Harvesting Insights:

A Drought-Focused Agriculture Data Analysis

Multivariate Analysis

Texas Tech University

December 7th, 2023

Jonathan Busch, Joshua Corry, Emily Spector, James Parker

Table of Contents

Introduction

Section authored by Josh Corry

Agriculture is a vital business, not just because it supplies food, but also because it provides jobs for millions of individuals (Belvidere & Boone County, IL). According to Numbeo's cost of living calculator, the cost of food accounts for approximately 41.9% of the standard cost of living ("Cost of Living in United States"). Agriculture, however, is not perfectly predictable, as it is affected by several different variables, ranging from weather conditions to economic support. In this paper, weather conditions and various economic factors will be analyzed.

Throughout time, the real price of food has varied, and there are many use cases for understanding the underlying causes for these fluctuations. This project seeks to examine the changes in US droughts to see if there is a significant relationship between droughts on agricultural employment and the price index of food. We obtained our data from two different sources, the Federal Reserve Economic Data (FRED) for most of our financial information, and the drought data came from the University of Nebraska's drought monitor (University of Nebraska).

The data includes the weekly measurements of droughts both measured by area and by percent population affected.

We recognized that our data is time series, which is not typically used for multivariate analysis, however, we still felt that we can draw meaningful information from it. The dataset was adapted for multivariate analysis to facilitate this. Both percent and area drought data were reported at 6 different levels: no drought, level 1, level 2, level 3, level 4, and level 5 of which the levels were determined by the extremity of the drought. For example, level one would represent a lower level of a drought and level 5 represents the highest level of drought. Additionally, datasets were merged to these drought datasets that included the producer price index, subsidies from the

government for cattle ranches, the employment in agriculture industry, as well as the yearly inflation rate (St. Louis FED, "Producer Price Index by Industry: Food Manufacturing") (St. Louis FED, "Government Subsidies: Federal: Agricultural") (St. Louis FED, "Employment for Agriculture, Forestry, Fishing and Hunting: Cattle Ranching and Farming (NAICS 1121) in the United States") (St. Louis FED, "Inflation, Consumer Prices for the United States"). The data starts at the beginning of the millennium and concludes at the end of 2021. A table of our variables with their abbreviations and description will be provided below in table 1.

Table 1: An Explanation of Variables

Abbreviation	Variable Name	Description
PAD1	Percent Area Drought	% of United States area considered to be in a drought
FCET	Farm Cattle Employment in Thousands	The employment in thousands in farm or ranching related industries
AGSBillions	Agricultural Subsidies in Billions	Government spending on the agricultural sector of the economy in Billions
YIR	Yearly Inflation Rate	Change in value of money as a percent of the delta of the producer price index of all goods versus a base year
PPIF	Producer Price Index of Food	What percentage of the land is affected by drought at the given time (US)
PPD1	Percent Population Drought	What percentage of the population is affected by drought at the given time (US)

The motivation behind our choice of dataset lies in the fact that it could help identify relationships between agricultural economic variables and current and future drought conditions. For example, we could determine whether the agricultural employment rate is likely to decrease, stay the same, or increase during times of drought.

