Data Analysis Report

Machine Learning Project:
Predicting Canada Crops Production using ACI climate data

Jing Xie

I. Introduction

The Actuaries Climate Index (ACI) dataset contains the frequency of extreme weather and extent of sea-level change for twelve different subregions in the United States and Canada (Actuaries Climate Index, 2020). The six components of the ACI are 1) High temperatures; 2) Low temperatures; 3) Heavy rainfall; 4) Drought (measured in consecutive dry days); 5) High wind; and 6) Sea level (Actuaries Climate Index, 2020).

II. Data Description

This Actuaries Climate Index (ACI) dataset contains climate data from 1961 to 2019 for 15 distinct climate regions in North America.

The following is a summary of important attributes in the ACI dataset:

Attribute	**Variable**	**Description**
Sea level	sealevel	measures the sea level relative to the land
Drought	CDD	measures the max number of consecutive days in a year with less than 1mm of daily precipitation
Rain	Rx5Day	measures the maximum 5-day rainfall in the month
Wind	WP90	measures the wind speed
High temp.	T90	measures the change in frequency of warmer temperatures above the 90^{th} percentile
Low temp.	T10	measures that of colder temperatures below the 10^{th} percentile
ACI combined	ACI	Is the combined score of all six attributes

The ACI dataset contains the attributes Sea level, Drought (measured in consecutive dry days), Rain (measured in maximum 5-day rainfall), wind power, frequency of high temperatures

and low temperatures. All these values are standardized to an index, and the ACI combined score is a combination of the index of all six attributes represented by the formula:

$$ACI = mean(T90_{std} - T10_{std} - WP90_{std} + Rx5Day_{std} + CDD_{std})$$

The dataset is in the format of an Excel workbook with 20 sheets of data, where each sheet contains the time series data for one of the above attributes. Each attribute has sheets for monthly, seasonal, and unstandardized, except for combined ACI score which only has monthly and seasonal. Each record contains time series data for a climate region, and columns show time in year and month or quarter.

III. Exploratory data analysis

The preprocessing for the ACI dataset was extensive given the original format of the dataset. A significant part of this involved transposition and grouping the climate region data for the numerous sheets into a single cleaned dataframe. Dates were converted to the datetime format. The resulting cleaned data frame consists of 10,606 records and 10 attributes: combined, date, dry, hightemp, lowtemp, rain, region, sealevel, wind, and coastal. Coastal is an indicator for whether the climate region has a significant portion of its boundary at an oceanic coastline. Complete python code for the preprocessing and exploration can be found in **Appendix A1 and Appendix A2**.

Comparing climate attributes

To get an idea of the dataset, nine graphs were plotted using different combinations of variables and climate regions to identify patterns and trends.

Standardized change in frequency of temperature was plotted against time for warm temperatures [T90] and cool temperatures [T10] and shown in **Appendix B Figure 1**. This graph indicates that, starting from 1980, the warm temp change goes above zero and continues an upward trend as time progresses. T90 and T10 are almost perfectly mirrored about the line y = 0.

Standardized change in frequency of wet [Rx5day] and dry [CDD] days was plotted against time in **Appendix B Figure 2**. This graph indicate a negative correlation between the two, although the fluctuation in magnitude is much greater for the dry line.

Standardized change in sea level [sealevel] and warm temperatures [T90] was plotted against time in **Appendix B Figure 3**. Sea levels exhibit an increasing trend very similar to that of warm temperatures.

Visualizing data by country: USA vs. Canada

Standardized change in frequency of temperatures by country are plotted against time in **Appendix B Figure 4**. USA has slightly but consistently higher frequency of warm temperatures from the late 1990s onwards in comparison with Canada.

Standardized change in sea level by country are plotted against time in **Appendix B Figure 5**. USA has higher sea level change from 1985 onwards in comparison with Canada.

Standardized change in combined ACI score by country are plotted against time in **Appendix B Figure 6**. USA has slightly but consistently higher frequency of warm temperatures from the late 1990s onwards in comparison with Canada.

Visualizing data by coastal indicator

Standardized change in frequency of warm temperatures [T90] by coastal indicator were plotted against time in **Appendix B Figure 7**. Coastal regions tend to have greater frequency of warm temperatures since the early 1990s.

Appendix B Figure 8 shows the standardized change in frequency of sea level by coastal indicator vs. time. Coastal regions exhibit higher sea level change since 1980, and the difference has been increasing consistently since then.

IV. Research questions and analysis

1. Canadian Crop

Research Question

The second research question is: is there a predicting model that will achieve the best performance of predicting wheat production using the ACI climate data over the years?

Research Motivation

Canada is one of the top crops exporting countries globally, and Wheat is the most important staple crop grown in Canada. However, in a changing climate, the increased frequency and severity of adverse weather events, which are often localised, are considered a major threat to wheat production. Therefore, a timely and reliable wheat production prediction analysis is important for regional and global food security.

Data Description

Research questions 2 introduced a new dataset of Area, Yield, and Production Report for Canadian principal field crops, including grains, oilseeds, and some pulse and special crops. The

dataset uses the AIMIS for area, yield and production (AYP) data by Canada, Western Canada, Eastern Canada or by province for principal field crops in Canada during 1908 to 2019.

