**Course:** **Assignment Submission for Advanced Environmental Economics: Modelling and Empirical Approaches**

**Name:** **Hyunseok Son (7490201, M.Sc. Economics)**

**Assigned Topic:** **E3 Climate change impacts**

# Original Analysis

## Original Analysis Introduction

Burke et al., 2016 tries to find causal impacts of climate change on agricultural yields. The paper mainly focuses on corn as it is one of the most major, important grains in the USA agricultural industry. To find out the impacts of climate change, Burke et al., 2016 addresses the ideal but impossible experiment to explain the motivation of the research. The ideal experiment is to observe two identical Earths and see if outcomes diverge when a climate gradually changes in one Earth.

Previous literature has been using cross-sectional approaches or panel approaches. But cross-sectional approaches have been criticized for omitted variable concerns since different areas may have different technology level, proportion of well-informed and skilled farmers and so forth. So, the existing literature has preferred panel approaches.

‘Long difference’ approach of Burke et al., 2016 was an attempt to improve the literature that had been led by panel estimates. It addresses that panel estimates overestimate the short-term effects of weather, and it has concerns that it does not capture farmers’ adjustment ability that may take a longer time than the short-term effects of weather. Because the long difference approach observes the long-term differences, technological developments to mitigate the climate change risk would be also captured. These kinds of adjustments may be omitted in panel approaches or distort the results of panel estimates since these can be lagged but bring a rapid change in the adjustments of the agents.

To closely replicate the ideal experiment, the paper employs ‘Long Difference Approach’. It selects GDD (Growing Degree Days) and precipitation variables as independent variables to determine if they have significant impacts on long-term corn yield differences, which is the dependent variable. If climate differences over or below thresholds do not make outcomes diverge, then it cannot be rejected that climate change risk is successfully mitigated by the farmers, and on the other hand, if the outcomes do diverge, then farmers are struggling to overcome the struggles derived from the climate change.

### Limitation of Original Analysis

While long difference method may mitigate noises of original panel estimates approach, it has some concerns. First, it ignores the short, mid-term fluctuations in the climate and focuses only on the differences between long-term differences of the independent variables and the dependent variables. Then it confounds the farmers' lagged effort as it is unclear when these climate changes happened and how rapidly they took place. Second, it includes all data points that might be irrelevant for the research. Although a huge dataset that Burke et al., 2016 collected may mitigate this concern due to the law of big number, depending on the size of the irrelevant data points, it may lead to a substantial noise in the analysis. Third, the model puts both precipitation and temperature as independent variables with the same weight. In extension analysis where K-means clustering method is performed to classify 4 clusters, it will be shown that, in the relevant data points, precipitation and temperature variables have negative correlation to each other. Independent variables having such correlations can cause multicollinearity issues. Having multicollinearity issues may lead the analysis to display unstable estimates and makes it difficult to verify individual impacts of the independent variables. This negative correlation between precipitation may be the reason that the panel estimate,and the long difference approach suggest different thresholds.

In the extension analysis, K-means clustering method will be used to verify only relevant datasets and it will be investigated if systematic differences are found between the clusters.

### Assumptions of Original Analysis

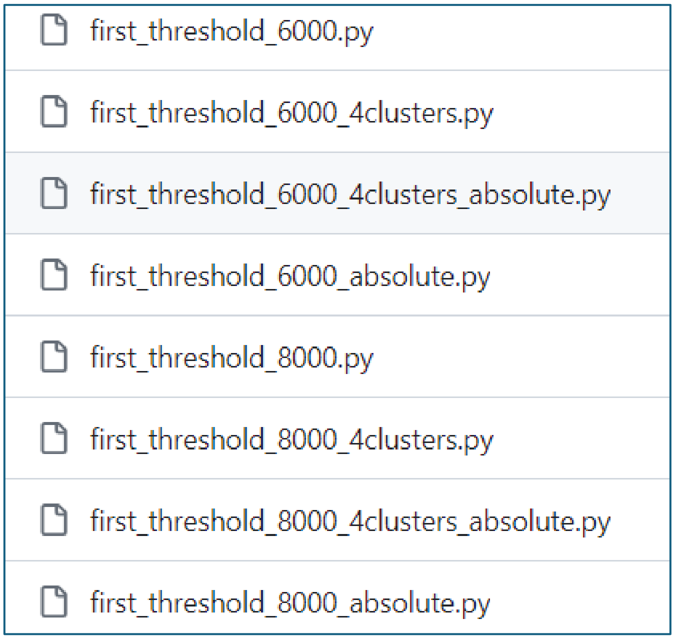
First, the long difference approach assumes that same fluctuation trends happened in between the investigated different periods and farmers made a same effort to overcome the change. Second, the long difference approach assumes that all data points are relevant for the analysis. It maps all long fluctuations in the independent variables and dependent variables. Coincidental changes in irrelevant data points will all be taken into account in the long difference approach.

## Extension Analysis

### Table and Data that will be extended

In the extension, Table 1 of Burke et al., 2016 will be extended. Table 1 of Burke et al., 2016 investigates causal impact of temperature, precipitation exposure over or below thresholds on corn yield differences. Therefore, this extension will focus on temperature, precipitation and corn yield data.

### Code and data connection explanation



first_threshold_6000.py
first_threshold_6000_4clusters.py
first_threshold_6000_4clusters_absolute.py
first_threshold_6000_absolute.py
first_threshold_8000.py
first_threshold_8000_4clusters.py
first_threshold_8000_4clusters_absolute.py
first_threshold_8000_absolute.py

The codes follow a specific pattern. 'Xth_threshold_Y000_***'. First threshold is 28, 29 Celsius degree GDD, second threshold is 30 Celsius degree and so forth. 6000 means that the period of 1960 to 2000 is used and 8000 means that the period of 1980 to 2000 is used. When there is a term 'absolute' at the end of the code file name, absolute number differences of corn yield are compared and when there is no 'absolute' in the file name, percentage differences of corn yield are compared. The different codes that are based on the same time period and threshold use the same datasets in the same folder, so please do not change the file names of the datasets.

### Motivation of the extension

The long difference method that Burke et al., 2016 used in the paper is improving the research on the exposure to temperature over a certain threshold's impact on agricultural yield substantially by employing a huge dataset that covers the period from 1960 to 2000 on a county, state level. Instead of investigating the short-term impacts as panel estimates did, the long difference method investigates the long-term impacts and enables the analysis to reduce noises that are caused by omitted variables or short-term fluctuations.

But it still has some potential flaws that it may be including a certain amount of noise due to its large dataset. In other words, it is possible that it is including the areas that are not plausible to be a part of this analysis. For example, if an area is too cold to inhabit active and prominent corn farms, corn yield differences in this area would depend greatly on exogenous factors such as policy, technology or industrial changes. Although the long difference approach is trying to mitigate this concern by regressing the data on a state level, depending on the proportion of these corn-farm-inappropriate areas, result may include substantial noise which is worth checking.

To check if the original dataset includes unsuitable datapoints, and to verify suitable and comparable sub-datasets, K-means unsupervised clustering method will be used on the county-level data. The goal is to mimic the ideal but impossible experiment that Burke et al., suggests. Burke et al., mentions that a perfect ideal experiment to analyze the causal impact would be to observe two identical Earths and see if there are any differences when only in one Earth is applied with gradually changing temperature. By using K-means clustering method, we will see if we can extract two comparable 'sub-Earths' where average temperature is similar but exposure levels to temperature over certain thresholds are different.

