TORNADO INTENSITY PREDICTION
BASED ON ENVIRONMENT ELEMENTS AT TORNADO EVENTS STARTING POINTS

A Thesis presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by

XINWEI DU
Dr. Grant Scott, Thesis Supervisor
May 2016
The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

Tornado Intensity Prediction Based on Environment Elements At Tornado Events Starting Points

Xinwei Du,
a candidate for the degree of Master of Science and hereby certify that, in their opinion, it is worthy of acceptance.

____________________________________
Dr. Grant Scott

____________________________________
Dr. Jianlin Cheng

____________________________________
Dr. Curt H. Davis
ACKNOWLEDGMENTS

I would first like to express my sincere gratitude to my advisor Dr. Grant Scott for the continuous support of my Master study and research. His guidance helped me in all the time of research and writing of this thesis. Whatever I am getting in trouble with the technique problem or can not express the idea accurately in my thesis, he is always there to provide help.

And I would like to thanks the rest of my thesis committee members: Dr. Jianlin Cheng and Dr. Curt H.Davis, for their time, encouragement, and insightful comments.

I thank my graduate program advisor: Jodie Lenser, thanks for her help during my whole study time.

Last but not the least, I would like to thank my family: Jun Du and Xiaoxia Zhang for supporting me in my whole life. And my friends: Muxi Chen, Xiaohui Yang, and Rui Xie for encouragement and valuable suggestions they gave for my thesis.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................ ii  
LIST OF TABLES ........................................ vi  
LIST OF FIGURES ........................................ vii  
ABSTRACT ................................................. ix  

CHAPTER  
1 Introduction ........................................ 1  
2 Background ........................................ 8  
3 Problem Definition ................................. 23  
   3.1 Purpose of Research ........................... 23  
   3.2 Data Sets ....................................... 24  
      3.2.1 Tornado data sets ....................... 24  
      3.2.2 Missouri Terrain Data .................. 25  
      3.2.3 Temperature Data Sets ................. 27  
      3.2.4 Precipitation Data Sets ................ 28  
   3.3 Challenges ................................... 28  
4 Methodology .......................................... 31  
   4.1 Research Tools ................................ 31  
   4.2 Data Preprocessing ............................ 32  
      4.2.1 Tornado Data Sets ..................... 32
7.2.2 SVM ......................................................... 72
7.2.3 SVM Prediction & Visualization Based on Raster Image . . . 74

BIBLIOGRAPHY .................................................... 76
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Original F-scale</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Enhanced F-scale</td>
<td>4</td>
</tr>
<tr>
<td>3.1 Tornado data sets attributes table</td>
<td>25</td>
</tr>
<tr>
<td>4.1 z-scores, p-value and confidence level from[29]</td>
<td>48</td>
</tr>
<tr>
<td>4.2 Missouri Tornadoes Amount in decades</td>
<td>49</td>
</tr>
<tr>
<td>4.3 Correlation matrix including slope, elevation</td>
<td>55</td>
</tr>
<tr>
<td>4.4 Correlation matrix including slope, elevation and temperature</td>
<td>55</td>
</tr>
<tr>
<td>4.5 Correlation matrix including slope, elevation and precipitation</td>
<td>56</td>
</tr>
<tr>
<td>4.6 partial correlation results for the slope and width</td>
<td>56</td>
</tr>
<tr>
<td>4.7 partial correlation results for the elevation and width</td>
<td>57</td>
</tr>
<tr>
<td>5.1 Error rate for different predictors choose using decision tree</td>
<td>59</td>
</tr>
<tr>
<td>5.2 Error rate for different predictors choose using SVM</td>
<td>60</td>
</tr>
<tr>
<td>5.3 Error rate for significant tornado prediction while using different predictors</td>
<td>63</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 An example for SVM from[9]</td>
<td>9</td>
</tr>
<tr>
<td>2.2 An example of a decision tree from[17]</td>
<td>14</td>
</tr>
<tr>
<td>2.3 An example of a fully connected neural network from[19]</td>
<td>17</td>
</tr>
<tr>
<td>3.1 U.S. Tornadoes between 1950-2013</td>
<td>25</td>
</tr>
<tr>
<td>3.2 Missouri DEM</td>
<td>26</td>
</tr>
<tr>
<td>3.3 Temperature Data Sets</td>
<td>27</td>
</tr>
<tr>
<td>3.4 Precipitation Data Sets</td>
<td>28</td>
</tr>
<tr>
<td>4.1 Missouri Tornado Events</td>
<td>34</td>
</tr>
<tr>
<td>4.2 Missouri Tornado Startpoint</td>
<td>35</td>
</tr>
<tr>
<td>4.3 ArcGIS slope computation neighborhood from[28]</td>
<td>36</td>
</tr>
<tr>
<td>4.4 Missouri slope map in degree</td>
<td>37</td>
</tr>
<tr>
<td>4.5 Missouri DEM map &amp; Tornado starting points</td>
<td>39</td>
</tr>
<tr>
<td>4.6 Add slope data into tornado data sets</td>
<td>39</td>
</tr>
<tr>
<td>4.7 Temperature weather station</td>
<td>40</td>
</tr>
<tr>
<td>4.8 Missouri tornado events and temperature weather station</td>
<td>41</td>
</tr>
<tr>
<td>4.9 Tornado events and Temperature weather station buffer</td>
<td>41</td>
</tr>
</tbody>
</table>
4.10 Precipitation station buffer and tornado events ........................................ 43
4.11 MO tornado F-scale HotSpot ................................................................. 45
4.12 MO tornado starting point F-scale HotSpot ............................................. 46
4.13 z-scores and p-value distribution from[29] ............................................... 47

5.1 Decision Tree Model Example ............................................................... 60
5.2 Missouri Possible Significant Tornado Occurring Place Based on Elevation and Slope Using SVM Prediction Model .......................... 64
ABSTRACT

In this research, we analyze the behavior of tornado events which occurred in the State of Missouri between 1950-2013. The formation of tornadoes is a complex meteorological phenomenon which is still not fully understood to the level of accurate predictability based on current atmospheric sensor networks, e.g., Doppler Radar. Environmental conditions at tornado start points may play a key role in determining the intensity; and therefore we explore the potential of using environment features as tornado intensity predictors. We analyze both the spatial and temporal features of historical Missouri tornado events. Environmental features at tornado starting point; such as ground slope, elevation, temperature and precipitation are explored as the predictors of tornado intensity through application of Decision Tree and support vector machines (SVM) models.
Chapter 1

Introduction

Tornadoes are a kind of a damaging weather condition which is pretty common in U.S. The United States has the most frequent tornado in the world. Every year, more than 90% of tornadoes occur in U.S. The United States receives more than 1,200 tornadoes annually—four times the amount seen in Europe. Violent tornadoes—those rated EF4 or EF5 on the Enhanced Fujita Scale—occur more often in the United States than in any other countries.[1] The United States receives over 80 deaths and 1,500 injuries associated with tornadoes each year.[2] These events cause more than 1.2 billion dollars in losses every year since this tornado statistic during 1995-2014.[3] 2011 is one of the most serious years in the history, it caused 9.49 billion dollars losses and 553 deaths. Therefore, research continues in the interest of better prediction tornado occurrence and strength; in hopes of reducing the damage caused by tornadoes.

So, firstly, what exactly is this kind of damaging weather condition? According to Glossary of Meteorology (AMS 2000), a tornado is "a violently rotating column of air, in contact with the ground, either pendant from a cumuliform cloud or underneath a
cumuliform cloud, and often (but not always) visible as a funnel cloud". In fact the wind is invisible, but the reason why we actual can see the tornado in real life is "it forms a condensation funnel made up of water droplets, dust and debris. Tornadoes are the most violent of all atmospheric storms".

The brutal truth is that we actually do not fully understand how tornadoes form. "Tornado formation is believed to be dictated mainly by things which happen on the storm scale, in and around the mesocyclone." The most common explanation right now is the tornado relies on strong convective weather condition. According to National Weather Service, "once a mesocyclone is underway, tornado development is related to the temperature differences across the edge of downdraft air wrapping around the mesocyclone." This means because of the temperature difference between the air near ground and high altitude, it would produce convective draft. In the interaction of warm draft and cold draft, the faster moving air begins to spin and roll over the slower wind. At the same time the horizontal wind begins spinning and rolling the vortex like a cylinder. On other hand, this cylinder causes more warm air to join the updraft, making the vortex become stronger and more powerful.

So in summary, there are two necessary requirements for tornado formation, first would be instability, it refers to unusually warm and humid conditions in the lower atmosphere; second would be wind shear, refers to the wind direction changing, and the wind speed increasing, with height.

These two factors also determine the tornado intensity. In general we use Fujita-Scale to represent the tornado intensity. Dr. T. Theodore Fujita developed a damage scale (Fujita 1971, Fujita and Pearson 1973) for winds, including tornadoes, which was supposed to relate the degree of damage to the intensity of the wind. This F-Scale
is a representative of wind speed. But this wind speed on original F-scale have never been scientifically tested and proven. The original scale evaluate form is as table1. The original F-scale should not be used anymore, because it has been replaced by an enhanced version. The Enhanced F-scale[5] is a much more precise and robust way to assess tornado damage than the original. It classifies F0-F5 damage as calibrated by engineers and meteorologists across 28 different types of damage indicators (mainly various kinds of buildings, but also a few other structures as well as trees).[6] The detail information about the enhanced F-scale is as table 2. The enhanced F-scale is still is a set of wind estimates, it use uses three-second gusts estimated at the point of damage based on a judgment of 8 levels of damage to the 28 indicators. The enhanced F-scale took effect 1 February 2007, and there are no plans to systematically re-evaluate historical tornadoes using the Enhanced F-scale. So all of the data from before 2007 would uses the original F-scale.

In modern tornado forecast, people are using Doppler radar to monitor the atmosphere. They look for the development of temperature and wind flow patterns in the atmosphere which can cause enough moisture, instability, lift, and wind shear for tornadic thunderstorms. But the atmosphere is very complex, it is hard to get a threshold which could be used as standard for the tornado.

