Predicting Sweetpotato Yield and Production in Areas Susceptible to Climate Change Effects

Lavanya Chawla (lc3601)

December 15, 2022

Abstract

The agricultural industry is greatly affected by climate change. As the climate changes, the quantity and nutritional quantity of crops wavers. This has a dire impact on global food production and food security. However, some communities are more at risk of these effects. Similarly, some crops are better suited for future climates. Both these communities and crops are underrepresented in research. This paper presents how Machine Learning can be used to predict yield and production of under researched crops grown as staple foods in developing countries such as Sweetpotato using open source and easily accessible agricultural and climate data. We found that a Random Forest model using agricultural data and annual climate aggregates perform the best at predicting yield while adding climate variables to agricultural data does not help predict production quantity. We also used our model to predict how the yield of the top Sweetpotato producers would change with temperature increases, assuming all other variables remain constant, and found that for most countries, the yield would increase.

1 Introduction

Crop yields are affected by weather and climate patterns. With rising temperatures, rising sea levels and high probability of extreme events, the agriculture industry will struggle to meet food demands, escalating food insecurity. Given this, in this paper, we built a Machine Learning model to predict Sweetpotato yield and production for increasing temperatures. In the following sections, we describe the importance of Sweetpotato in underrepresented communities and developing countries and summarize our data collection methods. We also describe the models we used in our analysis and lastly outline our results and next steps.

2 Background

Climate change brings with it rising temperatures, rising sea levels and the high chance of extreme events such as droughts, floods and heatwaves throughout the globe. While these changes will not affect all regions, nations or communities equally, they would have a dire impact on the global food system and food security. These changes, along with an increasing global population, will further burden a food supply chain that already leaves 2.3 billion food insecure. As temperatures rise, rainfall becomes erratic and atmospheric carbon increases, the quantity and nutritional quality of our current agriculture production will falter. The lack of access to good, nutrient-rich food would aggravate health problems. The risk of food insecurity and increase in illnesses will affect communities unevenly, with those in Asia and Africa, being affected the most. In these communities, lower-income families, women would feel the effects the most. Moreover, farm owners and workers would loose their livelihood, thus further slipping into food insecurity (FAO, 2020).

As the quantity and quality of food products decreases, and demand for food increases, we may face the need to clear more forests for croplands and use more intensive irrigation systems and chemical fertilizers. This would increase the agricultural industry's contribution to global carbon emissions, which would further contribute to climate change. In order to avoid this situation, we need to institute more efficient and sustainable agricultural practices and turn our attention to nutrient-rich, resilient crops. Manner and Etten (2018) highlight how nutrient-rich crops that are better suited to changing climates, such as Sweetpotato, Potato and Wheat are under-researched. Much of the agricultural research done requires long and expensive experiments using soil samples or sensor data that is not easy to gather. Moreover, much of the research conducted focuses on either a few crops that are commonly used in developed countries or only considers and studies growing areas in developed countries.

Some of the communities most at risk of climate change impacts and underrepresented in research are lower-income communities in Least Developed Countries in Africa, Asia and Pacific (United Nations, n.d.). In our analysis, we wanted to focus on an under researched crop that is a staple in these communities and is well suited to future climate. Our chosen crop, the Sweetpotato is considered a vital crop in developing countries but is a supplementary food in developed countries. In many of these developing countries, Sweetpotato is grown at a small or subsistence level and has been beneficial in times of famine and food insecurity. The nutrient-rich crop has a quick maturation period, can be grown in a varied range of climates and is one of the first crops planted following a natural disaster, The crop can easily be grown in fertile tropical solid, without the need for fertilizers, irrigation and much constant care, These communities also follow different growing and harvesting periods and plant a large range of varieties (Loebebstein and Thottappilly, 2009).

In this paper, we have attempted to build a Machine Learning model that predicts the yield and production of Sweetpotato in the top 20 producing countries based on climate variables. Since to our knowledge, much of the research related to Sweetpotato production dates either 1980s-90s and late 2000s, and given the nature of Sweetpotato production, we wanted to investigate whether it is possible to reasonably predict Sweetpotato yield and production using open source

production and climate data. While we focus on Sweetpotato here, we also wanted our analysis to be replicable across different products and regions. The subsequent sections detail our data sources, methodology and results.

