Using spatiotemporal relational probability trees to investigate drought morphology and tornadogenesis

Matthew W. Collier
Department of Geography
University of Oklahoma
Norman, OK 73072
mwc@ou.edu

Nathan C. Hiers
School of Meteorology
University of Oklahoma
Norman, OK 73072
Nathan.C.Hiers-1@ou.edu

0. Abstract

A new algorithm is explored which adds spatiotemporal capabilities to Relational Probability Trees. This algorithm is tested against two applications in the fields of meteorology and climatology.

1. Introduction

Currently, probability trees exist which may investigate static relational data (Neville, et.al. 2003). However, in most fields of science exploration, many objects exist that vary in both space and time, so to maximize the usefulness of Relational Probability Trees (RPT) in scientific investigation the trees must be able to consider Spatiotemporal (ST) data. A Spatiotemporal Relational Probability Tree (SRPT) acts like a standard decision tree, with probabilities of an event occurring at the leaves. The novelty of our approach adds the capability of handling ST data to the data mining algorithm. In general, the input data may consist of various mixtures of spatial and temporal data. The spatial data may be composed of grid points or possibly other objects existing in space. The temporal data is currently composed of 'snapshot' type time steps of the spatial data. The ST view of this mixture can reveal hidden knowledge and subtleties of phenomenon that are difficult or improbable for human discovery. The SRPT is applied to two problem domains: Tornadogenesis in the field of meteorology, and drought morphology in the field of climatology.

Tornadoes cause, in part, the $13B (Pielke and Carbone, 2002) of economic impact due to mesoscale storms annually in the United States, yet currently remain poorly forecasted and warned. The current primary instrumentation for observing mesoscale convection is the WSR-88D NEXRAD radar system, yet it is insufficient for observing tornadoes. Current WSR-88D radars are unable to see 76% of the volume between the ground and two kilometers (Brotzge et al., 2006), which is where a tornado resides, due to the beam of the radar remaining straight as the earth curves away. Furthermore, the national average for the probability of detection of a tornado once it forms is near 75%, while a respectable average, the false alarm ratio also remains constant at around 75%. Like all weather phenomenon, tornadoes are the result of many processes the change in both space and time, and the only way to learn about the processes is to be able to model and analyze them in both space and time, making spatiotemporal relational probability trees an ideal candidate to use for exploring tornadogenesis. Using the spatiotemporal relational probability trees should give meteorologists a better understanding of the important dynamic processes that result in tornadogenesis, which should lead to better understanding and prediction of tornadoes.

Drought morphology is an excellent candidate for the SRPT algorithm. First, droughts are dynamic, sporting changing characteristics in both space and time. Second, drought impacts cause 6 to 8 billion dollars in economic losses annually in the U.S. alone (Goddard, et. al. 2003). It is only with further research into the nature of drought that we can better understand how its impacts can be minimized through mitigation, planning, and adaptation. Goddard further recognized that, with regards to drought, "…the key to future progress lies in the new computer science research." Thus, the characteristics of drought along with the immense economic costs make drought an ideal problem domain.

2. Problem Definition and Algorithm

2.1 Task Definition

The data being used in the tornadogenesis exploration was created using Advanced Regional Prediction System (ARPS), which is a top weather forecasting system for mesoscale data (Xue et al., 2003). A total of 250 simulations lasting 3 hours long each that use supercell-favorable conditions are ran using grid spacing of 500m in the horizontal and a stretched hyperbolic tangent is used in the vertical, which places smaller grid intervals near the surface. Each grid point contains 84 derived and observable quantities that are important for understanding storm morphology and tornadogenesis. Each simulation can contain multiple storms, and each storm can be abstractly broken down into high level features that are dynamically important to its morphology. Unfortunately, the grid spacing currently being used is too large for a definitive resolution of a tornado, so all features matching the current meteorological theories on what a tornado is are considered 'extreme low level rotations' and can be used interchangeable throughout this paper. The schema that is passed to the dynamic relational probability tree contains high level features, relations, and attributes. Attributes are a part of the high level feature, and describe meteorological quantities and statistical distributions of quantities that exist within that feature. Relations describe how multiple high level features interact with one

another, examples of which are 'nearby', 'contains,' and 'overlaps.' In this experiment, six relations, nine high level features, and an average of thirteen attributes per high level features are used.