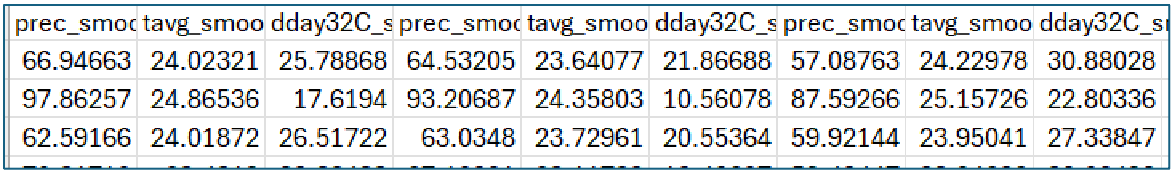
**Why K-means clustering method is suitable for the analysis**

K-means clustering method is an unsupervised clustering method. It calculates the Euclidean distances of data points and determines their similarity. It assigns centroids to random points of a dataset and clusters the dataset based on the proximity to the centroid. This assignment of centroid is repeated until some conditions set by the researcher are met. K-means clustering method is one of the most popular methods not only due to its computational efficiency and easy interpretability but also due to its adaptability to large datasets and its very strong tendency to convergence.

Although K-means method has disadvantages, in this research, the disadvantages may even work as reasons that K-means method is an appropriate approach for this analysis. First, K-means method cannot identify outliers of datasets, but our dataset includes temperature and precipitation data across counties. Therefore, the data is gradually changing across the data points, and it is reasonable to exclude the possibility of outliers' existence. Second, K-means method assumes that each cluster has similar size, and with the same reason mentioned in the first disadvantage, it is reasonable to consider that the clusters have similar sizes. Third, K-means method requires researchers to select the 'K' which is a number of clusters. But in our research, we are trying to verify comparable sub-Earths based on criteria (similar average temperature but different level of exposure to temperature over thresholds) so this rather works as a flexibility in this case.

## How K-means method is used in the analysis

By using K-means cluster, dataset will be clustered into certain numbers of clusters based on the independent variables used in Burke et al., 2016. The independent variables used for clustering are average temperature, precipitation, and GDD (Growing Degree Days) over threshold.



| prec_smoc | tavg_smoo | dday32C_s | prec_smoc | tavg_smoo | dday32C_s | prec_smoc | tavg_smoo | dday32C_si |
|---|---|---|---|---|---|---|---|---|
| 66.94663 | 24.02321 | 25.78868 | 64.53205 | 23.64077 | 21.86688 | 57.08763 | 24.22978 | 30.88028 |
| 97.86257 | 24.86536 | 17.6194 | 93.20687 | 24.35803 | 10.56078 | 87.59266 | 25.15726 | 22.80336 |
| 62.59166 | 24.01872 | 26.51722 | 63.0348 | 23.72961 | 20.55364 | 59.92144 | 23.95041 | 27.33847 |

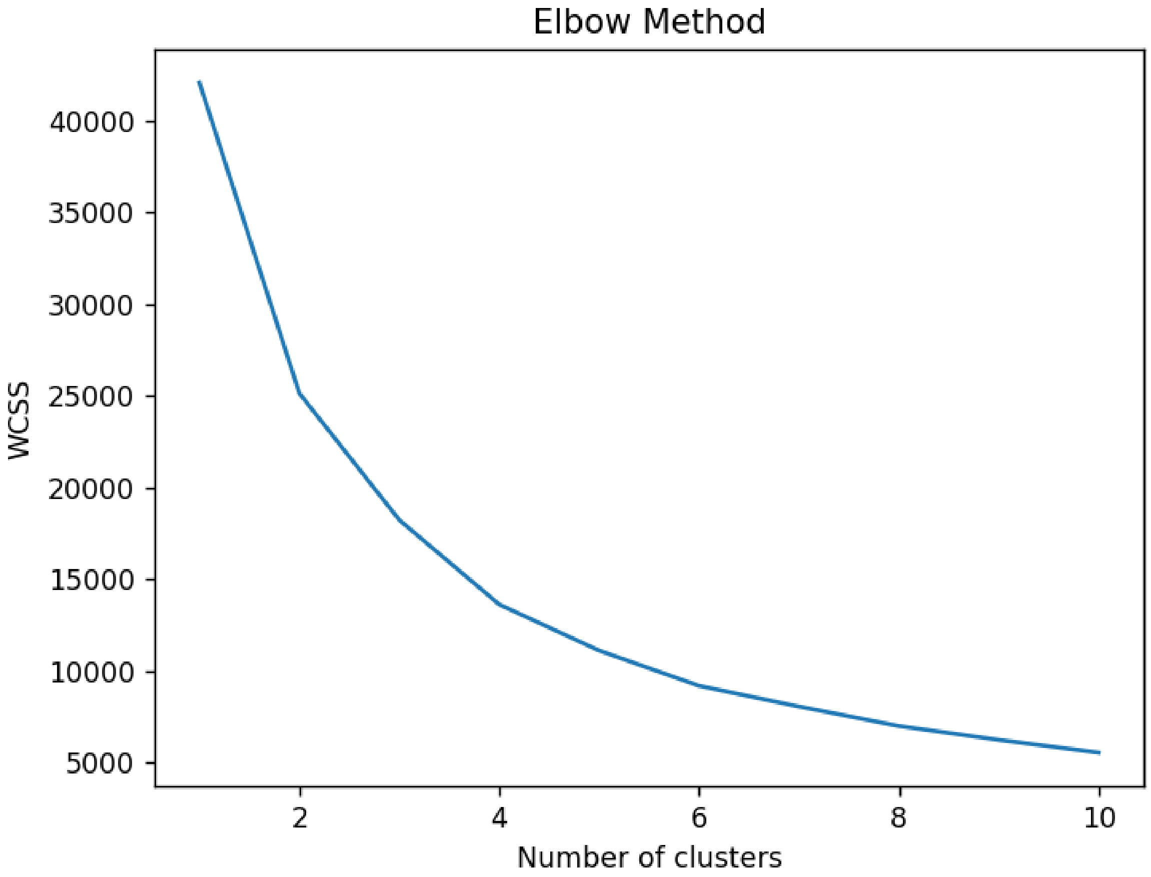
(Example of dataset where threshold is 32 Celsius degrees. Precipitation, average temperature, GDD over threshold of each period (1960-2000 or 1980-2000) is applied for clustering process.)

After the dataset is clustered, independent variables will be completely ignored and it will be studied if there are statistical differences in yield changes over long difference. That is, differences between corn yields in 2000 and 1960 or differences between corn yield in 2000 and 1980. Different time periods of from 1960 to 2000 and from 1980 to 2000 and different thresholds from 28 degrees to 32 degrees will be applied in each part of the analysis. As independent variables are used only for clustering the sub-Earths and ignored afterwards, it is a much stricter approach to verify causal impacts of climate differences in areas to corn yields.

