Lamont Internship 2022-2023

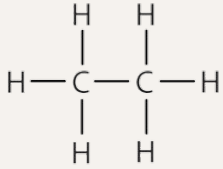
Anya Ganeshan

01

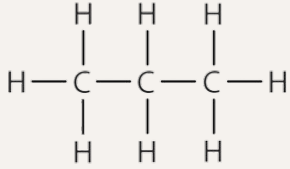
n-Alkane Chain Length Distributions

Background

- *n*-alkanes: straight, saturated hydrocarbon chain with only single bonds
  - ex) C25 = 25 carbon atoms in the chain
- Relevance
  - Long chain (C21-C37) *n*-alkanes = most long-lived and commonly used plant biomarkers
  - Applications: paleoecology and paleoclimatology
    - Understand more about past ecosystems, environments, and plant communities
- MAP: mean annual precipitation
  - Central Park: about 1140.5 mm
  - Death Valley: about 57 mm
- MAT: mean annual temperature
  - Central Park: about 12.2℃
  - Death Valley: about 32℃



Ethane



Propane

$$\mathrm{ACL} = \frac{(21\mathrm{C}_{21} + 23\mathrm{C}_{23} + \cdots + 39\mathrm{C}_{39})}{(\mathrm{C}_{21} + \mathrm{C}_{23} + \cdots + \mathrm{C}_{39})}$$

n-Alkane Chain Length

Question

Does the climate (MAT and MAP) affect a plant's ACL? Can an ML model predict climate factors based on the chain length distribution?

Data Collection

Research Papers & Lab's Collection

Data Cleaning

Find MAT & MAP from latitude + longitude

Modeling

Use Python ML Models to predict

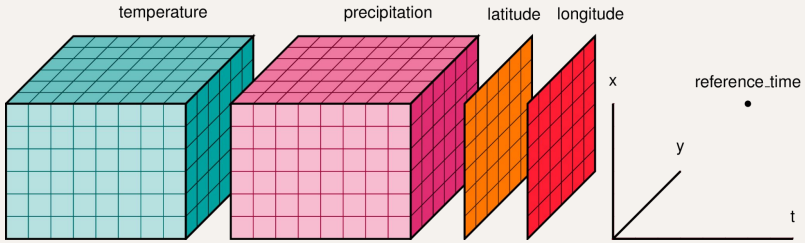
Results

Analyze results + conclude

Data Preparation

Asia & North America

- These datasets had information about the site's monthly temperature and precipitation (most accurate climate data)



Africa & South America

Dataset mapping collection site to latitude & longitude

Latitude and Longitude → MAT and MAP of closest weather station

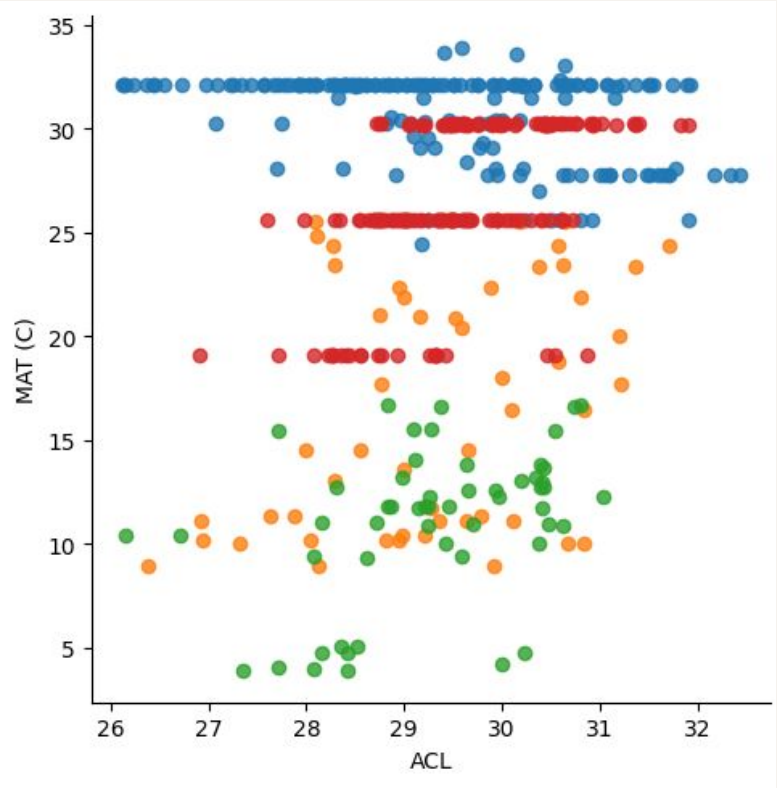
National Centers for Environmental Information