As such, forecasters use no single threshold or criteria for either a watch or a warning. Watches and warnings instead are fast-action judgment calls, based on numerous factors. SPC watch forecasters look for favorable overlaps of moisture, instability, lift, and vertical wind shear, for at least a few hours, over a concentrated area the size of a typical watch. For tornado warnings, some storms with little or no radar-detected rotation can produce weak tornadoes, while other storms with frightening-looking
Table 1.1: Original F-scale

<table>
<thead>
<tr>
<th>SCALE</th>
<th>WIND ESTIMATE *** (MPH)</th>
<th>TYPICAL DAMAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>&lt;73</td>
<td>Light damage. Some damage to chimneys; branches broken off trees; shallow-rooted trees pushed over; sign boards damaged.</td>
</tr>
<tr>
<td>F1</td>
<td>73-112</td>
<td>Moderate damage. Peels surface off roofs; mobile homes pushed off foundations or overturned; moving autos blown off roads.</td>
</tr>
<tr>
<td>F2</td>
<td>113-157</td>
<td>Considerable damage. Roofs torn off frame houses; mobile homes demolished; boxcars overturned; large trees snapped or uprooted; light-object missiles generated; cars lifted off ground.</td>
</tr>
<tr>
<td>F3</td>
<td>158-206</td>
<td>Severe damage. Roofs and some walls torn off well-constructed houses; trains overturned; most trees in forest uprooted; heavy cars lifted off the ground and thrown.</td>
</tr>
<tr>
<td>F4</td>
<td>207-260</td>
<td>Devastating damage. Well-constructed houses leveled; structures with weak foundations blown away some distance; cars thrown and large missiles generated.</td>
</tr>
<tr>
<td>F5</td>
<td>261-318</td>
<td>Incredible damage. Strong frame houses leveled off foundations and swept away; automobile-sized missiles fly through the air in excess of 100 meters (109 yds); trees debarked; incredible phenomena will occur.</td>
</tr>
</tbody>
</table>

Table 1.2: Enhanced F-scale

<table>
<thead>
<tr>
<th>FUJITA SCALE</th>
<th>DERIVED EF SCALE</th>
<th>OPERATIONAL EF SCALE</th>
</tr>
</thead>
<tbody>
<tr>
<td>F Number</td>
<td>Fastest 1/4-mile (mph)</td>
<td>3 Second Gust (mph)</td>
</tr>
<tr>
<td>0</td>
<td>40-72</td>
<td>45-78</td>
</tr>
<tr>
<td>1</td>
<td>73-112</td>
<td>79-117</td>
</tr>
<tr>
<td>2</td>
<td>113-157</td>
<td>118-161</td>
</tr>
<tr>
<td>3</td>
<td>158-207</td>
<td>162-209</td>
</tr>
<tr>
<td>4</td>
<td>208-260</td>
<td>210-261</td>
</tr>
<tr>
<td>5</td>
<td>261-318</td>
<td>262-317</td>
</tr>
</tbody>
</table>
circulations on radar displays still yield no tornado at all. Because of that variability, local NWS forecasters look at not only radar velocity, but any of many other radar products, spotter reports, analysis of the storm environment, history of existing storms, SPC guidance, short-fused weather models, and non-meteorological considerations such as potential human impact.[6] So the tornado forecast is a comprehensive judgement according to the reader, history contents, experience and etc. Currently we still cannot predict tornadoes as precisely as temperature or rain.

Other than the traditional climate forecasting method based on remote sensing, people have also begun to use machine learning methods to research tornado and other climate conditions. People are trying to use machine learning algorithms to generate a model which would fit history data well and then use this model to predict future situations. These methods often are not reliable compared to the analysis satellite nephogram. However, using machine learning methods to predict weather is becoming more and more common, and achieve increasingly interesting and useful results. These methods have become a useful supplement for the traditional weather research methods. Chapter 2 will provide more details.

Machine learning is one of the subfields of computer science, it explores the study and construction of algorithms and models that can learn from the data. Which means adjust the parameters that make the model could represent the data better and make more precise prediction for the future condition.

There are many different algorithms in machine learning which are suited to a variety of different types of problem. These methods can be separated by into two types; supervised learning and unsupervised learning. Supervised learning requires labelled data for algorithms to learn, which unsupervised data discover hidden pat-
terns within the data without the aid of labels. The supervised learning, for example, for the tornadoes; we know many tornado event information, like time, location and weather information at this time. Then we can use those information as predictors and labels to mark those tornado events. While we generate the tornado events–time, location, weather model, we can make prediction for the new tornado events, if we find some weather condition while appear in some place at a certain time, then it would be possible that here would occur a tornado. In regards to unsupervised learning, we just know there are many tornado cases, but those data sets do not provide any label to make classification predictions; so what we can do is trying to spatially cluster those tornado events to several groups. For example, if we do cluster analysis for the tornadoes according to the locations, we would find where the tornado are highly concentrated, so as to identify some place as a ”tornado valley”.

There is not exist best algorithms only exist the most suitable methods. Sometimes we can use many method to get the good results, another time might be only one algorithm could work on it. The weather system which is almost the most complex system, it is hard to develop a perfect model to simulating the whole weather system, what we can do is trying to pick one or several parts of the weather condition to analysis it.

In this manuscript I investigate methods to predict the tornado intensity based on environmental predictors for the tornado start point, including slope, elevation, temperature and precipitation. Firstly, after doing analysis for tornado events basic temporal and spatial distribution, it shows the tornado event patterns exist in some local cluster and temporal trends. I Then extract the four predictors (elevation, slope, precipitation and temperature) for the tornado start points from the topographical
maps using ArcGIS. I have chosen to focus my research area within the state of Missouri, USA. Tornado data is from the National Weather Service records of the tornadoes which have occurred between 1950-2013. After combining those spatial features and the tornado events together, I then analyze the relationships between these predictors and the tornado intensity—represented by tornado width, length and Fujita Scale. Correlations between indexes were also calculated. Prediction algorithms are performed in R statistical programming package.

The remainder of this manuscript is organized as follows. Chapter 2 introduce some research background for this topic. Chapter 3 provides the problem definition and introduces the data sets which have been used for this research. Chapter 4 details the methodology. Chapter 5 would discusses the results. Finally, Chapter 6 discusses conclusions and future work.
Chapter 2

Background

Researchers have produced good research results using machine learning methods for tornado prediction. These research methods include both supervised learning–like support vector machine (SVM) or artificial neural networks (ANNs)—and also unsupervised learning method–like hierarchical clustering (HC).

One of the most commonly used methods is support vector machines (SVM). Support vector machines belong to supervised learning models. Vapnik and Chervonenkis in 1963 first time mentioned the SVM. In 1992, Vapnik and other two people[7] suggest a way to create nonlinear classifiers through lead the kernel function to convert the problem to maximum-margin hyperplanes. And the current standard version was proposed by Corinna Cortes and Vapnik in 1993[8].

The SVM was designed to solve the classification problem which is the most common problem in machine learning. We know that if there exist several points, and we can found a line or hyperplane to divided those points into several parts, for the classification. For a given situation, there might exist multiple hyperplanes, all
of them are working for classifying the points. So we choose the hyperplane that the
distance from it to the nearest data point on each side is maximized as the maximum-
margin hyperplane or linear classifier. And the way to find the maximized-margin
hyperplane is called linear SVM. But we cannot ensue that every time the data set is
linearly separable, so to solve the nonlinear classification, the kernel machine or called
kernel function, was being introduced. That means; we can use the kernel function to
transfer original nonlinear object to high dimensions, which make it become a linear
classification problem in high dimensions. But the different between kernel function
and simply up dimension is kernel function is still computed in lower dimension
instead of in high dimensions. The Figure 2.1[9] show an example for how SVM
convert a nonlinear classification in low dimension into a linear classification in high
dimensions.

In [10], Trafalis, Adrianto, and Richman present their approaches to predict tor-
nado by using support vector machines. They are using SVM to help them labeling
data, like which one storm-scale circulations would produce tornadoes in the ground,
which is the most time consumable work in tornado prediction. They pick some fore-
cast correct rate evaluating standard or skill scores such as Critical Success Index
(CSI), Probability of Detection (POD), False Alarm Ratio (FAR), Bias, and Heidke Skill Score (HSS) to see the SVM predict correct rate. They set the data to a training set and a testing set and also make a comparison between different training set. They got the conclusion that active learning could significantly reduce the training data set size to get the same predict results compared to passive learning. The reduced training set means update classifier dynamically become possible.

Adrianto, Traflis, and Lakshmanan use SVM in another way. In [11], they trying to "estimate the probability of a tornado event at a particular spatial location within a given time window". They are using least-squares methodology to estimate shear, quality control of radar reflectivity, morphological image processing to estimate gradients and fuzzy logic to generate the compact tornado possibility, then use SVM to classify to get the final spatiotemporal probability field. And they also use Heidke's Skill Score and Critical Success Index to evaluate the prediction results.

Their idea is first according to MDA ground truth database to create the truth field. Which means this place has truly tornado happens with 20 minutes after observer the radar signal. Then they build the tornado possibility according to many other research results, like tornado are more likely to occur in the south-west region of the storm. They aggregating spatial fields of areas with tight gradients in the appropriate direction, areas proximate to high positive and negative shear and high reflectivity values using fuzzy logic weighted aggregate. The fuzzy possibility means some case is "only partly characterized by randomness and defines a pure probabilistic modeling with certainty due to a lack of trustworthiness or precision of the data or a lack of pertinent information." [12] And the breakpoint was being decided manually. Finally, they compare the tornado possibility region and truth region, to make a
classification for which region is tornadic and which are nontornadic, and this classifier was trained through SVM.

Their results were pretty good; they successfully find some high tornadic probability regions in future 30 minutes. But there are also some constraints because the method of labeling the tornado possibility region was manual, so it become pretty time consumable even not useful while the system receives the new data. Because it needs time to update the SVM classifier with new data points added in the training set.

Trafails, Santosa, Richman published a paper in 2004[13] to introduced their works. They developed a hybrid forecast system for the distinguish the tornadic from non-tornadic events by generating twenty rules for SVM classifiers. And they found this hybrid system could have 12.7 percent accurate improve compare to just use SVM directly.

