2.5 ANALYZING THE EFFECTS OF LOW LEVEL BOUNDARIES ON TORNADOGENESIS THROUGH SPATIOTEMPORAL RELATIONAL DATA MINING

David John Gagne II* School of Meteorology, University of Oklahoma

Timothy Supinie School of Meteorology, University of Oklahoma

Amy McGovern School of Computer Science, University of Oklahoma

Jeffrey Basara Oklahoma Climatological Survey, University of Oklahoma

Rodger A. Brown National Severe Storms Laboratory, Norman, OK

1. INTRODUCTION

When different air masses meet, such as along a warm front or a cold front, boundary regions exist. Given that air mixes continuously, the transition zone along the boundary is not instantaneous and includes regions of strong temperature and moisture gradients. In addition to fronts, boundaries also occur along drylines or due to outflow from thunderstorms. While boundaries are commonly associated with the generation of storms through the lifting of warm, moist air, their overall impact on the generation of tornadoes is not as well understood. While some theoretical and case studies have been performed on the connection (Maddox et al. 1980; Markowski et al. 1998; Rasmussen et al. 2000), an analysis of the tornadic supercell-front connection over a multiyear period has not been performed.

The purpose of this study is to examine the connection between frontal boundary zones and supercellular tornadogenesis over a ten-year period in Oklahoma. The study is based on a climatology of supercell thunderstorms matched with tornado reports and surface fronts. Due to the large amount of data associated with the number of supercells over that period of time, spatiotemporal relational data mining is being used to analyze the data. Spatiotemporal relational data mining has the ability to find significant patterns in large amounts of complex data that varies in space and time, such as weather data. By reducing the data to trends in the most important variables, it can determine if a given supercell and front setup can produce a tornado.

*Corresponding author address: 120 David L. Boren Blvd., Suite 5900, Norman, OK 73072; email: djgagne@ou.edu. This paper is an extension of a section from "Augmenting Spatiotemporal Relational Random Forests for Use in Real-world Severe Weather Applications", a paper submitted to the 2010 Knowledge Discovery and Data Mining Conference with permission from the authors.

2. BACKGROUND

The effects on frontal boundary zones on tornadogenesis has been explored through both theoretical and case studies. Through the analysis of a series of outbreak cases, Maddox et al. (1980) explains how enhanced moisture convergence and vertical vorticity in frontal zones can help cause the development of intense tornadoes . Markowski et al. (1998) and Rasmussen et al. (2000) elaborate on the physical model from Maddox et al. (1980) and describe how boundaries can also yield a zone of enhanced horizontal vorticity. A supercell thunderstorm with a strong updraft moving through the zone can vertically tilt and stretch the enhanced horizontal vorticity, which assists with the process of producing a tornado. Markowski et al. (1998) analyzed strong tornadic supercell thunderstorms over a one-year period and found that 70% occurred near frontal boundaries. However, due to the limited sample size and time period, further study was called for to quantify the relationship between boundaries and tornadoes over longer periods.

With a ten-year period, this study required more automated analysis to handle the greater volume of data. For this task, objective front analysis techniques were employed. The first objective front analysis techniques were developed by Renard and Clarke (1965). They located fronts in gridded fields by calculating the Thermal Frontal Parameter (TFP), as shown in Eq. 1, from θ, where θ is any thermal field, such as potential temperature, equivalent potential temperature, or thickness.

$$TFP(\theta) = -\frac{\nabla|\nabla\theta| \cdot \nabla\theta}{|\nabla\theta|} \quad (1)$$

Essentially, the TFP is the directional derivative of θ along its gradient (Renard and Clarke 1965). The relative maximums in the TFP field correspond to the warm boundaries of frontal zones, and since frontal boundaries are traditionally analyzed on the warm side of the zone,

the maximums in TFP are used as the basis for the objective fronts. Subsequent work on objective front analysis (Hewson 1998) has found the greatest success with minor variations on the analysis of TFP. Jenkner et al. (2009) improved the objective analysis of mesoscale fronts by adding a filter to their algorithm to remove areas with weak TFP gradients. This filter resolves the actual boundary zones better while removing most noise from the field.

Thorough analysis of the atmosphere requires data mining algorithms that can interpret spatial and temporal variations. While more traditional data mining algorithms can only analyze static attributes, spatiotemporal relational algorithms can analyze how objects, their attributes, and the relationships among the objects vary in both space and time. Two algorithms, the Spatiotemporal Relational Probability Tree and the Spatiotemporal Relational Random Forest, are used to for this analysis.