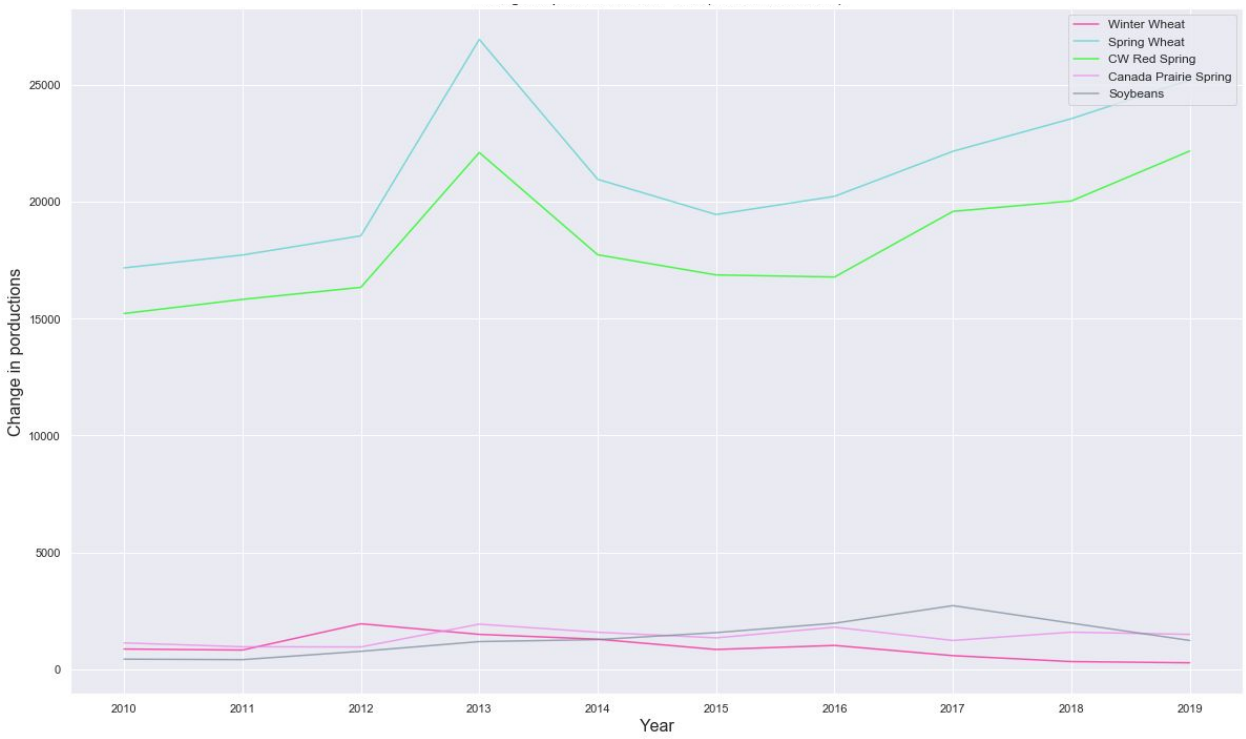
Some key variables from the dataset are listed as below:

Variable	Description
Seeded Area (000 ha)	Measures the seeded area (expressed in 1 000 hectares)
Harvested Area (000 ha)	Measures the harvested area of crops (expressed in 1 000 hectares)
Yield(t/ha)	Measures the yield of crops (expressed in tonnes per hectare)
Production (000 t)	Measures the production of crops (expressed in 1 000 tonnes)

Exploratory Results

Given that the Northern Plains region(Alberta, Saskatchewan, and Manitoba) dominated the cultivation of crops in Canada, we studied the data of the Northern Plains regions from 2010 to 2019 for a preliminary exploration. Interesting phenomena includes:

Different categories of crops are responding differently to certain climate indexes. In the area of Western Canada, spring wheat is mainly cultivated in the Northern Plains region(Alberta, Saskatchewan, and Manitoba). Seeing a mild descending of temperature in the Northern Plains throughout the decade(2010-2019), the production of spring wheat and CW Red Spring(wheat) reached a peak in 2013, while the production of Canada Prairie Spring(wheat) and Winter Wheat remained steady.



Change in Productions (Northern Plains Region 2010-2019)

The production of crops is not responding to a certain type of climate index radically, showing a result of a comprehensive effect of the climate indexes, while some climate indexes showed a greater effect on the production of crops.

Throughout the past decade in the Northern Plains region, the index of Consecutive Dry Days climbed to a peak in around 2013 and quickly descended to trough, as a reverse s-shape curve. The index of high-temperature frequency saw a mild increase, while the indexes of heavy rains and low-temperature frequency remained steady. As a result, the production of wheat didn't respond to a certain climate index directly but saw a mild increase after 2015 with a low index in consecutive dry days and steady indexes of temperatures and rainfalls.

Change in Climate index(Northern Plains Region 2010-2019)

Techniques

The analysis first conducted data cleaning and preprocessing via numpy dataframe to merge two datasets together and visualized the data using scatterplot to have an overview. In the stage of feature selection, correlations between variables were calculated, and visualized using heatmap to understand the features.

To experiment over machine learning models, we first used the Standard Regression model to check the performance of prediction, as the simplest form of regression for predictive modeling.

Given the possibility of overfitting caused by large coefficients of features in linear regression, we then consider applying regularization algorithms to penalize large coefficients. We conducted the Ridge Regression to diminish the complexity of the model by reducing the magnitude of coefficients without removing any variables. The Lasso Regression model was also used to check the performance of prediction. The model reduces the magnitude of coefficients as well as selects features among the highly correlated ones with reducing the coefficients of the rest to zero while the 10-fold cross-validation was also performed here to choose the best alpha, refit the model, and compute the associated test error.

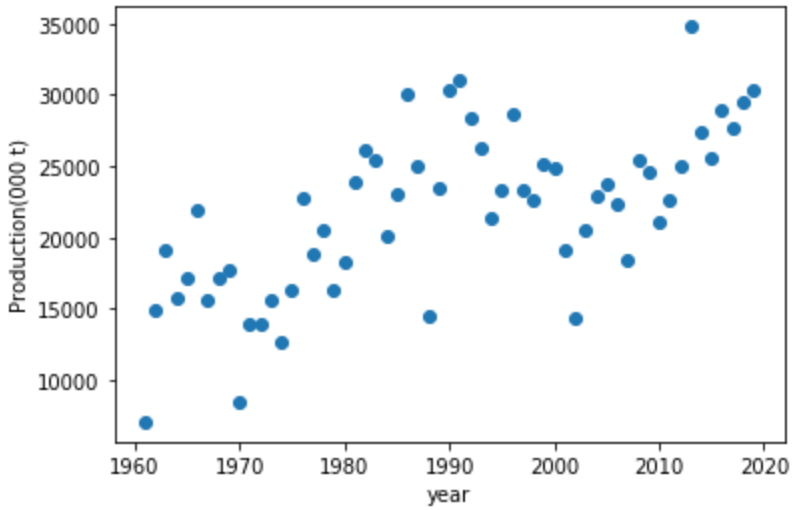
Data Analysis

- Processing:

We converted the [date] in the ACI dataset to [year] by applying an average on climate index to keep the two dataset consistent. And then we used the variable, [year], as a Foreign key, to concatenate the two dataset together as below.