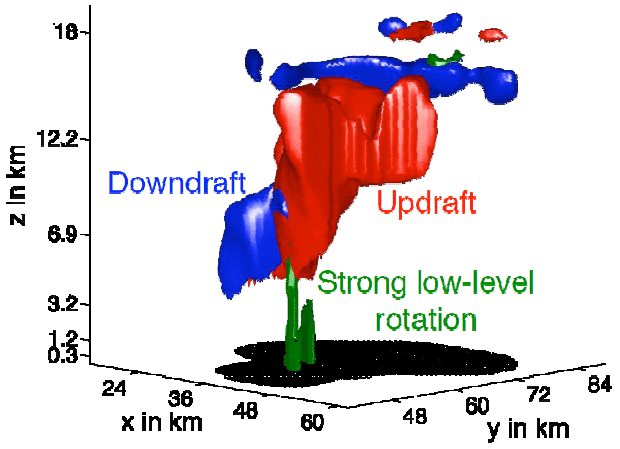

Figure 1 - An example of an abstracted supercell's features in the gridded model.

The set of questions that could be asked for the tornadogenesis data included asking if a single high level feature existed, if a relationship existed between two chosen high level features, and if a single high level feature existed for a chosen time that ranged from one minute to 15 minutes.

The drought data are freely available over the internet from the International Research Institute for Climate and Society (IRICS 2007). The data consist of gridded Palmer Drought Severity Indices (PDSI) for the world from 1870 to 2002. The spatiotemporal grid spacing is 2.5 degrees in both latitude and longitude, and every month in time (figure 2). The data were constructed by Dai, et. al (2004). Relationships in drought severity exist between both temporal and spatial steps.

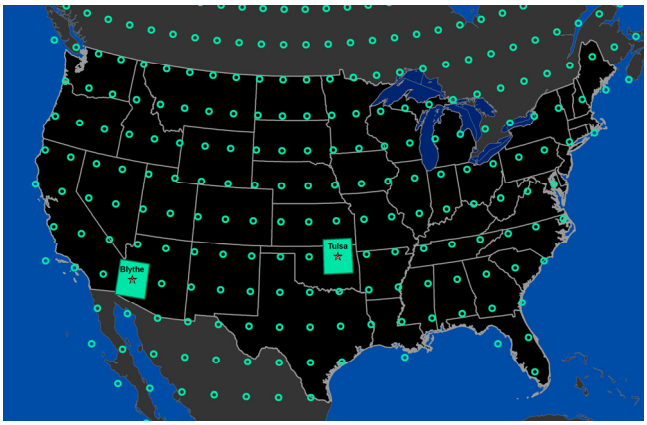

Figure 2 - Gridded, monthly, PDSI drought data from Dai, et. al., (2004).

The drought SRPT is constructed to answer two sets of queries based against positive labels that indicate dryness increased at the point in question (PIQ). The first second of questions asks whether or not drought will exist at the PIQ at the next time step given that drought exists in one of the four cardinal directions from the PIQ. The second set of questions asks whether or not dryness increases at the PIQ given that dryness increases in one of the four cardinal directions from the PIQ. In brief, the questions being tested against the labels include four questions of existence, and four questions of intensity change.

2.2 Algorithm Definition

A data set that contains high level features or objects, relations, and attributes are read into the program, and can then be accessed. The data set being fed in also must delineate all possible high level features or objects, relations, and attributes that are able to be used to create dynamic questions as well as the style of questions that are desired by the person wishing to investigate the data using this algorithm. Random questions are then asked, and a have a chi-squared value calculated using a confusion matrix with enough frequency to ensure within a desired p-value that the question with the highest chi-squared value is the most important question at that node in the SRPT. Once the best question has been determined, children nodes are created that act as 'yes' and 'no' branches and accept all cases that belong in the 'yes' or 'no' branches respectively. This is repeated for all nodes until the chi-squared value of a child node is no longer statistically significant in its chi-squared value, in this case the child node becomes a leaf node and the probability of the desired event occurring is recorded.

```
Main()
    initializeStructures()
    readData()
    generateQuestions()
    while nodesUnknown()
        makeConfusionMatrix()
        checkSignificance()
        if Significant χ²
            UpdateNode()
        else
            CreateLeaf()
```

Figure 3 - The Spatiotemporal Relational Probability Tree Algorithm.