The Spatiotemporal Relational Probability Tree (McGovern et al. 2008), or SRPT is a probabilistic classification decision tree that analyzes spatiotemporal relational graphs. To build its nodes, the SRPT selects a random sample of question types at each node and determines which question is most significant. At the leaves of the trees are probabilities based on the distribution of the graph labels. Because only a random subset of the possible questions can be sampled at each node, the resulting SRPTs are not always the near optimal result. Different runs of the SRPT algorithm can produce very different trees, especially at small sample sizes.

In order to account for and take advantage of this inherent variability in the SRPT, the Spatiotemporal Relational Random Forest (SRRF) was developed (Supinie et al. 2009). A SRRF is an ensemble of SRPTs based on the Random Forest (Breiman 2001) where the individual trees in the ensemble are grown on a bootstrap resampling on a subset of the training set in order to create a more robust ensemble. Supinie et al. (2009) showed significantly improved performance for the SRRF over the SRPT for the prediction of convective turbulence. This project makes that same comparison of algorithms on the fronts and supercells data.

3. DATA AND METHODS

Our data was created from a ten-year analysis of supercell thunderstorms and surface boundaries in the state of Oklahoma. The supercell data came from a climatology of 950 Oklahoma supercells from 1994-2003 by Hocker and Basara (2008). Surface frontal boundaries associated with each supercell were analyzed from Oklahoma Mesonet surface observations (McPherson et al. 2007). Tornado track data came from a database of tornado track reports from the NOAA Storm Prediction Center and National Climatic Data Center.

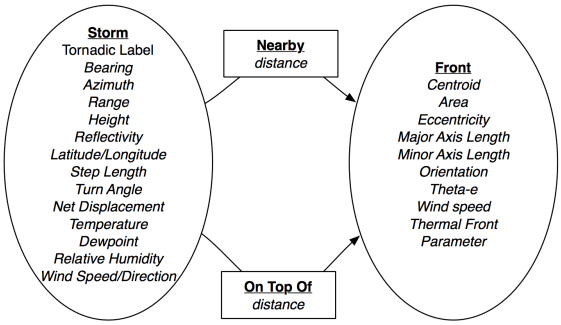

Figure 1: Schema for tornadogenesis data.

Fronts were located objectively in the Oklahoma Mesonet data. Equivalent potential temperature (theta-e), a commonly used measure for front analysis due to its ability to capture both variations in temperature and moisture, was calculated at each Mesonet station. Then a weighted linear interpolation scheme was performed to create a gridded theta-e field. The TFP field was calculated from the theta-e field, and the relative maximums in the TFP field that had a strong enough TFP gradient were designated as fronts initially. Filtering was then done to remove analyzed fronts not meeting a minimum size threshold. A simple time-tracking scheme was implemented and fronts lasting less than 15 minutes were removed. The compiled fronts were then matched with each supercell.

Each group of supercells and frontal boundaries was labeled based on whether or not the supercell produced a tornado. Tornado tracks were matched spatially and temporally with the supercell tracks to determine if a supercell produced a tornado. The front and supercell data were related using the schema shown in Figure 1, where Nearby relationships indicated storms and fronts less than 40 km apart and On Top Of relationships indicated a distance of less than 10 km apart, the typical diameter of a supercell thunderstorm updraft.

To fully explore the algorithms' ranges of predictive ability, multiple runs of each algorithm were performed. 30 runs of 10, 100, 500, and 1000 samples were performed for SRPTs, 10-tree forests, and 50-tree forests. Once the SRPTs and SRRFs are trained, their models are analyzed for skill and variable importance. Skill is measured based on area under the ROC curve, or AUC (Fawcett 2001), where 1 is perfect, .5 is random, and 0 is perfectly wrong. Variable importance estimation analysis was then performed to rank the variables in terms of how significant they were for correct classification of the storms.

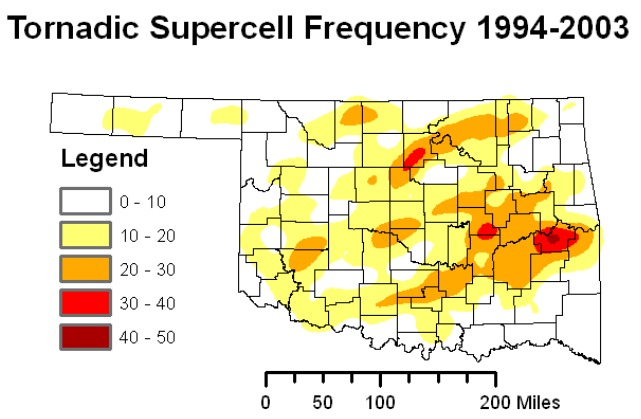

Figure 2: Number of tornadic supercells that have passed within 30 km of a point from 1994-2003.