Latitude and Longitude → MAT and MAP of nearest available data

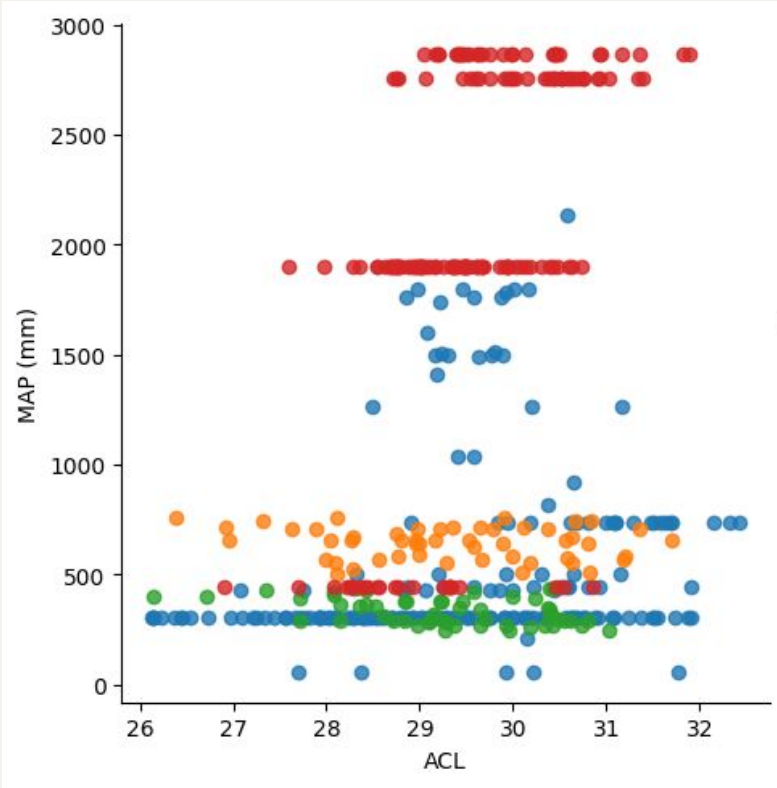
Terra Climate data

Despite shifting to the Terra dataset, it was still not granular enough to provide the MAT and MAP for the exact longitude and latitude.