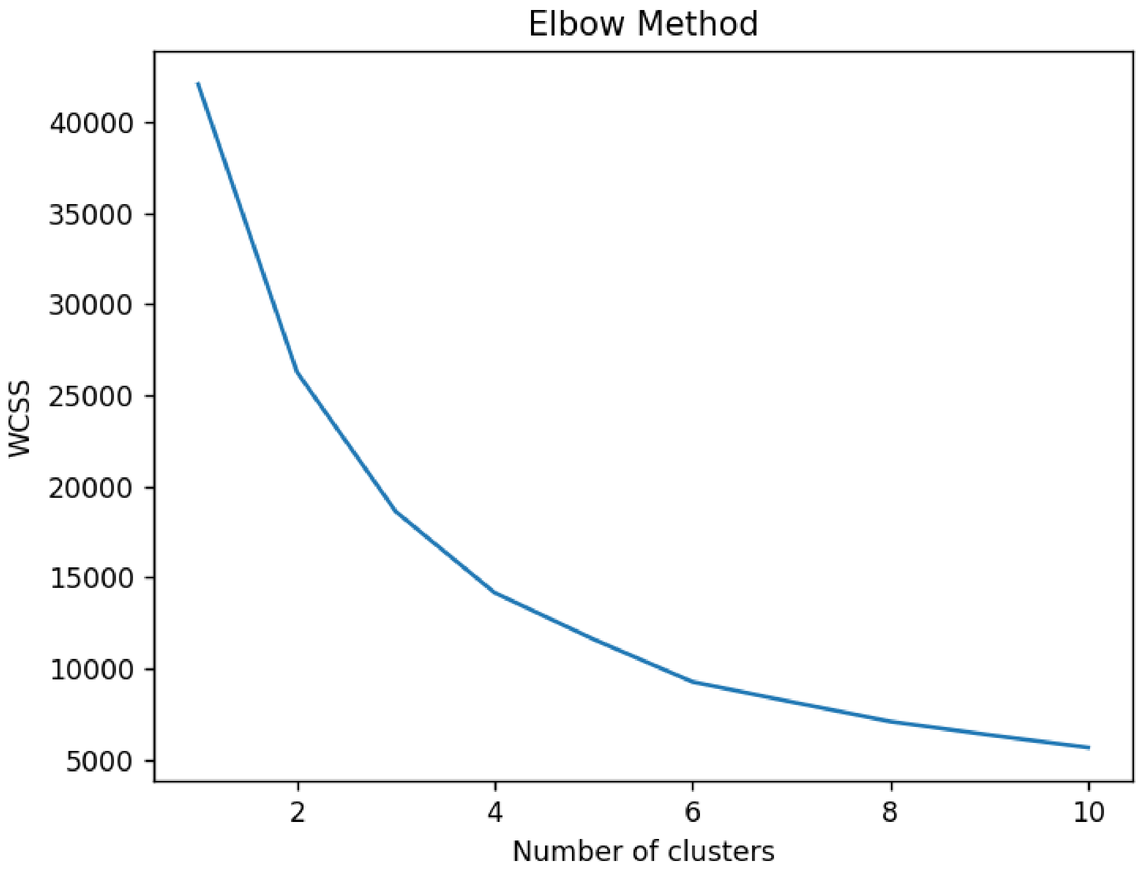
## Elbow method to choose 'K'

Since, K-means method requires researchers to select the 'K' which is a number of clusters, the researchers should be careful not to select an unreasonable 'K' that will make the clusters underfit or overfit the data. Therefore, employing the elbow method before selecting the 'K' is a useful measure to avoid such risks. The elbow method is a method that calculates, per 'K', distances between centroids and data points. This sum of distance within cluster typically declines rapidly in the smaller 'K's and slowly in the bigger 'K's. The point where the decrease of distance sum slows down can be considered as an effective point to cluster the datasets.

In all of the codes, the elbow method is applied at the beginning. Below are some examples of the elbow graphs of different time periods and thresholds. In all elbow graphs, K from 2 to 4 is shown as a reasonable range for selection.



(Threshold: 30 Celsius Degree, Time period: 1960 – 2000)



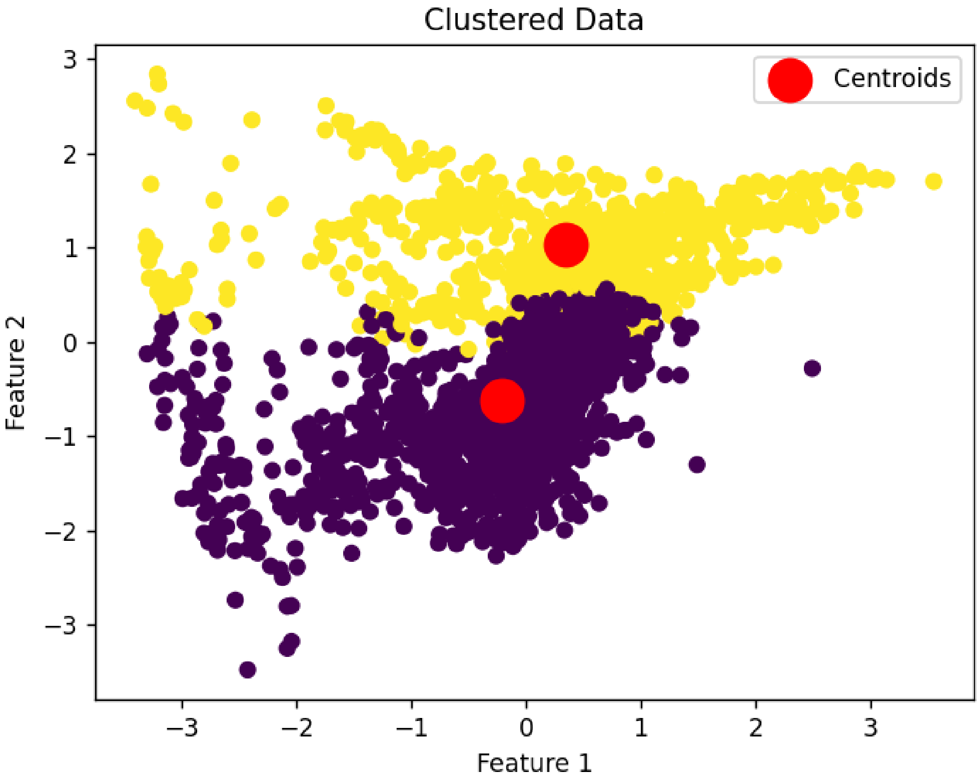
(Threshold: 31 Celsius Degree, Time period: 1980 – 2000)

**Result (K = 2)**

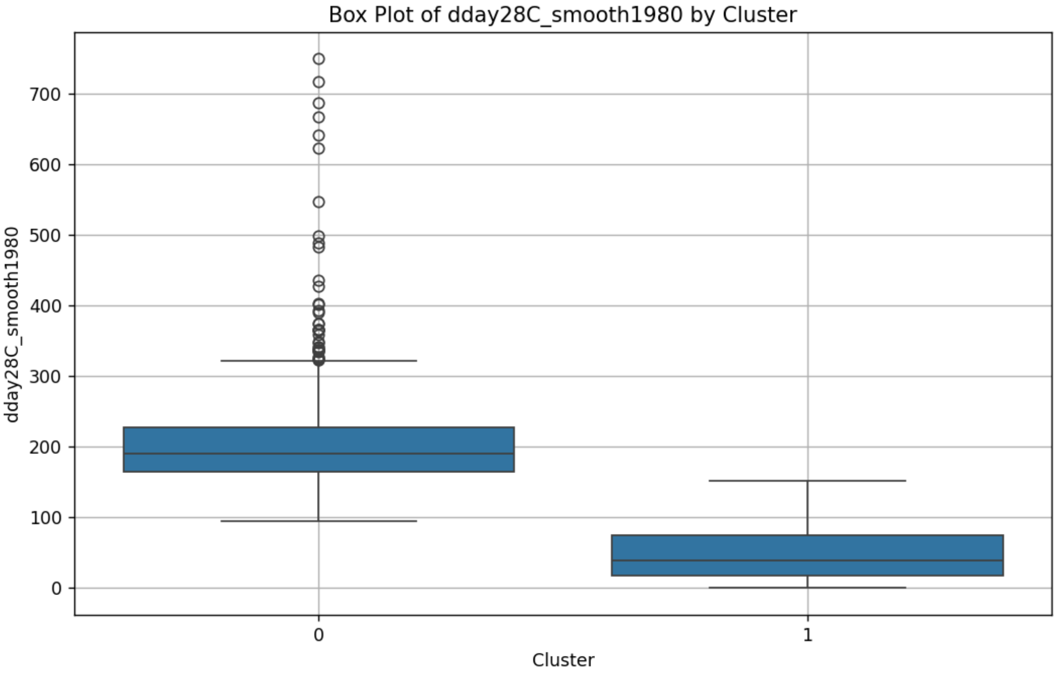
K-means method requires the researchers to select the ‘K’. The ideal experiment of the research is to observe two identical Earths, so in the first result, it will be investigated if there is useful information to be found when ‘K’ is set as 2.