By way of example, and following after figure 3, let one assume we are looking at the drought morphology data, there is only drought to the west of our current point of interest, and drought only gets worse to the north after there is drought of our point of interest. Begin by looping through all possible questions, keeping

track of true positives, false positives, true negatives, and false negatives in order to create a confusion matrix. Once the confusion matrix has been created, calculate the χ^2 values for each question, and using it to calculate the p-value. Assuming the designated p-value threshold is greater than the calculated p-value for the question, a node is created that stores the question chosen, the children nodes, and the probability dryness increasing in strength at the origin point. Two children nodes are also created, one to answer the question as 'yes' and the other to answer the question as 'no.' The children nodes receive all questions that have been used in that branch as well as all simulations that answered 'yes' or 'no' in their respective branches. The first node and two children have been created and the loop begins again. Assuming the best question for this loop was if drought exists to the west, it can no longer be used in subsequent nodes below it. The 'yes' branch node would then loop through all possible questions, except the ones used and choose the highest value χ^2 out of the remaining questions. Assuming there is no further questions whose p-value is below the threshold, the yes branch becomes a 'leaf' and is assigned the probability of drought increasing at the origin point using only the simulations that have been passed on to it. The next loop would now search the 'no' branch and the process would continue likewise until there were no more significant questions found and the tree would then terminate.

3. Experimental Evaluation

3.1 Methodology

For both problem domains, our hypothesis is that complex meteorological and climatological phenomenon both contain higher order structure that is discernible through data mining and knowledge discovery techniques.

As this experiment used a small, but expert chosen by importance, subset of all possible interesting high level features and relations that could have been chosen, the expected output of the tree differs from what one may expect when a full range of high level features and relations are used. As two of the high level features were used to determine if a storm was tornadic, they were excluded from the testing, therefore of the remaining features one would expect stretching of vertical vorticity to be the most important. Stretching of vertical vorticity is the quickening or slowing of wind turning in the vertical coordinate as it morphs itself while keeping a constant volume. Consider a right cylinder, the shorter the height of the cylinder, the larger the area of the circle will be on top and bottom of it. The taller the height of the cylinder, the smaller the area of the circle will be. Using the concept of conservation of angular momentum and applying to the right cylinder, spinning along its z-axis, the taller the cylinder is, the faster it spins, which is what a tornado mimics.

The expected outcome of the drought analysis is that drought propagates across a PIQ away from the continental interior of the United States, and away from the subtropical highs located in Mexico and in the southern reaches of the United States.

For both problem domains, respective data is read into custom Python scripts, following the pseudocode discussed above to create the trees. K-fold cross-validation is used to create a measure of confidence in our results. Each fold produces an "Area Under the Curve" (AUC) statistic that can be averaged to show the overall effectiveness of the tree in splitting the examples.

3.2 Results

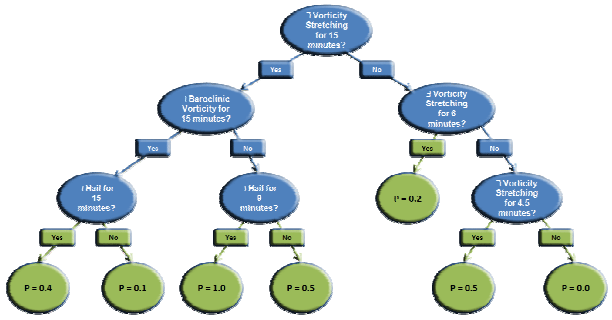

Figure 4 - An example tree from the 6-fold ensemble created with the SRPT algorithm. P-value of 3e-03 used to create the leaves.