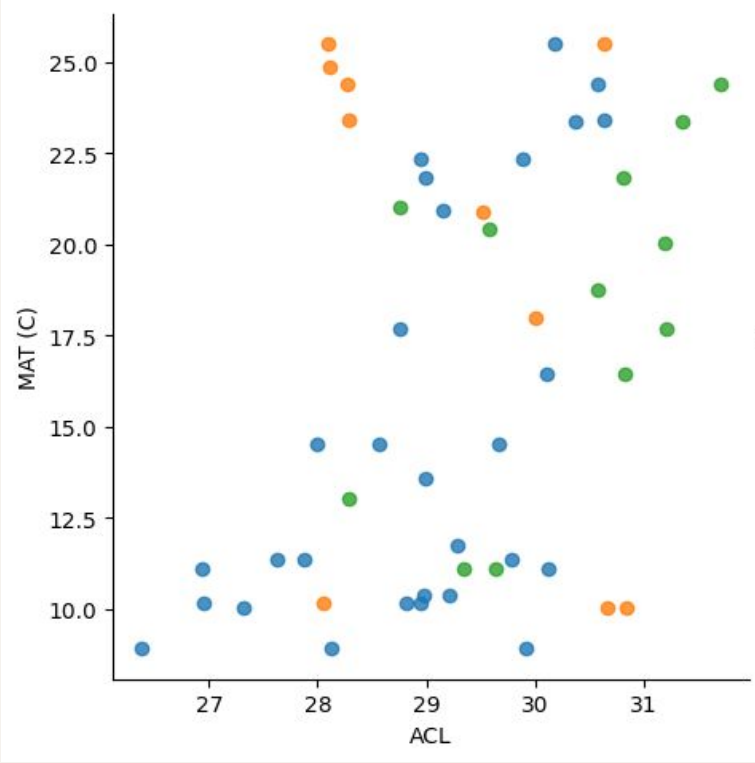
Global

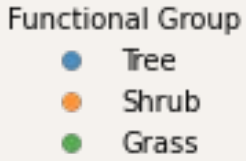

MAT (C)
ACL
35
30
25
20
15
10
5
26
27
28
29
30
31
32

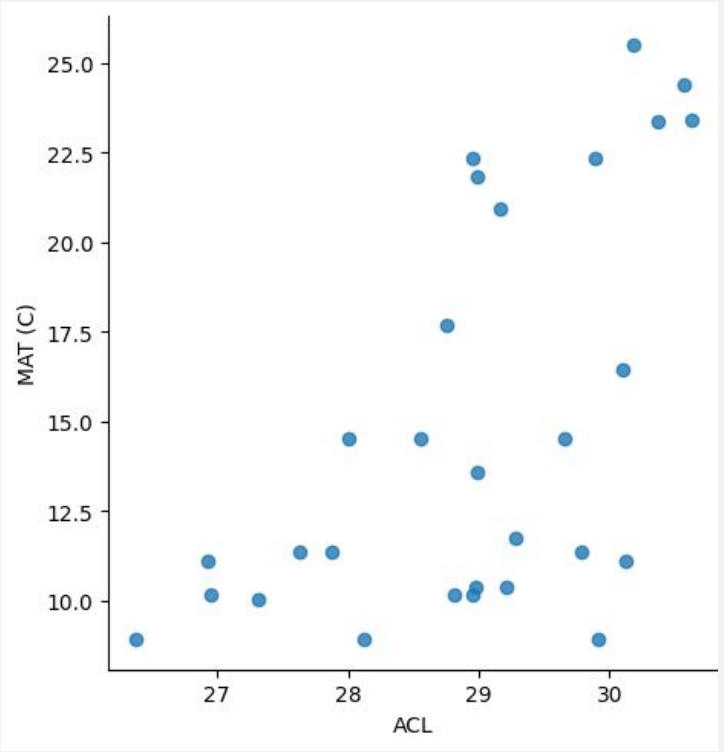

Continent
Africa
North America
Asia
South America


MAP (mm)
ACL
3000
2500
2000
1500
1000
500
0
26
27
28
29
30
31
32

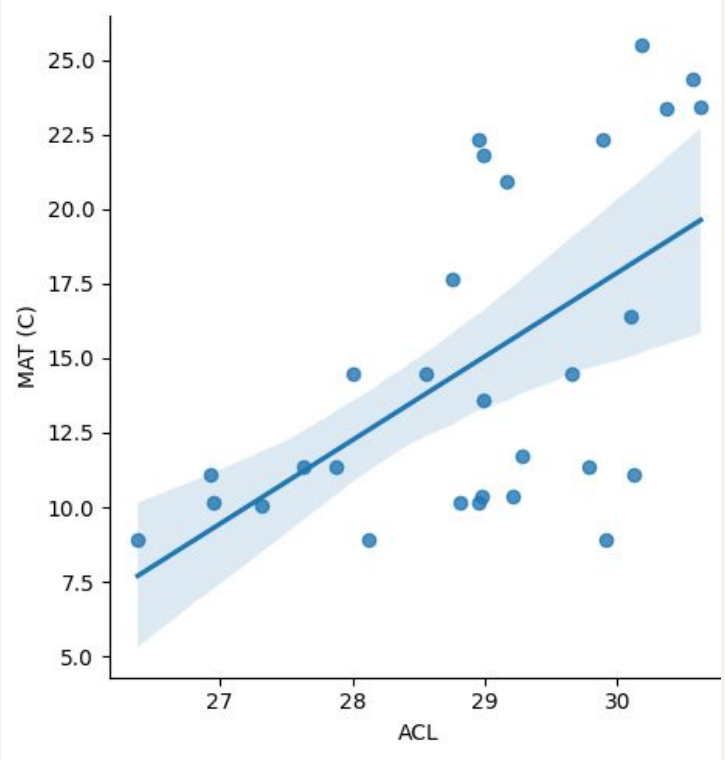
North America



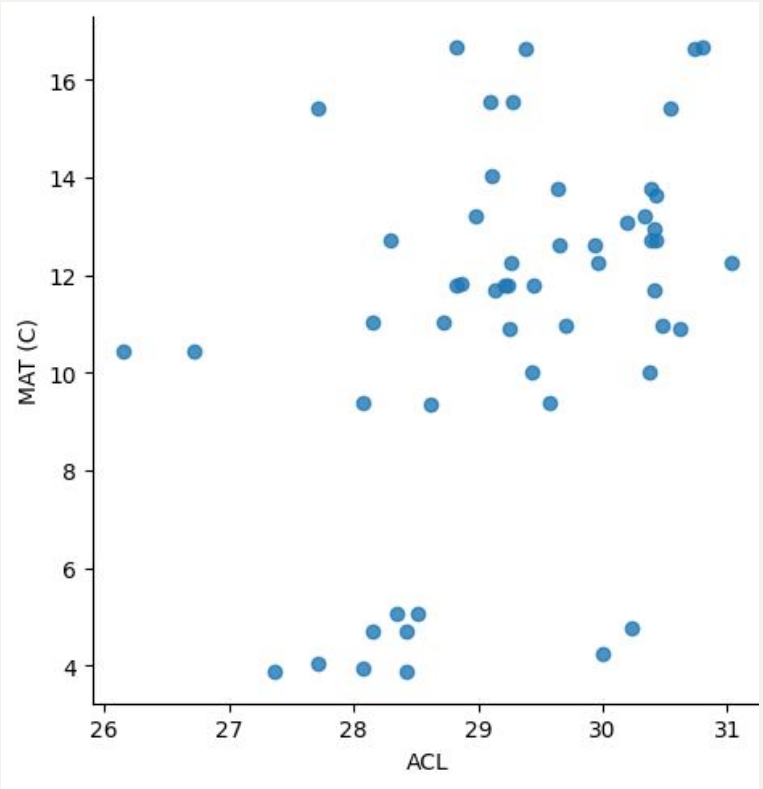


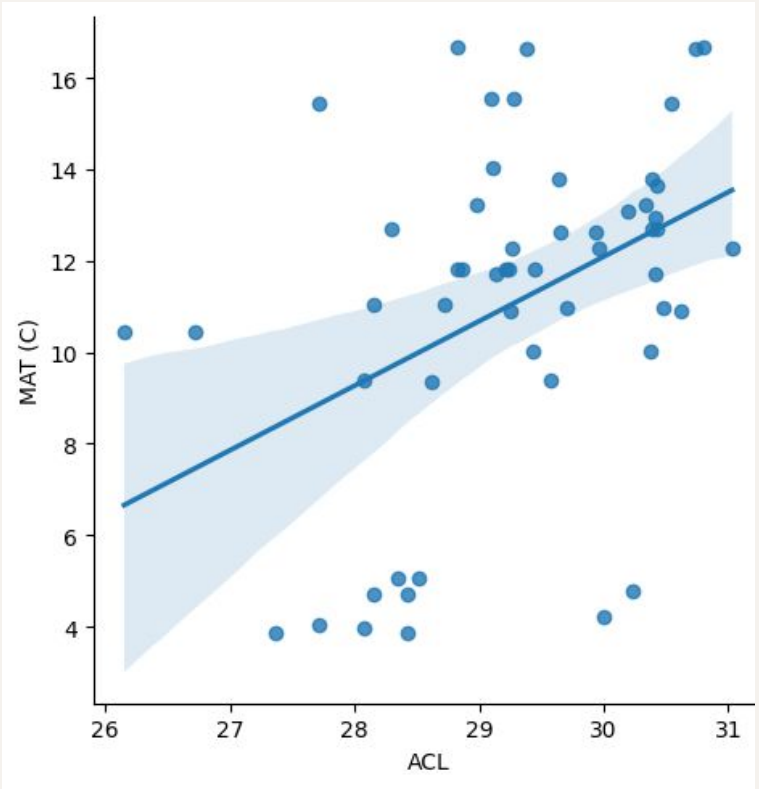


North America - Trees



Asia - Shrubs


MAT (C)
ACL
16
14
12
10
8
6
4
26
27
28
29
30
31


MAT (C)
ACL
16
14
12
10
8
6
4
26
27
28
29
30
31

Machine Learning

A subset of AI that describes machines that use experience/data to improve computer performance at a task

Supervised learning

- Use labeled datasets to predict outputs or classify inputs
  - Classification = predicting categorical values
  - Regression = predicting a discrete number
- Use **attributes/features** to predict a **target/label**
  - ex) predicting house prices based on square foot, material, and age