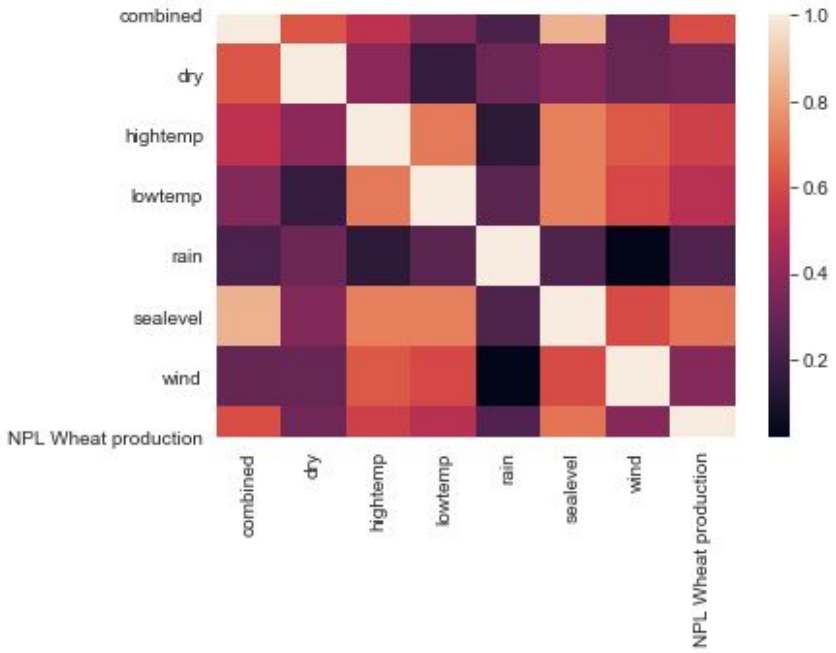
	year	combined	dry	hightemp	lowtemp	rain	sealevel	wind	NPL Wheat production
0	1961	0.211667	-0.188333	-0.095000	-0.028333	-0.211667	1.201667	0.535833	7077.0
1	1962	0.241667	0.006667	-0.080833	-0.004167	-0.154167	1.133333	0.537500	14860.0
2	1963	0.235000	0.049167	-0.043333	-0.030833	-0.115833	1.098333	0.396667	19132.0
3	1964	0.205833	0.030000	0.071667	-0.128333	-0.130833	1.044167	0.085000	15730.0
4	1965	0.155833	0.046667	0.023333	-0.085000	-0.078333	1.002500	-0.145833	17200.0

Using the visualization of scatterplot, we're able to have a glance of the data. From the plot, we can see that there's an escalating trend of wheat production in the NPL region over the years, while there are some outliers in the data. Given the limited data points, only 59 data points from 1961 to 2020 after the inner join, we decided not to consider removing the outliers.



- Features selection:

In the first step, the correlation between variables, including [hightemp], [lowtemp], [rain], [wind], [dry], [sealevel], [combined] and [NPL wheat production], is calculated and we also drew a heat map to show the correlations. None of the attributes demonstrate a strong correlation and the highest number was around 0.7, lower than 0.8. Therefore all the variables are kept.



- StandardRegression Model:

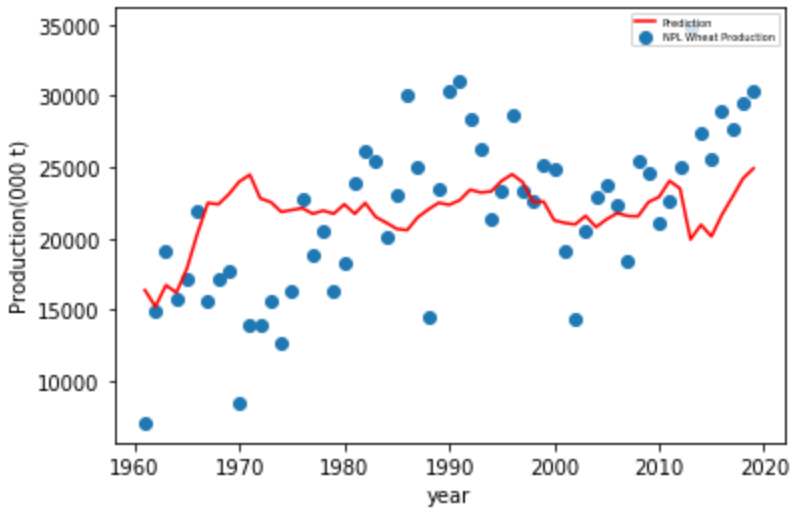
First, we conducted a Standard Regression using all the features. Below is the result of estimated coefficients.

year	1.41161783e+01
combined	-6.38790257e+04
dry	1.22929443e+04
hightemp	2.11561649e+04
lowtemp	-6.73902075e+03
rain	2.80985887e+04
sealevel	2.03173081e+03
wind	9.29318444e+03

Our Standard Regression model has an r-square of 0.5854 when applied on the dataset, which represents that around 58.54% of the data are fitted to the regression line. However, the result of MSE and RMSE are relatively high and we attribute it to that there are many outliers, especially the one for 1961 and around the 1990s. What's more, the predictions around the 1970s

unexpectedly went toward the opposite direction of the data. Thus, we would not deem the r-square score to be solid proof for the goodness-of-fit for our model.

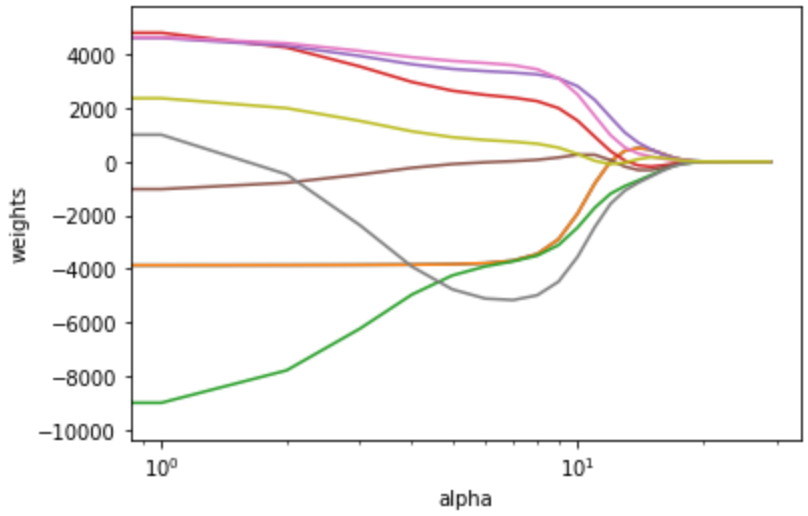
R Square	0.5853969464008362
MSE	30155462.432886325
RMSE	5491.398950439344



- <u>Ridge Regression Model:</u>

Next, we conducted the Ridge Regression to explore the data and modify the model, using the concept of regularization to reduce the magnitude of coefficients without removing any of the features.

In order to determine the alpha, we plot the relationship between alpha and the weights, a line for each feature. The plot shows that as we increased the value of alpha, coefficients were approaching zero.