Data Cleaning and Visualization

Section authored by Josh Corry

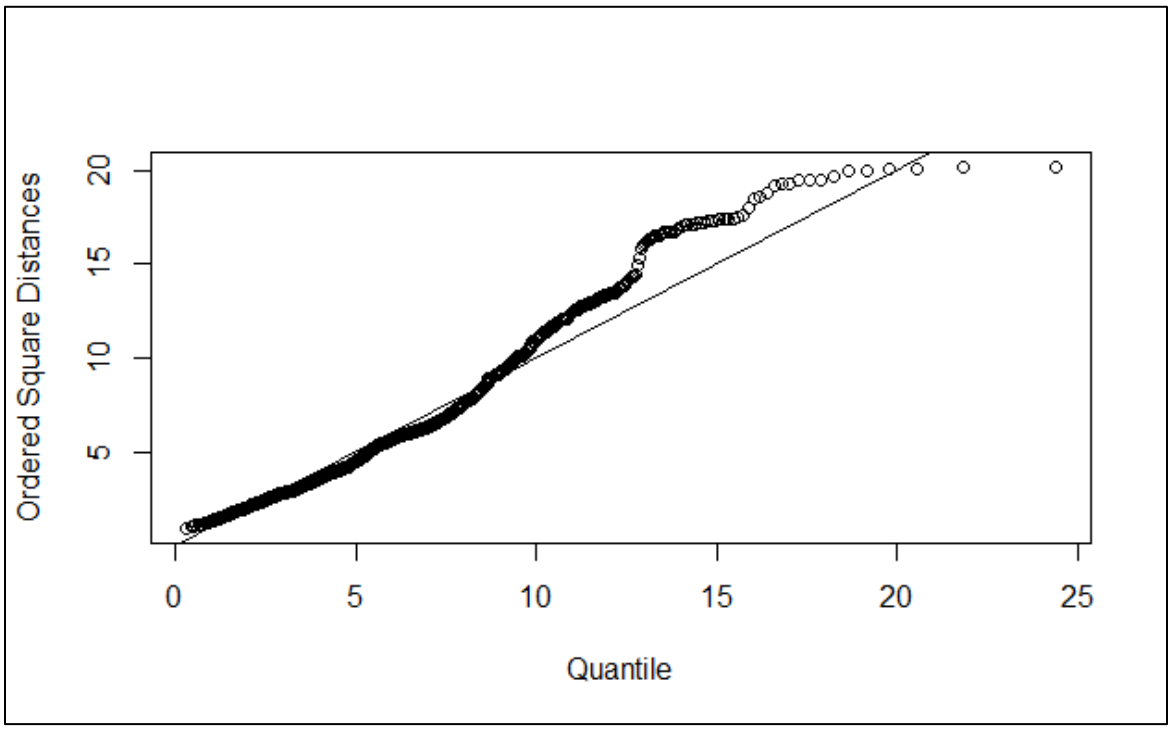
As we discussed in the introduction, our data includes 12 different variables that describe the drought levels that affect area and population. However, once we began performing analysis, we realized that these variables would all be highly correlated, as each subsequent drought level is a subset of the level before it. Furthermore, the percent of population/area with drought level 1 and the percent of population/area with no drought level are essentially the same. This is due to the fact that the percent of population/area in no drought is simply 1 minus the percent population/area in level 1 drought. This was necessary to compute the Mahalanobis distances for our outlier detection.

Additionally, we elected to turn the week, month, and year variables into row-names in order to adapt the data set for multivariate analysis. It follows that each row represents a specific week in the interval of time that our dataset encompasses. This is an essential understanding for further analysis, such as clustering.

Using the Mahalanobis distances, we were able to determine our outliers. Interestingly, every week in 2020 was considered to be an outlier. This would make sense, as the year started with an increase in US inflation rates, and quickly led into the Corona Virus pandemic. This would in turn lead to abnormal measures for things like employment rate, inflation, and producer price index. The only other outliers were two weeks in 2013.

The plot of the quantiles versus the ordered squared distances can be found below where it is clear that there are outliers within our dataset as reported in figure one.

Figure 1: Indication of Outliers



To find exactly which values were outliers within the dataset after confirming that outliers clearly existed through visualization, a chi squared test was done on the mahalanobis distances. The outliers were then reported which indicated that 2020 was the problem year within the dataset.

After the outliers were removed, we compared the correlation matrices of before and after, and noticed a significant increase in correlation, especially between Yearly Inflation Rate (YIR) and Agriculture Subsidies, which rose from 0.19 to 0.55.

In Figure 2, the correlation matrix generated before outliers were removed can be found. In Figure 3, the correlation matrix after the data was cleaned is reported.