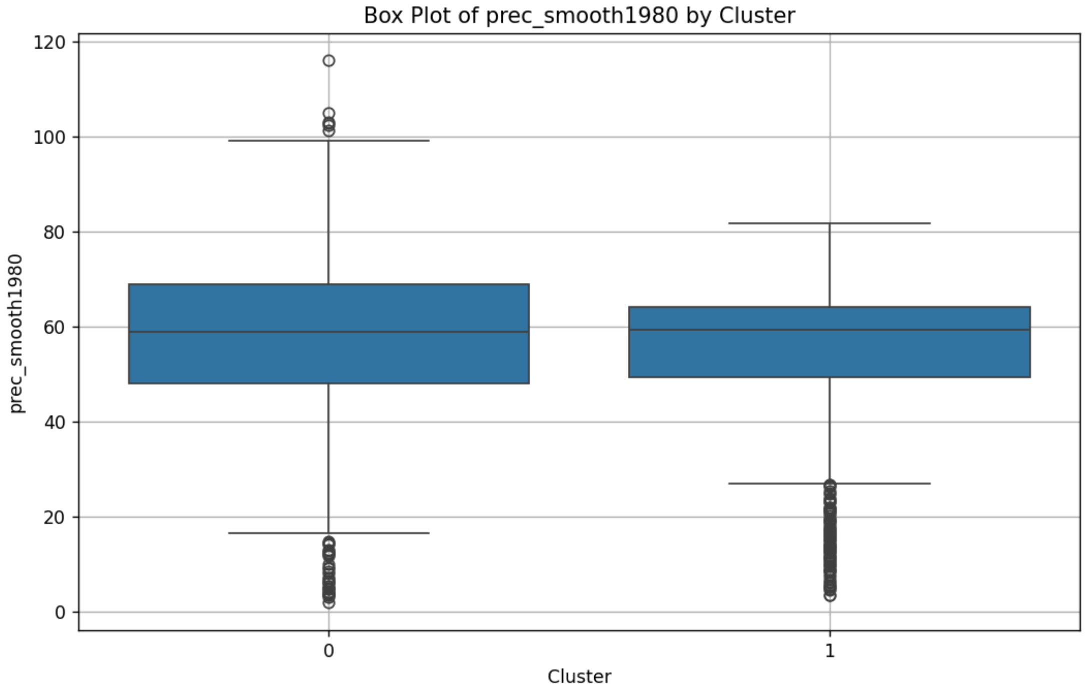
**Result (K = 2, Threshold = 28, 29 Celsius Degrees, Time period = 1980 – 2000)**

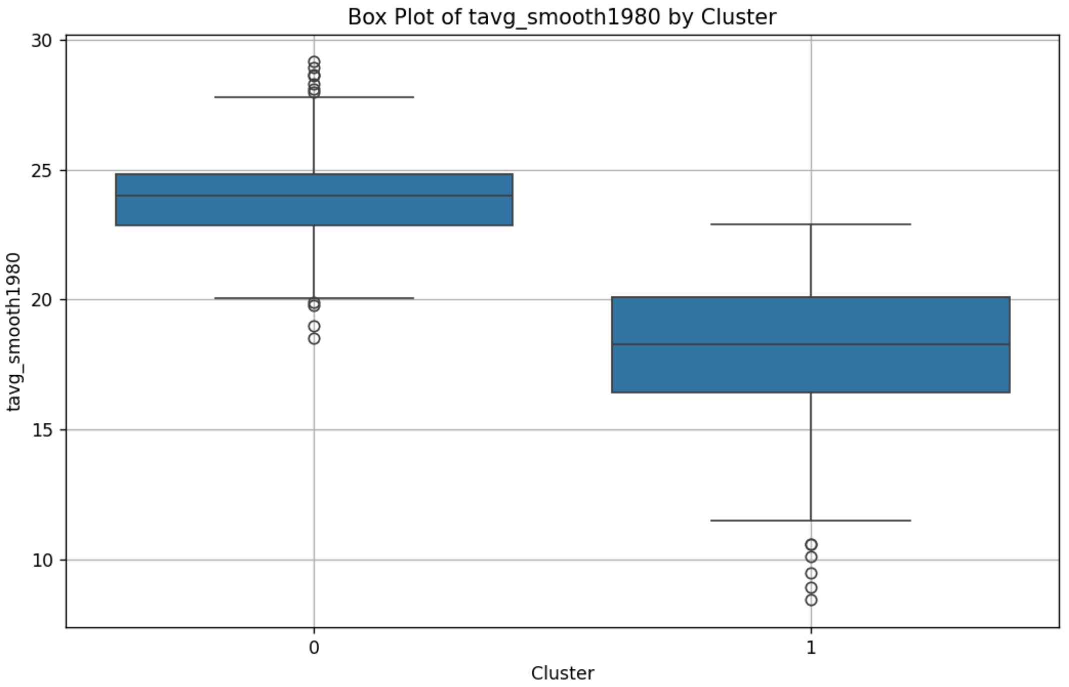
As Burke et al., 2016’s table 1 uses thresholds of 28, 29 Celsius degrees in the time period of 1980 to 2000, in this paper, analysis with same criteria will be shown. In the codes, threshold until 32 for both time periods 1960 -2000 and 1980 – 2000 are tested for absolute and relative differences of corn yields.



(Visualization of clusters when K = 2, Feature 1 = Precipitation, Feature 2 = Average temperature)







(Box plots for GDD over threshold, precipitation, and average temperature. The box represents the middle 50% percentile, the line above shows 75% percentile, and the line below shows the 25% percentile. A line in between the middle 50% percentile represents the mean value.)

Across the timeline, it is consistent that GDD over threshold is higher in the cluster 0 with mean value being higher than 75% percentile of cluster 1. Additionally, in cluster 0, precipitation is distributed with more variance with the mean value being lower or higher than cluster 1. In other words, the clusters classified clusters with higher average temperature and GDD over threshold while criteria in precipitation level is ambiguous. Therefore, in this case, we will name cluster 0 as 'Hotter Sub-Earth' and cluster 1 as 'Colder Sub-Earth'.

After Sub-Earths are clustered, only the differences in the corn yields of them will be compared.  When relative differences ( {2000_yield} / {1980_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in Cluster 0: 0.4240520029913708
- Mean of 'difference' in Cluster 1: 0.2573106438938082
- t-statistic: 10.51405740003395
- p-value: 4.427861327832881e-25

When absolute differences ( {2000_yield} – {1980_yield} ) are compared, the result is as follows:

- Mean of 'difference' in Cluster 0: 26.996434002583978
- Mean of 'difference' in Cluster 1: 22.806407241835146
- t-statistic: 3.939173556107524
- p-value: 8.51273192524965e-05

So, 'Hotter Sub-Earth' outperformed 'Colder Sub-Earth'.