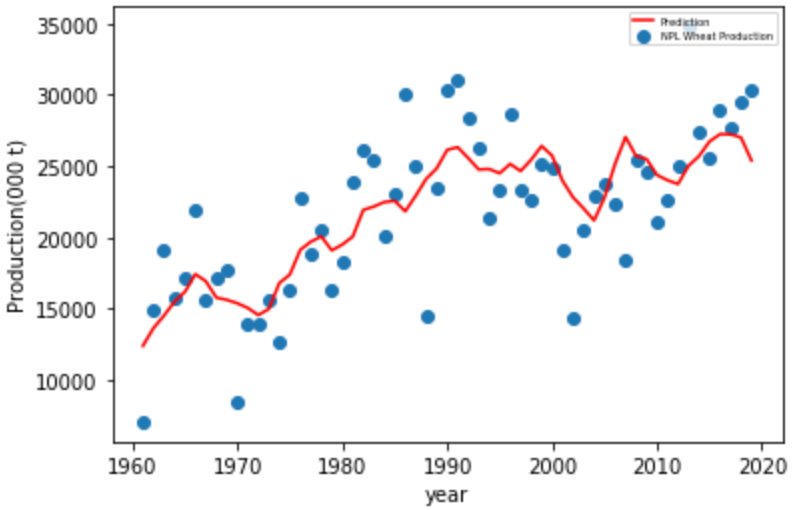
Using an alpha value of 0.01, the Ridge Regression was conducted. From this table of estimated coefficient scores, we can see that the Ridge Regression shrinks the parameters to varying degrees. Thus, we expected it to prevent the multicollinearity of the model.

year	13.38475221
combined	1508.18856446
dry	653.82137328
hightemp	7367.83093148
lowtemp	2064.42929443
rain	12866.18514019
sealevel	-8031.92187452
wind	-1684.30363055

The result for our Ridge Regression model has the same r-square of 0.5854 with the Standard Regression model when applied on the dataset, which represents that around 58.54% of the data are fitted to the regression line. Similarly, the results of MSE and RMSE are relatively high, although there is a slight improvement upon the Standard Regression. Thus, we would not deem this result to be solid proof for the goodness-of-fit for our model.

R Square	0.5853898187279886
MSE	13694356.45190883
RMSE	3700.5886628898425

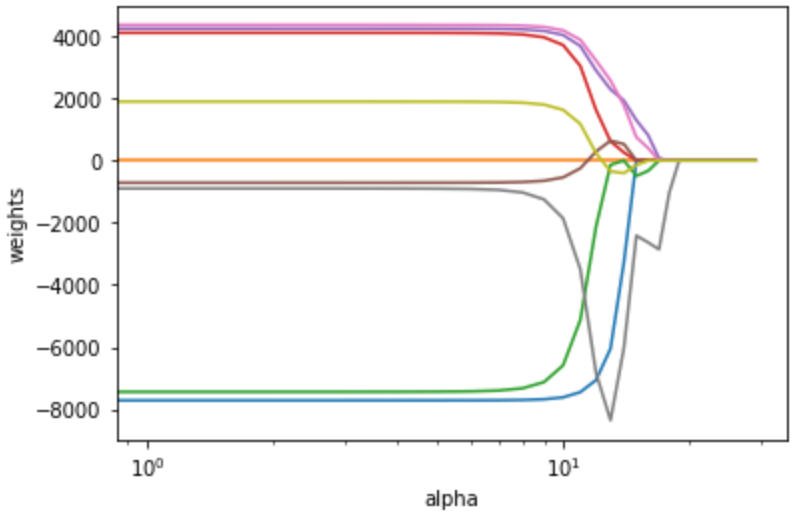
The performance of our model is then double-checked by this visualization showing the difference between the predicted values and the target values in the dataset. Despite the outliers, the graph shows that the model has a good performance in predicting our target, NPL Wheat production, and we can see that it performs better in fitting to the data in comparison with the previous plot of Standard Regression. However, it's notable that the trend is predicted to descend after around 2018, which is a complete reversal of the result of Standard Regression as well as the data. Thus, we considered it an over-fitting model regardless of the R square score.



- Lasso Regression Model(Cross Validated)

Next, we conducted the Lasso Regression(Cross Validated) to explore the data and modify the model, using the concept of regularization to reduce the magnitude of coefficients as well as to select features among the highly correlated ones with reducing the coefficients of the rest to zero.

In order to determine the alpha, again, we plot the relationship between alpha and the weights, a line for each feature. The plot shows that as we increased the value of alpha, coefficients were approaching zero. Plus, some coefficients are reduced to absolute zeroes even at smaller alpha values. Therefore, lasso selects only some features while reducing the coefficients of others to zero, which is known as feature selection and which is absent in case of Ridge Regression.



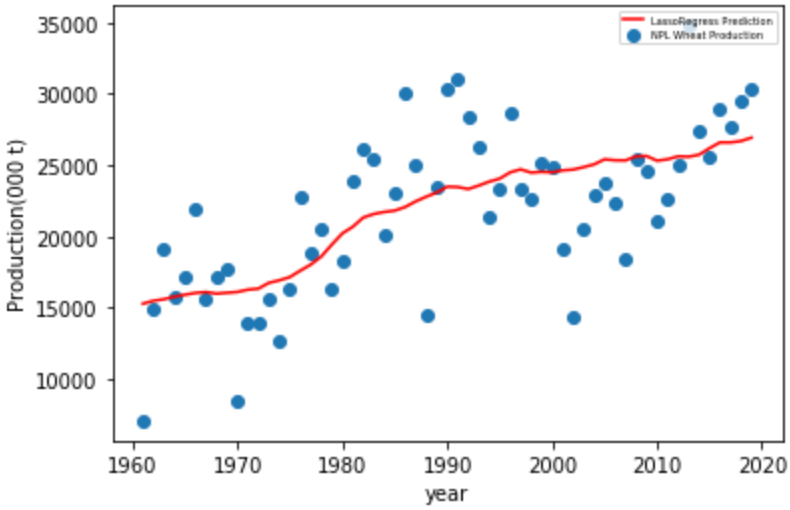
We then performed 10-fold cross-validation to choose the best alpha, refit the model, and compute the associated test error. From the results of the best model coefficients as listed below, we notice that 5 of the 9 coefficient estimates are exactly zero. The Lasso thus has a substantial advantage over ridge regression in that the resulting coefficient estimates are sparse.

year	9.85790556
combined	-0.
dry	-125.83376186
hightemp	-0.
lowtemp	-0.
rain	-0.
sealevel	-3384.54757536
wind	-0.

As we can see, the MSE and RMSE for our model have been increased but the value of R-square has been decreased, compared with the results of Standard Regression and Ridge Regression.

