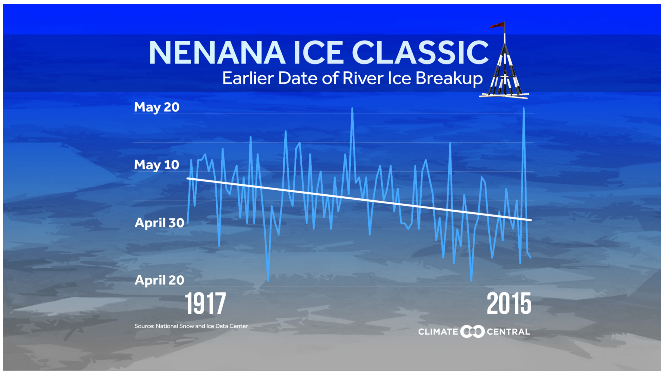
Project Proposal:

Predicting the Nenana Ice Classic


NENANA ICE CLASSIC
Earlier Date of River Ice Breakup
May 20
May 10
April 30
April 20
1917
2015
Source: National Snow and Ice Data Center
CLIMATE CENTRAL

Authors: Julian Benitez Mages, Aidan Johansson, Michael Lembck, Garrin Shieh

DS5110

Prof. Bemis

Fall 2021

Summary:

Every year since 1917, the Nenana Ice Classic has taken place on the Tanana River in Alaska, where bettors place bets on the exact date and time that the river ice will break. In the 2021 contest, the jackpot was split amongst 12 people who guessed correctly, with each taking home $19,465.92. We would like to build a model that predicts when the ice will break using over 100 years of weather data from the site of the contest. The deadline to send tickets to the contest is April 5 of each year, so we would like to get as accurate a prediction as possible from data up to that date. Ideally, the true test for our model will be to input this winter's data and use it to send in our own predictions for the ice classic this spring. Our plan is to create a regression model to predict the breakup time using a dataset that contains many years' worth of weather data, including min and max temperatures, snowfall and snow depth, as well as wind speed and direction.

Proposed Plan:

The first major process will be tidying and organizing the dataset to build the model. We would like to have each row of the final dataframe to represent one year, with the breakup time and all the possible predictor values for that year. We would like predictor values to consist of weather data such as average temperatures and snow depth, aggregated across months, weeks, or the entire winter. Another predictor we can use is the ice thickness, which is recorded on certain days throughout the winter and gives a clear idea as to when the ice may break.

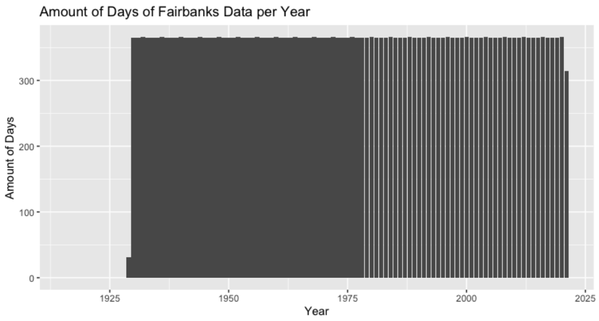
Once we have our dataframe in the proper configuration, we will start off by analyzing the effectiveness of various predictors for the breakup time, and use these to create an accurate model. We will use processes such as k-fold cross-validation and stepwise regression to create the most accurate model. We will assess a potential need to compute logarithmic transformations on the predictors as well as the predicted value, as well.

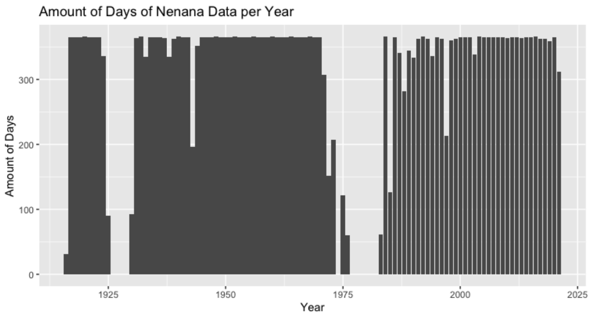
There will be many potential challenges and hurdles for this project. One main issue is that our datasets don't have complete data for many predictors, as the weather stations weren't able to measure such statistics 100 years ago. We will either have to only focus on years where all the predictors are present or use all the years' worth of data in which not all predictors are indicated. For example, data on the minimum and maximum temperature are consistent since 1917, but some of the values such as snowfall and wind speed have only been around for the Nenana Weather Station since the 1980s. This also applies to the ice thickness metric, which only started being reported by the contest organizers in 1989.

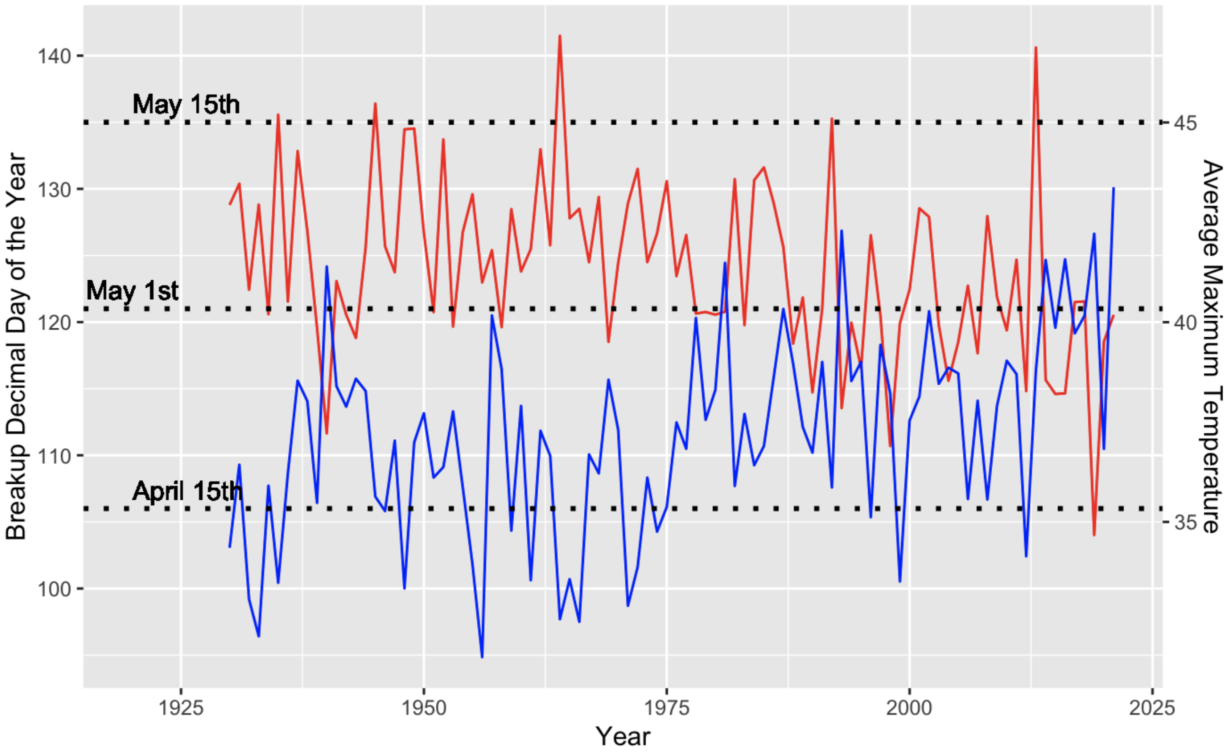
Another method we found to fix this issue of missing data is to use another weather station located in Fairbanks Alaska to supplement our missing data. It is the next best option as it is a large neighboring city only 55 miles away (quite close for Alaskan standards), and has much more complete historical data due to it being a much larger town. In total, we have two datasets, one from Nenana and one from Fairbanks which we will combine to ensure we get the most complete data.

Preliminary Results:

The first two grey bar graphs show each year's number of days with complete data in each prospective dataset. Each of these datasets has some holes in them due to extreme weather conditions so we will combine them to create the most accurate model possible. The last graph shows the day that the ice on the river broke each year in addition to the average maximum temperature of that year. It seems that there is generally a negative correlation between the day of the year and the average maximum temperature. If the average temperature is colder, then the ice will break later. This correlation makes logical sense as the ice melts when the temperature rises. Obviously, further into our project, we will break down various predictors to create a more accurate model but it is nice to see that there is a correlation at the start.


Amount of Days of Fairbanks Data per Year


Amount of Days of Nenana Data per Year



References:

https://www.nenanaakiceclassic.com/
https://nsidc.org/data/nsidc-0064/versions/2