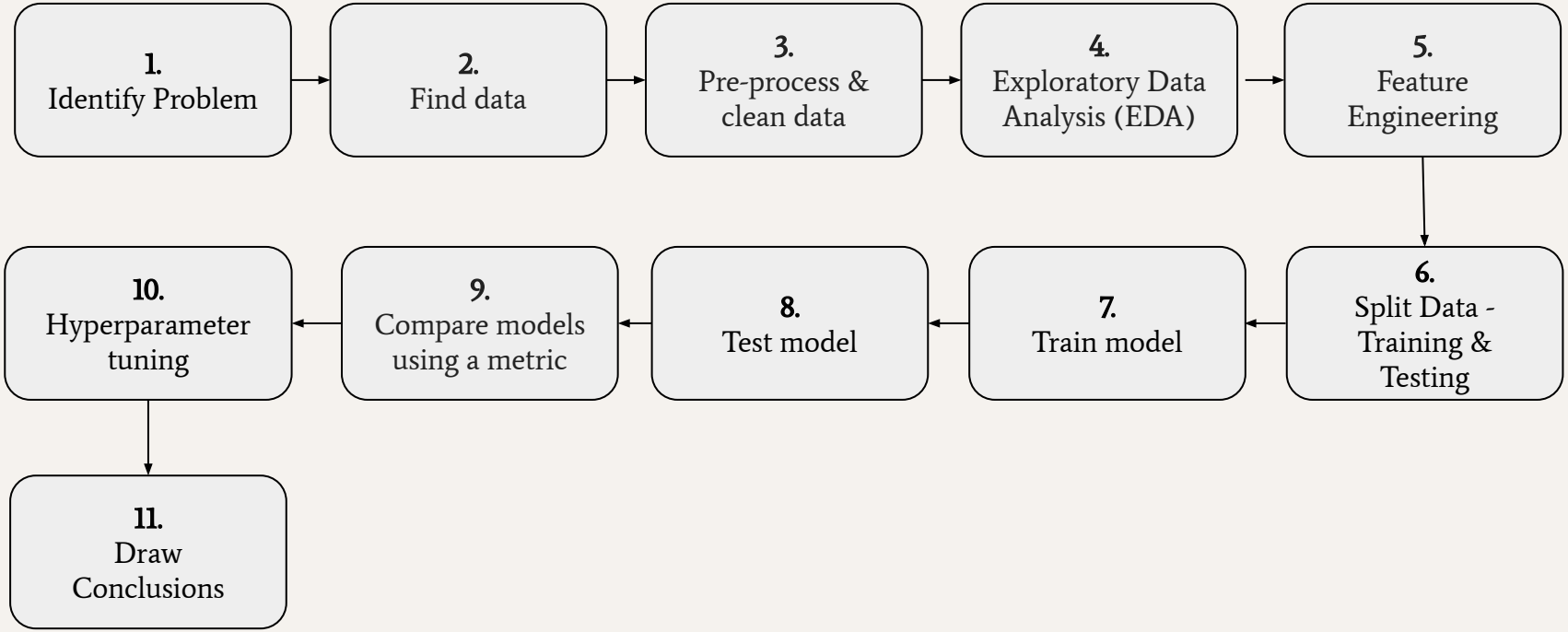
Unsupervised learning

- Discover patterns, categorize, analyze and cluster unlabeled data (not predicting)
- ex) Amazon recommends you with items to buy based on your purchase history

Reinforcement learning

- Method based on trial and error → reward good behavior & punish undesired behavior
- ex) self driving cars

Supervised Learning Process



Linear Regression

- Uses historical (training) data to predict an output variable
  - Input = continent, relative concentration of each chain length, ACL, plant functional type (tree vs grass), etc
  - Output = MAT (mean annual temp)

- Using the equation: $Y = \alpha + \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + ...$
  - y=predicted output value
  - x_i=each feature (input)
  - Goal: statistically find the best values of all of the parameters (line of best fit)

- Techniques to prevent overfitting (relying on training data too much)
