Cameron Sterling
Applied Machine Learning
Wayne Lee
May 8, 2025

Distinguishing between Routine and Conflict-Associated Vegetation Fires in North Syria

GitHub
Here's my github page for this project! https://github.com/cameronsterling/ML_Final/tree/main

Introduction

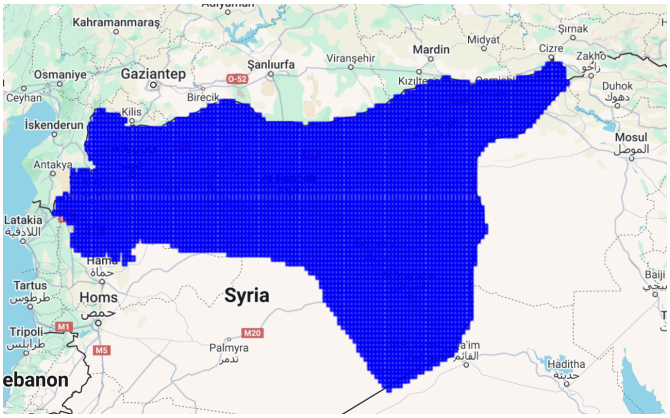
This project builds on existing literature, which notices that there has been a high number of fires in Northern Syria in the past years. Scientists do not know whether these fires are a result of a dramatic increase in agricultural production in areas that have been resettled or they are a result of conflict and instability. Some hypothesize that farmers or settlers in areas of Northern Syria that have been stabilized are burning brush to clear the way for new fields. Others note that non-state and state-backed armed groups may set fires to croplands to weaken certain groups economically, force certain groups to migrate away from their ancestral lands, or that fires may be caused by explosions or unexploded ordinance that prevents farmers from accessing their fields.

This project asks *can machine learning distinguish between routine and conflict-associated vegetation fires in Northern Syria?* The answer to this question is important for a variety of stakeholders. Many human rights groups have accused various armed factions of intentionally creating vegetation fires. Other groups oriented towards economic development might be interested in renewed agricultural burns.

Using this approach, I found that conflict-associated fire anomalies are consistently present over time but are too irregular to form dense clusters, while routine fire anomalies show strong spatial and temporal concentration. The clustering model achieved a silhouette score of 0.83, indicating clear separation between the discovered groups.

Data and Model

This project attempted to first identify areas with anomalous levels of fire, and then use clustering to find out if they had similar characteristics. Areas were defined as 5x5 km boxes in Northern Syria, the definition of which can be seen in the Javascript code. Below is an image that shows the area of interest. Each of the boxes is also temporally relevant, with its own month and year from 2017-2022. The reason that these dates were selected is primarily data availability-related, but also because they represent the peak years of conflict in North Syria.



Area Covered

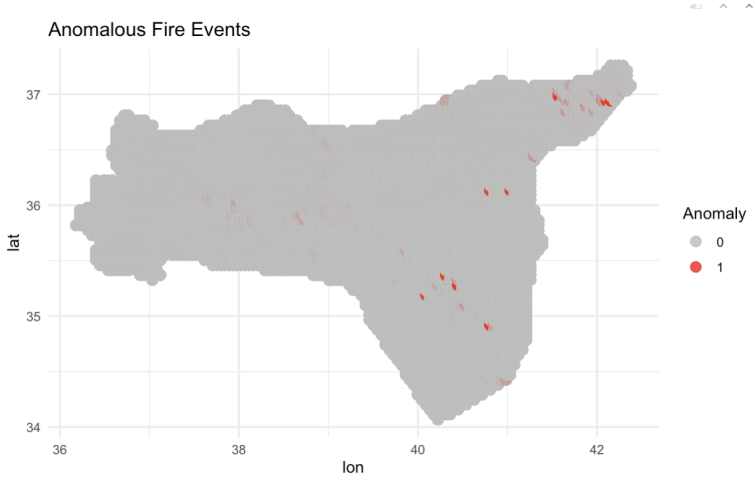
Fire Anomalies

Fire activity is captured using the [VIIRS Active Fire product], which detects thermal anomalies from space at 375 meter resolution. We aggregate the number of VIIRS fire detections in each grid cell by month, creating a spatiotemporal record of vegetation fires across Northern Syria. This serves as our primary outcome variable.

