**Final Project Analysis**

Introduction

For this project, our group decided to do our predictive analysis project on almonds crops within the United States. Almond production in the United States is a multi-billion-dollar industry generating over $21 billion in revenue each year and supporting over 104,000 jobs in California alone (California, 2016). California produces 80% of the world’s supply of almonds (California, 2016), and thus the United States is number one for almonds specialty crop export.

According to USDA’s 2012 Agriculture Census, ninety-one percent of almond farms are family-owned, and most growers and handlers are multigenerational. This means that the growers may be subject to more financial instability depending on the success of each year’s crops. They do not have a financial cushion that larger corporations might be able to supply, but they have a lot of experience to offer.

As costs may be getting tougher for producers, the climate is also shifting. Temperatures are slowly changing, and weather patterns are becoming more extreme, causing ripple effects in the growing season and rainfall timing and amounts (Climate Change Indicators: Weather and Climate, 2022). All these modifications could have serious impacts on productivity.

We are focusing on predicting almond crop yield and potential for profits as global climate changes. There are multiple questions we initially would like to try to answer with more sparking up along the way:

1. Are temperatures showing a positive trend supportive of global warming patterns within the project data time frame?
2. Are crop yields decreasing or increasing over time?
3. Are there signs that increased costs (i.e., water, fertilizers, pesticides, transportation) have cut into profit margins over time?

All these questions would be particularly important for smaller growers trying to ensure that their business models continue to be sustainable for future growing seasons.

Methods/Results

*Data Source*

The data for this project comes from the Food and Agriculture Organization (FAO) of the United Nations (Food and Agriculture Organization, 2023). This organization’s mission is to tackle world hunger by achieving food security and access to high-quality food for everyone. The data the organization collects measures important markers of food production, needs, and costs as well as the population characteristics of different countries.

The data our group is interested in is contained within several different datasets offered on the FAO website. We pulled variable data for United States almonds, in shell, from the following datasets:

* Production – Crops and livestock products (Crops and Livestock Products Data, n.d.)
  + *Crop Yield –* theamount produced per hectare*.*
  + *Area Harvested –* number of hectares harvested.
* Climate Change Indicators – Temperature change on land (Temperature Change on Land Data, n.d.)
  + *Temperature Change –* degrees Celsius of difference from the previous meteorological year.
* Prices – Producer Prices (Producer Prices Data, n.d.)
  + *Producer Price –* theamount paid to growers when selling crops.
* Producer Price Trade – Crops and Livestock Products (Producer Price Trade Data - Crops and livestock products, n.d.)
  + *Export Value –* total value of exported crop that year.
* SDG Indicators - SDG Indicators (SDG Indicators, n.d.)
  + *Level of Water Stress –* theproportion of water withdrawn from available sources.

The format that the FAO puts the data in is more descriptive than immediately useful for analysis.  Many columns contain repetitive filter data that is not relevant to the analysis and different variables often end up as rows within the same generalized table. The easiest way to download the data is to put one variable at a time into a CSV via the FAO website.

*Data Cleanup*

Each dataset was prepped, cleaned, and transformed into a single dataset. We extracted relevant columns from each CSV file, renamed columns to be more descriptive, and addressed any missing values within each file.

Next, the group merged the different target variable columns into a single primary dataframe based on the year value. This allowed us to look for trends over time. During data exploration, it was determined that some of the target variables varied on the periods in which they were collected. Only one variable had a value for 2022 when the data was originally pulled so the group decided to end the analysis in 2021. Two other variables (producer price and water stress) were not recorded as far back as 1961 like the others. The group decided to create subsets of the data to address this allowing for variables to remain in the overall analysis without fully limiting the scope of the others. The three data subsets are as follows:

* Set One runs from 1961 – 2021.
  + Contains all variables except producer price and water stress.
  + This data set will help us address how changing temperatures impact almond crops. It will also allow us to look at historical import and export trends.
* Set Two runs from 1991 – 2021 and adds the producer price variable to those in Set One.
  + This additional variable will help us look at the potential profits that an almond grower could take home from selling their crop.
* Set Three runs from 2000 – 2020 and adds the water stress variable to those in Set Two.
  + This last variable allows us to see if there are any significant impacts on harvest due to water scarcity.

*Graphical Analysis*

To explore our data, the group decided to use line plots to use for the preliminary analysis and was able to see some notable trends in some of the variables. Temperature, yield, hectares harvested, and export value all show steady overall climbs as time passes. Imports have a marked spike around 2013-2014 and other than that are almost non-existent with a slight uptick near the end of our data time frame. The producer price has an overall gain over time but also has more peaks and valleys than the other variables do, suggesting some market volatility. Notably, there is a peak in the same 2013-2014 period as the peak in imports suggests that almonds were in hot demand during that time frame. Water stress does not show a distinct overall trend with a dip in 2010 followed by a jump up to the highest levels around 2015 which it maintained until the end of the study period. Water stress also showed little variability in its values (9.57 to 11.17%).