In another paper[14], Trafalis, Ince, and Richman made a comparison four classification method Support Vector Machines (SVM), linear discriminant analysis (LDA), neural networks (NN) and Mesocyclone Detection Algorithm(MDA). And their results shows the SVM is the most accurate algorithm.

The LDA in some degree can be seen as an SVM lite edition. The main idea is trying to label the data and transforming those data points into the lower dimension, the more close points or objects would become more close after transforming. For instance, LDA transforming the one class of points into a line, others into another line, so we can separate those two kinds of points. But the LDA is only a linear classification approach; it cannot deal with the unlined situations. And it required some prerequisites like normality and homoscedasticity of the distribution, which are
not always valid in tornado prediction.

The MDA is a server weather detect algorithm based on Doppler radar data. The first step of MDA is to "search for a consistent increase of Doppler velocity in the azimuthal direction at a constant range[15]". Then use seven feature components: the slant range, the azimuth angles at both ends of the run, the Doppler velocities that correspond to those azimuth angles at the slant range, and the SHEAR (Tangential) and MOMENTUM (Angular) to consist a pattern vector. And all of the pattern vector that containing magnitudes of angular momentum and azimuthal shear typical of mesocyclones are from "features." If an element is too small it is discarded. If an element is sufficiently large and not symmetrical, it is classified as a shear region. Sufficiently large, symmetric shear regions are characteristic of mesocyclones. If these regions are in close vertical proximity, a mesocyclone is identified. Shear regions in close vertical proximity identify 3-D shear regions. The remaining features characterize uncorrelated shear.” So the MDA algorithm is a rule-based algorithm and not show the relationship between the tornado and the input variables. Those ruled base algorithms perform not as well as the supervised learning algorithms like SVM and ANN at most of the situations. Because not only the ruled might be wrong or not suitable for all situations and also those rules may not cover all possible contingency.

In[16], Holmes introduced his research results about using decision tree and neural net to identify sever weather Radar characteristics. He actually combining the decision tree and neural net together to get better results. The decision tree was used to help identify the more influential attributes for server weather classification. So just the important predictors were used for neural net. Through using decision tree reduce the number of inputs for the neural net could help to reducing overfitting and
providing faster training speed.

The decision tree is a popular tool in machine learning. The name comes from the way it displays the analysis conclusion, as an inverted tree. Figure 2.2[17] shows an example of a decision tree. The targets of the decision tree are to set up a classification or regression prediction model. A decision tree is a flowchart-like structure, includes root node, leaf nodes, and intermediate nodes. The root node is the node at the top of the tree, and there is one and only one root node per tree. The nodes without any child nodes are named leaf node, and the nodes beyond the root node and have child nodes are named intermediate node. Every node in the tree would include a specific number of samples. Root node includes all of the samples, and node in each level have fewer and fewer sample. Every path to leaf node would correspond to an inference rule. So for every new object predictions, just according to the input value enter the different branches until the leaf node.

For a classification and regression tree, the first step is to “grow” the tree from the root node, using sample data sets to train the model. The reduction process is the process dividing the data among nodes based on decision points. The branches in the tree are grown during the data division procedure. When a node has homogenous data samples, it is meaningless to continue reduction, so this node becomes a leaf.

To avoid overfitting, we need a satisfied measure to know when to end reduction. The standard for termination of a branch is: all of the samples in this node have the same class (i.e., homogeneity), or it satisfied threshold in achieved. A commonly used measure for decision tree reduction is Gini coefficient. Gini is used to measure
Figure 2.2: An example of a decision tree from[17]
the different degree of output variables in a node. It is defined as:

\[ G(t) = 1 - \sum_{j=1}^{k} P^2(j|t) \]  \hspace{1cm} (2.1)

where \( t \) is the node; \( k \) is the number of class as for the output variables; \( P(j|t) \) is the normalized density of the output variables in node \( t \) that belong to the class \( j \). The normalization ensures that each node’s probability is comparable. The classification decision tree uses the Gini coefficient to measure the decreasing degree of heterogeneity. So the best threshold to stop the tree growth is the point that makes the change rate of \( G(t) - \delta G(t) \) largest.

For regression trees, it is pretty similar to the classification tree. However, the decision variables change from Gini to variance. Since the regression tree output variables are numeric, the variance is a suitable indicator to show the heterogeneity. It is defined as:

\[ R(t) = \frac{1}{N_t - 1} \sum_{i=1}^{N} [y_i(t) - \bar{y}(t)]^2 \]  \hspace{1cm} (2.2)

In the formula, \( t \) is the node, \( N_t \) is the number of samples containing in the node \( t \); \( y_i(t) \) is the output value for the i input in node \( t \); \( \bar{y}(t) \) is the mean value of node \( t \) output variable. The decision point that make the change rate of \( R(t) \) largest would be the best decision threshold.

After the reduction of the tree is complete, the next step is pruning. Since the reduction process is the node partitioning, it means every subsequence node in a branch will have less samples compare to the parent nodes. So deeper the nodes, the more individuation the data characteristic. That is means this rule would be precise enough but not so generalized, which leads to overfitting and less re-usable for new
data prediction. So to avoid overfitting, we prune the tree, removing overly restriction rules and ensure the tree has a suitable generality.

Generally, the pruning include pre-pruning and post-pruning. The pre-pruning is trying to avoiding the tree over growth, and the post-pruning is pruning the tree after reduction is complete, then do the pruning. There are two ways to do the pre-pruning. The first one is set the limit on depth of the tree. The second is to set the minimum number of samples that one node should have; the node cannot branching if the number of samples in this node is less than this threshold. For the post-pruning, it just to computing the error rate while pruning the branches, it would stop pruning until the error rate reach a threshold.

The method used to post-pruning for class-regression tree is called MCCP (minimal cost complexity pruning). As discussed above, the post-pruning is using error rate as the threshold. So, using the error rate as the cost function, and the number of leaves in the tree as the complexity, for the decision tree $T$, the cost complexity is defined as:

$$R_{\alpha}(T) = R(T) + \alpha |T|$$  \hspace{1cm} (2.3)

$R(T)$ is the prediction error for testing data sets, for classification tree it is the false positive rate, for regression tree it is the normalized residual error. $|T|$ is the number of the leaves. $\alpha$ is the complexity parameter, or CP. It represent the increased complexity by increasing every one single node.

The neural networks is another commonly used machine learning algorithms for tornado prediction. It also is called as artificial neural networks (ANNs). ANN is a bionics product. It is inspired by the biological nervous systems-neuron, nervous network-processing information. The center element of ANN is the novel structure of
the information processing system.[18] It means how to set the ”transform” rules for each neuron. Like for a neuron set if it sees there exist any number 1 in the input than output number 2. There is an example of a fully connected neural network in Figure 2.3[9].

The structure of an ANN can be divided into three parts, or layers, one for ”input”, one for ”hidden layer”, and an ”output”. The input unit represent the raw data that is provided to the network; the hidden units indicate this units’ work are being hidden, it was the intermedia step of the processing, it determined by the activities of the input units and the weights on the connections between the input and the hidden units; the output, obviously provided the final results of this network, and the value of output units are determined by hidden units and the weights between the hidden and output units. Of course, this is just most typical and simple ANN structure, according to
demand, it has many variant. ANN is good at "capturing existing patterns and trends from noisy data"[18] which make it become a good approach to analysis tornado.

The ANN is applied to tornado prediction earlier than SVM. In 1995, Marzban and Stympf[20] introduced their method to use ANN in tornado prediction. In this paper, the authors divided the circulation into several different situation like "severe" and "tornadic", associated those circulations with the history "truthed" tornado data to consist the training sets for ANN. And then using the ANN to predict the new events. And they also mentioned their neural network was modified new version. The new version allowed them to adjust the network’s weights which are usually un-interpretable. Also, they make a comparison of the predict results between ANN and MDA. The results showed the ANN outperformed than MDA. The most interesting part of their works is after they find the optimal network or model, they no longer request the system give the answer for tornado prediction with "yes" or "no". "It is possible to transform the output of the network to probability, reflecting the level of confidence associated with a given outcome."[20] The probability way could make the tornadic prediction curve more close to the truth tornadic events curve.

In [21], Lakshmanan, Stumpf, and Witt introduce their ANNs which has added some variations that improve the performance of the network generate by Stumpf and Marzipan. According to their comments, the first upgrade they made is they compute the minimized weight as the sum of the cross-entropy and the sum of the square of weights which could decrease the overfitting. And they divided the training data case-wise instead of the pattern wise. Because they consider that the MDA detections from a single data tend to be highly correlated, the testing set would correlate to the training set in patter-wise situation and would bring spurious global
minimum. Besides, they pruned input features and outliers automatically using the prior threshold rather than manually.

In 2005, Trafails, Santosa and Richman published a paper[?] to introduce their work on using different types of algorithms in tornado prediction. Including artificial neural networks (ANNs), Bayesian neural networks (BNNs), support vector machines (SVMs) and Bayesian support vector machines (BSVMs). The BNNs is a kind of new network, not like the ANNs, it provide an objective for the network, and every point in the network has an actual meaning, the edge between the points represent the causality. BNNs avoid some problem may exist in ANNs, the most obvious one is the overfitting. Because the ANNs’ object is to find the minimum error, when the model going to too complex, it becomes easy to lead to an overfitting situation. The BSVMs is kind of an upgrade version for the traditional SVM. It is applied using trigonometric loss function[22] to "integrate Bayesian inference with SVM smoothly while preserving their individual merits. This trigonometric loss function could help to provide the desirable property of sparseness in sample selection. Their results show the BSVM is the most accurate learning network. BNN and SVM almost got the same accurate degree; both are better than ANNs.

Those research approaches have one thing in common. All of them are using machine learning methods to make classify. They are trying to recognize which region would occur tornado which are would not. Those approaches do get a lot of good results, but they are also existing some problems like local optimal, timing consuming and cannot complete the real time predict (because it need adjust the model in the new event). The classification algorithm is the most direct but not the only way to predict the tornado.
In [23], Lakshmanan, Adriano, Smith and Stumpf claim a new approach for tornado prediction. In this paper they consider the problem as spatiotemporal one, trying to estimate the probability of a tornado event at a particular spatial location within a given time window. The authors mentioned one of the motivation for this research is the previous statistic classification methods are not performed so well, one of the possible reason is that the non-tornadic area is much bigger than tornadic but the original data appear very small difference in space.