Figure 2: Correlation Matrix with Outliers

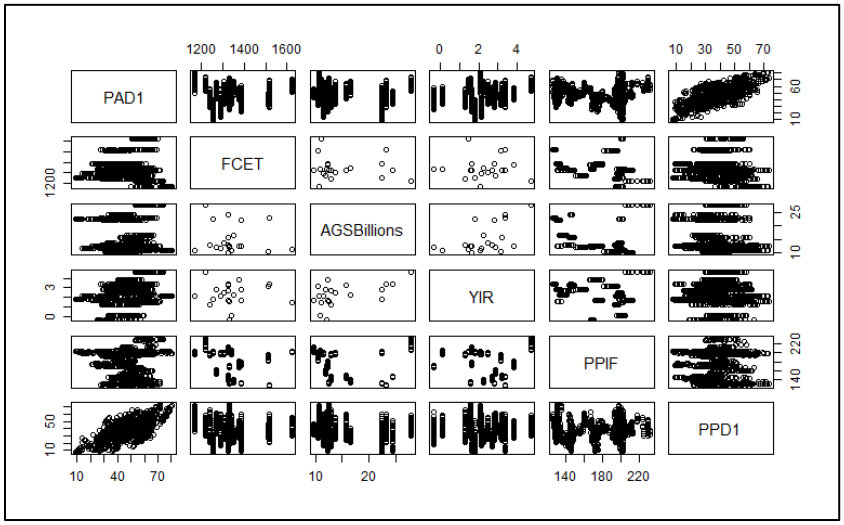
	PAD1	FCET	AGSBillions	YIR	PPIF	PPD1
PAD1	1.00	-0.01	-0.05	0.20	-0.10	0.70
FCET	-0.01	1.00	-0.12	0.02	-0.23	-0.05
AGSBillions	-0.05	-0.12	1.00	0.19	0.03	-0.25
YIR	0.20	0.02	0.19	1.00	-0.20	-0.04
PPIF	-0.10	-0.23	0.03	-0.20	1.00	0.04
PPD1	0.70	-0.05	-0.25	-0.04	0.04	1.00

Figure 3: Cleaned Data Correlation Matrix

	PAD1	FCET	AGSBillions	YIR	PPIF	PPD1
PAD1	1.00	-0.02	-0.03	0.21	-0.10	0.69
FCET	-0.02	1.00	-0.10	0.01	-0.22	-0.06
AGSBillions	-0.03	-0.10	1.00	0.55	-0.22	-0.23
YIR	0.21	0.01	0.55	1.00	-0.17	-0.06
PPIF	-0.10	-0.22	-0.22	-0.17	1.00	0.07
PPD1	0.69	-0.06	-0.23	-0.06	0.07	1.00

As can be seen in the figures, some correlations are slightly lower after cleaning the data, however this change is negligible. Other correlations did change noticeably, such as the aforementioned yearly inflation rate and the agricultural subsidies. The correlation between the producer price index of food and the agricultural subsidies in billions also had a more dramatic change, going from 0.03 to –0.22.

Figure 4: Scatterplot Matrix of Clean Dataset



To further explore the clean dataset a scatterplot matrix was created to identify where interesting correlations may be occurring. This can be found above in Figure 4 which displays a scatter plot for each combination of variables.

Though messy, these results make it clear that PPD1 and PAD1 are positively correlated, which was expected to be the case. It was expected that when performing exploratory factor analysis that the PDD1 and PAD1 would be combined into one factor. Another interesting point is the correlation between YIR and the AGSBillions where there seems to be a minor positive correlation which is confirmed by our correlation matrix.

Dimension Reduction

Section authored by Emily Spector

The principal component analysis (PCA) conducted in this study aimed to reveal relationships within our dataset and reduce its dimensionality. We were able to transform six variables into a set of three uncorrelated principal components, which collectively account for 76.9% of the total variance. Table 2 reports the principal component loadings.

