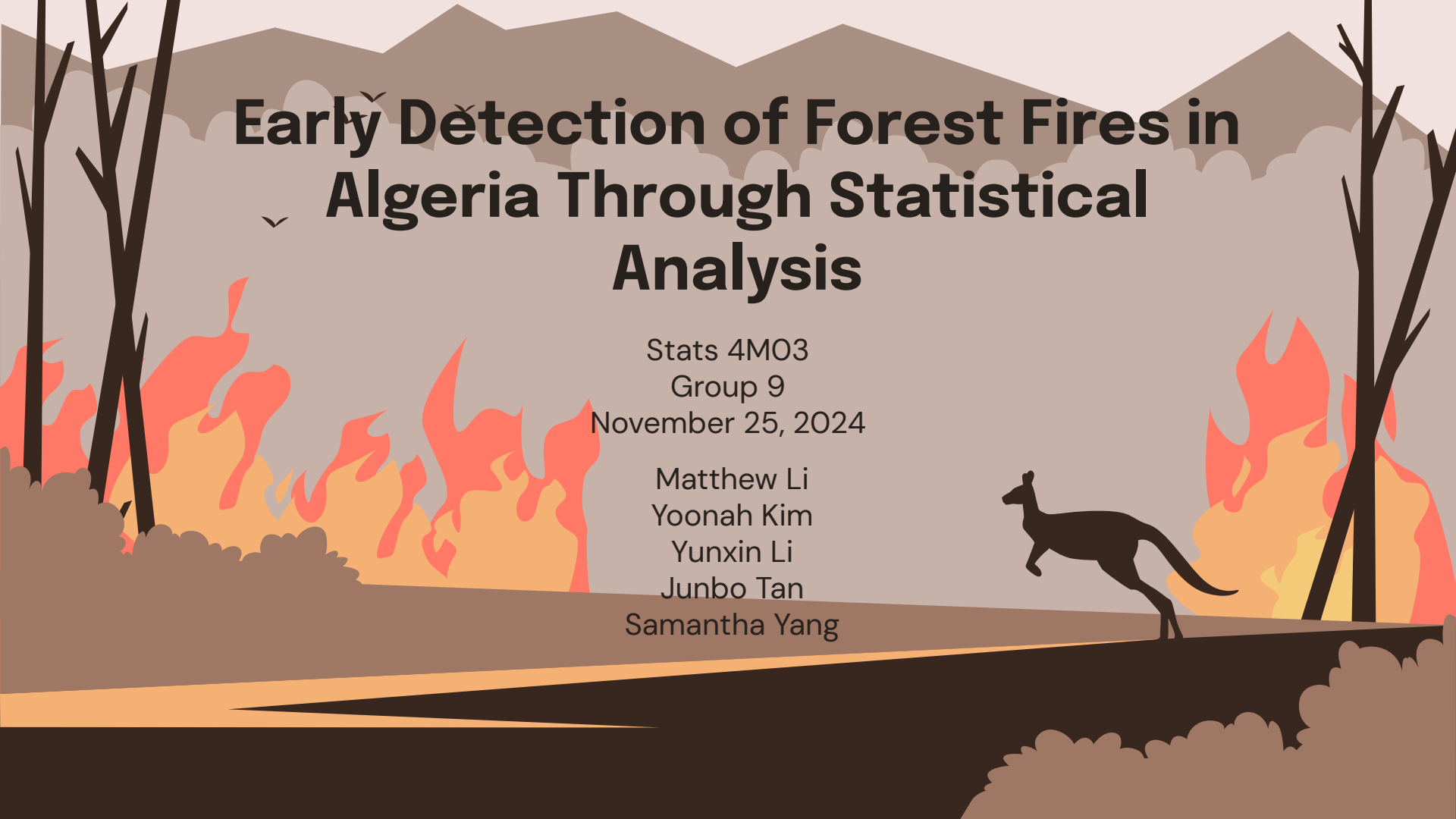
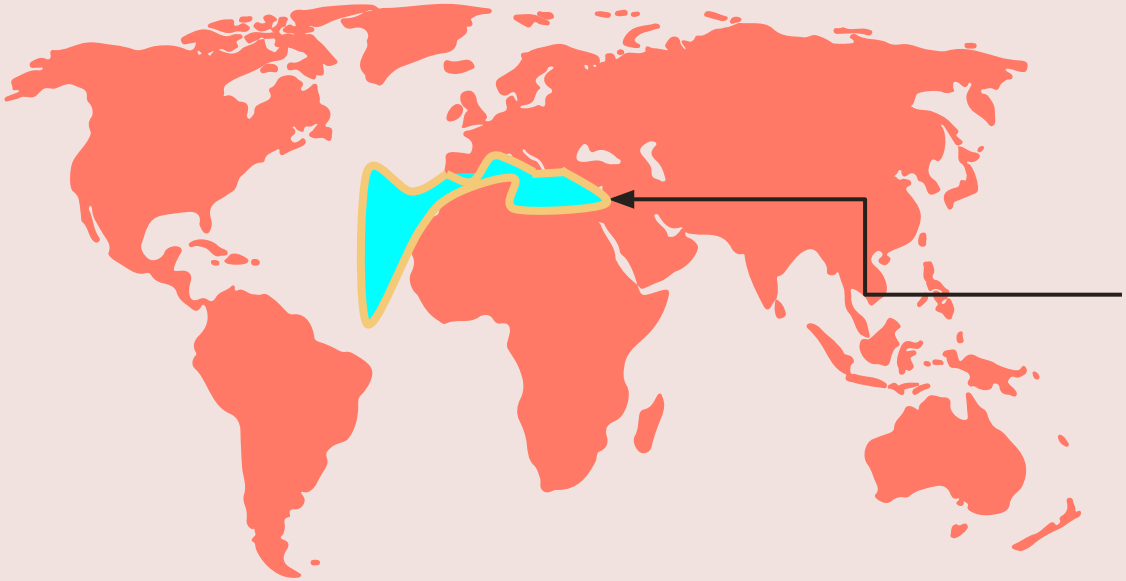
Early Detection of Forest Fires in Algeria Through Statistical Analysis