They were first forming the truth field which is quite likely for the previous researchers’ work. Then they design the different detecting method, which is setting the observed data would only be used to make prediction advected backward and forward corresponding to the movement of the observed data over time. And they were also using ANNs to help them do the quality control for the radar reflectivity-to clean the noisy. Besides, they use a new method named LLSD(linear least squared)[24] to computing the rotational and divergent shear from Doppler radial velocity data. Then they aggregated the fields with tight gradients with the fuzzy logic weighted aggregate as the tornado probability fields. The probability fields were then been clustered using region growing and the properties of each region were determined. The properties were computed form the values at each pixel in the region of any other spatial fields. Finally, they compared the regions and the ground truth, if a corresponding tornado was observed in the ground truth filed the region would be marked as tornadic. They trained a feed-forward neural network as the classifier.

Above are some current research results for using machine learning method for the tornado occurrence predictions. And except the tornado occurrence predictions, machine learning could also be used in tornado intensity analysis.
Most of the research achievements about tornado intensity prediction are in climatology area. Researchers are trying to explore how tornado intensity correlate to some meteorological features like velocity wind shear and helicity.

Like Colquhoun and Riley’s [25] analysis the relationships between the tornado intensity and various stability and wind-related parameters. They first analysis the relationships between tornado intensity and a wild range meteorological indexes, such as mean hodographs, mean temperature and dewpoint temperature soundings. And they found the tornado Fujita-Scale is significantly correlating with convective available potential energy (CAPE), no correlations of F-scale with 600 hPa wind shear (S6) or bulk Richardson number (BRi). Then they given a equations consist of lifted index (Li), S6, stream wise vorticity (SV), storm-relative environmental helicity (SREH), BRi, and energy helicity index (EHI) to express F-scale.

And recently, in 2015, March 19, Lopes and Machado publish a paper [26] provide a new angle for tornado prediction. This paper is kind of a general statistic analysis for the tornado happened on U.S. This paper consider the annual tornado as a time series sequence, they found the annual tornado intensity can be modeled as “time series of Dirac impulses with amplitude proportional to the index quantifying the size of the events”. They using the database containing the tornado events occurred in the US during 1950 to 2013 as research object, first using means of the Fourier transform to analyze the annual time series in frequency domain; then ”adopt the concept of circular time and compute the circular correlation between the annual time series”, finally use hierarchical clustering (HC) and multidimensional scaling techniques (MDS), which both are unsupervised learning method to visualize the patterns. So from the fist step work the results show that tornadoes collective behavior
exhibit correlations and characteristic patterns, but they also said the complementary analysis is needed to reveal deeper characteristics of this phenomena. In step two, they are trying to use hierarchical clustering and multidimensional techniques to cluster the every year during the 1950-2013 period. The hierarchical clustering is one of the cluster methods that using the distance between the objects to judge whether two or more items could be cluster into one cluster. "The MDS is a means of visualizing the level of similarity of individual cases of a dataset."[27] And for the results, they fond the data are failed to fit into a simple type of periodicity. This can be interpreted as the time period length is insufficient to grab some period or some long run behavior is present and needs to be further analyzed. They also said ”Embedding other variables, namely trajectories, pressures, temperatures and humidity may help to get a clear picture of the global pattern.” So in summary Lopes and Machado’s creative work–consider the tornado intensity as time series object–showing the tornado existing time concentration. But it also shows the tornado is too complicated and chaotic, which make it almost impossible to be exhibit by a single model, we need ”generate more approaches and yield important insights to better understand the phenomenon".
Chapter 3

Problem Definition

3.1 Purpose of Research

The purpose of this research is to analyze the relationships between the tornado intensity and the geospatial features at the tornado event start points. A goal is to explore a machine learning model to simulate and predict the tornado intensity based on the outset point environmental features. This analysis is using the state of Missouri as the research area. We limit our focus to the State of Missouri, because the data scale is too large for whole U.S.

The tornado intensity is represented by three variables: tornado length, tornado width, and Fujita-Scale. The Fujita-Scale can represent the tornado wind speed. The geospatial features includes elevation, slope, temperature and precipitation.

This research attempts an initial exploration of how the terrain and the environment condition influence the tornado intensity; trying to find another way to analyze
and predict the tornado intensity besides the traditional meteorology methods. Since the terrain and environment condition would be much easier to catch and analyze compared to the sophisticated atmosphere system, we believe this worths investigation. This research could become a useful supplement for the traditional tornado predict analysis based on professional meteorological knowledge.

3.2 Data Sets

The research data sets are consisting of four parts. Tornado records from 1950-2013; spatial terrain map layer for Missouri; average temperature data records for Missouri and the hourly precipitation data sets.

3.2.1 Tornado data sets

Tornado data sets come form the National Weather Service. Total number of tornado events in this data sets is 58959, counting all of the tornado records from 1950-2013 (right now it update to a new version, which contains 1950-2014 tornadoes, data change details can be seen in www.spc.noaa.gov/gis/svrgis/). The data is provide as a ESRI Shapefile. The ESRI Shapefile is one of the most commonly use GIS file type. Developed by ESRI company. It is a vector data format used to store the location, shape and attributes of geographic features in vector formate. The tornado data set is like Figure 3.1, which was displayed extracted from ArcMap.

The attribute table for this tornado layer is shown in Table 3.1. This tornado data set includes about 59000 tornado events between 1950-2013; each event including more than 20 detail of information, such as occurred time, start location, end location,
tornado length, width, Fujita-Scale, fatalities, injuries and property loss.

The terrain data sets (discussed in section 3.2.2) are too large to analyze for the entire U.S. as the research area. Therefore, I use Missouri State as the research focus area; only the tornadoes happened within Missouri State would be utilized in this research. The Missouri tornado data is extracted from the full US tornado data sets.

### 3.2.2 Missouri Terrain Data

The terrain data is a digital elevation model data (DEM). DEM use an ordinary matrix to represent the ground elevation. Through using some analysis tools, like
ArcGIS, we can get DTM from the DEM. DEM is one kind of DTM (digital terrain model data), DTM would including all of the terrain features like slope and abstract. The terrain data used in this research is from Missouri Spatial Data Information Service (MSDIS http://www.msdis.missouri.edu/data/dem/index.html). It provides entire Missouri’ DEM map. Figure 3.2 shown how Missouri DEM map looks like.
3.2.3 Temperature Data Sets

Real-time temperature data sets for the moments while the tornado happened were not found for use in this research. Therefore, I use the normal temperature data sets from National Centers For Environmental Information (NOAA, ftp://ftp.ncdc.noaa.gov/pub/data/normals/1981-2010/products/). The temperature data is a normals data set for temperatures measured between 1981-2010. It including two separate data sets; one is the temperature station information, including name, location and latitude/longitude coordinates. Another one is the temperature measured data by each station for every hour in every day of the year.

The structure of temperature data sets is show in Figure 3.3. The first column is the weather station id, and then the month and date, and then 24 columns for hourly temperature on that data. Since each station has at least 365 records per years, between 1981-2010, and we have 329391 temperature records.
3.2.4 Precipitation Data Sets

The precipitation data also comes from NOAA. The hourly precipitation data sets tend to be fairly large if we want to cover the same period with the tornado events. So I just use the 1990-1995 Missouri State precipitation data sets for this research. The precipitation data sets also consist of the two parts, one is the measured precipitation data, another is the precipitation measurement station data sets.

The precipitation data sets are organized as Figure 3.4. The precipitation data set have 101493 records.

![Figure 3.4: Precipitation Data Sets](image)

3.3 Challenges

The challenges for this research mainly appears in two aspects. First one is how to integrate the multiple environment data—like the spatial features, slope, elevation, temperature and precipitation—to every single tornado events. In other words, I need to add those features value to every single tornado event. For example, the raw tornado data sets only containing the tornado intensity information, but we need to add the slope and elevation value to the tornado events based on the tornado starting
point location, add temperature and precipitation value to tornado events based on both location and time.

It is necessary to standardize all of the data sets in same time range and spatial consistency. Spatial consistency, ensures data points across data sets are locate in the same area; and time consistency, ensures the same period of the time. The four environment features are, slope, elevation, temperature, precipitation; the fist two require spatial consistency only, but the last two require both spatial and time consistency, which make them more complicated to pairing with the tornado events. That is the reason temperature and precipitation data sets are form weather station data sets. We need this station information to filter the tornado events that could fit the specific temperature and precipitation data.

In order to reach the needed consistency, the data sets sizes become the largest barriers. The data size is the reason why just choose Missouri instead of the whole U.S as the research area. If the spatial terrain data sets are too large, then it would impossible to complete the analysis and extract the value of environment features using regular resources and ArcGIS. In order to aggregate all counties' DEM into a whole Missouri State DEM, it takes about a whole day on my PC. The final results of Missouri DEM is about 2.68GB. This is just first step of processing the geospatial data; the following would need to do slope analysis which requires even more computational resources.

Another challenge is how to generate the prediction model. My goal is to predict the tornado intensity (expressed by tornado width, length and Fujita-Scale) based on the environmental information (elevation, slope, temperature, precipitation) at the tornado occurring point. So firstly, the correlation analysis is essential, which make
the environment features are suitable predictors for the prediction algorithms. We explore multiple prediction models to get the best prediction results.
Chapter 4

Methodology

4.1 Research Tools

There are several primary software tools in this research. First one is ArcGIS; an umbrella name for a series professional geography problem analysis tools, like ArcMap and ArcCatalog. In this research, I mainly use ArcMap which is the core software for the ArcGIS tool sets. ArcMap provides strong abilities to deal with the geospatial data. ArcMap provides all functions base on the map. It could help to do the map design, producing, and analysis. In this research, ArcMap helps to produce the Missouri State DEM through those Missouri counties’ DEM. We then use ArcMap to perform analysis this state DEM to get the slope data. Afterwards, we combine the slope and elevation data to the tornado events. Finally we filter the weather data information based on the weather station data, and then combining the tornado events to the corresponding weather data.
The second software tool is MySQL. MySQL is one of the most popular databases in today's world. In this research, the MySQL is used to do the time consistency between the weather data (temperature and precipitation) and tornado events. It also used to reorganize the data sets to produce the final data sets, used for model generation.