R Square	0.4931041923478293
MSE	16558150.669237135
RMSE	4069.170759409973

The performance of our model is then double-checked by this visualization showing the difference between the predicted values and the target values in the dataset. Although the accuracy has been slightly diminished, the model succeeded in avoiding overfitting the data, without being affected by the outliers in general, showing an ascending trend of wheat production over the years.



Therefore, we can conclude that the Lasso Regression is the best model on performing the prediction of wheat production in the NPL region using the ACI climate index data among these three models. With a slightly low score of R square, 49.31% of the data are fitted to the regression line, without overfitting to the data, especially the outliers in the dataset, while given limited data points.

The corresponding code for this section can be found in **Appendix D**.

Limitations

The are several limitations of this research are discussed in this section:

- Dataset

  The ACI dataset contains climate data that are collected based by month while the Canada Crops data are collected by year. Plus, in terms of the dimension of geography, the Canada Crops contains production data that are collected based on regions while the ACI dataset climate data are combined into regions. When merging the two datasets for this research, the accuracies perceived at both the monthly level of climate index and the province level of production are diminished to some extent.

  Moreover, given that the Northern Plains region(Alberta, Saskatchewan, and Manitoba) dominated the cultivation of crops in Canada, it's the only region with valuable data for this study, as another reason resulting in the limited size of the dataset.

  Thus, a future direction for this research can be using a crop production dataset of the United States, where crops are cultivated geographically more widely, so as to improve the accuracy with more data at the geographic level.

- Method

  Although the Lasso Regression proved to be the best model among the three models in this study, there's still limitation of this method: it retains only one variable among correlated ones and reduces the coefficients of the rest to zero, which may lead to a loss of information, resulting in lower accuracy of the model, for example, the index of [sealevel] and [high-temp] are highly correlated with a score of -0.726416 while the latter was removed from the model.

- The complexity of reality

  Due to the complexity of crop cultivation, the model may not be able to perform well without considering some realistic factors. First of all, the climate conditions required for growing crops may vary in different seasons over the year, simply applying average on the monthly indexes will result in a loss of accuracy, which is a possible cause of the outliers. Plus, considering the spans of time in the data(from 1961 to 2020), the rapid development of modern agricultural technology has contributed significantly to the crops yield and production, which is unquantifiable in this study.

  Last but not least, the dramatic climate change throughout the past several decades, leading to a higher frequency of extreme weather, is also an unignorable factor that is worth being considered.

Appendix

Appendix A1: See file aci_cleaning.ipynb
Appendix A2: See file aci_exploration.ipynb
Appendix B: ACI dataset graphs