  - Lasso: penalty = absolute value of the sum of the coefficients (L1-norm)
  - Ridge: penalty = sum of the squares of the coefficients (L2-norm)
  - **higher values of coefficients → bigger penalty → magnitudes of coefficients are reduced**

Validating Regression Models

R^2 (coefficient of determination)

- Statistical measure of how close the data is to the best fit line (close to 1 is best)

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)}$$

  - SSR = sum of residuals squared (residual = actual - predicted)
  - SST = sum of distances of each data point from the mean

MSE (mean squared error)

- Measure of the amount of error in models (close to 0 is best)

$$MSE = (average\ of\ residuals)^2$$

$$= \frac{(sum\ of\ residuals)^2}{\#\ of\ data\ points}$$

Results

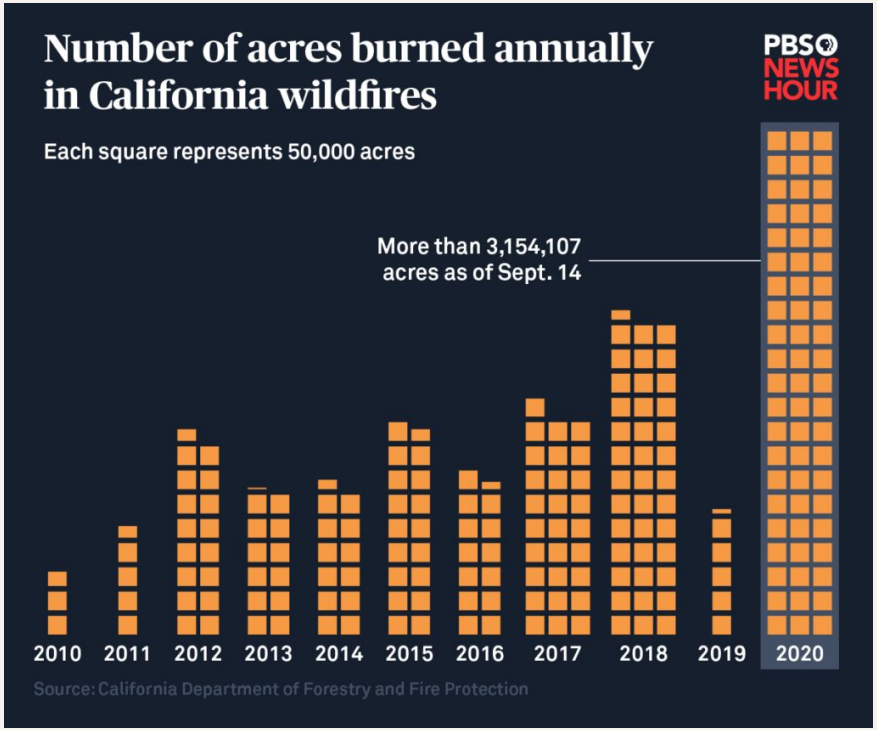
Category	R^2	MSE	# of Data Points
Trees - North America	0.95	1.92	29
Trees - Africa	0.89	0.58	66
All - Global	0.87	6.58	437
Trees - Global (Ridge)	0.89	4.7	252