The next tool is SPSS. The SPSS is a typical statistic analysis software provided by IBM company. The SPSS provide not only the basic statistic description but also sophisticated inferential and multivariate statistical procedures like factor analysis, cluster, and regression. Through SPSS, we can easily access the appropriately model for the data sets. It provides a user-friendly interface to adjust the model parameters. In this research the SPSS has been used to analyze the relationships between the tornado intensity and environment factors.

Finally, R used to generate the prediction models. R is a statistical programming language and software environment. Through R, with the support of the enormous set of libraries, users can analyze the data and apply many most machine learning algorithms quickly.

### 4.2 Data Preprocessing

#### 4.2.1 Tornado Data Sets

As discussed above, this research focus on Missouri State as the research region. However, the original tornado data sets include tornado events for the entire U.S. So first step is to reduce to only the tornadoes while occurred within Missouri State.
ArcGIS can help to complete this process. While the original tornado data sets was a shape file, it can be imported into ArcMap directly. Moreover, the map looks like Figure 3.1. So from here we can see this tornado data set is a polyline vectors layer. Each line segment represents a tornado event. The two terminal vertexes are tornado starting point and ending point. Each line segment has some attributes, such as state, starting point and ending points latitude and longitude coordinates, Fujita-Scale, tornado path width, length, occurrence time, facilitates, property loss in dollars. So according to the attribute "state" we can filter all of the tornado belong to Missouri. Just simply write a filter equations in ArcMap, set "state=MO" it can take out all of the tornado events happened within Missouri. The Missouri tornado data sets would like the Figure 4.1.

Until now the tornado events still expressed through the polyline, but in this research, I want to analyze the relationships between the tornado intensity and the environment factors in the tornado starting point. So the following work is to get the starting points from the tornado event polyline layer. The ArcMap provides a tool named "polyline to point", which could help to get the starting point of the input polyline layer, and keep all of the attributes for the each tornado events. After this step, the tornado events now are represented by the tornado starting point. The polyline layer has been converted to a point layer. This point layer is shown in Figure 4.2.

4.2.2 Extract Slope From Missouri DEM

The slope for every point in the ground means the included angle between the horizontal plane and tangent plane that pass through this point. In regards to ArcGIS,
Figure 4.1: Missouri Tornado Events
Figure 4.2: Missouri Tornado Startpoint
the slope is the maximum rate of elevation change in value from that cell to its eight neighbors. The slope value represents the terrain ups and down. The lower the slope the flatter the terrain, the higher the slope value the steeper the terrain. According to the ArcGIS Resource documents[28], the slope degrees is computed as

\[
slope\_degree = \text{ATAN}(\text{rise} / \text{run}) \times 57.29578
\]

(4.1)

where

\[
\text{rise} / \text{run} = \sqrt{([dz/dx]^2 + [dz/dy]^2)}
\]

(4.2)

The values of the center cell and its eight neighbors determine the horizontal and vertical deltas. The neighbors are identified as letters from a to i, with e representing the cell for which the aspect is being calculated[28]. The neighborhood raster is shown Figure 4.5[28].

Using ”slope” spatial analysis tool in ArcGIS, I generate the Missouri slope data from the Missouri DEM. Slope map in degree is shown in Figure 4.6.
Figure 4.4: Missouri slope map in degree
4.2.3 Add elevation and slope value for each tornado events starting points

This step is a fusion operation between three layers: Tornado starting point, Missouri DEM and Missouri slope. Figure 4.7 show the overlay effects for the tornado starting points and Missouri DEM map. However, they are the different type of layers (the tornado starting points is a vector layer and the Missouri DEM is a raster layer) so it can not be directly fused togetger. There are several ways to solve it. One is to convert the layers into same type (the raster layer and vector layer can be converted to each other) and then intersect. Another one is using the analysis tools provided by ArcGIS named ”extract to points”. After this operation, the slope and elevation value have been added to each tornado events. Like the "RASTERVALU" column in Figure 4.8.

4.2.4 Combining the temperature data to tornado events

4.2.5 Join the tornado events and temperature data together

Combining tornadoes to temperature data is a two dimension pairing. One is the spatial attributes, another one is time. In other words, the tornado event can be assigned the temperature value if this tornado event occurs within a certain distance of the temperature weather station and the occurrence time is within the time period of the temperature records’ time period.

   The first step is to filter down to the tornado events which happened within a valid spatial closeness of temperature weather stations. Therefore, we import the temperature weather station data into ArcGIS, shown in Figure 4.9. Moreover, the overlay
Figure 4.5: Missouri DEM map & Tornado starting points

Figure 4.6: Add slope data into tornado data sets
Figure 4.7: Temperature weather station

map for tornado events and weather stations is shown in Figure 4.10. For this study, I use 15 miles as the valid range for the weather station. It means only the tornadoes which occurred within 15 miles of the weather station would be chosen for time pairing. This filtering process is performed using ArcGIS ”Buffer” and ”intersection” analysis tools. The ”Buffer” could produce a buffer polygon for each weather station (input point layer) with a specific buffer distance (15 miles). The overlay layer for the weather station buffer layer and Missouri tornado events is shown in Figure 4.11. The ”intersection” tools would pick the tornado events that are within this polygon buffer layer.

After this process, the new tornado events layer’s attributes table would containing both the tornado information and the corresponding weather station information.
Figure 4.8: Missouri tornado events and temperature weather station

Figure 4.9: Tornado events and Temperature weather station buffer
4.2.6 Add temperature information for each tornado event

Using the resulting tornado terrain data (from Section 4.4.1), we can now link the events with the corresponding temperature weather station data. This allows us to have a temperature value for each tornado event based on the weather station information.

First, need to output the attributes table for 4.3.1’s results layer and import into Mysql database. Since the temperature data sets are organized as 24 hours data in 24 columns, it must be transformed. Once transformed, then we can use the standard join operation in the database. Therefore, we need to write scripts to identify the tornado time and transform the corresponding hour data in the temperature data sets. I developed a database stored procedure to do this. After this process, each tornado events would get a temperature attribute to show the temperature at the tornado occurred time at them occurred place. However, due to the constrains of the source data, I am not using the real-time temperature data for the tornado occurrence time. This is a average temperature data for the tornado occurrence time.

4.2.7 Combining the precipitation data sets with tornado events

4.2.8 Tornado events geospatial filtering

The tornado events filtering operation for the precipitation is quite similar to filtering based on temperature data sets. Also using the ”Buffer” and ”intersection” tools in ArcGIS. After import the precipitation weather station layers. The overlay effects for tornado events and precipitation weather station buffer layers are like Figure 4.12.
Figure 4.10: Precipitation station buffer and tornado events

This time, I use 20 miles as the buffer distance.

4.2.9 Add precipitation value to each tornado events

Unlike the temperature data sets, the precipitation value are more straight forward to combining with tornadoes. We use a standard update operation in MySQL, which can add the precipitation value to the tornadoes. However, in the precipitation data there are lots of events could not get a valid value. This is due to incomplete source data for the original precipitation data sets (in original precipitation data sets, NOAA use 99999 to represent the lack or invalid of input data).
4.3 Tornado events distribution features analysis

The first step is to observe the tornadoes’ primary statistics information and their spatial-temporal distribution characteristics. Specifically, we seek to find if the tornado belongs to some distributions, such as concentrations in some regions or the tornado appear on a periodic cycle. This might reveal some hidden properties for the tornado, and provide a direction for the predictors to choose and the appropriate models.

4.3.1 Tornado spatial characteristic

After loading the tornado shapefile into ArcGIS, we use the spatial statistical tool—Hot Spot Analysis (Getis-Ord Gi*) to generate a Hot Spot map for the tornado intensity. This is shown in Figure 4.13. The red features means there existing a high F-Scale tornadoes cluster. And the blue features means existing low F-Scale tornadoes cluster. It would have a more clear view if using tornado starting points as the object. The tornado events starting points F-Scale Hot Spot map is shown in Figure 4.14. For those two maps, the random distribution means those parts of tornado events do not appear to clustering in F-Scale.

Before explaining how the Hot Spot Analysis tool works, it is necessary to explain the spatial autocorrelation. The autocorrelation means the correlation of a variable with itself through space. Alternatively, it can be described as the measure of variables which appears to have a significant cluster or dispersion base on their location. Like the correlation in statistics, spatial autocorrelation also existing significance. In ArcGIS, this significant value called z-score and p-value. The p-value is a measure
Figure 4.11: MO tornado F-scale HotSpot
Figure 4.12: MO tornado starting point F-scale HotSpot
to determine if a random process created the probability of the observed spatial pattern. Where the small p-value means the pattern appear to a cluster (for example, not random), and high p-value leads to a random distribution. However, these "small" value still quite obscure; so we need a confidence level, the z-scores. The z-scores are standard deviations, both z-scores and p-values are associated with the standard normal distribution as shown Figure 4.15[29]. So very high or very low (negative) z-scores, associated with very small p-values, are found in the tails of the normal distribution. This indicates it is unlikely that the observed spatial pattern reflects the theoretical random pattern, or it tends to be a cluster. The corresponding table of z-scores, p-value and confidence level is shown in Table 4.1[29].

The Hot Spot Analysis tool calculates the Getis-Ord Gi* statistic (pronounced G-i-star) for each feature in a dataset. The returned results are the z-scores and
Table 4.1: z-scores, p-value and confidence level from[29]

<table>
<thead>
<tr>
<th>z-score (Standard Deviations)</th>
<th>p-value(probability)</th>
<th>Confidence level</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;-1.65 or &gt;+1.65</td>
<td>&lt;0.10</td>
<td>90%</td>
</tr>
<tr>
<td>&lt;-1.96 or &gt;+1.96</td>
<td>&lt;0.05</td>
<td>95%</td>
</tr>
<tr>
<td>&lt;-2.58 or &gt;+2.58</td>
<td>&lt;0.01</td>
<td>99%</td>
</tr>
</tbody>
</table>

p-value. The calculation formula from[30] is given as:

\[
G^*_i = \frac{\sum_{j=1}^{n} \omega_{i,j}x_j - \bar{X} \sum_{j=1}^{n} \omega_{i,j}}{S \sqrt{\frac{\sum_{j=1}^{n} \omega_{i,j}^2 - (\sum_{j=1}^{n} \omega_{i,j})^2}{n-1}}} \tag{4.3}
\]

where \(x_j\) is the attribute value for feature \(j\), \(\omega_{i,j}\) is the spatial weight between feature \(i\) and \(j\), \(n\) is equal to the total number of features and:

\[
\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \tag{4.4}
\]

\[
S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \tag{4.5}
\]

\(G^*_i\) statistic is a z-score so no further calculations are required.