Stats 4M03
Group 9
November 25, 2024

Matthew Li
Yoonah Kim
Yunxin Li
Junbo Tan
Samantha Yang



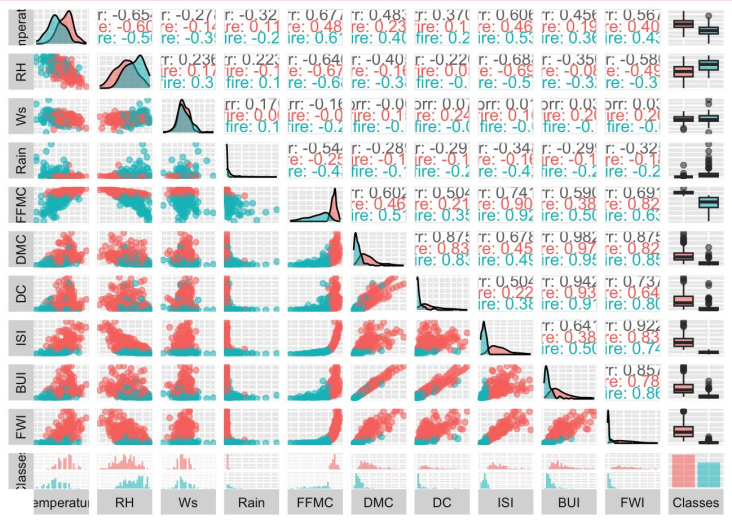
Introduction & Abstract



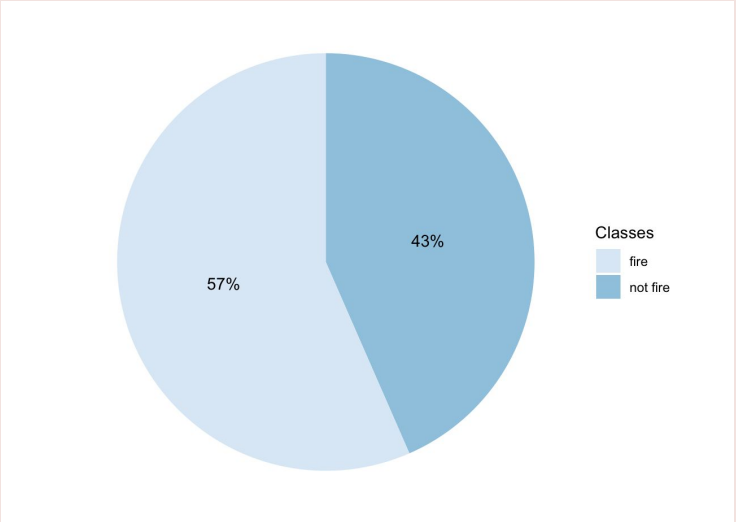
Mediterranean Basin

Bordered by Algeria, Morocco, Spain, etc.