Table 1: The distribution of tornadic and non-tornadic supercell durations.

	Tornadic	Non-Tornadic
Count	223	727
Proportion	0.235	0.765
Median Duration (hr)	2.71	1.71
Mean Duration (hr)	2.90	1.96
Std. Dev. Duration (hr)	1.48	1.09
Max. Duration (hr)	9.33	7.06
Min. Duration (hr)	0.32	0.08

4. RESULTS

Table 1 shows the class distribution of the supercell thunderstorms and Figure 2 shows the spatial distribution of tornadic supercells in Oklahoma. Most supercells in the data were found to be non-tornadic. Tornadic supercells were found to last an hour longer on average than non-tornadic supercells, a significant (p=0.01) difference. Although duration is well correlated with tornadic supercells, it is not a predictive variable. Fully examining the array of variables in the data required the data mining algorithms.

We first examined AUC as a function of the number of trees in the forest and the number of distinctions sampled at each level. Increasing the tree size resulted in statistically significant improvement for all sample sizes from 1 to 50 trees. No significant gain was made by increasing the number of trees from 50 to 100. Varying the sample size had little effect on the AUC of the forests beyond 100 samples. Although the AUC's indicated some skill, the lack of a larger increase in AUC related to the large increase in number of trees indicated additional issues with the algorithm's interpretation of the dataset. Investigation of the contingency tables revealed that increasing the number of trees had the primary effect of increasing the number of non-tornadic storms correctly classified while the number of correctly classified tornadic storms remained small. The most likely cause of this was the unbalanced class distribution of the tornadic and non-tornadic supercells. Increasing the number of trees in the forest would make the forest more likely to vote for the majority class and less likely to vote for the minority class. To increase the likelihood that the forests would select tornadic storms, we performed a resampling of the training data where we undersampled the non-tornadic supercells to match the number of tornadic supercells.

Figure 3 shows AUC as a function of the same variations but on the resampled data. The 1-tree and 10-tree forests made the largest gains in AUC while the 50-tree forest's AUC improved for lower sample sizes but remained unchanged for higher sample sizes. The 50-tree forest most significantly outperformed the 10-tree forest and single tree at sample sizes less than 500. At larger sample sizes the 10 and 50 tree forests had very similar skill. Inspection of the contingency tables revealed that a much larger number of tornadic storms were selected correctly by the forests, but there was a slight decrease in the number of correctly selected non-tornadic storms. Since selecting tornadic storms correctly is the greater priority, this shift in performance is desirable. Likewise, the performance as measured by the True Skill Statistic (Woodcock 1976) improved dramatically with the resampled data.

The importance of each variable was calculated for the resampled data, as shown in Figure 4. Many of the most important variables dealt with how the storms