Figure 1: Standardized change in frequency of temperatures vs. Year

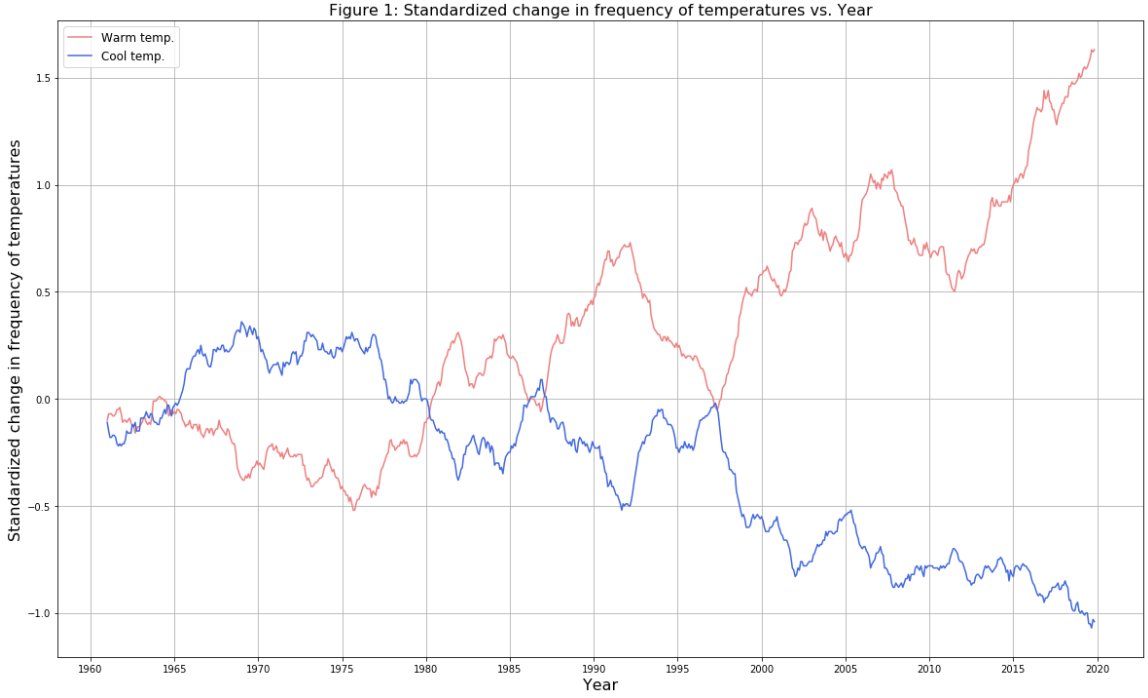


Figure 2: Standardized change in frequency of wet and dry days vs. Year

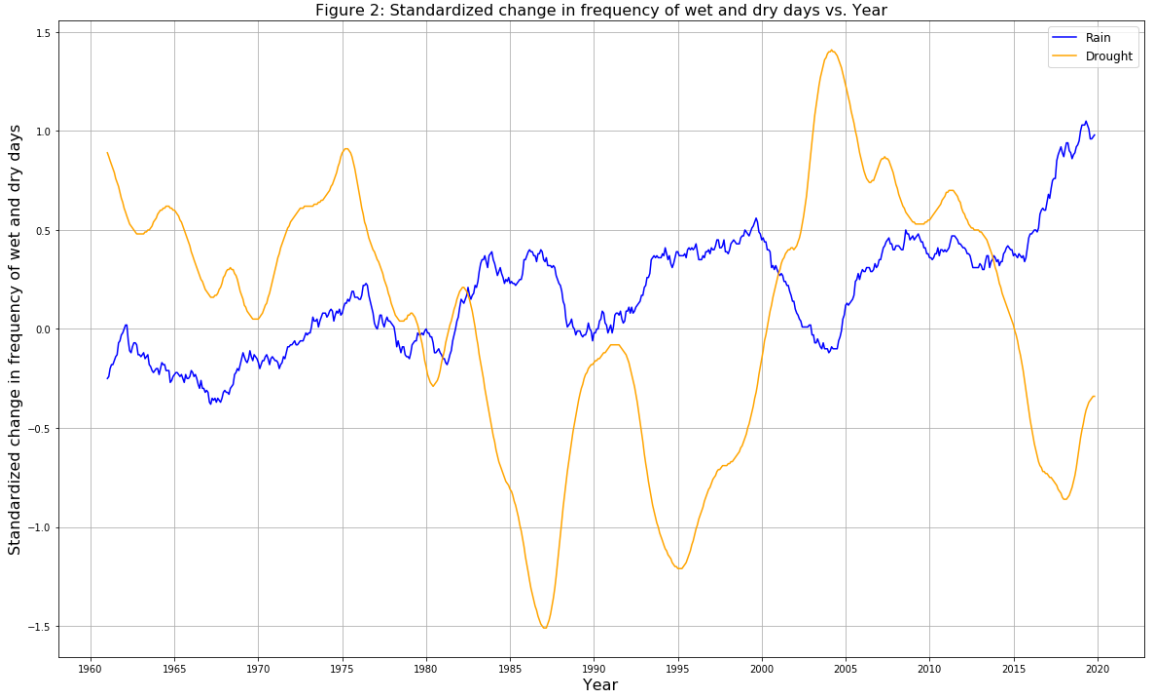


Figure 3: Standardized change in sea level and warm temp. vs. Year

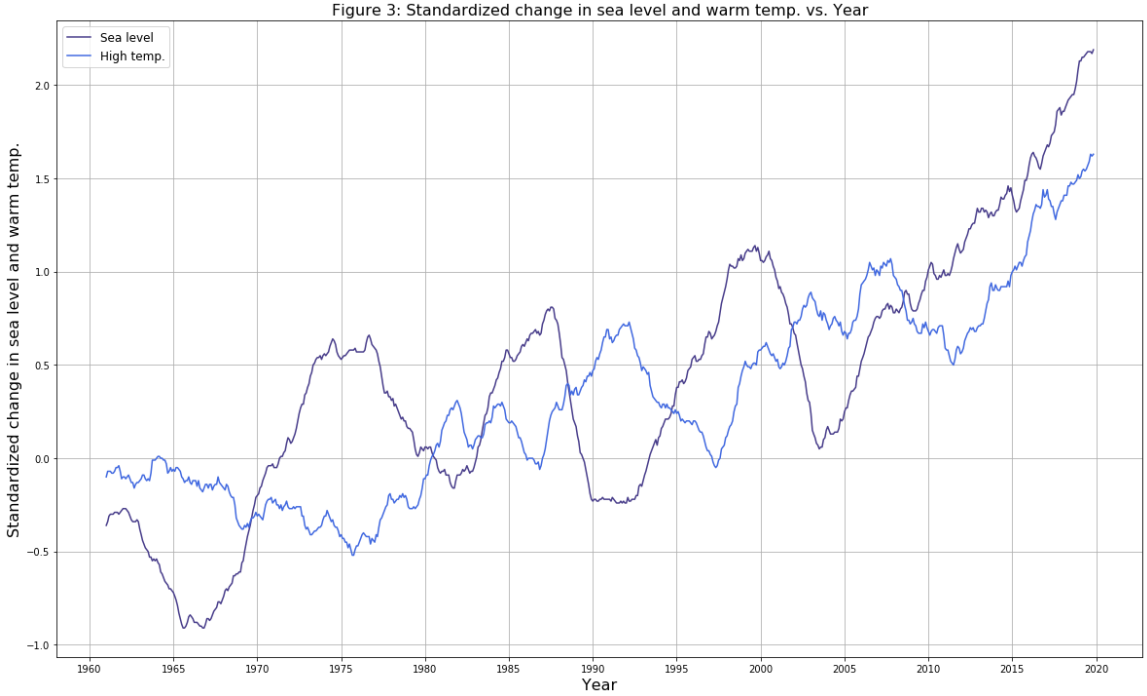


Figure 4: Standardized change in frequency of temperatures by country vs. Year