The key findings from our principal component analysis are as follows:

- PC1 explains the reduced amount of government subsidies with increased population drought (or vice versa) and accounts for 29.3% of the total variance.
- PC2 explains an increase in the yearly inflation rate along with an increase in area drought (or vice versa) and accounts for 27.9% of the total variance.
- PC3 explains farm and cattle employment increases and a decrease in the consumer price index of food (or vice versa), accounting for 19.7% of total variance.

Table 2: Principal Component Loadings

Loadings	Comp.1	Comp.2	Comp.3
PAD1		0.559	
FCET			0.787
AGSBIllions	-0.501		
YIR		0.549	
PPIF			-0.540
PPD1	0.610		

These principal components can be further analyzed to understand what exactly is happening within these relationships. The goal is to not only to highlight observed relationships among variables, but also to dive into the fundamental drivers that shape these connections which could be done with more research and analyses. Of particular interest would be the exploration of why drought levels inversely coincide with government subsidies in PC1.

Cluster Analysis

Section authored by James Parker

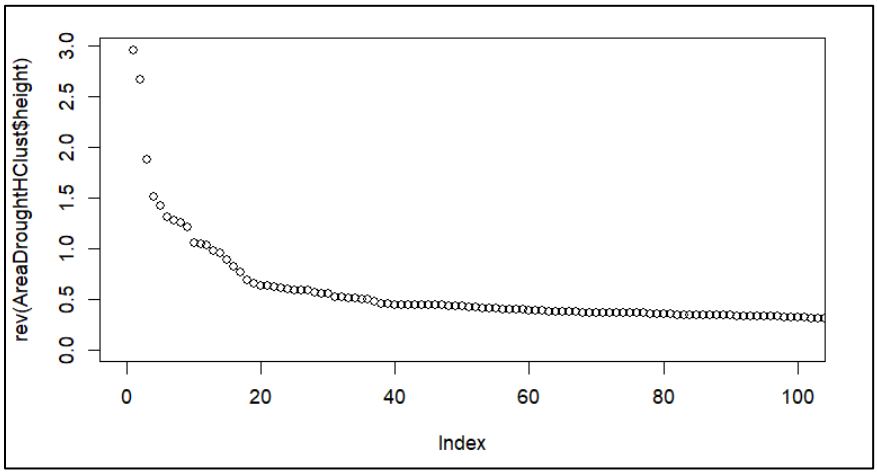
As discussed earlier, our data was inherently time series and, as a result of this, each row represents a point in time. It was for this reason that our group was interested in how the various clustering techniques would group our data. Logically, the algorithms attempted to cluster based on the years. For example, after visual analysis of the scree plot when performing single linkage hierarchical clustering, the “elbow point” seems to be at or around 20 clusters, and our data consisted of just over 20 years. These “obvious” insights, however, were not the focus of our

analysis. We wished to find more subtle and interesting insights from our data. For this reason, we chose 3 or 4 clusters in most cases to see if these numbers of clusters could provide any interesting findings.