David Gottlieb

02
Fire Data

Background

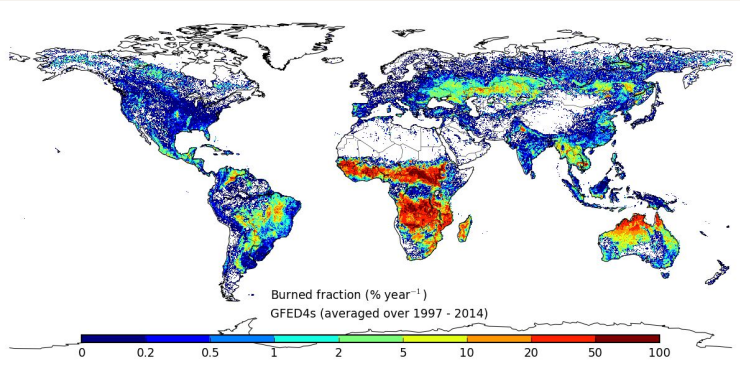
- Relevance
  - Climate change → fires are expected to increase in intensity and frequency
  - Want to understand how the distribution of precipitation might impact fire frequency
  - How can we reduce fire frequencies and adapt to this future threat?

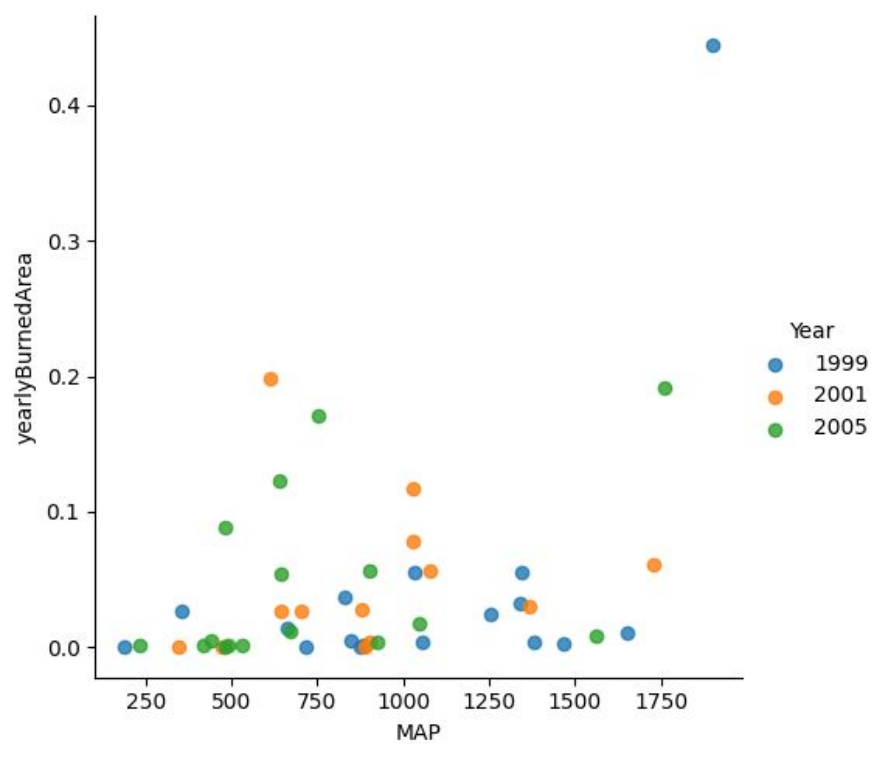


Yearly Burned Area & MAP

- Question: how is annual burned area in different locations affected by the mean annual precipitation in the region?
  - Is the dry season length a factor?