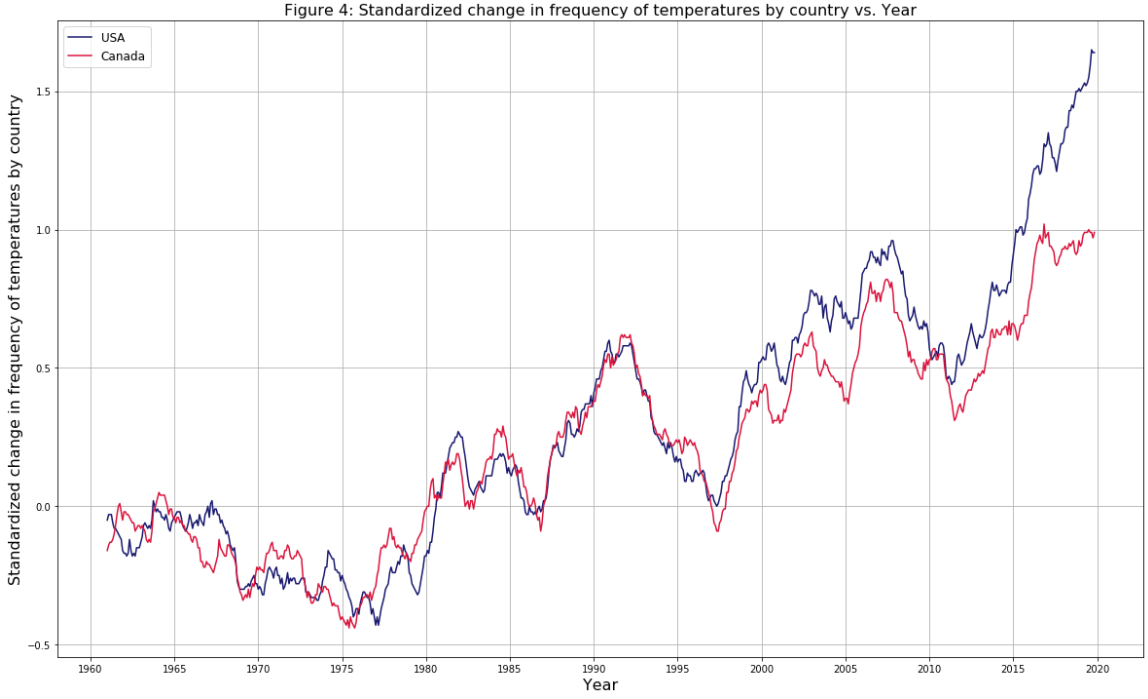


Figure 5: Standardized change in sea level by country vs. Year

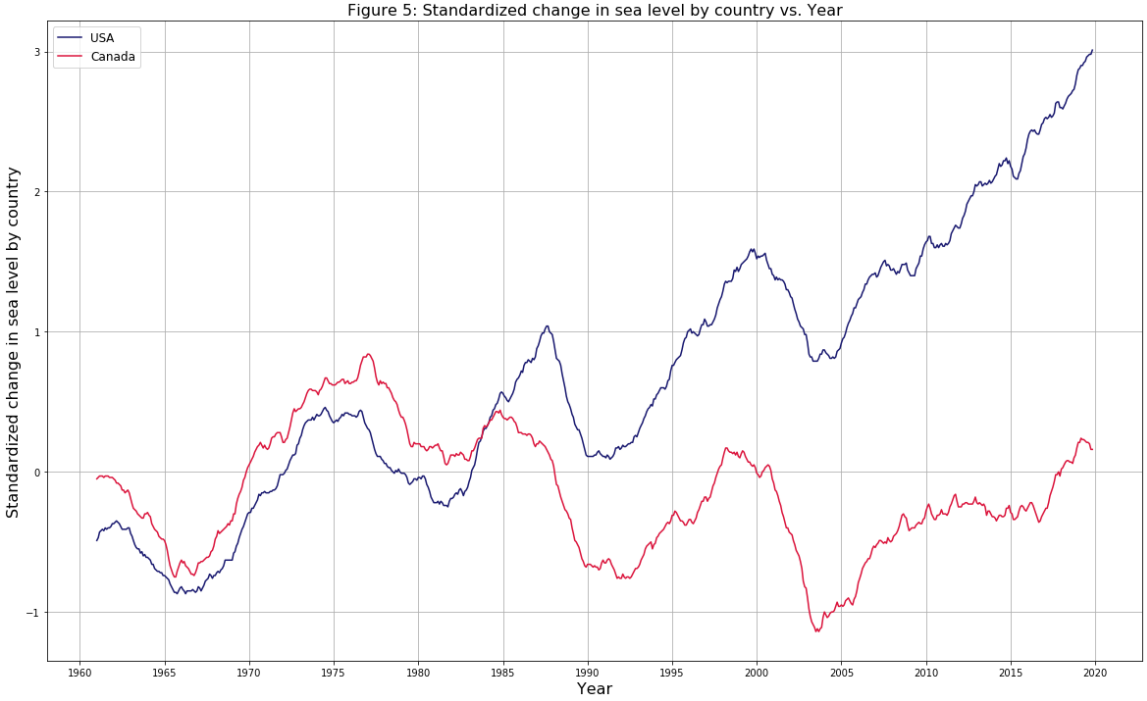


Figure 6: Average standardized ACL combined score by country vs. Year

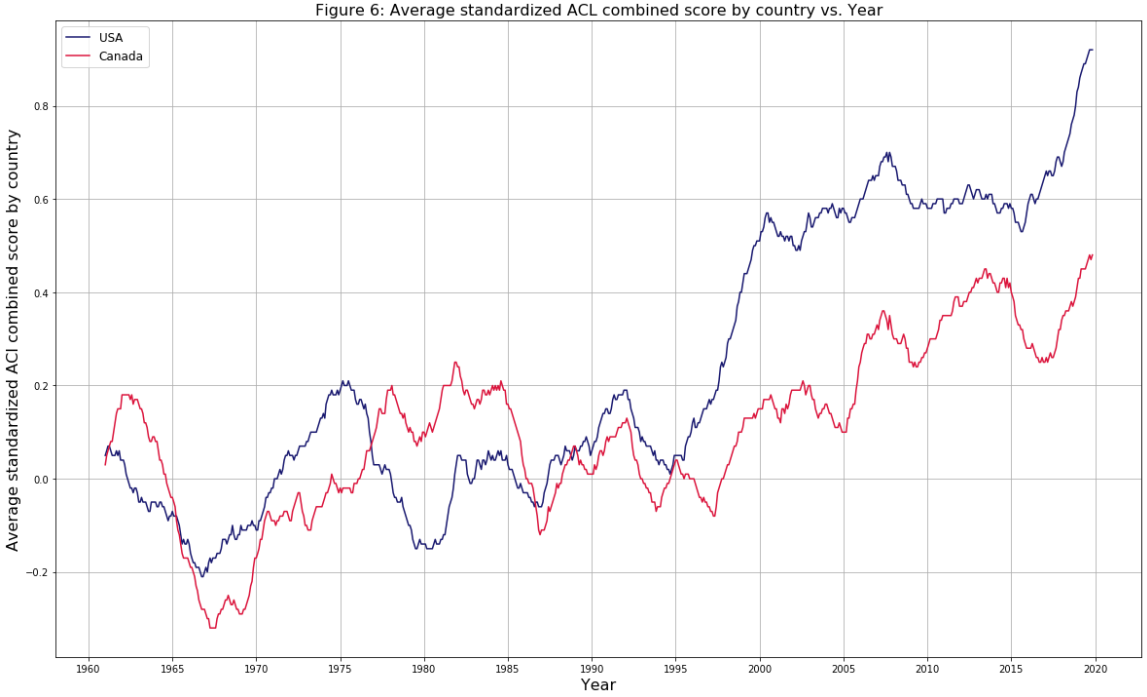


Figure 7: Standardized change in frequency of temperatures by coastal indicator vs. Year