The group decided to use boxplots to see if notable outliers exist in some of the variables. The boxplots indicate outliers in hectares harvested, export value, and import. The ones of most potential concern are in the import variable as these data points are quite extreme when compared to the rest of the data. However, some suggestions within other variables such as producer price suggest this extreme value may be realistic despite its extremity.

Based on the data subsets previously described, the group created a correlation matrix heatmap for each subset. Since each subset adds a new variable and changes the time frame over which each variable is considered, each subsequent matrix contains repeated variables from the previous one, but with slightly different correlation values. Set One showed high correlations between many of the variables: year/hectares harvested (0.94), hectares harvested/export value (0.92), year/yield (0.88), and hectares harvested/yield (0.84). Set Two showed similar patterns with the previous variables but slightly weakened in most cases. The most highly correlated variables were combinations of year, hectares harvested, and export value. The highest correlation the added producer price variable had was with export value (0.73) and year (0.72). Set Three shows the same pattern as Set Two is the same most highly correlated variables and the added variable of water stress only reached a correlation value of 0.62/0.63 when paired with export value and hectares harvested, respectively. There was a slight negative correlation with yield (-0.24).

After the results from the line plots, boxplots, and correlation matrixes, the group decided to drop Set Three from the analysis due to the small time period and little variability in the water stress variable.

*Modeling*

After researching previous methods used for crop predictions, a time series model was chosen as the best model to use. We built and evaluated an ARIMA model, providing the SARIMAX results. To get the results for the ARIMA model, the group started by splitting each subset into training and test datasets as follows:

* Train-Test Split for Set One (1961 - 2021, 60 years)
  + Train 1961 – 2010, 50 years
  + Test 2011 – 2021, 10 years
* Train-Test Split for Set Two (1991 - 2021, 30 years).
  + Train 1991 – 2015, 25 years
  + Test 2016 – 2021, 5 years

After splitting the datasets, the group created a function to look at both datasets and determine the best order to use in the Arima Model. Based on those function results, it was determined that the best ARIMA orders were (2, 1, 1) for Set One and (2, 1, 2) for Set Two. From these orders, the group created an ARIMA model for Set One and Set Two and printed out the SARIMAX results. Next, the group plotted each variable’s actual vs forecasted values.

*Model Results*

We visually determined in most of the plots Set Two’s model was closer in its predictions than Set One. We then compared the performance metrics and found that the Set One ARIMA model was overall more accurate in its predictions than the Set Two ARIMA model as shown in the lower MAE, MSE, and RMSE metrics. To determine which SARIMAX result is the best for both Set One and Set Two, you would need to compare the AIC values of the fitted models. The set with the lower AIC value would indicate a better fit. After looking through all the results in Set One and Set Two, the temperature change variable has the lowest AIC with results of 169.987 for Set One and 70.112 for Set Two. The SARIMAX also shows the difference in observations made between Set One (50) and Set Two (25).

Conclusions

Overall, our model did not do an excellent job of predicting many of the variables. From the error metrics, the export value variable error is extremely high in both sets meaning this variable is difficult for either model to predict. In terms of trying to provide a useful prediction to almond growers, the variables our model was best able to predict were temperature change and yield. Although the temperature change is not particularly helpful, the yield might be. A caveat is that the model often overestimates the crop yield so this must be considered when interpreting model results to make business decisions.

Going back to our original questions, the data and model predictions did show a positive trend in temperatures supportive of global warming patterns within the project data time frame. Crop yields are generally increasing over time both in the area harvested and, in the amount harvested per area unit. However, despite there being an overall increase in export value, the recent decreases in the amount the producer gets when selling the crop suggest lower profit margins. This might be due to increasing production costs such as labor, water, fertilizers, pesticides, and transportation. Overall, the trends and model predictions suggest that a producer might be able to get roughly the same crop yield as the past couple of years, but cost increases may make it less feasible to turn a profit.

*Ethics and Other Concerns*

One of the concerns with using models to make any sort of prediction is the possibility of error. If a business bases important decisions solely on the model’s predictions and they turn out to be misguided, then it can lead to detrimental impacts on the business with the extreme being going bankrupt. Especially in this case where the model results do not give a clear picture of future expectations. If a business decides to invest in more property and equipment to increase their production and thus potentially profits based on the model’s potentially incorrect predictions, this money could have been unwisely spent.

The other issue is any market is volatile, especially when it comes to non-essential goods. Although almonds are nutritionally excellent and provide many health benefits (Zelman, 2022), they are not always affordable to everyone. In some ways, they are a luxury snack in comparison to many other options. This means that if the economy gets tighter, it is possible that fewer people will be able to purchase them, and producers will have a smaller market in which to sell or will have to sell at reduced profit margins. This type of market swing is difficult to predict and can lead to great uncertainty. All model results should be used carefully.