Exploratory Data Analysis & Data Preparation



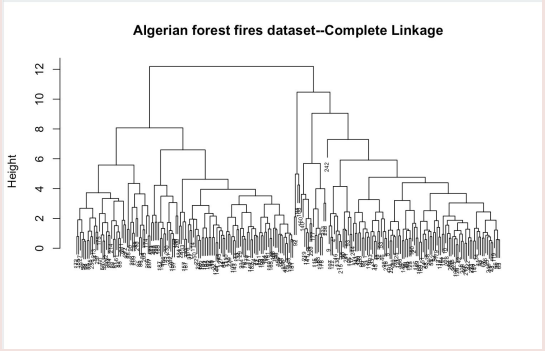
Pairs Plot: All Variables

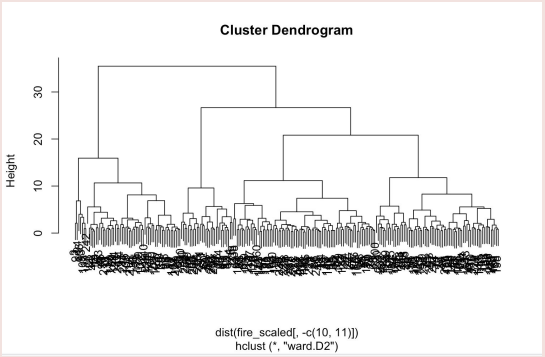


Pie Chart: Fire vs. Not Fire

- Categorical variables such as “Region” and “Date” were removed
- No missing values
- All analyses followed a 70%/30% split for training and testing

Hierarchical Clustering





Methods Attempted/ARI

- **Single Linkage**: 0.005081677
- **Complete Linkage**: 0.280948
- **Average Linkage**: 0.02707615
- **ward.D2**: 0.3939026

Best model

: ward.D2

Centroid Clustering Methods: k-Medoids

Clusters

- Euclidean dissimilarity
- Centers are actual data points

K-value

- Number of clusters
- Run algorithm with different k (2 to 5)

