Wildfire Prediction Using Weather Pattern Analysis

Matthew Presti
Computer Science
University of Colorado
Boulder, CO, USA
mapr2282@colorado.edu

Holly Schwecke
Computer Science
University of Colorado
Denver, CO, USA
hosc2215@colorado.edu

Haoye Tang
Computer Science
University of Colorado
Denver, CO, USA
hata6503@colorado.edu

Purpose Statement

Wildfires represent a growing threat to ecosystems, property, and human lives across the western United States. As climate change intensifies, understanding the relationship between weather patterns and wildfire activity becomes increasingly important for effective prediction and management. Our project aims to mine the complex relationship between comprehensive weather data from NOAA and NASA MODIS satellite wildfire data to identify specific weather pattern signatures that precede wildfire events.

The primary motivation for this research stems from the rising frequency and intensity of wildfires affecting urban areas in recent years. By leveraging data mining techniques to extract meaningful patterns from historical weather and wildfire data, we hope to discover the following: specific weather pattern signatures that precede wildfire ignition, weather sequences that create conditions conducive to rapid fire spread, and historical weather indicators that could forecast fire season severity. Our focus is primarily on identifying general weather trends preceding fire events, as this represents the most actionable insight for fire management agencies. If time and resources permit, we will extend our analysis to patterns leading to extreme wind events and other severe weather conditions that exacerbate fire behavior.

These insights could provide valuable guidance for resource allocation, early warning systems, and mitigation strategies, potentially reducing the devastating impacts of wildfires on communities. By mining the data to uncover these specific patterns, we aim to contribute practical, implementable knowledge to wildfire prediction and management efforts.

Literature Survey

Several researchers have explored the relationship between weather conditions and wildfire activity, employing various data mining and machine learning approaches. Cortez and Morais (2007) applied neural networks and support vector machines to predict forest fire spread based on meteorological data[1]. Jaafari et al. (2019) compared multiple machine learning methods for wildfire susceptibility mapping, finding that ensemble approaches typically outperform individual models [2].

More recently, Jain et al. (2020) combined satellite information with weather data to predict wildfire risk, highlighting how important it is to properly organize and prepare weather pattern data for analysis [3]. Yu et al. (2022) built upon this approach by adding information about drought conditions and plant health. Their study showed better prediction results, especially when forecasting major wildfire events. They found that long-term drought measurements and seasonal vegetation changes were particularly useful indicators for predicting fire behavior [4].

These studies offer helpful approaches, though most struggle to effectively combine weather data and wildfire information across both location and time. Our project seeks to close this gap by developing better methods to integrate data from different geographic areas and by creating testing approaches that account for how wildfire conditions change over time.

Proposed Work

Data Mining Approach

Our data mining methodology follows a case-study scaling approach. We will begin by focusing on individual, well-documented wildfire events and mine the weather data preceding these events to identify potential patterns. This focused approach allows us to develop our pattern recognition techniques on a manageable scale before expanding to larger datasets.

For each case study, we will extract weather data from the 30-day period preceding the fire ignition, applying various temporal aggregation techniques to identify weather sequences that may have contributed to fire conditions. We will search for anomalies in temperature, precipitation deficit, relative humidity, and wind patterns compared to historical averages for the same location and season. Once we have established effective pattern mining techniques at the individual fire level, we will scale our approach to analyze multiple fires within specific regions.

This incremental scaling approach ensures we develop robust data mining techniques tailored to wildfire-specific questions before attempting to process the entire dataset. If constrained by time or resources, we will prioritize identifying general weather trends preceding fire events over specialized analyses of extreme wind or other exceptional weather patterns.

Data Preprocessing and Integration

We will preprocess NOAA weather and NASA MODIS wildfire datasets, handling missing values and standardizing measurements to create a unified analysis structure. For missing weather data, we will apply interpolation methods such as K-Nearest Neighbors to ensure consistency.

Geographical Data Integration

Our initial approach for geographic integration required complex processing due to the original space-delimited format of NOAA data. However, we have now secured a CSV version of the NOAA GSOD dataset that includes explicit geographical coordinates (latitude and longitude). This significant improvement aligns perfectly with our wildfire data format, enabling us to directly mine the relationships between specific weather station readings and nearby wildfire events.

Based on feedback and further investigation, we recognized that NOAA weather station data is point-based, while wildfire data may cover polygonal areas or be represented as points. This spatial mismatch requires careful integration to properly mine the relationship between weather conditions and fire occurrences.

