MA678 Final Project Proposal

Personal Statement:

If it is possible, I hope to find a job that can relate to geography and statistics. My research project is to derive the probability of fire from various factors. I think this project has some connection with my career goals. (But I have no direction at all for this career goal, so I guess it may still be difficult to take this direction in the future...)

Question:

1. How to organize and prepare data, and how to establish connections between each database?

2. What factors are related to the duration of fire?

3. What factors are related to fire severity (fire size)?

4. What factors are related to the number of fires?

The data scores:

I want to use 4 datasets to do the project:

1. condition: From [Predict Droughts using Weather & Soil Data (kaggle.com)](https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data)’s test\_timeseries.csv, which includes 21 columns, from 2012-2020. It can connect to other datasets by using the key of FIPS(Federal Information Processing System Codes). Also, it contains information on precipitation, humidity, temperature, etc.=

2. drought: From U.S Drought Monitor [Current Map | U.S. Drought Monitor (unl.edu)](https://droughtmonitor.unl.edu/CurrentMap.aspx). This dataset has different drought areas in different counties, from 2012-2020. It also has FIPS Codes to connect other datasets.

3. population: From [Index of /programs-surveys/popest/datasets/2010-2020/counties/totals (census.gov)](https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/counties/totals/). It has the population in each county from 2010-2020. This dataset only has a county name column to connect to other datasets.

4. fire: From [2.3 Million US Wildfires (1992-2020) 6th Edition (kaggle.com)](https://www.kaggle.com/datasets/behroozsohrabi/us-wildfire-records-6th-edition?select=data.csv)’s data.csv. It has 39 columns, from 1992-2020, I will only use 2012-2020. I will use DISCOVERY\_DATE, DISCOVERY\_TIME, CONT\_DATE, CONT\_TIME to calculate the average duration of the fire in each year, use FIRE\_SIZE to calculate the average fire size in each year, and use FIPS\_CODE to connect to other datasets.

Proposed Timeline of Work:

Data Processing: to Nov 15.

EDA: from Nov 15 to Nov 20.

Modeling and Validation: Nov 20 to Nov 25.

Write up: Nov 25 to Dec 7.

Excepted Outcome:

Since many fires are caused directly or indirectly by factors, when the population of a county increases, its number of fires should be more. In addition, geographical factors including precipitation and land conditions will also affect the number and duration of fires to a certain extent. The ultimate hope is to build a model to describe the results. If the data of this project cannot establish a relationship with fire, then it can be considered that some factors have not been included in the model, such as the level of economic and social development, infrastructure (mainly fire protection), etc.