The tornadogenesis SRPT, as seen in figure 4, shows that the most significant question is the existence of stretching vertical vorticity for at least 15 minutes. The yes branch then uses the existence of baroclinic generation of vertical vorticity as its most significant question. After baroclinic generation of vertical vorticity, the existence of hail for various times is used for the subsequent branches. Going down the 'no' branch for the root node, the existence of stretching of vertical vorticity for various times is seen in the subsequent child and grandchild nodes. The probabilities range from a maximum of 1 to a minimum of 0. The p-value threshold for significance was set to 3e-03. The six-fold cross validation produced an AUC of 0.77 with a standard deviation of 0.14. A visual representation of the AUC graph can be seen in figure 5. A total of 1023 storms were used in this experiment, with 38 of them being classified as tornadic.

Results for drought in Tulsa are displayed in figures 5 and 7 below. In the example shown, the most important question to ask in the Tulsa tree is, "Does drought exist to the West?" While, "Does drought exist to the East?" become the next most important question for both the yes and no branches of the first node. Finally, "Does drought exist to the South?" is the last significant question down the yes branch. This particular tree used 1e-08 as the p-value for significance in the questions. The average AUC for this series of folds was 0.69 with a standard deviation of 0.20.

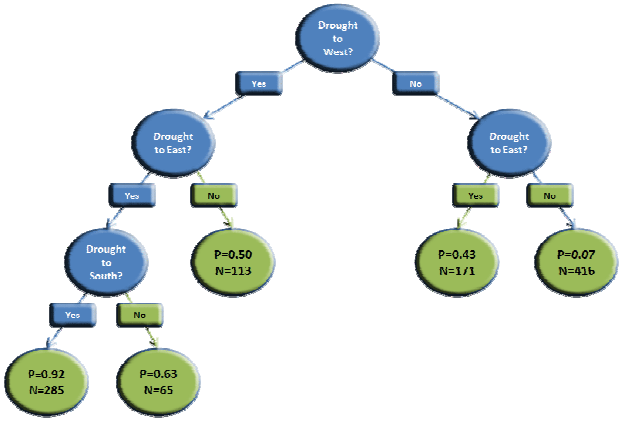

Figure 5 – An example tree from the 4-fold ensemble created with the SRPT algorithm. P-value of 1e-08 used to create the leaves.

The example tree from the Blythe, AZ run (figure 6) finds that the most important question for the first node is, "Does drought exist to the East?" The second most important question is , "Does drought exist in the south?" The leaves were produced at a p-value of 1e-07. The average AUC is 0.55 with a standard deviation of 0.03.

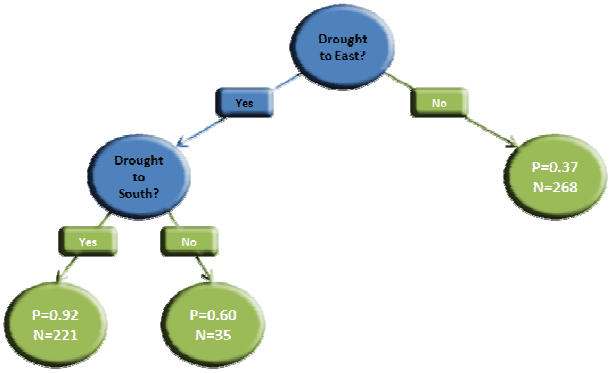

Figure 6 – Example tree for the 4-fold ensemble created with the SRPT algorithm. P-value of 1e-07 used to create the leaves.

3.3 Discussion

The tornadogenesis hypothesis was supported by the results in that both stretching of vertical vorticity was found to be the most significant question, and that this type of data was able to be properly investigated using a SRPT. Those familiar with the current theories of tornadogenesis may be surprised that neither updraft nor downdraft were found in the SRPT, as a strong rear flank downdraft is central to the current theories of supercellular tornadogenesis. This can be explained for two reasons; one being that the limited set of questions being asked in this experiment had no inquiry of the attributes of the high level features except for time in existence. The other explanation is that supercell thunderstorms are defined by having a rotating updraft, and are expected to have at least one downdraft. Since all storms used in this experiment were of a supercellular nature, all would be expected to have a sustained updraft and downdraft, meaning that tornadic and non-tornadic storms couldn't be differentiated using these types of questions. The AUC value being above 0.50 shows that the SRPT was able to perform better-than-random, and that SRPT's can be successfully used to find meaningful knowledge about spatiotemporal data sets that potentially have higher order objects related to tornadoes that are currently not well understood.