Areas were then marked as "anomalous" as identified through an isolation forest model with 100 trees. The top 95 percent of anomaly scores were considered as areas with anomalous numbers of fires per month in any given year the distribution of anomaly scores are as follows:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.4050	0.4050	0.4050	0.4119	0.4050	0.9001

In all, 12271 of 326237 rows were marked as anomalous fires. The figure below shows a map of anomalous fires in Northern Syria collapsed over time.



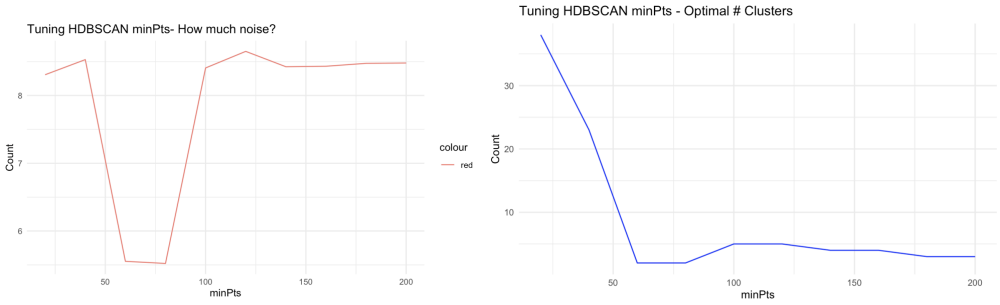
Clustering Algorithm

After identifying anomalous fire events, I used the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm to group these anomalies based on their shared characteristics. Unlike k-means, HDBSCAN does not require pre-specifying the number of clusters and is particularly well-suited for discovering irregularly shaped or unevenly sized clusters, a common feature in real-world geospatial data like conflict zones.

I used several sources of data to form my clusters:

1. Environmental Data: The core of the environmental data draws from MODIS-derived NDVI (Normalized Difference Vegetation Index), which serves as a proxy for vegetation health. MODIS NDVI products, typically available at 250 meter resolution and composited every 16 days, are widely used to monitor ecosystem dynamics. In our study, we aggregate NDVI values monthly and spatially to a 5×5 km grid, allowing us to quantify seasonal greenness patterns and detect deviations that might reflect ecological stress, degradation, or recovery. MODIS NDVI was accessed through GEE, which can be replicated through the javascript code I uploaded to my git repository. There were 65210 NA's for this dataset, which were imputed with K-NN. To account for weather-related drivers of fire activity, we incorporate monthly precipitation data from CHIRPS. Precipitation is a key contextual variable: both excess rainfall and drought can shape fire risk and vegetation fuel loads. Alongside NDVI, precipitation helps define the environmental "normal" within each grid cell and month, allowing for the detection of fires that occur under unexpected conditions. CHIRPS was accessed through GEE, which can be replicated through the javascript code I uploaded to my git repository. There were 65210 NA's for this dataset, which were imputed with K-NN. Temporal features were encoded using sine and cosine transformations of the month to account for cyclic seasonality.
2. Boundaries: I used the *GAUL/FAO* dataset. This, through the GEE script, provided geographic boundaries. You can see how to access it through the Java Script in the repository.
3. Conflict: To assess whether anomalous fires may be linked to political violence or war, I integrate data from the Armed Conflict Location & Event Data Project (ACLED). ACLED provides geocoded records of political violence, protest, and military activity reported by local and international sources. We aggregate ACLED events by type, actor, and sub-event category into the same 5×5 km grid system, allowing us to test whether unusual fire events align with conflict presence or intensity. ACLED data is available from 2017 onwards. Data can be accessed through this API. It is vectorized, so I coerced it into the grid that I created. It was included in the dataset in the form of event counts per month-gridbox for different types of events such as attacks, airstrikes, and IED explosions.

To run HDB scan optimally, I engaged in hyperparameter tuning. Through this approach, I found that the best minimum number of points to include was 60 in order to minimize the number of clusters (better to explain the model) and noise.



Of the anomalous fires, the model returned two clusters with 94 and 11919 points respectively; 258 points were “noise”. The silhouette score was 0.83, which indicates that the model has a strong fit.