These maps show the tornado events F-scale value appear in local clusters. There is several places that are more frequently experiencing the tornadoes with high F-scale compared to other places; and also some area is concentrated with low F-scale tornadoes.

4.3.2 Tornado temporal trends

Since the research data sets of tornadoes are includes the tornado events occurring during the time period 1950-2013; it is also necessary to analyze the tornado temporal
As I introduced in chapter 2, Lopes and Machado are doing many works in the analysis the Tornadoes in time series. They do find the tornado intensity (they using tornado path length and width) exhibits some temporal features.

The total tornado event amounts and significant tornado amounts in Missouri for every decade are computed to show the tornado events’ temporal trends. The results are shown in Table 4.2. The significant tornado means the tornadoes with F-scale is at least 2. The final row is not a decade since it just counting the tornadoes amount from 2010 to 2013.

The reason why the 2000s decade appear to have an explosive increase is due to the emergence of Doppler radar use by the National Weather Service, Doppler help to report many more insignificant tornadoes and the tornadoes that happened in remote area. The significant tornado amounts are not increasing as rapidly as compared to the total tornado amounts. From this tornado quantity analysis, it shows the tornado events differ year by year. Therefore, we can expect that some time-different features would affect the tornadoes.

<table>
<thead>
<tr>
<th>Decades</th>
<th>Tornado Amounts</th>
<th>Significant Tornado Amounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950-1959</td>
<td>213</td>
<td>75</td>
</tr>
<tr>
<td>1960-1969</td>
<td>303</td>
<td>93</td>
</tr>
<tr>
<td>1970-1979</td>
<td>240</td>
<td>99</td>
</tr>
<tr>
<td>1980-1989</td>
<td>245</td>
<td>52</td>
</tr>
<tr>
<td>1990-1999</td>
<td>287</td>
<td>37</td>
</tr>
<tr>
<td>2000-2009</td>
<td>557</td>
<td>80</td>
</tr>
<tr>
<td>2010-2013</td>
<td>215</td>
<td>27</td>
</tr>
</tbody>
</table>
Based on the tornado events spatial and temporal features analysis, it clearly shows the tornado events exist in local clusters and have a temporal differential. So I have chosen environment factors (slope, elevation, temperature and precipitation) that I think would affect the tornado intensity, trying to generate the tornado intensity prediction model using those four factors.

4.4 Relationship analysis

Because of the data matching challenge (tornado events are occurred far away from the weather station, or the weather records not available while the tornado occurred time), the data records that contain all of the slope, elevation, temperature and precipitation only have 53 records. This is too small, so I have to analysis and model the four factors separately. Therefore, I generate three intermedia data sets. First is the tornadoes with slope and elevation information, containing about 2060 records. Second one is tornadoes with slope, elevation, and precipitation that containing 355 records. Finally, the tornadoes with slope, elevation and temperature, which just containing 105 records. In order to use the maximum number of objects for the relationships analysis–between the tornado intensity (represented by tornado width, length, and Fujita-Scale) and environment variables (slope, elevation, temperature, and precipitation) in the tornado starting point. I use those three intermedia data sets to analyze the relationships between tornado intensity and different environment variables. For instance, the first dataset would be used to analyze the relationships between the slope, elevation and tornado intensity.

In this research, the relationship is meant as the correlation. Correlation is indi-
cating how two or more variables may be. The correlation coefficients represent the degree of correlation.

There are many types of correlation coefficients. The most commonly used one is Pearson correlation coefficient. The definition of Pearson correlation coefficient is dividing the covariance of the two variables by the product of their standard deviations. To compute the Pearson correlation coefficients require to compute the expected values and standard deviations for the two variables. So for two random variables \( x \) and \( y \), and with their expected values \( \bar{x}, \bar{y} \), and standard deviations \( \sigma_x, \sigma_y \), the Pearson correlation coefficients \( \rho_{x,y} \) can be defined as:

\[
\rho_{x,y} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{E[(x - \bar{x})(y - \bar{y})] }{\sigma_x \sigma_y}
\]

(4.6)

where \( \text{cov} \) is the covariance, \( E \) is the expectation.

However, there is a necessary condition for Pearson correlation coefficient: the variables must follow a normal distribution, or at least the data sets must pick from the data table with the same equal interval. If the data sets do not satisfy this condition, then it needs to use other correlation coefficient, like Spearman, which is utilized in this research (because the data sets value are not follow the normal distribution).

The Spearman correlation coefficient called as Spearman’s rank correlation coefficient or Spearman’s rho. It is a nonparametric measure of statistical dependence between two variables. Spearman’s rank correlation coefficient defined as the Pearson correlation coefficient between the ranked variables. In other words, the Spearman will firstly rank the original data sets, using \( d_i \) to represent the difference between ranks, if all \( n \) ranks are distinct integers, it can be computed using the following
formula

\[ \rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \]  \hspace{1cm} (4.7)

The Spearman coefficient value range is [-1,1], larger than 0 means positive correlation, smaller than 0 means negative correlation, equal to 1 means completely positive correlation, equal to -1 means completely negative correlation. The more reach to 0 the less the correlation between the variables.

So, even though the correlation coefficient shows the degree of the correlation, it still needs to prove that the correlation is not caused by the casual sample, but is real property. This re-check process is called hypothesis testing.

The first step is to specify the null hypothesis \( \rho = 0 \) and an alternative hypothesis \( \rho \neq 0 \). The second step is to choose a significance level. Normally it is 0.05. The third step is to compute the sample value of the correlation coefficient. The fourth step is to compute p, the probability of obtaining a difference between and the value specified by the null hypothesis (zero) as large or larger than the difference obtained in the experiment. We can test this hypothesis through Student-testing (t-testing)[31]. The formula is shown below.

\[ t = \frac{r \sqrt{N - 2}}{\sqrt{1 - r^2}} \]  \hspace{1cm} (4.8)

After computing this probability value, compare it with the significance level setting in step 2. If it is smaller than the setting level then reject the null hypothesis. This means the correlation is significant, otherwise, if the computed value larger than the setting level it is not significant.

So in summary, to check whether two variables are correlated or not, there
are two values need to check. One is the correlation coefficient, generally using r to represent. Another one is the hypotheses testing computed value, usually called p value. So the p show the two variables are correlated or not, and the r value shows the degree of the correlation.

The correlation analysis discussed so far is only for two variables. However, for multiple variables, such as three, four or even more, we must examine the partial correlation.

The partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed[32]. Similar to the correlation, the degree of partial correlation is also measured by a partial correlation coefficient. Generally, there are three ways to compute the partial correlation. The first one is using linear regression, for example here are three variables, x, y, z. Then we can generate two regression equations, one is using z to predict x, and the other one is using z to predict the y. The resulting residual values represent the variation in x and y that is unexplained by z. And the correlation of these residuals represents the correlation of x and y while controlling the effect of z. Second way is using recursive equation. Since the nth-order partial correlation (i.e., with —Z— = n) can be easily computed from three (n − 1)th-order partial correlations. The zeroth-order partial correlation[32]. For every $Z_0 \in Z$ we have

$$
\rho_{XYZ} = \frac{\rho_{XYZ\{Z_0\}} - \rho_{Z_0XZ\{Z_0\}}\rho_{Z_0YZ\{Z_0\}}}{\sqrt{1 - \rho_{XZ_0Z\{Z_0\}}^2}} \sqrt{1 - \rho_{Y_0Z_0Z\{Z_0\}}^2}
$$

(4.9)

Here the zeroth-order partial correlation is defined to be the regular correlation coefficient. The final method is using matrix inversion, where the correlation matrix (or
alternatively covariance matrix) \( \omega = \rho_{X_i, X_j} \), is positive definite and therefore invertible. If we define \( P = \omega^{-1} \), we have:

\[
\rho_{X_i, X_j \setminus \{X_i, X_j\}} = -\frac{p_{ij}}{\sqrt{p_{ii} p_{jj}}}
\]  

(4.10)

where \( \rho_{X_i, X_j \setminus \{X_i, X_j\}} \) represent the partial correlation coefficient for \( X_i \) and \( X_j \) while control all other variables. \( p_{ij} \) represent the \( i \)th row and \( j \)th column value in the correlation matrix.

While loading the data sets into SPSS, we can compute the correlation matrix for four factors since as discussed, the data sets used are three different data sets: one is containing the tornado intensity (F-scale, width, length), slope, and elevation; second one containing tornado intensity and precipitation; the last one containing the tornado intensity and temperature. The reason to use three different data sets is I want to use the maximum data samples for each analysis. However, the tornado records which contain all four factors only have 53 records. Therefore, the following correlation results appear to have different results for the same sets of variables, because they are computing under different data scale.

The correlation matrix for the slope, elevation is shown in Table 4.3. Moreover, the correlation matrix for slope, elevation and temperature is in Table 4.4. The correlation matrix for the slope, elevation, and precipitation is in Table 4.5. The MAG variables in the results tables are actually the F-scale.