One note of particular interest in the tornadogenesis SRPT was the 'no' branch beginning with the root node. All subsequent significant questions along the 'no' branch dealt with the stretching of vertical vorticity for varying time intervals, but each time interval became smaller the further down the tree it was. This could be signs of an implicit cyclical pattern in the relative strength of a supercell thunderstorm. A supercell thunderstorm could begin as strong, weaken and reorganize, and strengthen again in order to produce a tornado.

The results of the drought study appear to verify the hypotheses. The AUCs, which are above 0.50, indicate that the algorithm is performing better than random at spitting the examples. This shows that there is significant, high-order structure in the data representing the drought. Furthermore, both the Tulsa, OK, and Blythe, AZ case studies indicate that drought begins in each respective location by moving away from the continental interior and subtropical highs.

4. Related Work

The tornadogenesis research extends and compliments the experiments done in Rosendahl (2007) in two important ways: changing the way storm cells are defined and tracked and extending the data mining techniques. This experiment uses SRPTs in conjunction with storm features being abstracted in order to extract more meteorologically useful information about the formation of extreme low level rotations. This experiment also redid the way a storm was defined into a more physically and thermodynamically accurate version which is less prone to losing a track on a storm, and more prone to getting more and better grid points of the storm as well as more storms in general.

Much work has been done in characterizing the spatial, temporal, and spatiotemporal extents of drought and its effects. However, the bulk of this work is all static, based upon the 'snapshot' approach. Real spatiotemporal integration of the data has been lacking. However, one group (Shin, et. al. 2000) has begun the process of integrating space and time in droughts through their neural net study. The current work attempts to add another point of view to the spatiotemporal investigation of drought by machine learning techniques through the use of SRPTs.

