Executive Summary: Today's Texas Might be Tomorrow's Ohio: Building a Geographic Climate Change Predictor

Group Members: Alison Duck, Jack Neustadt, David Pochik, and Tawny Sit

Introduction:

From the dawn of industrialization to today, the average global temperature has shifted upward by ~2.7 degrees Fahrenheit (~1.5 degrees Celsius) due to increased greenhouse gas emissions. The effects of this temperature increase have led to, among many other complications, more extreme weather patterns, increased sea levels, and hotter days. If emissions are left unchecked and temperatures continue to rise at their current (or projected) rate, then this will lead to drastic shifts in regional climate. For example, today's annual average temperature in Ohio will increase to that of today's annual temperature in Texas in Y years. This project explores and analyzes geographical climate change data in the contiguous United States from 1950 to the current year. The objective is to predict regional features, e.g., temperature, precipitation, or snowfall, for a given year based on historical data, i.e., if I want to live in an area Y years from now that has roughly the same temperature or climate as region X today, where would I go?

Dataset (Acquisition, Cleaning, and Preprocessing):

We acquired weather station data for the United States from the National Oceanic and Atmospheric Administration's (NOAA) online climate database. For model features, we focused on minimum/maximum/mean temperature, snowfall, precipitation, and heating degree days. Due to increased weather station infrastructure in 1950, we limited our timeline with a lower bound at this year and an upper bound at 2024, thus providing us with ~1 million data points. We eliminated NaN entries from the data and created two geographical grouping routines (spatial cells of either (1) latitude/longitude structured grids or (2) K-Means cluster grids) to reduce computational overhead (convert ~1 million points to ~5000 points by averaging model features within each cell at each year) while maintaining descriptive climate properties.

We acquired climate zone maps from the USA Department of Energy Building America (BA) Climate Zone boundaries. These data are stored in geojson, which we import using the Geopandas package. The individual BA climate zones are given as polygon type objects – the classification of each weather station is then done by checking the longitude and latitude of each station and seeing which polygon object/climate zone it belongs to. There are two main caveats to this matching: (1) the climate zone boundaries are county-based and depend on data from a certain range of years, and thus a weather station in a given year and a given location may be classified as one climate zone even though its corresponding climate data may better match another climate zone type; and (2) the climate zone classifications are done using seasonal data that is not included in the NOAA weather station data. We attempt to account for these issues in our feature selection stage while training the classifier.

Model Selection and Results

We created k-fold cross-validation time-series models with a 94-6 train-test split for the climate data using (1) simple linear regression (SLR) with the structured latitude/longitude grid approach, (2) SLR, and (3) quadratic regression (depending on RMS error at final cross-validation step) with the K-Means clustering approach. The latitude/longitude grid approach attained RMS errors with respect to the final validation step on the order of 10-50% of the training set mean, where the snowfall data produced the highest errors because some regions had little to no snowfall. The K-Means clusters used the final validation set to select quadratic vs. linear models

cluster-by-cluster and had median RMSE values over all clusters of 0.6-0.7°C for all temperature features (~4-11% of the test set median over all clusters), 186 mm for snow (~41%), 129 mm for precipitation (~15%), and 186 for heating degree days (~7%).

We then designed a classification model that predicts the future geographical shift in climate zones based on the regression model data. We tested this classifier with four schemes (logistic regression, linear discriminant analysis, k-Nearest Neighbors, and Random Forest) using 10 k-fold validations, where we found that Random Forest provided the best inferential results with an accuracy of ~90%, F1 score of ~90%, precision of ~90%, and the smallest log-loss compared to other models. When we settled on the Random Forest classifier as our model and performed analysis on the test set, we found that our performance metrics were effectively the same as those in the validation sets, which indicated that the model was not overfitted and was generalizable.

We applied our climate change trends from our time-series analysis to weather station data to predict the expected climate in the year 2050, then we re-ran the classifier to see if these weather stations would be labeled a different classification after accounting for climate change. Our results are what we expected - many regions will become warmer, especially around the border regions of each climate zone. We note that some of the re-classifications are likely classifier errors. For instance, the confusion matrix of the test set classification showed that the classifier incorrectly labeled some “Marine” climates as “Cold,” thus we knew the classifier was likely incorrect when it indicated that a Marine climate zone would become Cold after accounting for climate change.

Future Steps and Potential Improvements.

The latitude/longitude structured grid approach is highly parameterized (selection of grid refinement, k-fold cross validation inputs, and minimum weather stations points per grid all play a role in determining model accuracy), thus it may benefit from parameter optimization. We also had to make certain assumptions in our classifier analysis – especially in propagating our climate change trends from the gridded and clustered time-series analysis onto individual weather stations – that we could have probed further to see how or if it affected our results. One future extension would be to use our predicted climate changes to build a predictive model for household energy costs.