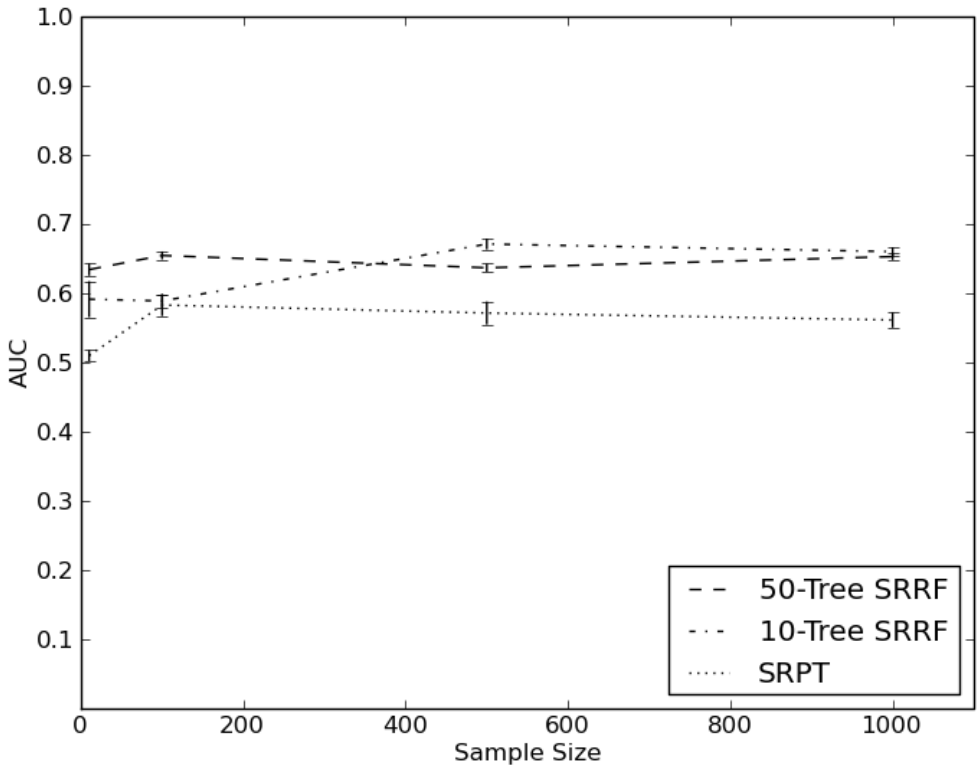

Figure 3: AUC for the resampled Fronts and Tornado Data as a function of sample size for 10- and 50-tree SRRFs and a single SRPT. Error bars indicated 95% confidence intervals.

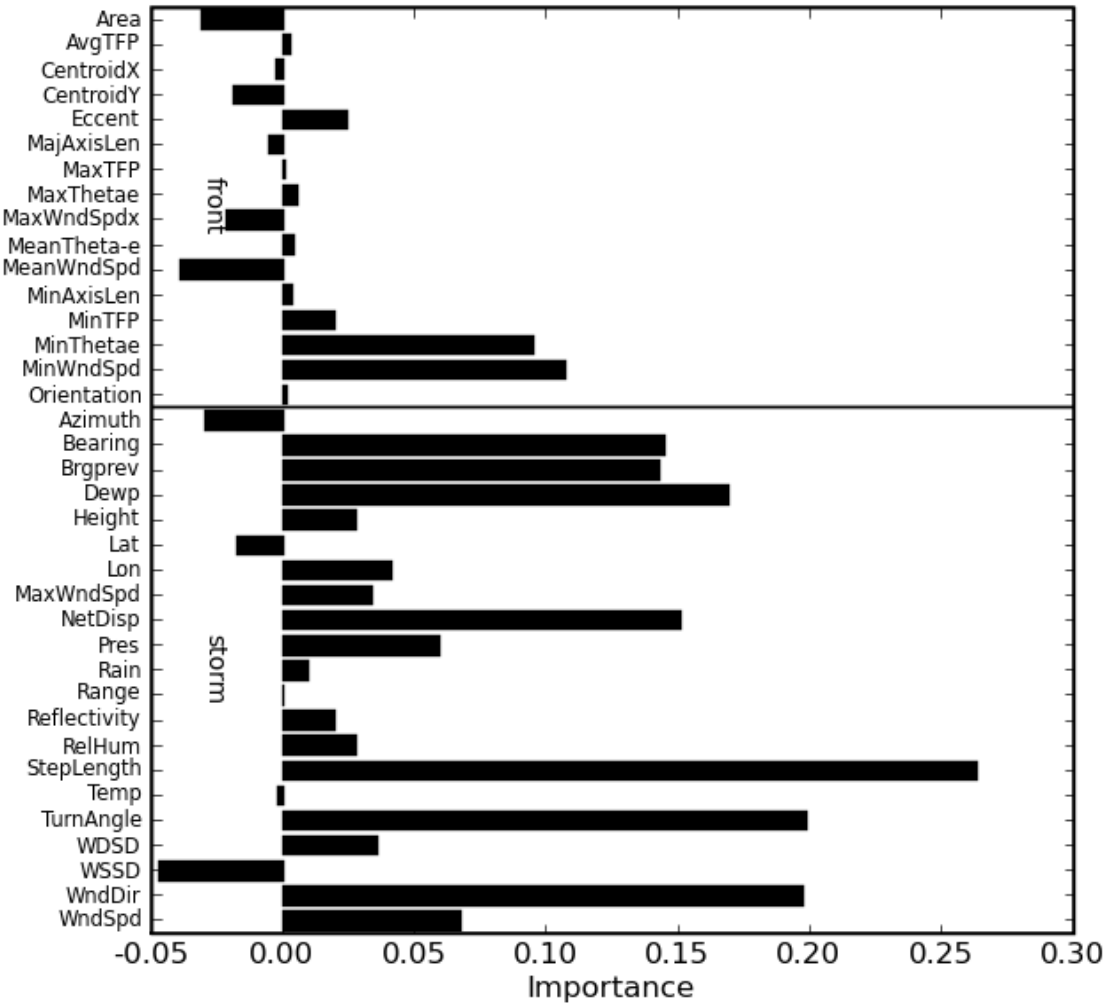

Figure 4: The importance of attributes for the front and tornado data, averaged over 30 runs of a 50-tree SRRF with a sample size of 500.

