Final Project Data Memo

Callum Weinberg

January 19, 2022

Overview of Dataset

For our final project, we plan on using the Global Environmental Indicators dataset provided as a dataset on Kaggle. The dataset can be accessed here: https://www.kaggle.com/ruchi798/global-environmental-indicators [1]

The dataset was put together by UNSD (United Nations Statistics Division) Global Environmental Statistics. The Kaggle auther is Ruchi Bhatia. The dataset was compiled by surveying nations' "stastistical offices and/or ministries of environment (or equivalent institutions) in response to the biennial UNSD/UNEP Questionnaire on Environment Statistics".

The dataset is a compilation of datasets, which correspond to different environmental indicators at national levels. The main categories are Air and Climate, Biodiversity, Energy and Minerals, Forests, Governance, Inland Water Resources, Land and Angriculture, Marine and Coastal Areas, Natural Disasters, and Waste. There are roughly 50 comma separated files between all the categories, each with a corresponding Excel file that operates as the data dictionary. Each file has a number of metrics related to environmental indicators, and the nation in question. The files reviewed so far all have one observation per nation. However this varies a bit depending on the file, as some datasets have portions of sub-areas broken out into their own observations, and in some datasets are missing nations. Therefore most of the datasets have 150-250 observations. Some of the datasets have a country ID, but it is missing in some of the csv files. There will need to be some data cleaning done to merge the files.

We plan on exploring most metrics available as predictors (so long as they make sense in a temporal manner). Therefore there will likely be over 50 predictors initially considered. Some of the csv files have more than one metric worth considering. As mentioned above, there are missing observations for some of the metrics. Some smaller countries are often missing data. We will have to take an approach for which we both exclude predictors with a lot of missing information, but will likely include some predictors with small amounts of missing data, and implement a methodology to handle the missing data.

Overview of Research Questions

Proposed Timeline

Any Questions or Concerns

Given the approach of this project will include considering a relatively large number of predictors, we anticipate a few issues may arise. We will likely need to consider some shrinkage or subset methods for a regression based approach, given n might not be that much larger than p. Additionally, we will likely need to consider some missing-data related strategies other than dropping observations, given the variation in which predictors are missing for which countries (most countries will likely be missing from at least one predictor). We don't have any questions at the moment, but appreciate the instructor support.

[1] Ruchi Bhatia; UNSD Global Environment Statistics. (2021; June). [GLobal Environmental Indicators, Version 1. Retrieved January 21, 2022 from https://www.kaggle.com/ruchi798/global-environmental-indicators.