Hierarchical Clustering

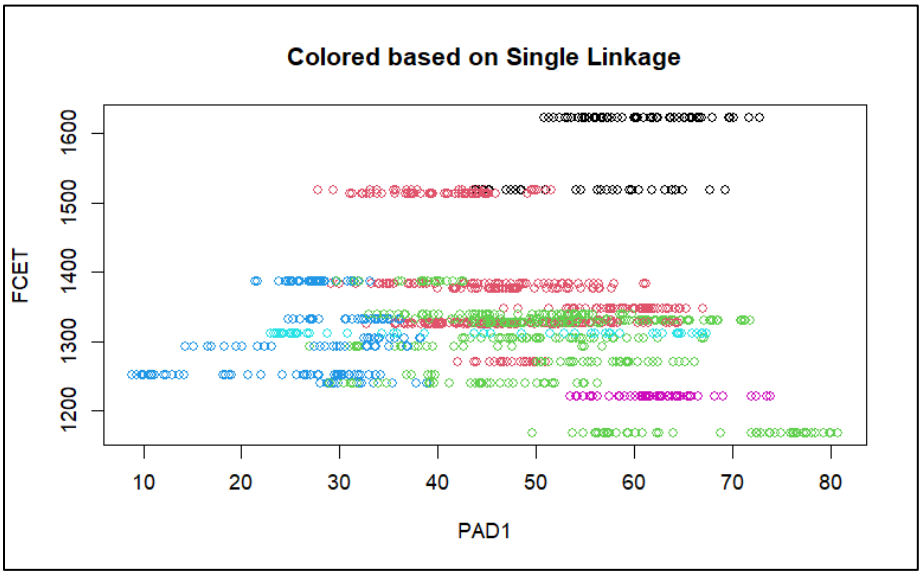
The first method we used for cluster analysis was hierarchical clustering. When creating the scree plot, specifically for single linkage, the elbow point suggests that the number of clusters should be about 20. This can be seen in Figure 5 below which shows the scree plot of the single linkage.

Figure 5: Scree Plot for Single Linkage



When conducting the analysis, however, we decided to use fewer groups. We experimented with groups of three to six, all of which were underwhelming and provided little insights into our data. This was primarily due to the amount of overlap that occurred when performing this type of clustering. As can be seen in Figure 6, the clusters that are created from single hierarchical clustering have significant overlap. We found that the results of complete and average hierarchical clustering were similar to single linkage. Interestingly, complete linkage struggled just as much with the overlap problem as the other linkage methods.

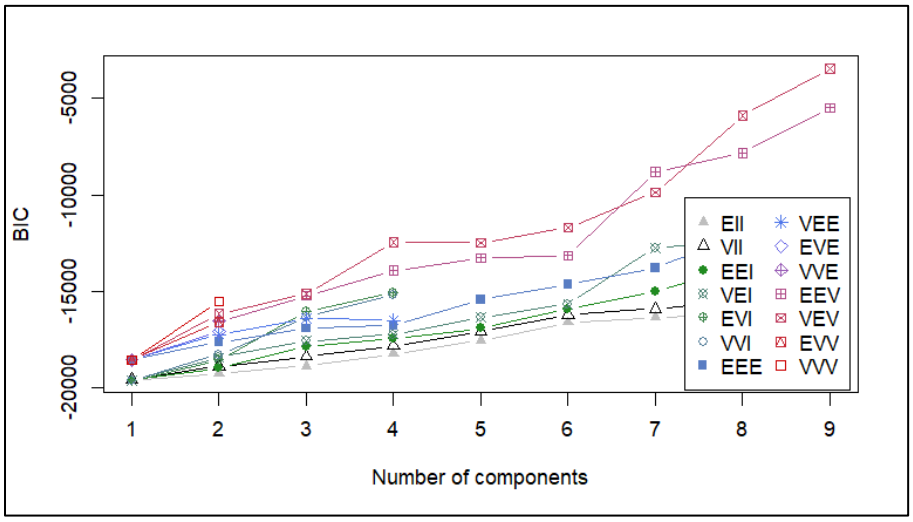
Figure 6: PAD1 and FCET Single Linkage



Model Based Clustering

The last method that was checked was model based clustering. As we did not have predefined clusters, we decided to let the mclust function choose the number of clusters for us. Below is Figure 7, a report on the BIC values for determining the best number of clusters.