This result was consistent in all other domains of threshold and time periods. That is, 'Hotter Sub-Earth' always outperformed 'Colder Sub-Earth' in every threshold from 28 to 32 Celsius degrees and in both time periods of 1960 - 2000 and 1980 – 2000.

**Limitations of K = 2 approach**

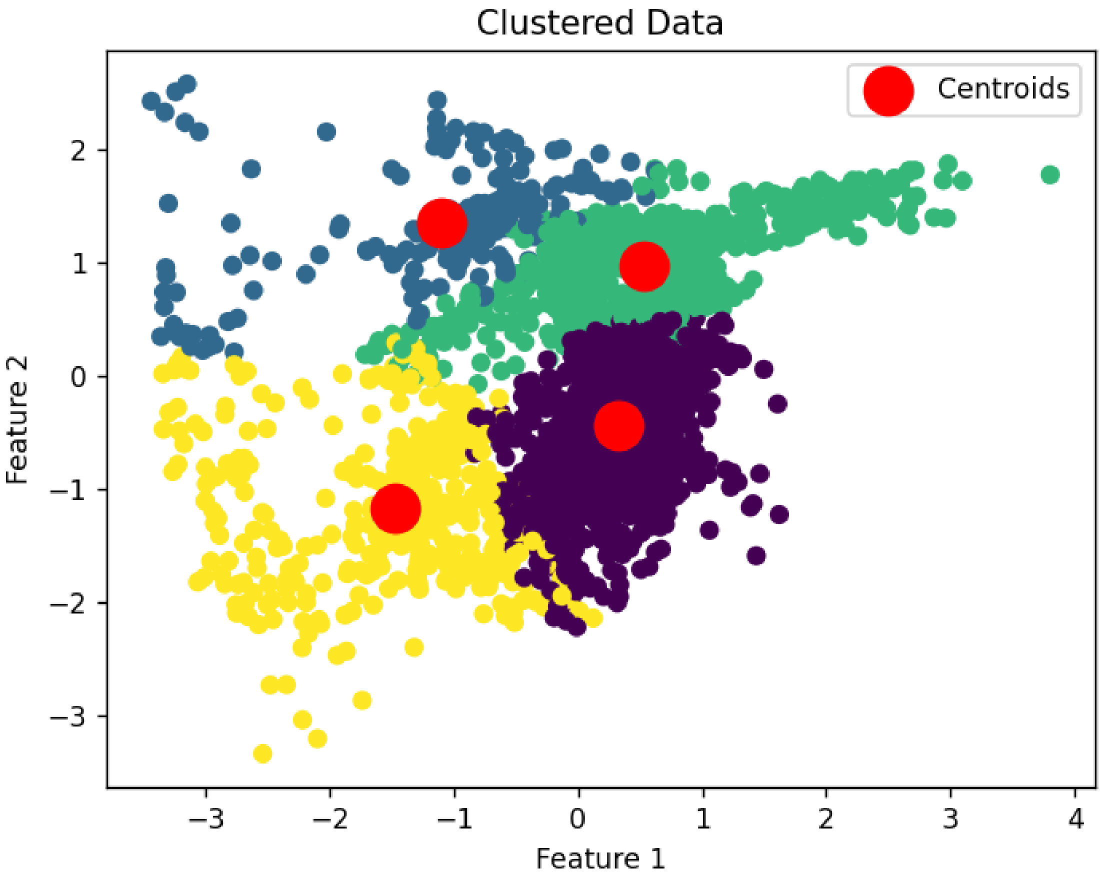
As shown in the box plots, it should be suspected that 'Colder Sub-Earth' might be too cold be a feasible comparison candidate. Burke et al., 2016 suggests 29 Celsius degrees as an optimal temperature for corn yields. But 'Colder Sub-Earth' is far from this optimal temperature. Its average temperature is lower than 20 while the average temperature of 'Hotter Sub-Earth' lies between 20 to 25 with its 75% percentile going above 25 Celsius Degrees. It is possible that 'Colder Sub-Earth' is too cold to be compared to

‘Hotter Sub-Earth’, therefore, ‘Colder Sub-Earth’ does not make substantial efforts to increase corn yields and neither to overcome climate struggles. If this is the case, then it makes sense that ‘Hotter Sub-Earth’ always outperforms ‘Colder Sub-Earth’ in every case.

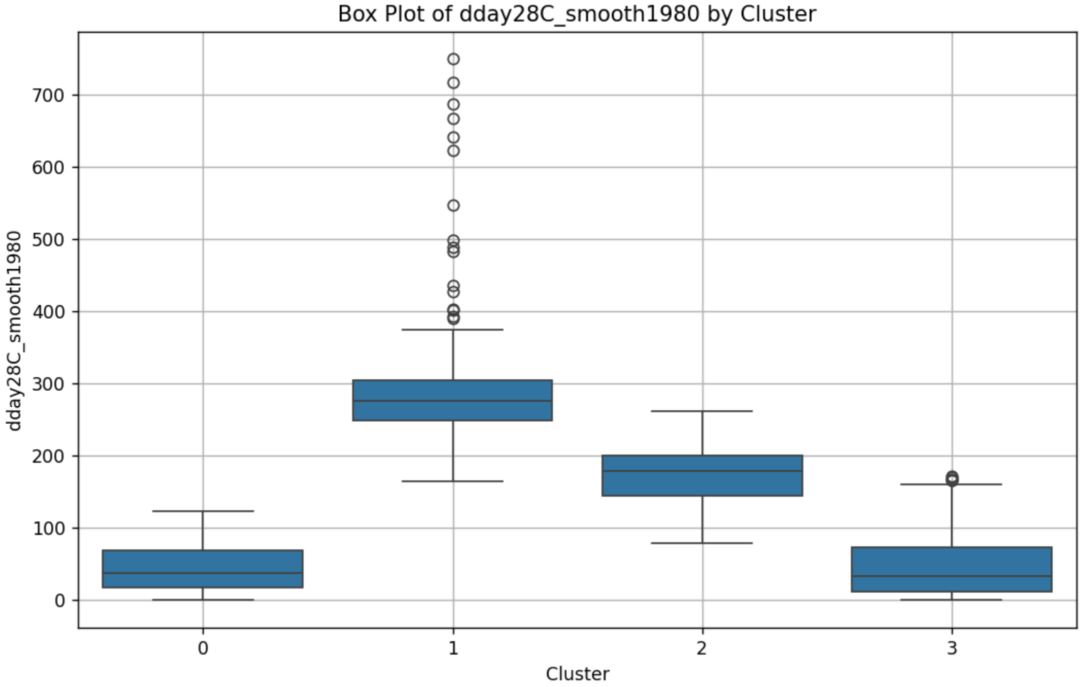
**Result (K = 4)**

The same structured analysis will be further performed by selecting 4 as ‘K’ since the elbow graphs suggest that 4 lies within a reasonable range of ‘K’s.