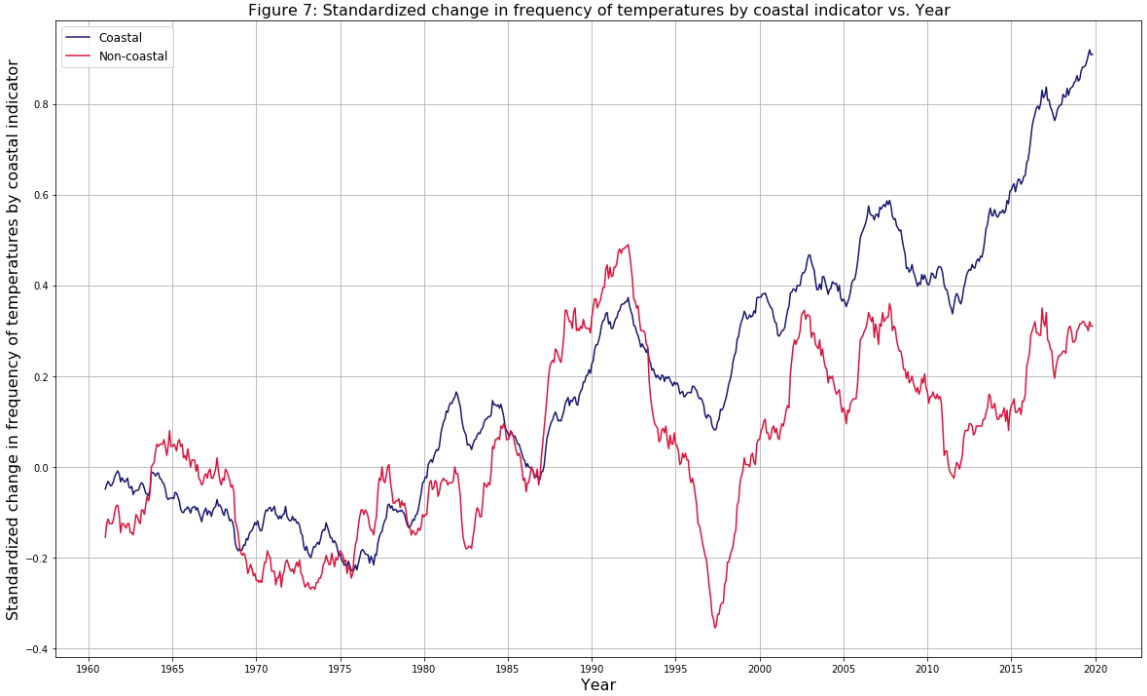
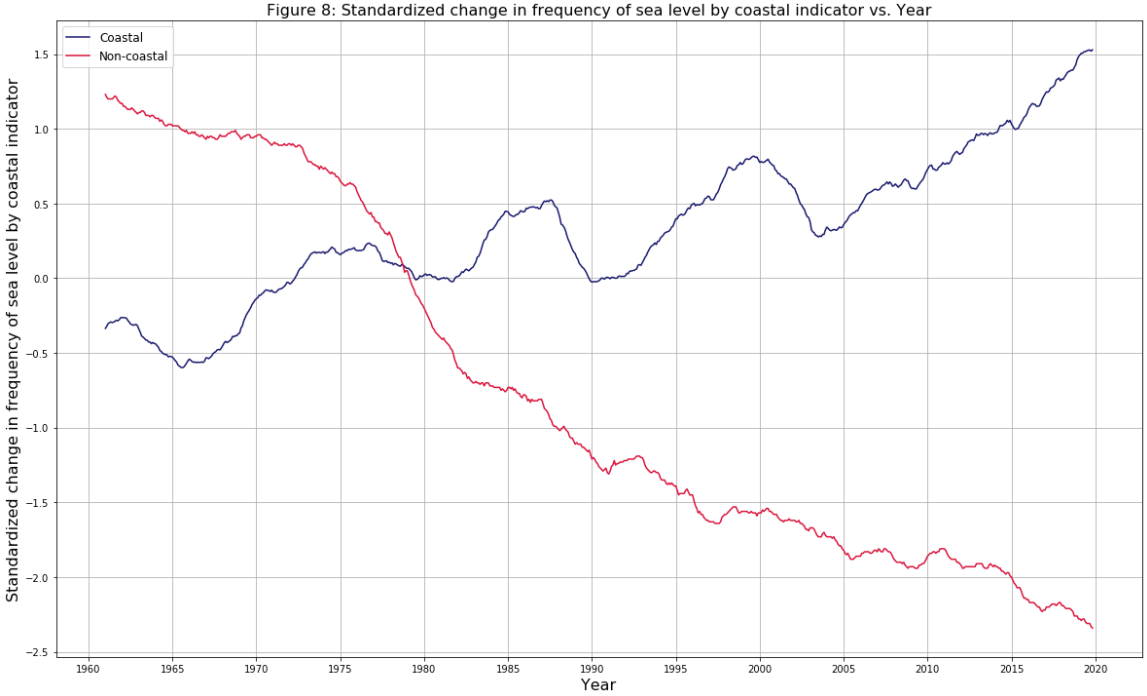
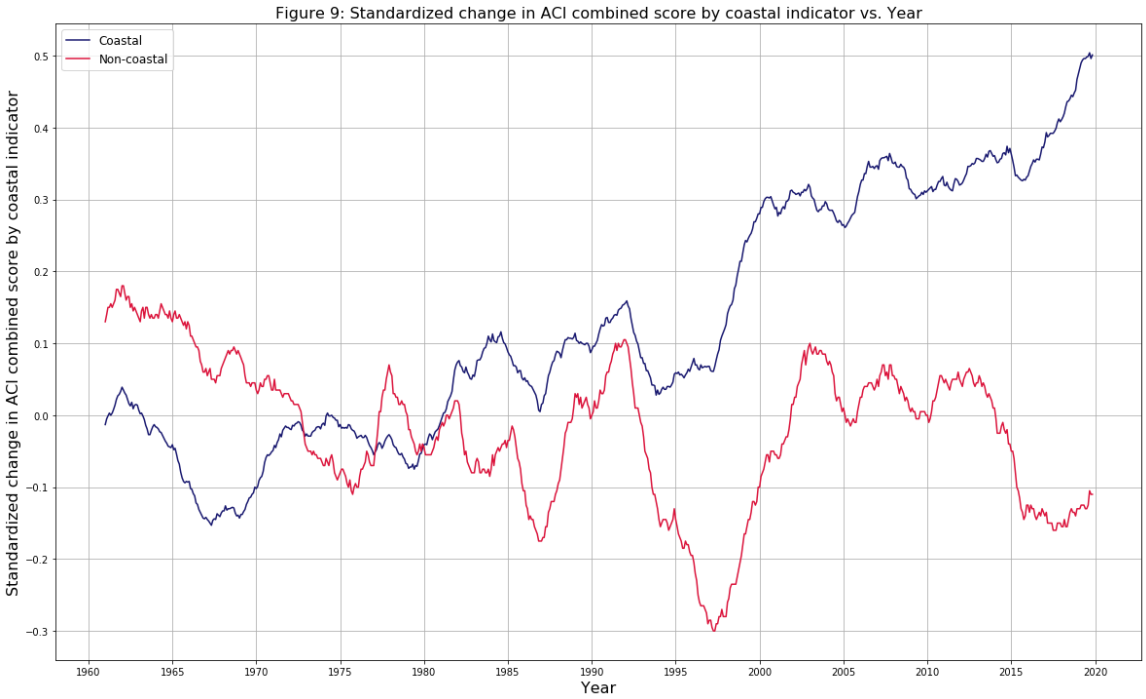


Figure 8: Standardized change in frequency of sea level by coastal indicator vs. Year





Appendix D: See file Canada Crop Analysis.ipynb