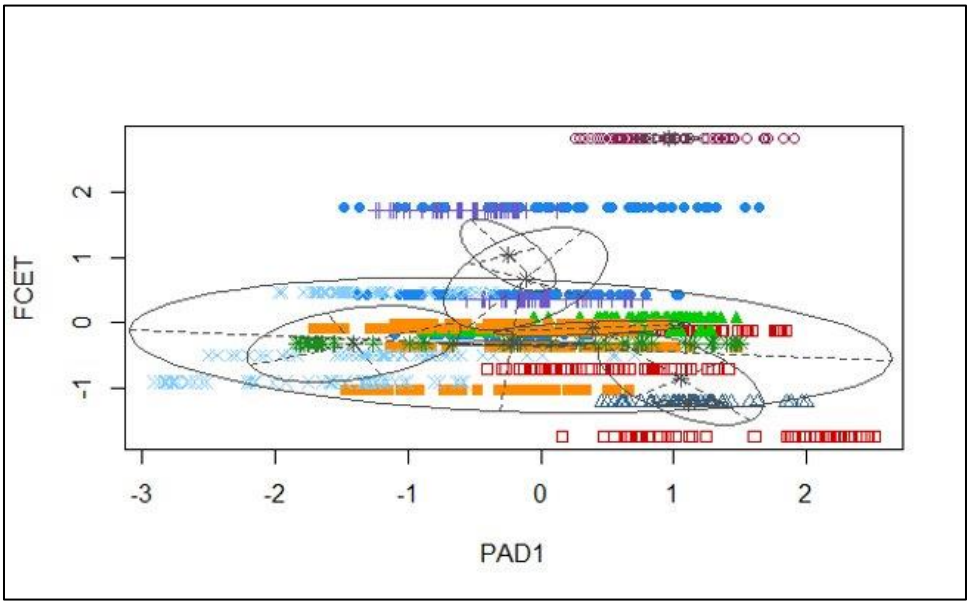
Figure 7: Model Based Clustering BIC



From Figure 7, it is clear that in terms of BIC, the best number of clusters for Model Based Clustering was nine. Within Figure 7 the VEV line bumps up at 4 clusters then drops down again

indicating it may be possible that less clusters are better as the penalty component within the BIC is not strong enough to keep our clusters from becoming over-complex. However, we chose to go with the reported number of clusters that the Mclust function chose. Figure 8 reports the classification for our model-based clustering. There is major overlap within this model as well as results that do not lead to any meaningful interpretation or insight.

Figure 8: FCET PAD1 Model Based



K-Means Clustering

Lastly, we chose to perform K-means clustering which had some of the most interesting insights that could be concluded from the analysis. Once again, we chose to use 3 clusters. Table 3 below shows the count for each cluster after performing K-means clustering.

Table 3: K-means Clustering Counts

Cluster 1	Cluster 2	Cluster 3
227	540	381

From the cluster counts alone, it is clear that K-means is a better fit of our data as the clusters are better spread out. We then sought to look at the column means for each cluster. By looking at the mean values of each cluster for each column, we can see the characteristics of each cluster. As can be seen in Table 4, cluster 1 represents times where the percentage of area and population in drought was above average, while agricultural employment was relatively low (or vice versa). Cluster 2 includes times where employment was high and the producer price index of food was low (or vice versa), and cluster 3 includes times when drought was low, and the inflation is also low (or vice versa).

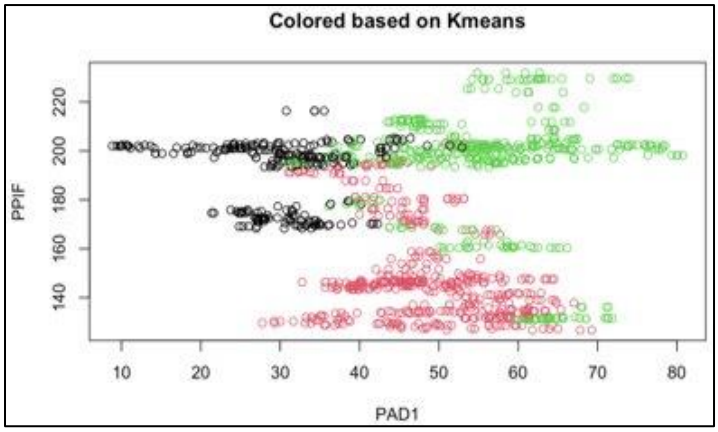
Table 4: Column Means for K-means Clusters