moved, including bearing and turn angle. The dewpoint temperature and wind direction in the storm environment were also very important. High dewpoint temperatures indicate a large amount of low-level moisture, which is a key ingredient for strong supercell thunderstorms and tornadoes. Wind direction is strongly tied to many storm mechanisms. Front variables also had some effect. The two most important of these were a high theta-e (a combined measure of temperature and moisture) on the cool side of the front, and minimum wind speed. The significance of the wind and storm motion variables potentially indicates that the SRRFs have found patterns differentiating supercells in different flow patterns. By promoting dew point and theta-e, the SRRFs potentially found some connection between the environment's moisture profile and the likelihood of tornadoes.

5. CONCLUSIONS

This paper has shown how the connection between frontal boundaries and supercellular tornadogenesis can be analyzed with spatiotemporal relational data mining algorithms. By learning from a 10-year climatology of supercell thunderstorms and fronts objectively analyzed from surface data, Spatiotemporal Relational Random Forests are able to distinguish tornadic and non-tornadic supercells with some skill. The highest skill scores were found by resampling the data to increase the percentage of tornadic storms and by increasing forest and sample size. The most important variables for detecting tornadic supercells were related to the movement of the storms, the wind direction, and the moisture amounts near the storms and in the frontal zones. The distance of the supercell from the fronts was not determined to have a significant impact on tornadogenesis.

Acknowledgments The authors thank Nathaniel Troutman, Matthew Collier, and James Hocker. This material is based upon work supported by the National Science Foundation under grants NSF/IIS/CAREER 0746816 and NSF/IIS/0938138. The Oklahoma Mesonet is funded by the taxpayers of Oklahoma through the Oklahoma State Regents for Higher Education and the Oklahoma Department of Public Safety.

References

Breiman, L., 2001: Random forests. *Machine Learning*, **45**, 5–32.

Fawcett, F. P. T., 2001: Robust classification for imprecise environments. *Machine Learning*, **42**, 203–231.

Hewson, T., 1998: Objective fronts. *Meteorological Applications*, **5**, 37–65.

Hocker, J. and J. Basara, 2008: A geographic information systems-based analysis of supercells across oklahoma from 1994-2003. *J. Appl. Meteor. Climatol.*, **47**, 1518–1538.

Jenkner, J., M. Sprenger, I. Schwenk, and C. S. S. D. D. Leuenberger, 2009: Detection and climatology of fronts in a high-resolution model reanalysis over the Alps. *Meteorol. Appl.*.

Maddox, R., L. Hoxit, and C. Chappell, 1980: A study of tornadic thunderstorm interactions with thermal boundaries. *Monthly Weather Review*, **108**, 322–336.

Markowski, P., E. Rasmussen, and J. Straka, 1998: The occurrence of tornadoes in supercells interacting with boundaries during VORTEX-95. *Wea. Forecasting*, **13**, 852–859.

McGovern, A., N. Hiers, M. Collier, D. Gagne, and R. Brown, 2008: Spatiotemporal relational probability trees. *Proceedings of the IEEE International Conference on Data Mining*, 935–940.

McPherson, R. A., C. A. Fiebrich, K. C. Crawford, R. L. Elliott, J. R. Kilby, D. L. Grimsley, J. E. Martinez, J. B. Basara, B. G. Illston, D. A. Morris, K. A. Kloesel, S. J. Stadler, A. D. Melvin, A. J. Sutherland, H. Shrivastava, J. D. Carlson, J. M. Wolfinbarger, J. P. Bostic, and D. B. Demko, 2007: Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *J. of Atmos. and Oceanic Technology*, **24**, 301–321.

Rasmussen, E., S. Richardson, J. Straka, P. Markowski, and D. Blanchard, 2000: The association of significant tornadoes with a baroclinic boundary on 2 June 1995. *Mon. Wea. Rev.*, **128**, 174–191.

Renard, R. and L. Clarke, 1965: Experiments in numerical objective frontal analysis. *Mon. Wea. Rev.*, **93**, 547–556.

Supinie, T., A. McGovern, J. Williams, and J. Abernethy, 2009: Spatiotemporal relational random forests. *Proceedings of the IEEE International Conference on Data Mining (ICDM) workshop on Spatiotemporal Data Mining*, IEEE.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.*, **104**, 1209–1214.