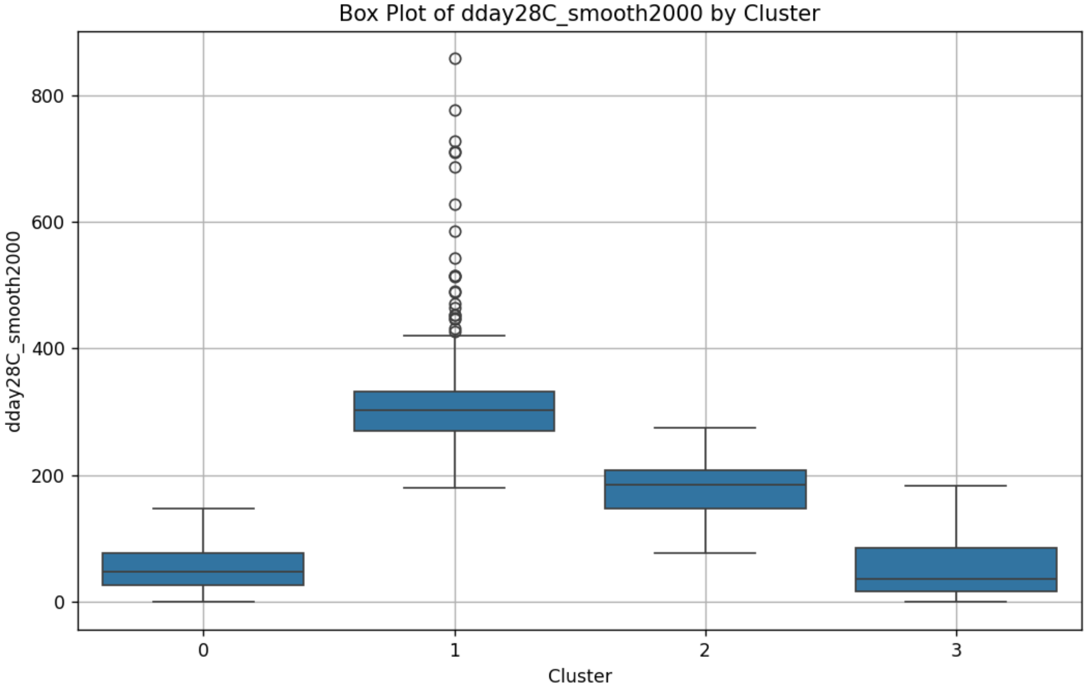
**Result (K = 4, Threshold = 28, 29 Celsius Degrees, Time period = 1980 – 2000)**



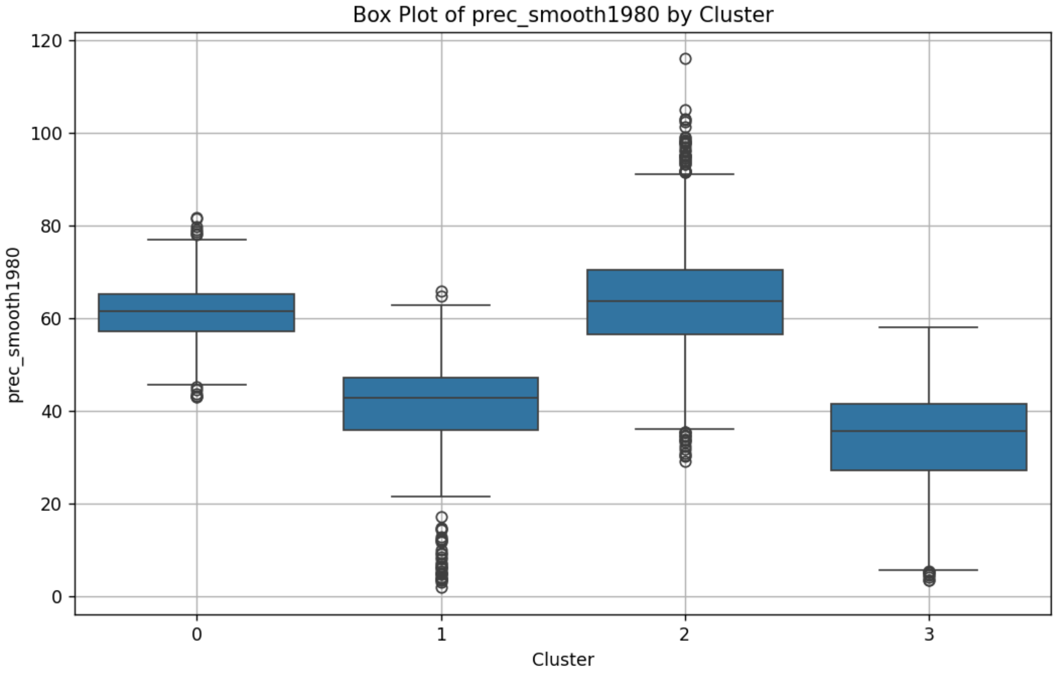
(Visualization of clusters when K = 4, Feature 1 = Average temperature, Feature 2 = Precipitation)