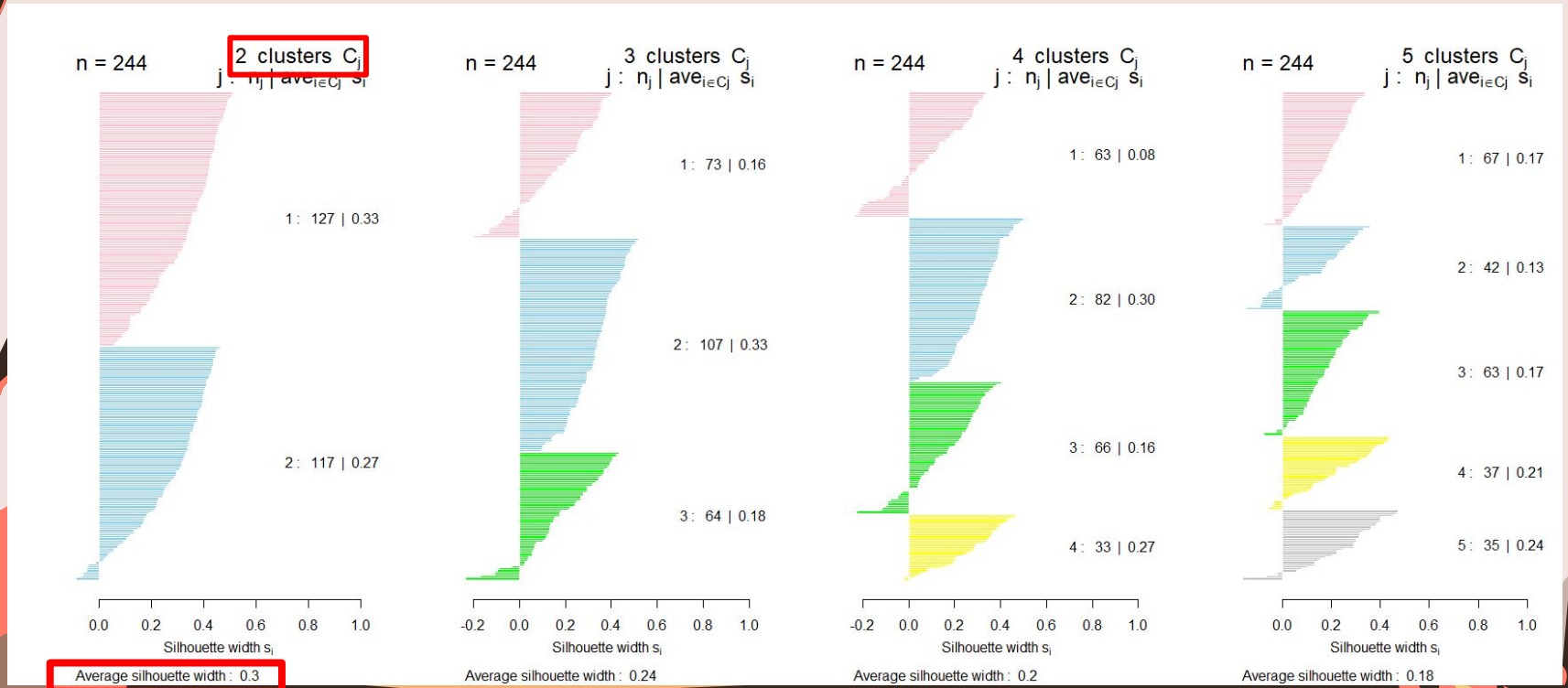
Best k-value

- Silhouette Method
- K=2

ARI

- 0.530272 < 0.6
- Moderate accuracy

Silhouette width of K-medoids from k=2 to 5



Logistic Regression

	Estimate	Std. Err.	z value	Pr(>|z|)
(Intercept)	2.657e+01	5.296e+04	0.001	1
Temperature	2.187e-09	4.818e+04	0.000	1
RH	-5.410e-09	4.834e+04	0.000	1
Ws	-7.828e-10	3.180e+04	0.000	1
Rain	4.926e-10	3.126e+04	0.000	1
FFMC	-8.302e-09	6.273e+04	0.000	1
DMC	-7.215e-09	2.095e+05	0.000	1
DC	1.267e-08	1.208e+05	0.000	1
ISI	3.839e-08	6.280e+04	0.000	1
BUI	-1.824e-08	2.982e+05	0.000	1
Classesnot fire	-5.313e+01	1.037e+05	-0.001	1

Table 1: Logistic Regression Results for full model

Logistic Regression

	Estimate	Std. Err.	z	p	Odds Ratio
intercept	0.9056	0.2865	3.160	0.002	-
Temperature	0.8675	0.2396	3.620	< 0.01	2.380846
DC	2.2660	0.4714	4.807	< 0.01	9.640362
p-value=1.825222e-20					

Temperature:
(20℃ ~ 44℃)

DC:
Drought Code Index
(7 ~ 220.4)

$$\log\left(\frac{\pi}{1-\pi}\right) = 0.9056 + 0.8675x_1 + 2.2660x_2$$

MCR = 0.166667

Discriminant Analysis: LDA vs. QDA

Two models:

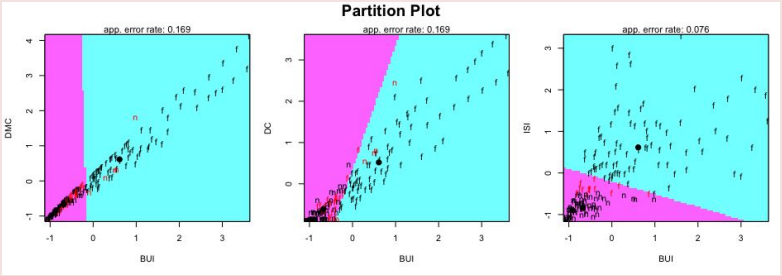
- Full model: all variables
- Reduced model: Class~Temperature+DC

Full model results:

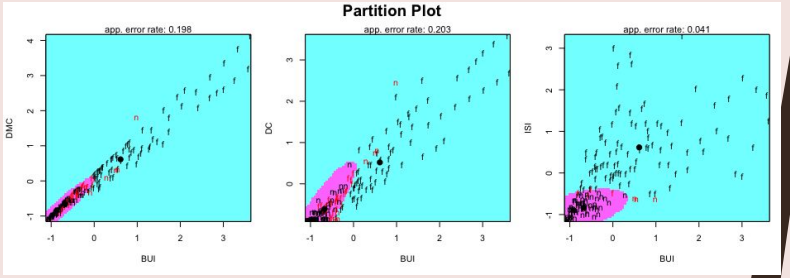
- LDA MCR = 0.08333
- QDA MCR = 0.04167

Reduced model results:

- LDA MCR = 0.16667
- QDA MCR = 0.25



Partition plot for LDA



Partition plot for QDA

Discussion & Conclusion

- Hierarchical vs. Centroid-based clustering
  - Ward.D2 vs. K-medoids
  - ARI = 0.3939026 vs. ARI = 0.4776753
- LDA/QDA vs. logistic regression
  - Reduced model
  - MCR = 0.16667 LDA, logistic regression
- Next Steps
  - Attempting other methods
    - Clustering - ARI over 0.6
    - Classification - potential improvement