For our spatial integration methodology, we will implement a Buffer Analysis approach. For each wildfire detection point, we will create circular buffers at multiple radii (10km, 25km, and 50km) and extract weather data from stations falling within these zones. This approach allows us to mine weather patterns at different spatial scales to determine which geographical scope provides the most predictive insights for wildfire events. For instance, we can investigate whether localized weather anomalies (10km radius) or broader regional patterns (50km radius) are more strongly associated with fire ignition.

We will include a visualization diagram illustrating this buffer methodology in our final report and validate our approach by comparing the predictive strength of weather patterns across different buffer sizes. This approach will enable us to mine both local and regional weather patterns and assess which scale reveals the most significant fire-preceding weather signatures.

Feature Engineering

For feature engineering, we'll develop temporal aggregations at multiple scales to mine weather

patterns preceding fires. Our focus is on discovering specific weather sequences that reliably precede wildfire events, rather than just building predictive models. Based on our recognition that conditions prior to wildfire detection may be more significant than conditions on the detection day itself, we will mine the data using:

- **Temporal sequence mining**: Identifying specific sequences of weather patterns (e.g., hot-dry-windy conditions) that frequently precede fires

- **Anomaly detection:** Finding unusual deviations from seasonal norms that correlate with increased fire likelihood

- **Threshold identification:** Discovering critical temperature, humidity, or wind thresholds beyond which fire risk increases substantially

- **Pattern duration analysis:** Determining how long specific weather patterns must persist before wildfire risk escalates

We'll mine the historical data to identify specific drought signatures by examining temperature and precipitation patterns, discover the particular characteristics of high-risk wind events, and determine the weather pattern transitions that most frequently lead to fire ignition. By focusing on mining these specific patterns rather than general feature engineering, we aim to extract actionable insights about the weather conditions that create heightened wildfire risk.

Analysis Progression

Our data mining workflow begins with pattern discovery in case studies of significant wildfires, examining weather data preceding ignition to identify recurring sequences and anomalies. We'll analyze these patterns across varying time lags, geographical contexts, and seasons to determine their consistency and predictive value. This approach allows us to mine both temporal dimensions (how far in advance patterns appear) and spatial dimensions (how patterns vary by region and ecosystem), efficiently identifying the most reliable fire-preceding weather signatures.

Through this mining process, we'll prioritize extracting interpretable, actionable insights about specific weather sequences rather than black-box predictions. By focusing on pattern discovery and validation rather than complex modeling, we aim to deliver practical knowledge that fire management agencies can readily implement. If resources permit, we'll extend this foundation to develop predictive models, but our core focus remains identifying specific weather signatures that reliably precede wildfire events.

Datasets

NOAA Global Surface Summary of Day (GSOD)

The NOAA GSOD dataset originates from the National Climatic Data Center (NCDC) and USAF Climatology Center. It is available at https://www.kaggle.com/datasets/noaa/noaa-global-surface-summary-of-the-day. This comprehensive dataset covers over 9,000 weather stations worldwide, spanning from 1929 to the present with daily updates. At multiple millions of records and approximately 3.3GB in size, it provides extensive temporal and spatial coverage for our analysis. The dataset includes daily summaries of temperature (mean, max, min), wind (speed, gusts, sustained), precipitation and snow depth, pressure, visibility and dew point, and various weather events including fog, rain, snow, hail, and thunder.

We initially worked with the original space-delimited .op format of the GSOD data, which required complex parsing and lacked explicit coordinates. We have now obtained a more compatible CSV version that includes latitude and longitude coordinates for each weather station. This significantly simplifies our spatial integration approach and improves the

precision of our weather-to-wildfire data mapping.

NASA MODIS Satellite Wildfire Data

The NASA MODIS satellite wildfire dataset is sourced from NASA MODIS satellite data, available via Kaggle at https://www.kaggle.com/datasets/avkashchauhan/california-wildfire-dataset-from-2000-2021. This dataset covers the period from 2000 through March 25th, 2022, focusing on a geographical range that encompasses the western United States. With 7.8 million data points, this dataset provides comprehensive geographic wildfire data including fire detection points, dates, and confidence levels, allowing for detailed spatial-temporal analysis of wildfire patterns.

Evaluation Methods

Evaluation Criteria

We'll assess our discovered weather patterns on criteria specific to wildfire prediction needs:

Pattern Consistency & Precision: How consistently and reliably the pattern precedes wildfire events across regions and years.
Lead Time: How far in advance patterns appear before ignition, thus providing sufficient warning time.
Specificity & Recall: Whether patterns primarily occur before wildfires rather than in non-fire situations, and what percentage of fires they capture.
Regional & Seasonal Variation: How pattern reliability differs across landscapes and seasons
Key Variables: Which weather factors contribute most significantly to fire risk.