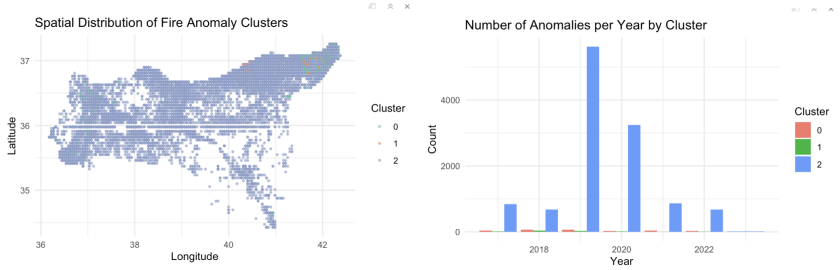
The clustering algorithm returned two substantive clusters and a set of points labeled as noise. Summary statistics for each group are provided below:

Cluster	**count**	**mean_attack**	**mean_shelling**	**mean_airstrike**	**mean_ndvi**	**mean_precip**
0 (noise)	258	0.027131783	0.91472868	0.104651163	0.3284641	101.489940
1	94	0.0000000	0.00000000	0.000	0.3265926	83.471665
2	11919	0.002097491	0.03213357	0.003523786	0.1926603	7.146412

Cluster 2 comprised the majority of anomalous fire observations (n = 11,919). This group was characterized by low NDVI and low average precipitation, and had middling levels of recorded conflict activity. These observations are likely associated with low-level conflict driven fires, including displacement or abandoned areas.

Cluster 1 was much smaller (n = 94), with similarly high NDVI and precipitation values but no recorded conflict activity. While the size of this cluster limits strong conclusions, it may represent purposeful /controlled burns.

Observations not assigned to a cluster by HDBSCAN (cluster 0, n = 258) exhibited the highest levels of conflict intensity across all indicators, particularly shelling and airstrikes. These data points also had high NDVI and precipitation, suggesting that they occurred in vegetated, well-watered areas. The fact that these points were labeled as noise by the algorithm likely reflects their relatively low frequency and irregularity in feature space, rather than a lack of distinctiveness. Given their elevated conflict metrics, these may represent directly conflict-associated fire anomalies (such as arson) that were too sparse or spatially diffuse to form a high-density cluster under the HDBSCAN framework.



These notions are supported by the above figures. These, together, suggest that most fire anomalies are widespread and routine, with clear spatial and temporal concentration in 2019 and 2020. Conflict-related anomalies appear more scattered and rare, failing to form dense clusters but remaining consistently present across years.

To evaluate the added value of unsupervised clustering, a rule-based classifier was constructed as a baseline. Under this approach, any grid-month with at least one recorded conflict event (across any ACLED sub-event type) was labeled as “Conflict,” while all others were labeled as “Other.” This produced a binary grouping of the same anomalous fire events used in clustering.

Category	**count**	**mean_attack**	**mean_shelling**	**mean_airstrike**	**mean_ndvi**	**mean_precip**
Conflict	334	0.09580838	1.853293	0.2065	0.246	6.72
other	11973	0.0000000	0.00000000	0.000	0.195	9.80

The rule-based classifier captures a small subset of anomalies (334 out of 12,271) as conflict-associated. These observations have higher NDVI values and slightly lower precipitation on average, with non-zero mean attack rates. However, this binary classification fails to capture variation within the remaining 97% of anomalies labeled as “Other.” It does not distinguish between anomalies due to environmental stress, agricultural burns, or other non-conflict processes. In contrast, the HDBSCAN clustering identified multiple subgroups among the anomalous fire events.

Conclusion

This project applied an unsupervised clustering approach to identify patterns among anomalous vegetation fires in Northern Syria. HDBSCAN was used to cluster spatial-temporal anomalies derived from VIIRS fire data, with environmental and conflict-related contextual variables. The algorithm returned two substantive clusters and a set of outliers labeled as noise. These clusters corresponded to ecologically distinct fire anomalies, with varying levels of precipitation, NDVI, and conflict intensity.

Notably, directly conflict-related fire anomalies were captured almost exclusively in the noise component of the model, highlighting their irregularity and sparsity. A rule-based classifier flagged conflict events with a clear threshold, but failed to distinguish among the broader range of non-conflict anomalies, reinforcing the value of clustering in uncovering structure among the majority of events. These results suggest that while most fire anomalies are routine or environmentally driven, conflict-associated anomalies are consistently present but defy easy categorization due to their variability. Future work could improve model performance and interpretability by incorporating additional covariates, testing different anomaly detection techniques, and integrating higher-resolution fire severity data.