contexts.

Acknowledgements

Acknowledgements go to Danny Lee and Steve Norman at the USDA-Forest Service EFETAC for their BBN expertise. Thanks also to Rick Dubrow at EPA Region 4 for his leadership on the SENRLG project.

References

Aitkenhead, M.J. & Aalders, I.H. (2008). Predicting land cover using GIS, Bayesian and evolutionary algorithm methods. *Journal of Environmental Management* 90 (2009) 236-250

Ames, DP, Anselmo, A., 2008. Bayesian Network Integration with GIS, in: Shekhar,S., Xiong, H. (Eds.), Encyclopedia of GIS, Springer, New York, pp. 39-45.

Berg, M., Kreveld, M., Overmars, M. & Schwarzkopf. O. (2008). *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin.

Bolstad, Paul (2012). *GIS Fundamentals*, 3rd Edition, Atlas Books.

Caquard, S. (2011). Cartography I: Mapping Narrative Cartography. Progress in Human Geography. Sage Publications.

Cormen, T. Leiserson, C.E., & Rivest, R.L. (1990). *Introduction to Algorithms*. MIT Press, London.

Krieg, M.L. (2001). *A Tutorial on Bayesian Belief Networks*. Surveillance Systems Division Electronics and Surveillance Research Laboratory

Lauritzen S.L. & Spiegelhalter D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society* 50:157–224.

Lee, D.C. and B.E. Rieman. (1994). *Bayesian viability assessment module: a tool for assessing the population viability of resident salmonids*. USDA Forest Service Intermountain Research Station.

Oxford English Dictionary. (2012). OED Online. Oxford University Press. <http://dictionary.oed.com/>.

MacEachren, A.M., (1995). How Maps Work: Representation, Visualization and Design. Guilford Press, New York, London, 513 pp.

National Parks Service. (2012). National Register of Historic Places retrieved from www.nps.gov/nr/

Norman, S.P., Lee, D.C., Jacobson, S. & Damiani, C. (2012). *Assessing Risk to Multiple Resources Affected by Wildfire and Forest Management Using an Integrated Probabilistic Framework.* Advances in Threat Assessment and Their Application to Forest and Rangeland Management

Norsys Software Corporation (2012). Netica Bayesian Belief Network Software. Acquired from http://www.norsys.com/

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and Techniques in Modern Geography* 38: 41.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo.

Taylor, K.J. (2003) *Bayesian Belief Networks: A Conceptual Approach to Assessing Risk to Habitat*. Utah State University, Logan, Utah, USA.

Tuan, Y-F. (1975). *Place: An Experiential Perspective*. Geographical Review, Vol. 65, No. 2

Unwin, D.J. (1996). *GIS, spatial analysis and spatial statistics*. Progress in Human Geography 20(4): 540-441.

Walker, A., Pham, B. & Maeder, A. (2004). *A Bayesian framework for automated dataset retrieval in Geographic Information Systems*, 10th International Multimedia Modeling Conference, 2004. Brisbane, Australia.