- Data Collection:
  - GFED files - yearly burned area



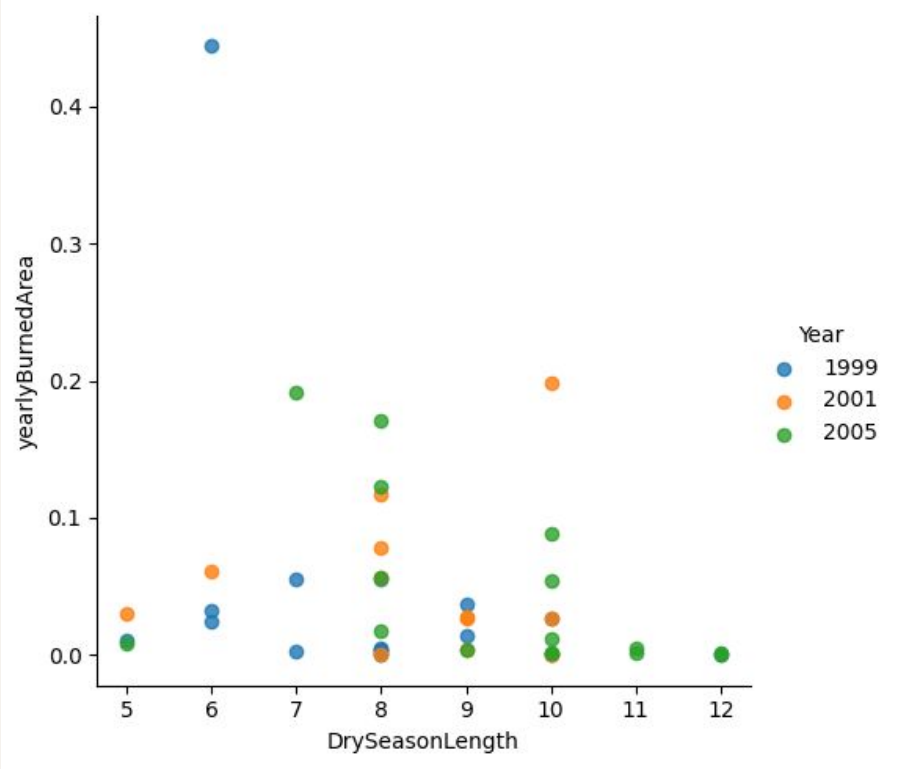


Dry Season Length

Dry season length:

< 5% of MAP OR

< 100 mm/month (Nix 1983)



Summary

- MAT v ACL: correlation between MAT & ACL and we can use ACL (along with other attributes) to predict MAT

- Yearly Burned Area v MAP:
  - Too little precipitation = not enough water for vegetation to grow
  - Too much precipitation = less likely to ignite and will burn slower

- Next Steps
  - More accurate climate data
  - More data

- Thank you to Troy, Ruth, and the entire lab for your support and mentorship

Works Cited

https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily

http://thredds.northwestknowledge.net:8080/thredds/catalog/TERRACLIMATE_ALL/data/catalog.html

Wang, J., Xu, Y., Zhou, L., Shi, M., Axia, E., Jia, Y., ... & Wang, G. (2018). Disentangling temperature effects on leaf wax n-alkane traits and carbon isotopic composition from phylogeny and precipitation. *Organic Geochemistry*, *126*, 13-22.

Bush, R. T., & McInerney, F. A. (2015). Influence of temperature and C4 abundance on n-alkane chain length distributions across the central USA. *Organic Geochemistry*, *79*, 65-73.

Feakins, S. J., Bentley, L. P., Salinas, N., Shenkin, A., Blonder, B., Goldsmith, G. R., ... & Malhi, Y. (2016). Plant leaf wax biomarkers capture gradients in hydrogen isotopes of precipitation from the Andes and Amazon. *Geochimica et Cosmochimica Acta*, *182*, 155-172.

Struck, J., Bliedtner, M., Strobel, P., Schumacher, J., Bazarradnaa, E., & Zech, R. (2020). Leaf wax n-alkane patterns and compound-specific δ 13 C of plants and topsoils from semi-arid and arid Mongolia. *Biogeosciences*, *17*(3), 567-580.