**EIB 2 - Climate Risk Scores**

**Progress report 3, Oct 22 - Nov 4**

**Progress:**

* Methods of downsampling:

1) Moving Average - Used last year, effective

2) Window -

**Progress report 2, Oct 22 - Nov 4**

**Progress:**

* Data Cleaned and updated to include 2019
  + No web scraping was needed
* Generally, KNN and random forests are the best imputation methods [1]. However, our data is not missing at random.
  + <1% of data is NaN (number even lower in the ~60 CCPI countries)
  + Also, different sources have different meaning of 0, ‘ -’,NaN so we will not apply a blanket imputation method,but rather a mix.
    - Special cases such as Kosovo/Montenegro/Serbia, or Timor-Leste, and lesser scale ex-Soviet Republics, active war zones etc.
* For a small amount of data missing, interpolation is a reasonable imputation method (polynomial interpolation can be very effective)

**Plan for next 2 weeks:**

* Focus on modeling:
  + Different approach from last year, no weighted moving averages

**Questions:**

* Need to clear out data files. Are the ones mentioned in project description pdf the only important ones? (Rents, Emissions, Renewable Energy, and Energy Efficiency)
* Is Energy Efficiency and Energy Intensity the same?
  + For Renewables we used World Bank instead of EIA data
  + Last year, why did they use Coal/petr rent and consumption

- Why are energy rents and consumption both included as data sources in the last years project?

- Last year, used moving average to condense all the years into one? Any particular reason, it seems that this loses a lot of data. (there is no point to impute using this way)

**Stochastic interpolation, brownian bridge sigma 0**

**Github link:** https://github.com/FabioHodo/EIB2\_ClimateRisk

**Progress report 1, Oct 7 - Oct 21**

**Task for next meeting:**

* Literature review of autoimpute models (no black box models)
* Use clustering by PPP/ capita,regions to impute data for missing countries
* Potentially explore use of unstructured data (i.e. world bank data)

**Progress:**

* Reviewed previous work and code, rough recap below: (note this year 1 more meeting

Meeting 1 - Web scraped last 10 years from eia.gov, collected the rest

Meeting 2 - Finish data collection, report on CCPI methodology

Meeting 3 - Literature review on forecasting time series and imputation

Meeting 4 - Fit LSTM, ARIMA, Varma models. Use brownian bridge for monthly

Meeting 5 - Predict out to 2030, classify 2022 (while missing some variables)

* Found out priorities to work on from past work
* **Gather more data:** Forecast (from 2017 to 2030) using 28 years worth of data (1990-2017).
* **Try more imputation methods (e.g. autoimpute):**
* **Optimizing the LSTM model:** stacked LSTM structure
* **Model hyperparameter tuning:** e.g. using Bayesian optimization
* **Control overfitting:** PCA done, add lasso and ridge regularization.
* Identified three main parts of the project

1. **Imputing missing data**

Improvement: autoimpute & using countries with similar traits

1. **Prediction of future data**

Problem: prediction data cause inaccuracy for classification model

Solve: add more explanatory variables

Improvement: parameter tuning & stacked LSTM

1. **Classification model**

Improvement: parameter tuning & overfitting control

* Explore more explanatory variables
* Created Github Repo

**Plan for next 2 weeks:**

* Learn further about pros and cons of EWMA & KNN compared with autoimpute
* Use autoimpute to impute missing data
* Look into different imputation methods other than EWMA and KNN, impute data on similar countries for example
* Gather more data, potentially from different sources (to be discussed)
* Learn some basic information about LSTM and get some insights to improve it

**Questions:**

* Discuss exact deliverables of the project. Where to focus:
  + Input data (e.g. Feature selection)? 2
  + Data imputation? 3
  + Model Performance ? 1
* Add year 2020 to the project? (some of the databases may be updated)
* Look into different data sources?
* Why should the data set used in classification model range to 2030, why not up to now?