	PAD1	FCET	AGS Billions	YIR	PPIF	PPD1
Cluster 1	0.97	-0.96	0.36	0.25	0.72	0.96
Cluster 2	0.21	0.60	-0.06	0.48	-0.73	-0.05
Cluster 3	-0.88	-0.27	-0.12	-0.84	0.61	-0.49

The column Means above can be visualized when looking at scatterplots of the data. For example, in Figure 9 the scatterplot of PPIF and PAD1 are shown. In this visualization, three distinct clusters are formed, with the green cluster representing cluster 1, the red cluster representing cluster 2, and the black cluster representing cluster 3.

From this visualization, we can begin to make assumptions about our clusters. For example, cluster 1 might represent times when drought was bad, and the producer price index of food remained high. Cluster 2 could represent times when drought was average to slightly above average, but the PPIF was relatively low. Cluster 3 could represent specific times when, despite there being low drought, the PPIF was still high.

Figure 9: Scatterplot of PPIF and PAD1



It remains possible to gather even further insights from this data. If our group were to continue analysis on this data, we could dig deeper into the times that are clustered together and begin to find similar weeks, months, or years. Once we found these specific groups, it would be interesting to see if any different variables could be affecting this clustering.

Exploratory and Confirmatory Factor Analysis

Section authored by Jonathan Busch

Exploratory factor analysis was completed before attempting to perform confirmatory factor analysis. The scaled data-frame that was created earlier was used as an input into the factanal function with the number of factors set to three. Both the number of factors set to four and two were attempted as well. Two factors resulted in a lower percentage of the variance explained and four factors did not run. The resulting loadings created by factors set to three can be found below in Table 5. Note that any missing values in the table indicate that the loading value was below the designated cutoff point of 0.4.

Table 5: EFA Factors

	Factor One	Factor Two	Factor Three
PAD1	0.776		
FCET			-0.455
AGSBillions		0.451	
YIR		0.542	
PPIF			0.537
PPD1	0.959		

From these results we determined that the first factor represented an overall level of drought that the United States was experiencing as it was the combination of both the percent area drought and percent population drought variables. The second factor was determined to be Inflation as it combined both the agricultural subsidies and the yearly inflation rate variables. We suspected that as inflation devalues each dollar, more dollars needed to be spent on subsidies to get the same real value spent. The last factor is the relationship of the producer price index of food increasing as the total employment in agriculture decreases. Therefore, we determined this to be representative of the food supply as if employment falls, we can expect less food output and therefore higher food prices. This interpretation requires that efficiency of farming does not increase enough to offset the decrease in employment.

With these three factors, 46% of the variance was explained. We did not consider this especially ideal but given our inability to add more factors we decided if the RMSE was acceptable then these factors still hold if not strongly. The root mean square error was calculated which

resulted in an RMSE of 0.09 which showed that our factors were in fact sufficient even though only a small part of the data-sets variance could be explained in only three factors. We considered the reason for this to be the lower correlations between many of the variables while other variables such as the PPD1 and PAD1 were very highly correlated as they are similar. It should be noted that there is a major difference between the results of our factor analysis and that of our PCA analysis. PCA had the PAD1 variable on principle component 2 and the AGSBillions had a negative loading in principle component one. This provides evidence that the effect of percent population in drought and percent area in drought may have different relationships with our variables as the first 3 principal components represented 77% of the variance but the factor analysis represented much less variance at only 46%.