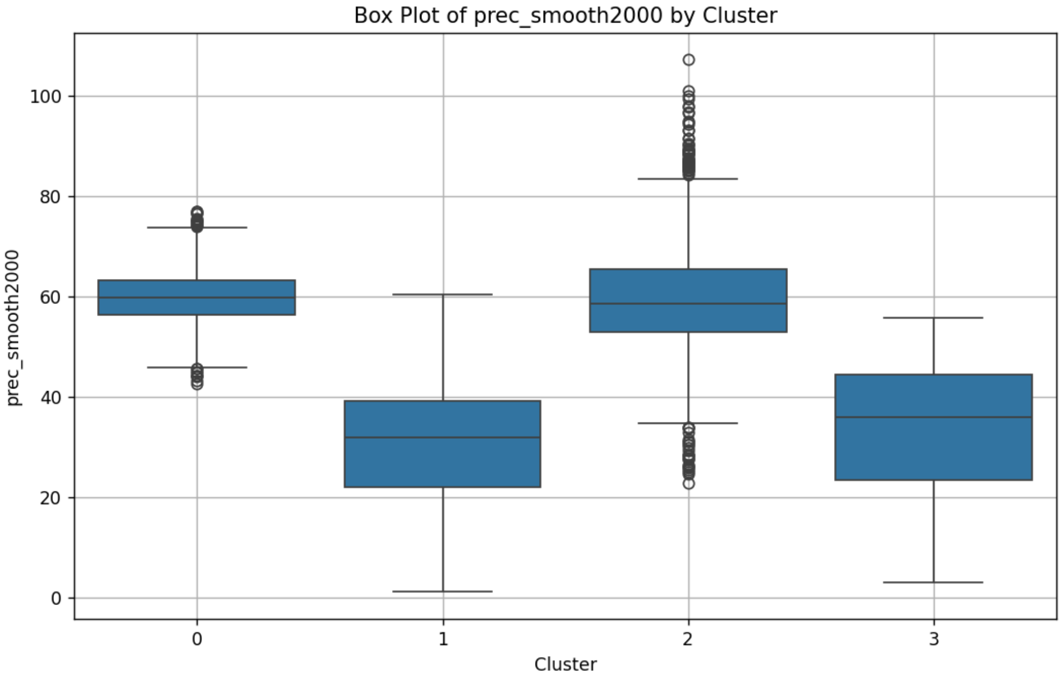


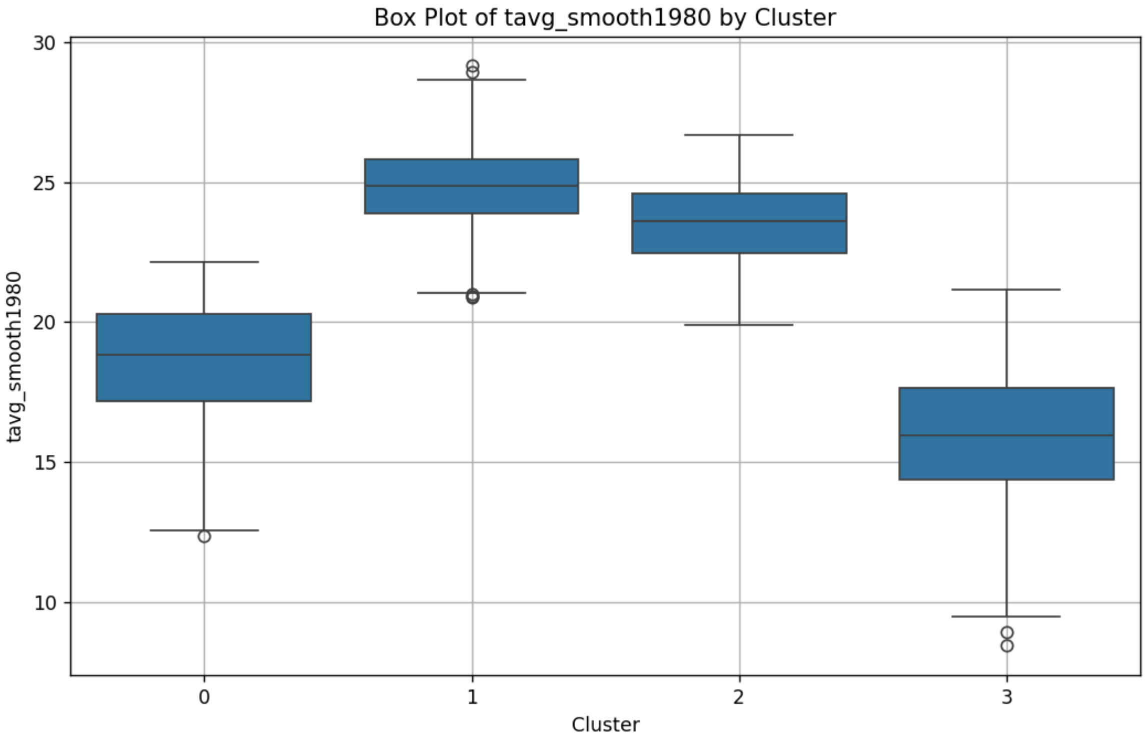
Box Plot of dday28C_smooth1980 by Cluster

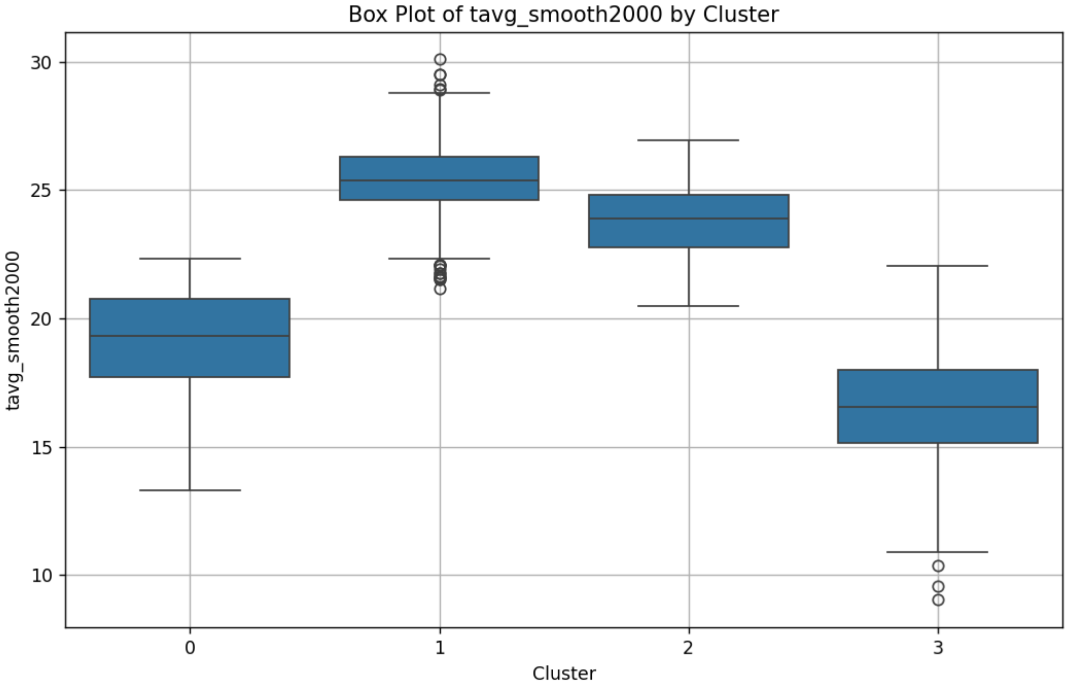


Box Plot of dday28C_smooth2000 by Cluster


Box Plot of prec_smooth1980 by Cluster
prec_smooth1980
0
20
40
60
80
100
120
Cluster
0
1
2
3


Box Plot of prec_smooth2000 by Cluster
prec_smooth2000
0
20
40
60
80
100
Cluster
0
1
2
3





In the analysis we can focus on cluster 1 and cluster 2. In terms of GDD, cluster 1 and cluster 2 are away from 0 than other two clusters but still show substantial differences.

Cluster 1’s mean is higher than cluster 2’s 75% percentile. And interestingly, between cluster 1 and cluster 2, precipitation differences reverse compared to temperature differences. These two clusters where they exceed 0 in GDD threshold meaningfully would have been major drive of coefficients suggested in the original study. In the original study, however, precipitation and temperature are both independent variables with the same weight. It can cause multicollinearity issues as they show negative correlations in a persistent way.

Besides, these clusters are even more meaningful since they have similar average temperatures. Cluster 1 will be named ‘Hot Sub-Earth’ and Cluster 2 will be named ‘Warm Sub-Earth’.

With exposure to temperature threshold being different and precipitation showing corresponding patterns across different timelines and thresholds, this analysis with K =4 will bring much more meaningful results than K = 2 analyis.

When relative differences ( {2000_yield} / {1980_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in Cluster 1(Hot Sub-Earth): 0.24325248128547589
- Mean of 'difference' in Cluster 2(Warm Sub-Earth): 0.4422410870102109
- t-statistic: -3.7540023853555025
- p-value: 0.00019794850787644024

When absolute differences ( {2000_yield} – {1980_yield} ) are compared, the result is as follows:

- Mean of 'difference' in first Cluster 1: 19.373219745762707
- Mean of 'difference' in second Cluster 2: 27.789544227882036
- t-statistic: -2.6683263293367796
- p-value: 0.007911200562665154

In this analysis, unlike the K = 2 analysis, ‘Warm Sub-Earth’ outperformed. But an interesting pattern will show with comparison with time period of 1960 to 2000. The result will be explained in detail in the remainder.