These criteria focus on extracting actionable insights for fire management rather than general model performance metrics, directly addressing questions relevant to fire management agencies.

Model Selection

While our primary focus is on mining specific weather patterns preceding wildfire events, we'll employ Gradient Boosting, Random Forests, and LSTM Networks as complementary tools to reveal complex variable interactions and temporal sequences that might not be evident through direct pattern mining. These approaches will help extract interpretable insights rather than serve as standalone predictive systems.

Time-Series Cross-Validation

When evaluating our discovered patterns, we'll use a simple Forward Chaining approach that respects time progression. We'll identify patterns using historical data and validate on subsequent years, ensuring our validation periods cover complete fire seasons to account for seasonal variations. This straightforward approach will help verify pattern reliability across multiple fire seasons without introducing unnecessary complexity.

Tools

Python will serve as our primary programming language, with Pandas and NumPy providing core data manipulation capabilities. For spatial data operations, we will employ GeoPandas, which extends the functionality of Pandas to geospatial data. Our modeling will leverage Scikit-learn for traditional machine learning implementations, XGBoost for gradient boosting, and TensorFlow/Keras for LSTM implementation.

For visualization purposes, we will use Matplotlib and Seaborn to create informative graphics that communicate our findings effectively. Version control will be maintained through GitHub, and Jupyter Notebooks will facilitate exploratory analysis and documentation throughout the project.

Milestones Completed

We have successfully completed the following milestones:

1. Data Acquisition and Initial Exploration

- Downloaded both NOAA GSOD and NASA MODIS wildfire datasets
- Conducted initial exploration of data structures and quality
- Identified format challenges in the original NOAA data
- Successfully located and acquired an enhanced CSV version of NOAA GSOD data that includes geographical coordinates

2. Case Study Selection

- Identified several significant wildfire events as initial case studies
- Extracted weather data for the 30-day periods preceding these events
- Performed preliminary analysis of weather patterns preceding these specific fires
- Documented interesting weather anomalies that appeared consistently before ignition

3. Data Preprocessing and Cleaning

- Developed cleaning pipelines for both datasets
- Standardized date formats for temporal alignment
- Addressed missing values in weather metrics through appropriate imputation techniques
- Validated data quality and consistency across years

4. Initial Spatial Integration Framework

- Designed buffer analysis methodology for spatial integration
- Developed and tested coordinate-based matching algorithms
- Validated spatial relationships between weather stations and wildfire points
- Created prototype integrated dataset for a subset of regions and time periods

5. Preliminary Pattern Mining

- Identified several promising weather sequences that frequently preceded fire events in our case studies
- Discovered regional variations in significant weather patterns
- Documented threshold values for temperature and humidity that appeared most relevant to fire risk
- Created initial visualizations of pattern frequency in fire vs. non-fire situations

Milestones Todo

The following milestones remain to be completed:

Complete Spatial-Temporal Integration

Finalize buffer radius optimization for all regions

Complete full integration of datasets across entire temporal range

Validate integrated dataset for completeness and accuracy

Document integration methodology in detail

Advanced Feature Engineering

Develop comprehensive lagged features (1-day to 30-day)

Create advanced drought and heat indices

Engineer wind pattern features specifically for fire spread prediction

Implement seasonal context features to capture regional variations

Model Development and Evaluation

Implement and train all selected model types

Set up time-series cross-validation framework

Conduct preliminary model comparisons

Analyze feature importance across models

Model Optimization and Ensemble Creation

Fine-tune model parameters for optimal performance

Develop ensemble prediction approach

Evaluate models across different spatial and temporal contexts

Document model performance and limitations

Final Analysis and Documentation

Complete comprehensive analysis of results

Prepare final visualizations and interpretation

Document findings, methodology, and implications

Finalize project report and presentation

References

[1] Cortez, P., & Morais, A. (2007). A data mining approach to predict forest fires using meteorological data. In Proceedings of the 13th Portuguese Conference on Artificial Intelligence (pp. 512-523).

[2] Jaafari, A., Zenner, E. K., & Pham, B. T. (2019). Wildfire spatial pattern analysis in the Zagros Mountains, Iran: A comparative study of decision tree based classifiers. Ecological Informatics, 50, 201-221.

[3] Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. Environmental Reviews.

[4] Yu, L., Jiang, P., Wang, S., & Lai, J. (2022). A machine learning framework for wildfire susceptibility mapping integrating satellite remote sensing and meteorological data. Remote Sensing of Environment.