Confirmatory factor analysis was then attempted using the "sem" library's specifyModel function. However, the model failed to converge within the allotted iterations given to try and create a model. Different combinations of variables and models were attempted to see if a valid CFA could be generated. We believed the error was due to not enough variables with high enough correlations between each of them to create a proper CFA model as no attempted combinations worked. Furthermore, the time series nature of the data may also be causing issues with an attempt at CFA. It may be possible to get a working CFA model if additional variables such as total food output are added to the dataset. Brian, Everitt, and Hothorn within the textbook "An Introduction to Applied Multivariate Analysis with R" explains "…it should be remembered that both principal components analysis and factor analysis are similar in one important respect–they are both pointless if the observed variables are almost uncorrelated In this case, factor analysis has nothing to explain and principal components analysis will simply lead to components that are similar to the original variables." This quote shows that if variables with higher, but not too high they are

effectively the same variable, were included within the dataset more interesting information could have been garnered from the exploratory factor analysis and the SEM model may have functioned.

Conclusion

Section authored jointly

The PCA and the EFA both reported different results when grouping the variables into principal components and factors respectively. This generated an interesting insight in that the droughts in PCA are included into different principal components while in factor analysis the two drought variables were included within the same factor. Within the cluster analysis the K-means clustering had three clusters as the most intuitive number of clusters. As we had time series data these three clusters may represent three different eras in which the variables have different relationships.

As to be expected, most of the issues we encountered in our analysis came from the fact that our data was inherently time series. This led to the suggested number of clusters being influenced by seasonality, and there was also an absence of true clusters to base our cluster analysis on. We also initially thought we had more variables than we ended up actually using in our analysis due to redundancy between the drought variables. With the lack of variables, we had less than ideal correlation coefficients that made some models not run very well. Our weekly data was not optimal to perform multivariate analysis on, and perhaps aggregating on a broader time scale such as monthly or yearly data would have been better.

Despite the challenges of using an unconventional dataset for multivariate analysis, we were able to experiment with many different multivariate analysis techniques. This allowed us to

further develop our understanding of these techniques and serve as a learning experience going forward.

For a more comprehensive analysis of this data in the future, we should implement the addition of more variables. Total output of food within the United States and regional drought data would both be good variables to add to our current data set and would allow us to create better clusters, and possibly allow us to create a SEM plot. Another change would be the addition of region labels which would allow us to fully transform the data into proper multivariate data. This would allow for better cluster analysis as we could tell which regions are most similar to each other. With these additions, it would be likely that the data would report a better correlation matrix which would lead to a better principal component analysis and factor analysis.

Works Cited

Belvidere & Boone County, IL. Importance of Agribusiness in Supporting Economic Development. 25 July 2023, https://www.growthdimensions.org/news-and-events/p/item/51747/importance-of-agribusiness-in-supporting-economic-development

Brian, Everitt, and Torsten Hothorn. An Introduction to Applied Multivariate Analysis with R. 1st ed., 2011.

"Cost of Living in United States." Numbeo, 2023, https://www.numbeo.com/cost-of-living/country_result.jsp?country=United+States.

St. Louis FED. "Employment for Agriculture, Forestry, Fishing and Hunting: Cattle Ranching and Farming (NAICS 1121) in the United States." *Federal Reserve Economic Data*, 1 Jan. 2022, https://fred.stlouisfed.org/series/IPUAN1121W200000000.

St. Louis FED. "Government Subsidies: Federal: Agricultural." *Federal Reserve Economic Data*, 1 Jan. 2022, https://fred.stlouisfed.org/series/L312041A027NBEA.

St. Louis FED. "Inflation, Consumer Prices for the United States." *Federal Reserve Economic Data*, 1 Jan. 2022, https://fred.stlouisfed.org/series/FPCPITOTLZGUSA.

St. Louis FED. "Producer Price Index by Industry: Food Manufacturing." *Federal Reserve Economic Data*, 1 Jan. 2022, https://fred.stlouisfed.org/series/PCU311311.

University of Nebraska. *U.S. Drought Monitor*. 3 Oct. 2023, https://droughtmonitor.unl.edu/DmData/DataDownload/ComprehensiveStatistics.aspx.