We see from the correlation coefficient in Table 4.3 that the slope and tornado path width, elevation and tornado path width are significant correlated (significant at 0.01 level). To examine whether this correlation actually exists within these two variables or produced by the interactions of other variables; I performed a partial
Table 4.3: Correlation matrix including slope, elevation

<table>
<thead>
<tr>
<th></th>
<th>slope</th>
<th>elevation</th>
<th>MAG</th>
<th>LEN</th>
<th>WID</th>
</tr>
</thead>
<tbody>
<tr>
<td>slope</td>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td>.240**</td>
<td>.017</td>
<td>.036</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.000</td>
<td>.445</td>
<td>.098</td>
</tr>
<tr>
<td>elevation</td>
<td>Correlation Coefficient</td>
<td>.240**</td>
<td>1.000</td>
<td>.010</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.</td>
<td>.650</td>
<td>.146</td>
</tr>
<tr>
<td>MAG</td>
<td>Correlation Coefficient</td>
<td>.017</td>
<td>.010</td>
<td>1.000</td>
<td>.552**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.445</td>
<td>.650</td>
<td>.</td>
<td>.000</td>
</tr>
<tr>
<td>LEN</td>
<td>Correlation Coefficient</td>
<td>.036</td>
<td>.032</td>
<td>.552**</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.098</td>
<td>.146</td>
<td>.000</td>
<td>.</td>
</tr>
<tr>
<td>WID</td>
<td>Correlation Coefficient</td>
<td>.099**</td>
<td>.058**</td>
<td>.469**</td>
<td>.619**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.009</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).

Table 4.4: Correlation matrix including slope, elevation and temperature

<table>
<thead>
<tr>
<th></th>
<th>MAG</th>
<th>WID</th>
<th>LEN</th>
<th>temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAG</td>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td>.438**</td>
<td>.407**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>WID</td>
<td>Correlation Coefficient</td>
<td>.438**</td>
<td>1.000</td>
<td>.578**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.</td>
<td>.000</td>
</tr>
<tr>
<td>LEN</td>
<td>Correlation Coefficient</td>
<td>.407**</td>
<td>.578**</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.</td>
</tr>
<tr>
<td>temperature</td>
<td>Correlation Coefficient</td>
<td>.041</td>
<td>-.121</td>
<td>-.157</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.679</td>
<td>.220</td>
<td>.109</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).
Table 4.5: Correlation matrix including slope, elevation and precipitation

<table>
<thead>
<tr>
<th></th>
<th>MAG</th>
<th>WID</th>
<th>LEN</th>
<th>pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAG</td>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td>.648**</td>
<td>.611**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>WID</td>
<td>Correlation Coefficient</td>
<td>.648**</td>
<td>1.000</td>
<td>.618**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.</td>
<td>.000</td>
</tr>
<tr>
<td>LEN</td>
<td>Correlation Coefficient</td>
<td>.611**</td>
<td>.618**</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.</td>
</tr>
<tr>
<td>pre</td>
<td>Correlation Coefficient</td>
<td>-.87</td>
<td>.28</td>
<td>-.17</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.113</td>
<td>.609</td>
<td>.109</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).

Table 4.6: Partial correlation results for the slope and width

<table>
<thead>
<tr>
<th>Control Variables</th>
<th>slope</th>
<th>WID</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEN &amp; elevation &amp; MAG</td>
<td>Correlation</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
</tr>
<tr>
<td>WID</td>
<td>Correlation</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>0.174</td>
</tr>
</tbody>
</table>

correlation analysis for the slope and width; elevation and width. The slope and width partial correlation results is shown in Table 4.6. Elevation and width partial correlation results is shown in Table 4.7.

Unfortunately, from the partial correlation analysis, it shows those two variables sets are not correlated; so it will not be able to build the regression equations to compute the tornado intensity directly. However, since the F-scale value comes from 0 to 5, it can be used as a class variable. Therefore, we can explore classification algorithms to predict the new events belong to which kind of class (F-scale) of tornado intensity.
Table 4.7: partial correlation results for the elevation and width

<table>
<thead>
<tr>
<th>Control Variables</th>
<th>slope</th>
<th>WID</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEN &amp; slope &amp; MAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>slope</td>
<td>1.000</td>
<td>0.55</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>WID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>0.55</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.13</td>
<td></td>
</tr>
</tbody>
</table>

In this research, I generate the decision tree and support vector machine (SVM) in R for the classification prediction model.

### 4.5 Model generation

#### 4.5.1 Decision Tree

In R the package named Rpart, which is a CART (classification and regression tree) tree, and could builds both classification and regression trees. Before running the algorithms on the data sets, first, we divide the data sets into training set and testing set. Since the data scale is not big enough to directly divided into two parts, we use cross-validation techniques.

Cross validation divides the data sets into several non-intersect groups, each time pick one of the group as the testing data sets, and others be the training data sets. Repeat this process to change the training group and testing group.

As discussed above there are three different intermedia data sets, containing different factors. So we need to apply the mode into those three data sets separately. First, uses slope and elevation to predict F-scale, which has 2060 data records. Second uses slope, elevation and temperature to predict F-scale, containing 105 records.
Third, uses slope, elevation and precipitation to predict F-scale, contains 335 records. For each kind of data set I applied a 5 fold cross validation.

4.5.2 Support vector machine (SVM)

The basic theory of the support vector machine is introduced in chapter 2. In this research I use a standard SVM tool LIBSVM[33]. LIBSVM is a software for support vector machine, including classification, regression and distribution estimation. And it also support multi-class classification. This tool also available in R, it is the package ”e1071”. Also set the same dividing threshold (80%) with the decision tree model for the training set and testing set.
Chapter 5

Results

The F-scale prediction error rate using decision tree with different predictors combination is shown in Table 5.1. Figure 5.1 show one of the example of the decision tree model generated within the 200 times experiments while using slope, elevation and temperature as the predictors. And the Table 5.2 show the F-scale prediction error rate using SVM with different predictors.

We see that the models explored lead to a poor prediction accuracy for 0-5 F-scale prediction. The minimum error rate and the average error rate has a huge difference

<table>
<thead>
<tr>
<th>Predictors Combination</th>
<th>Average Error Rate</th>
<th>Minimum Error Rate</th>
<th>Data Sets Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope+ Elevation</td>
<td>0.7003</td>
<td>0.6105</td>
<td>2060</td>
</tr>
<tr>
<td>Slope+ Elevation+ Temperature</td>
<td>0.7521</td>
<td>0.5833</td>
<td>105</td>
</tr>
<tr>
<td>Slope+ Elevation+ Precipitation</td>
<td>0.8968</td>
<td>0.6957</td>
<td>335</td>
</tr>
</tbody>
</table>
Figure 5.1: Decision Tree Model Example

Table 5.2: Error rate for different predictors choose using SVM

<table>
<thead>
<tr>
<th>Predictors Combination</th>
<th>Average Error Rate</th>
<th>Minimum Error Rate</th>
<th>Data Sets Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope+ Elevation</td>
<td>0.6278</td>
<td>0.5663</td>
<td>2060</td>
</tr>
<tr>
<td>Slope+ Elevation+ Temperature</td>
<td>0.5524</td>
<td>0.2857</td>
<td>105</td>
</tr>
<tr>
<td>Slope+ Elevation+ Precipitation</td>
<td>0.4791</td>
<td>0.3939</td>
<td>335</td>
</tr>
</tbody>
</table>
for the last two predictors combinations, both for decision tree and SVM. For instance, while using slope, elevation and temperature as the predictors, the minimum error rate for decision tree is 0.5833 but the average error rate is 0.7521; the SVM perform better but still has a 0.2857 minimum error rate and a 0.5524 average. The data scale is limited so the division of the training set and testing sets significantly affects the results. For example, there is only one tornado in the 5, in F-scale, but if this record must be assigned to testing sets instead of training sets, this one is therefore, an expected error prediction. Since the training sets not even containing any records with a 5 F-scale. When using only slope and elevation it has larger data scale, 2060, which could get more stable prediction error rate, it has a 0.6105 minimum error rate and 0.7003 average error rate while using decision tree; 0.5663 minimum error rate and 0.6278 while using SVM. However, those two predictors can not support the model to get an accurate enough prediction. While using SVM, the minimum error rate for slope and elevation combination is much larger than the other two three predictors combinations, 0.5663 versus 0.2857/0.3939.

There are serval possible reasons for the poor prediction results for the F-scale. First one is the lack of data. Since the temperature data sets only have 105 records, and precipitation only has 335 records, both are too small compare to all tornadoes–2060 records. And what’s worse, there are only 53 records of tornado events that have a valid value in all four predictors. A further reason is the temperature data are not a real time temperature. It is the normal temperature in that hour. It is also not so easy to collect the real-time temperature data for specific tornado occurrence time and place. Furthermore, the predictors are not sufficient. There are many other possible predictors which could be used in the tornado intensity prediction. These features
include the land cover, radiation, wind speed, and some sophisticated meteorological features like velocity wind shear and helicity can be used in the model generation. But those variables have a common problem—they are even more difficult to collect for the particular tornado occurrences.

Since the two models both not performing well in predict the F-scale, we simplify the problem and improve the prediction precision. We re-classify for the tornado events based on their F-scale into two class problem: significant tornado or insignificant tornado. The tornado events with F-scales 2,3,4,5 are marked as significant, and scales 0 and 1 are insignificant.

Applying the decision tree and SVM model to the data sets again, and it return the following results show in Table 5.3. If we just predict the future tornado would significant or not, the models work much better. Both decision tree and SVM show improvements in the minimum error rate. But the decision tree still has a significant difference between the average error rate and minimum. For example, the slope, elevation and precipitation predictors combination while using decision tree has a 0.3779 minimum error rate and a 0.4178 average error rate. SVM appears to have better prediction accuracy than the decision tree while there are three predictors. For example in the slope, elevation and precipitation predictors combination, the SVM has a 0.2121 minimum error rate versus 0.6129 for decision tree; and 0.4366 average error rate verses 0.7964. But the SVM model still not doing well while only use slope and elevation compared to use three predictors in SVM model. That shows the two predictors are not enough to get an accurate prediction.