*Wishlist*

There is at least one dataset that we wish we had and that would be the yearly rainfall totals. While we had access to the level of water stress data highlighting the scarcity of water, that data was extremely limited in timespan and did not show much variability. It also did not only consider climate change but also includes other changes such as fluctuating demands from human needs. As our populations increase, we require more water to sustain our bodies and our lifestyles. Having rainfall data more directly shows rainfall trends rather than indirectly accessing it as we attempted here.

If we were to run this analysis again, we would try to incorporate this variable, but also attempt to look at the interactions of variables when making predictions rather than looking at each variable on its own.

# References

California, A. B. (2016, June). *California Almond Industry Facts.* Retrieved from almonds.com: https://www.almonds.com/sites/default/files/2016\_almond\_industry\_factsheet.pdf

*Climate Change Indicators: Weather and Climate.* (2022, August 1). Retrieved from www.epa.gov: https://www.epa.gov/climate-indicators/weather-climate

*Crops and Livestock Products Data*. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/QCL

*Food and Agriculture Organization*. (2023). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data

*Producer Price Trade Data - Crops and livestock products*. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/TCL

*Producer Prices Data*. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/PP

*SDG Indicators*. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/SDGB

*Temperature Change on Land Data*. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/ET

Zelman, K. M. (2022, November 10). *Health Benefits of Almonds.* Retrieved from www.webmd.com: https://www.webmd.com/diet/health-benefits-almonds

*Appendix*

***1: Variable Trends Over Time***

A graph showing the temperature change

Description automatically generated A graph showing the growth of the stock market

Description automatically generated

A graph showing the growth of the year

Description automatically generated A graph showing export value over time

Description automatically generated

A graph of a number of years

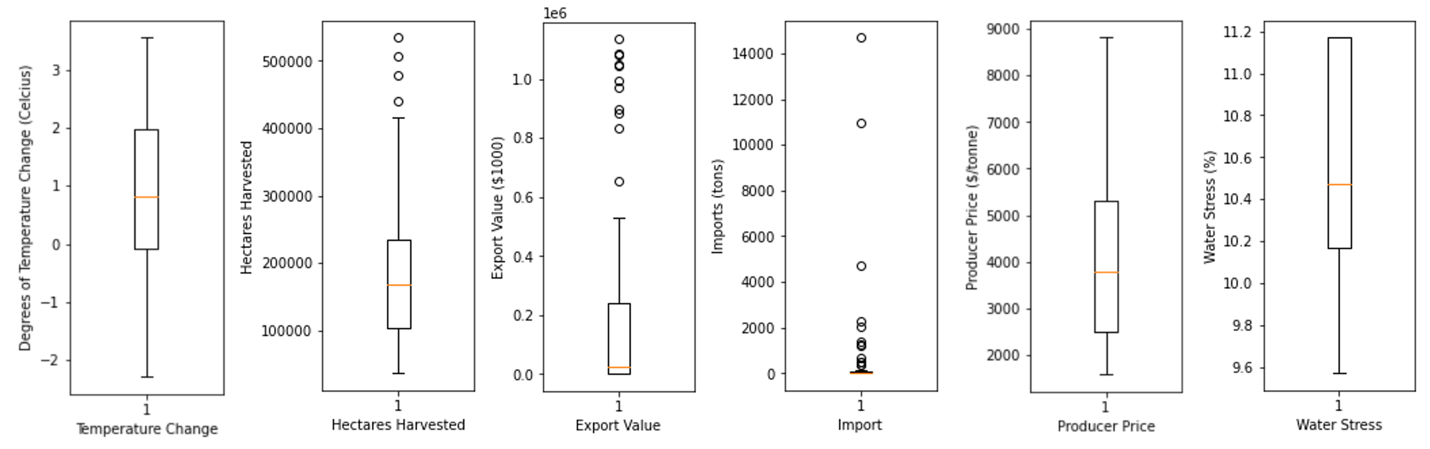
Description automatically generated

A line graph with numbers and text

Description automatically generated A graph of water stress

Description automatically generated

***2: Boxplots for Examination of Potential Outliers***



***3: Correlation Matrixes for Data Subsets***

A graph of heatmap

Description automatically generated A graph of a graph with different colored squares

Description automatically generatedA graph of a heatmap

Description automatically generated

***4: SARIMAX results for each variable.***

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

***5: Comparing each variable’s actual values to the models’ predicted values.***

A graph with lines and text

Description automatically generated A graph with lines and text

Description automatically generated

A graph of a graph with different colored lines

Description automatically generated A graph with a line and a black and orange line

Description automatically generatedA graph with lines and text

Description automatically generated A graph with lines and numbers

Description automatically generated

***5: Comparing each variable’s error metrics from each model.***

|  |  |
| --- | --- |
| Set One ARIMA Model | Set Two ARIMA Model |
|  |  |

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated