CSCI 4622 Project Milestone 3

Alex Ojemann

Andrew Floyd

What were the key pieces from your Pitch Feedback?

Stationarity: A requirement for the ARIMA model is that the data is stationary, meaning that the mean is not changing over time. In our case, we expect the data to fluctuate within each year because SWE is much higher during certain seasons than others, but we don't expect it to vary much across years.

Adding other models: LSTM (Long Short Term Memory Neural Network) and Prophet were mentioned as additional models that could be explored for our time-series data and Random forests and Gradient Boosting for our regression data.

Model Evaluation: We are considering mean-squared error (MSE) as our metric to evaluate the performance of our model and to control for the signs of errors. To cross validate the time-series model performance we can cross-validate on subsets of the data, ensuring that the test data is always the most proximal in time.

Type of time-series forecasting: Our reviewer wondered which type of time-series forecasting we will use. We will use seasonal time-series forecasting because our data is expected to greatly vary by season given our knowledge of SWE.

Dataset size: Our reviewer asked if our sample of data, 10 years of data at each station, would be a sufficient sample size to avoid overfitting. We think it is a good place to start as it allows us to keep the total number of stations high, and we hope we can sufficiently control for this effect by using cross-validation.

What are the key discoveries you've made so far in your month of progress?

'NA' values are prevalent in the time-series data even after the sites where start and end date requirements were not met have been excluded. We will filter these sites out because these values will disrupt seasonally-dependent time-series projections.

The data for the planned linear regression has been harder to come by than anticipated. SNOTEL provides monthly and semi-monthly measurements of snow depth, snow water equivalent, and snow density, and other data such as temperature, humidity, precipitation, etc can likely be gotten but it would require us to link the two datasets and I don't know that we can easily do that. So, we may have to just focus our efforts on the time-series analysis since that data is more forthcoming and has greater granularity.

What are the updates you're making to your project?

- We will filter out sites where the maximum yearly SWE mean is 20% higher than the minimum yearly mean SWE to ensure relative stationarity.
- We will add the additional models specified as stretch goals.
- We will use RMSE to evaluate the relative strength of our models and use rolling window cross validation to validate our time-series models.
- If we are under the impression that the model is significantly overfitting the data, we may try creating time-series models using multiple sites that are close together geographically rather than one for each site.

Confidential Feedback

- **Any issues to report?**

  None