Since the slope and elevation are kind of static information, we can plot the possible location for Missouri that would appear significant tornado. Based on the SVM
Table 5.3: Error rate for significant tornado prediction while using different predictors

<table>
<thead>
<tr>
<th>Predictors Combination</th>
<th>Average Error Rate</th>
<th>Minimum Error Rate</th>
<th>Data Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Decision Tree</td>
<td>SVM</td>
<td>Decision Tree</td>
</tr>
<tr>
<td>Slope+Elevation</td>
<td>0.4178</td>
<td>0.4457</td>
<td>0.3779</td>
</tr>
<tr>
<td>Slope+Elevation +Temperature</td>
<td>0.6288</td>
<td>0.3522</td>
<td>0.2728</td>
</tr>
<tr>
<td>Slope+Elevation +Precipitation</td>
<td>0.7964</td>
<td>0.4366</td>
<td>0.6129</td>
</tr>
</tbody>
</table>

model using slope and elevation as the predictors. In R, using "Raster" package to extract the raster layer value (elevation and slope), then send those value into SVM model, and finally the "Raster" package could generate the prediction results raster layer. This map is shown in Figure 5.2. The green parts are predicted place for significant tornadoes, and grey parts are the predicted place for insignificant tornadoes.
Figure 5.2: Missouri Possible Significant Tornado Occurring Place Based on Elevation and Slope Using SVM Prediction Model
Chapter 6

Conclusion

This research explored the influence of the environment variables applied to the tornado intensity. I have applied decision tree and support vector machine approaches to the tornado intensity prediction, based on the environment features in the tornado start point. The models do not predict the details tornado intensity (F-scale) in high accuracy, due to the small scale of data sets and lack of predictors. The tornado is a very complex phenomena; happening in a dynamic system. Too many factors affect the tornado intensity; my selected four variables are not enough to make an accurate prediction. Since the tornado comes with strong airflow, if is clear we need more atmospheric features like air pressure, humidity and velocity wind shear. Additionally, we need to be sure enough data samples are available, it can expect to have a better result. Since we still not fully understand this complex dynamic phenomenon, we do not know which kinds of factors would lead to best prediction model. Much more research is need in future work.

Additionally, SVM model shows a better performing than decision tree in this
research. The SVM model does good prediction accuracy for determining the new tornado events would be a significant tornado or not while using three predictors (temperature or precipitation). Though this kind of significant prediction is less meaningful. But overall this procedure still provide an interesting angle to analyze and predict the tornado intensity. Since the tornado events do appear a kind of local cluster in spatial distribution and an amounts change very by time. The environment features should have an influence on the tornado intensity. We may be able to improve the prediction accuracy after getting enough data points and add more predictors.
7.1 MySQL Store Procedure for Paring the Temperature and Tornado Events

```sql
CREATE DEFINER='root'@'localhost' PROCEDURE `temp_paring`() BEGIN
    set @num=0;
    while @num<105 do
        set @tisql = CONCAT("select T from msproject.temp LIMIT ",
            @num, ",1 into @ti;" );
        prepare stmtti from @tisql;
        execute stmtti;
        deallocate prepare stmtti;
        if date_format(@ti, '%T') <= '01:30:00' then set @coll='one'
```
else if date_format(@ti, '%T') <= '02:30:00' && date_format(@ti, '%T') > '01:30:00' then set @coll='two';
else if date_format(@ti, '%T') <= '03:30:00' && date_format(@ti, '%T') > '02:30:00' then set @coll='three';
else if date_format(@ti, '%T') <= '04:30:00' && date_format(@ti, '%T') > '03:30:00' then set @coll='four';
else if date_format(@ti, '%T') <= '05:30:00' && date_format(@ti, '%T') > '04:30:00' then set @coll='five';
else if date_format(@ti, '%T') <= '06:30:00' && date_format(@ti, '%T') > '05:30:00' then set @coll='six';
else if date_format(@ti, '%T') <= '07:30:00' && date_format(@ti, '%T') > '06:30:00' then set @coll='seven';
else if date_format(@ti, '%T') <= '08:30:00' && date_format(@ti, '%T') > '07:30:00' then set @coll='eight';
else if date_format(@ti, '%T') <= '09:30:00' && date_format(@ti, '%T') > '08:30:00' then set @coll='nine';
else if date_format(@ti, '%T') <= '10:30:00' && date_format(@ti, '%T') > '09:30:00' then set @coll='ten';
else if date_format(@ti, '%T') <= '11:30:00' && date_format(@ti, '%T') > '10:30:00' then set @coll='eleven';
else if date_format(@ti, '%T') <= '12:30:00' && date_format(@ti, '%T') > '11:30:00' then set @coll='twelve';
(@ti, '%T')> '11:30:00' then set @coll='twelve';
elseif date_format(@ti, '%T')<='13:30:00' && date_format(@ti, '%T')> '12:30:00' then set @coll='thirteen';
elseif date_format(@ti, '%T')<='14:30:00' && date_format(@ti, '%T')> '13:30:00' then set @coll='fourteen';
elseif date_format(@ti, '%T')<='15:30:00' && date_format(@ti, '%T')> '14:30:00' then set @coll='fifteen';
elseif date_format(@ti, '%T')<='16:30:00' && date_format(@ti, '%T')> '15:30:00' then set @coll='sixteen';
elseif date_format(@ti, '%T')<='17:30:00' && date_format(@ti, '%T')> '16:30:00' then set @coll='seventeen';
elseif date_format(@ti, '%T')<='18:30:00' && date_format(@ti, '%T')> '17:30:00' then set @coll='eighteen';
elseif date_format(@ti, '%T')<='19:30:00' && date_format(@ti, '%T')> '18:30:00' then set @coll='nineteen';
elseif date_format(@ti, '%T')<='20:30:00' && date_format(@ti, '%T')> '19:30:00' then set @coll='twenty';
elseif date_format(@ti, '%T')<='21:30:00' && date_format(@ti, '%T')> '20:30:00' then set @coll='twenty_one';
elseif date_format(@ti, '%T')<='22:30:00' && date_format(@ti, '%T')> '21:30:00' then set @coll='twenty_two';
elseif date_format(@ti, '%T')<='23:30:00' && date_format(@ti, '%T')> '22:30:00' then set @coll='twenty_three';
else set @coll='twenty_four';

END if;

set @sensql=CONCAT("select avg(msproject.temperature.*) ,
@coll," from msproject.temperature where wid=(select wid_1 from msproject.temp LIMIT" ,@num," ,1) and MONTH(DATE_FORMAT((select T from msproject.temp LIMIT" ,@num," ,1) ,'%Y-%m-%d-%T'))=msproject.temperature.'month' and DAY(DATE_FORMAT((select T from msproject.temp LIMIT" ,@num," ,1) ,'%Y-%m-%d-%T'))=msproject.temperature.'date' into @outvar;" );

prepare stmt from @sensql;
execute stmt;
deallocate prepare stmt;

set @upsql=CONCAT("update msproject.temp set msproject.temp.temperature=" ,@outvar," where msproject.temp.T='" ,@ti," '");

prepare stmttwo from @upsql;
execute stmttwo;
deallocate prepare stmttwo;

set @num = @num+1;
7.2 R code for model generation

7.2.1 Decision Tree

```r
library("rpart")
library("rpart.plot")
library("DMwR")

# import the data file
tornado=read.csv(file="~/Desktop/data.csv", header=TRUE, sep="", )
tornado$Sig <- factor(tornado$Sig)
#tornado=scale(tornado)

tornado <- SMOTE(Sig~slope+elevation+temperature, tornado,
                 prec.over=100, prec.under=100)

e <- rep(0,10)
k<-10
tornado$id <- sample(1:k,nrow(tornado),replace=TRUE)
list <- 1:k
```
for (i in 1:k) {
    trainingset <- subset(tornado, id %in% list[-i])
    testset <- subset(tornado, id %in% c(i))

    fit <- rpart(Sig ~ slope+elevation+temperature, data=
                   trainingset, method="anova")
    pred <- predict(fit, testset)
    ConfM1 <- table(testset$Sig, pred=pred)
    err[i] <- (sum(ConfM1) - sum(diag(ConfM1)))/sum(ConfM1)
}

7.2.2 SVM

library("e1071")
library("DMwR")
# import the data file

tornado=read.csv(file="~/Desktop/data3.csv", header=TRUE, sep="",""

tornado$MAG <- factor(tornado$MAG)
# tornado=scale(tornado)
# tornado <- SMOTE(MAG~slope+elevation, tornado, prec.over=10,
#                   prec.under=30)
err <- rep(0,5)
k <- 10
tornado$<-$id <- sample(1:k, nrow(tornado), replace=TRUE)
list <- 1:k

for (i in 1:k) {
  # split the data sets into testing and training
  # training.indices <- sample(nrow(tornado), 1800)
  # training <- rep(FALSE, nrow(tornado))
  # training[training.indices] <- TRUE
  
  # tornado.input <- tornado[training,]
  # tornado.input = data.frame(tornado.input)
  # tornado = data.frame(tornado)
  #
  trainingset <- subset(tornado, id %in% list[-i])
testset <- subset(tornado, id %in% c(i))

  model <- svm(MAG~slope+elevation+prec, data=trainingset)
pred <- predict(model, testset)
ConfM1 <- table(testset$MAG, pred=pred)
err[i] <- (sum(ConfM1) - sum(diag(ConfM1)))/sum(ConfM1)
}

73
7.2.3 SVM Prediction & Visualization Based on Raster Image

library("e1071")
library("DMwR")
tornado=read.csv(file="~/Desktop/new.csv", header=TRUE, sep=" , ")
tornado <- SMOTE(Sig~slope+elevation, tornado, prec.over=400, prec.under=400)

err<- rep(0,5)
m<-0

for (i in c(1:5)) {
#split the data sets into testing and training
training.indices <- sample(nrow(tornado), 700)
training <- rep(FALSE, nrow(tornado))
training[training.indices] <- TRUE

tornado.input<- tornado[training ,]
tornado.input=data.frame(tornado.input)
tornado=data.frame(tornado)

tornado$Sig <- factor(tornado$Sig)
model <- svm(Sig ~ slope + elevation, data = tornado.input)

}

library("raster")
library("rgdal")
elevation <- raster("~/Desktop/elevation.tif")
slope <- raster("~/Desktop/slope.tif")
#plot(elevation)
#plot(slope)
#create the raster object including all predictors
logo <- brick(elevation, slope)

r1 <- predict(logo, model)

plot(r1)
}
Bibliography


mesocyclone detection algorithm description.

Ron Holmes; NWS State College PA. Using a decision tree and neural net to identify severe weather radar characteristics.

University of Pennsylvania computer science and information science. Decision trees.

Neural networks.


Travis M. Smith; Kimberly L. Elmore. The use of radial velocity derivative to diagnose rotation and divergence.


[28] ESRI. How slope works. [ArcGIS Resources].

[29] ESRI. What is a z-score? what is a p-value? [ArcGIS Resources].