## Results with different thresholds and different time periods:

### Result (K = 4, Threshold = 28, 29 Celsius Degrees, Time period = 1960 – 2000)

When relative differences ( {2000_yield} / {1960_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in ‘Hot Sub-Earth’: 1.7515123121025793
- Mean of 'difference' in ‘Warm Sub-Earth’: 2.2458425986657202
- t-statistic: -1.2315938813798992
- p-value: 0.21900231104415327
- No statistical significance

When absolute differences ( {2000_yield} – {1960_yield} ) are compared, the result is as follows:

- Mean of 'difference' in ‘Hot Sub-Earth’: 61.52911940880503
- Mean of 'difference' in ‘Warm Sub-Earth’: 61.056000600000004
- t-statistic: 0.04022379245575934
- p-value: 0.9679397150295705
- No statistical significance

**Result (K = 4, Threshold = 30 Celsius Degrees, Time period = 1980 – 2000)**

When relative differences ( {2000_yield} / {1980_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 0.45193806327055563
- Mean of 'difference' in ‘Hot Sub-Earth’: 0.24447639034894994
- t-statistic: 5.11202235339909
- p-value: 4.594598026205592e-07

When absolute differences ( {2000_yield} – {1980_yield} ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 27.638256446153843
- Mean of 'difference' in ‘Hot Sub-Earth’: 23.032988804123715
- t-statistic: 1.8860071982906994
- p-value: 0.05989125790098806
- p-value < 0.1 (Weakly significant)

**Result (K = 4, Threshold = 30 Celsius Degrees, Time period = 1960 – 2000)**

When relative differences ( {2000_yield} / {1960_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 1.6638465054578842
- Mean of 'difference' in ‘Hot Sub-Earth’: 2.4195721259580947
- t-statistic: -4.918022408107486
- p-value: 1.3741145690005164e-06

When absolute differences ( {2000_yield} – {1960_yield} ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 57.81215956478405
- Mean of 'difference' in ‘Hot Sub-Earth’: 92.48057014285716
- t-statistic: -8.226261936571253
- p-value: 4.360469825003679e-15

**Result (K = 4, Threshold = 31 Celsius Degrees, Time period = 1980 – 2000)**

When relative differences ( {2000_yield} / {1980_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 0.44641564932758376
- Mean of 'difference' in ‘Hot Sub-Earth’: 0.2416628783895887
- t-statistic: -5.837755518333717
- p-value: 1.1696092925646784e-08

When absolute differences ( {2000_yield} – {1980_yield} ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 27.454765432098768
- Mean of 'difference' in ‘Hot Sub-Earth’: 22.406665802083335
- t-statistic: 2.0857530224469536
- p-value: 0.037508412060201286

**Result (K = 4, Threshold = 31 Celsius Degrees, Time period = 1960 – 2000)**

When relative differences ( {2000_yield} / {1960_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 1.6165916277830852
- Mean of 'difference' in ‘Hot Sub-Earth’: 2.5292743186240103
- t-statistic: -5.837755518333717
- p-value: 1.1696092925646784e-08

When absolute differences ( {2000_yield} – {1960_yield} ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 57.287701561194034
- Mean of 'difference' in ‘Hot Sub-Earth’: 94.3318735
- t-statistic: -8.680794379183903
- p-value: 1.3258123547472385e-16

### Result (K = 4, Threshold = 32 Celsius Degrees, Time period = 1980 – 2000)

When relative differences ( {2000_yield} / {1980_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 0.4224793555628994
- Mean of 'difference' in ‘Hot Sub-Earth’: 0.23320912267132857
- t-statistic: 4.734323154033357
- p-value: 2.799560344606275e-06

When absolute differences ( {2000_yield} – {1960_yield} ) are compared, the result is as follows:

- Mean of 'difference' in ‘Warm Sub-Earth’: 26.491130491304347
- Mean of 'difference' in ‘Hot Sub-Earth’: 21.78130332608696
- t-statistic: 1.9647146489385092
- p-value: 0.049950264582292514

### Result (K = 4, Threshold = 32 Celsius Degrees, Time period = 1960 – 2000)

When relative differences ( {2000_yield} / {1960_yield} – 1 ) are compared, the result is as follows:

- Mean of 'difference' in ‘Hot Sub-Earth’: 2.5601532918614684
- Mean of 'difference' in ‘Warm Sub-Earth’: 1.4742672135967942
- t-statistic: 7.476944087736316
- p-value: 3.333096924406263e-13

When absolute differences ( {2000_yield} – {1960_yield} ) are compared, the result is as follows:

- Mean of 'difference' in ‘Hot Sub-Earth’: 94.18599883333333
- Mean of 'difference' in ‘Warm Sub-Earth’: 56.16857145134576
- t-statistic: 9.508416401424244
- p-value: 7.547052804053872e-20

### Conclusion

The results from the above are summarized in the table below.

| Period | Threshold | Who Outperformed<br>(Hot VS Warm) |
|---|---|---|

| | | |
|---|---|---|
| **80 – 00** | **28, 29** | **Relative: Warm , Absolute: Warm** |
| **60 – 00** | **28, 29** | **Relative: (No Winner) , Absolute: (No Winner)** |
| **80 – 00** | **30** | **Relative: Warm , Absolute: Warm (Weakly Significant)** |
| **60 – 00** | **30** | **Relative: Hot , Absolute: Hot** |
| **80 – 00** | **31** | **Relative: Warm , Absolute: Warm** |
| **60 – 00** | **31** | **Relative: Hot , Absolute: Hot** |
| **80 – 00** | **32** | **Relative: Warm , Absolute: Warm** |
| **60 – 00** | **32** | **Relative: Hot , Absolute: Hot** |

In general, there is a trend that 'Hot Sub-Earth' outperformed in the period of 1960 – 2000 but 'Warm Sub-Earth' except for when the threshold is 28. 29 Celsius degrees where there is no winner in the period of 1960 – 2000. We can establish two hypotheses based on this result. First, agents are able to react in long term but are struggling to adjust to recent mid-term (20 years) impacts. Second, if we assume that climate is going higher over time in general for the areas studied, it could be that current 'Hot Sub-Earth' used to be the optimal environment for corns but is gradually shifting to suboptimal environment. In the meantime, 'Warm Sub-Earth' used to the suboptimal environment but is gradually shifting to optimal environment for corns. In the further analysis, it would be interesting to see fluctuations within clusters.

**Discussion**

- **New Insights**

First, it raised a doubt if only relevant sub-Earths can be extracted from the data and it found meaningful clusters that correspond to the research purpose which is to find out areas exposed in a different level to climate conditions over thresholds.

Second, it found out that in relevant subsets, precipitation level shows negative correlation to temperature. So, it is suggesting that regression in a conventional manner with both precipitation and temperature as independent variables can be problematic.

Third, it leads to two hypotheses that there might be long-term adjustments, but little mid-term adjustments, or corn-yield active area's dynamics are changing due to climate change.

- **Limitations and further possible extensions**

First, the analysis merely just defined 'which cluster is hotter' and not 'which one faced more changes in its climate'. Although K-means clustering method makes an effort to sort out only relevant data points, the extension analysis only focuses on the absolute values of the independent variables, not the fluctuations of them. It may be able to be further analyzed by multi-stage clustering and it will be explained in detail in ‘Further Possible Extension’ Section.

Second, it is a much stricter approach that does not include any independent variables in the end. Fluctuations within cluster including the independent variables could lead us to more detailed analysis.

- **Help Statement**

I got help from generative AI (Chat GPT) for code generation, for double checking the definitions of K-means, elbow method and for understanding the STATA code written by the author of the original analysis as I do not have knowledge in STATA. Code was around 80% generated by Chat GPT but I have python knowledge, so I checked if the code actually serves my research intentions and fixed parts where it did not correspond to my intentions. To follow the logic of the codes, please follow the comments of ‘first_threshold_6000.py’. Other .py files have Chat GPT’s comments as comments yet.

- **Reference**

Burke, Marshall, and Kyle Emerick. 2016. "Adaptation to Climate Change: Evidence from US Agriculture." American Economic Journal: Economic Policy, 8 (3): 106–40.