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 analyze the complex relationship between comprehensive weather data from NOAA and NASA MODIS satellite wildfire data to identify causative factors in wildfire activity and develop predictive models for fire season severity.

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: weather pattern signatures that precede elevated wildfire activity, methods to predict periods of extreme wind that pose elevated fire risks, and novel predictive indicators for forecasting fire season severity. These insights could provide valuable reference for resource allocation, early warning systems, and mitigation strategies, potentially reducing the devastating impacts of wildfires on communities.

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

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. A key innovation in our approach is how we combine data from different locations. Since weather stations and wildfire locations don't perfectly align, we'll create circular zones around each wildfire site (10km, 25km, and 50km across). This allows us to gather weather data from nearby stations within these zones. Using multiple-sized circles helps us determine the ideal distance for connecting weather patterns to wildfire events.

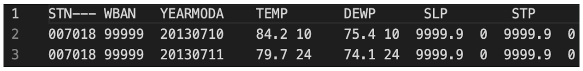
For feature engineering, we'll develop aggregations (1-day, 7-day, and 14-day averages) to capture weather trends preceding fires. We'll create drought measurements by examining temperature and rainfall patterns, identify extreme weather events like heat waves and dry spells, and track wind behavior crucial for fire spread. We'll also normalize measurements based on regional seasonal patterns to detect meaningful deviations.

Our analysis will progress from initial data exploration to identify weather-wildfire correlations, through time-lagged analysis to discover leading indicators, pattern mining to identify high-risk weather sequences, and finally to developing predictive models for fire risk assessment and extreme wind event forecasting based on the relationships discovered in our earlier analytical work.

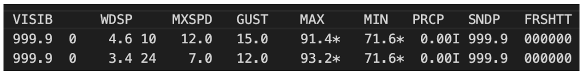
Datasets

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.



```
STN--- WBAN   YEARMODA    TEMP        DEWP      SLP        STP
007018 99999  20130710    84.2 10     75.4 10   9999.9  0  9999.9  0
007018 99999  20130711    79.7 24     74.1 24   9999.9  0  9999.9  0
```



```
VISIB       WDSP      MXSPD   GUST     MAX      MIN    PRCP   SNDP    FRSHTT
999.9  0     4.6 10    12.0   15.0     91.4*    71.6*  0.00I 999.9   000000
999.9  0     3.4 24     7.0   12.0     93.2*    71.6*  0.00I 999.9   000000
```

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

Model Selection

We will test several prediction approaches that each offer different advantages for our wildfire forecasting challenge. We'll use advanced tree-based methods like XGBoost that can handle different types of weather measurements together and can identify complicated connections between weather conditions and fire risks. These methods work well even when wildfires are rare events in our data, which is important for accurate predictions.

We'll also use Random Forest techniques that allow us to clearly see which weather factors are most important for predicting fires, even when there are unusual weather readings in our data. For capturing long-term temporal dependencies in weather patterns preceding fires, we will utilize Long Short-Term Memory (LSTM)

Networks, which can identify complex sequential patterns in time-series data. Finally, we'll combine these different approaches to create a more reliable prediction system, which helps improve accuracy when forecasting rare but dangerous wildfire events.

Time-Series Cross-Validation

To properly evaluate temporal predictions, we will implement forward chaining. This involves training on years 2000-2010 and validating on 2011, then training on 2000-2011 and validating on 2012, and continuing this pattern through the available data. This approach respects the temporal ordering of the data, preventing future information from being used to predict past events. We will ensure validation periods cover complete fire seasons to properly assess seasonal prediction performance.

Performance Metrics

We will assess model performance using multiple metrics. The Area Under ROC Curve (AUC) will measure classification performance, while Root Mean Square Error (RMSE) will evaluate continuous predictions such as fire risk indices. We will emphasize precision and recall, with particular attention to recall for high-risk events where missing a potential wildfire carries greater consequences than false alarms. Additionally, we will conduct feature importance analysis to enhance model interpretability and identify the most significant weather variables for predicting wildfire risk.

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

Our project will proceed according to the following timeline:

- Weeks 1-2: Data acquisition and cleaning - download and organize datasets, implement cleaning pipelines, document data quality issues
- Weeks 3-4: Develop spatial-temporal integration framework - test buffer analysis approach, implement spatial interpolation methods, create unified dataset
- Weeks 5-6: Feature engineering and exploratory analysis - create temporal aggregation features, develop drought indices, perform correlation analysis
- Weeks 7-8: Model development and initial evaluation - implement and train models, set up time-series cross-validation, compare preliminary results
- Weeks 9-10: Refinement and final evaluation - optimize model parameters, develop ensemble models, conduct comprehensive evaluation
- Weeks 11-12: Documentation and report writing - prepare visualizations, document methods and findings, finalize report

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.