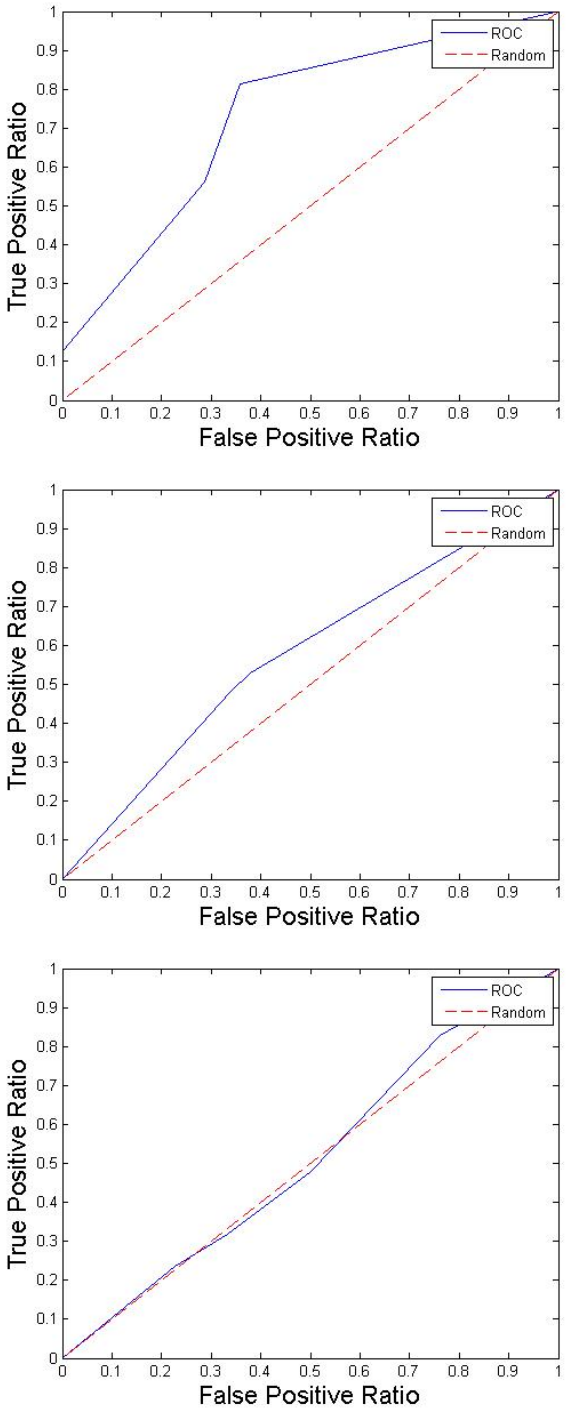

Figure 7 - Receiver Operator Curves for (upper graph) tornadogenesis, (middle graph) Blythe, AZ, and (lower grpah) Tulsa, OK.

5. Future Work

The current shortcomings of the tornadogenesis data is that there are only a few very basic high level features, relations, and attributes being used. Currently, there is work being done by meteorologists to suggest many more of each of these abstractions which should greatly increase the number of questions currently being asked, bring about new and interesting rules for tornadogenesis, and further increase the confidence of the accuracy of the rules that are being found.

Future work regarding drought would benefit greatly from a data set of higher spatial resolution. Furthermore, there are many concerns about the character of the PDSI index (Guttman 1998). Future work along these lines should utilize other indices such as the Standardized Precipitation Index which Guttman has shown to be more meaningful across geographic scales and regions. While the current drought work has focused on a couple of case study grid cells, this work will be applied to understanding ST patterns and processes across geographic space and time.

In both problem domains, a greater understanding of what potential questions exist must be more fully explored. Yuan & McIntosh (2002) developed a typology of queries that will provide a useful guide in this endeavor.

6. Conclusion

We conclude that spatiotemporal relational probability trees can successfully extend the functionality of relational probability trees into the spatiotemporal realm. We have shown three separate case studies from the fields of meteorology and climatology that demonstrate higher order structures can be revealed through this methodology. In these case studies, relationships were revealed that are consistent with expert knowledge.

7. Acknowledgements

The authors would like to thank Amy McGovern of the University of Oklahoma for many valuable conversations during the development and execution of this research.

Bibliography

Brotzge, J., Droegemeier, K. K., and McLaughlin, D. J. (2006). Collaborative adaptive sensing of the atmosphere (CASA): New radar system for improving analysis and forecasting of surface weather conditions. Journal of the Transportation Research Board, (1948):145–151.

Dai, A., Trenberth, K.E., and Qian, T., (2004). A global data set of Palmer Drought Severity Index for 1870-2002: Relationship with soil moisture and effects of surface warming, *Journal of Hydrometeorlogy*, **5**:1117-1130.

Goddard, S., Harms, S.K., Reichenbach, S.E., Tadesse, T., and Waltman, W.J., (2003). Geospatial Decision Support for Drought Risk Management, *Communications of the ACM*, **46**(1):35-37.

Guttman, N.B., (1998). Comparing the Palmer Drought Index and the Standardized Precipitation Index, Journal of the American Water Resources Association, **34**(1):113-121.

IRICS, (2007). IRI/LDEO Climate Data Library <http://iridl.ldeo.columbia.edu/index.html>. Last accessed 20Nov2007.

Neville, J., Jensen, D., Friedland, L., & Hay, M., (2003). Learning Relational Probability Trees, SIGKDD 2003, August 24-27, 2003, Washington, DC, USA.

Rosendahl, D.H., (2007). Identifying precursors to strong low-level rotation within numerically simulated supercell thunderstorms: A data mining approach. Master's Thesis, School of Meteorology, University of Oklahoma.

Shin, H., & Salas, J.D., (2000). Regional Drought Analysis Based on Neural Networks, *Journal of Hydraulic Engineering*, **5**(2):145-155.

Xue, M., Wang, D., Gao, J., Brewster, K., & Drogememeier, K., (2003). The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation, Meteorology and Atmospheric Physics, **82**:139-170.

Yuan, M., & McIntosh, J., (2002). A Typology of Spatiotemporal Information Queries, *Mining Spatiotemporal Information Systems*. K. Shaw, R. Ladner, and M. Abdelguerfi (eds.) Kuwer Academic Publishers. pp. 63-82.