3 Data Collection

For our analysis, we used easily accessible data from the following sources:

1. Food & Agricultural Organization of the United Nations (FAO) dataset on Crops and Livestock: Our data consisted of information on Country, Yield, Production and Area Harvested for Sweetpotato for the top 20 Sweetpotato producing countries from 1979 to 2020. We hope by including Country as a variable, we are able to account for differences in soil type and technological advancements (FAO, n.d.).
2. Climate data from AgERA5: ERA5 provides global, hourly estimates of a large number of climate variables. AgERA5 is a subset of this dataset focused on climate variables used in agricultural studies and provides daily estimates at a 10 km x 10 km spatial resolution from 1979 onwards (Boogaard et. al, 2020) From this dataset, we obtained information about
   (a) air temperature at a height of 2 m above the surface
   (b) daily cloud cover,
   (c) wind speed at a height of 10 m above the surface,
   (d) daily snow thickness,
   (e) water vapour's daily contribution to the total atmospheric pressure,
   (f) dewpoint temperature at a height of 2 m above the surface.

   We ensured that the climate data was collected for an approximate region surrounding the Sweetpotato producing regions in each country. Since no current geo-referenced database of Sweetpotato growing regions is available, we used Huaccho and Hijman's (2000) analysis from 1980s - 1990s to narrow down our production zones. We also assumed that these regions have not changed between 1979 and 2020. Appendix A provides a map with these Sweetpotato production zones.

For our analysis, we attempted to further divide our countries into smaller regions. However, due to the space and time complexity required to process and store the climate data for this task, we instead decided to perform our analysis at the country level.

4 Exploratory Data Analysis

Before building a prediction model, we performed EDA on our dataset. Our dataset contained 791 data points, covering 20 countries and a maximum of 42 years each (only Malawi - 7 years and Ethiopia - 28 years had less than 42 years). We had 5 non-climate variables (Year, Country, Production, Yield, Area harvested) and 6 climate variables (cloud cover, snow thickness, vapour pressure, 2m temperature, 10m wind speed, 2m dewpoint temperature). We had three type of aggregations for climate data - monthly, seasonal (groups of three) and annual. By plotting our monthly data for all time, we found every variable had a general seasonal trend. We also found that 2m dewpoint temperature and vapour pressure has a high correlation of 0.97 and thus, vapour pressure was dropped from our subsequent analysis. We also found that Production and Area harvested had a high correlation 0.98 and thus Area Harvested was dropped when predicting Yield. Figures from our EDA are available in Appendix B-D.

5 Modelling

In this section we summarize the models we built to predict both Yield and Production and the variables we choose. We used the sklearn package in Python to build and test our models. Due to the relatively small size of our dataset and the nature of our problem, we choose the following modelling techniques:

1. Linear Regression: fits a linear model to the training data such that the residual sum of squares is minimized. This was our base model that we compared our other models to.
2. Support Vector Regression: predicts target variable by fitting the error rate to a certain threshold. This method generally performs well on high dimensional datasets.
3. Random Forest: an ensemble model of multiple decision trees built on subsets of dataset sampled with replacement which helps reduce high variance inherent in decision trees.
4. Lasso Regression: minimizes a penalized residual sum of squares using the L1 norm and sets those variables that do not have any effect towards 0. This method is useful for feature selection.
5. Ridge Regression: minimizes a penalized residual sum of squares using the L2 Norm and shrinks the value of variables that do not have any effect towards 0. Thus, this method is useful for feature selection.
6. Grid Search Cross Validation: helps find optimal values for a given model by trying all combinations of hyperparameters.
7. Randomized Search Cross Validation: helps find optimal values for a given model by using a sample subset of all combinations of hyperparameters.

To predict Yield, we performed modelling techniques 1-3 on the following set of variables:

1. Country and Production - YM1
2. Country, Production and annual aggregates of climate variables - YM2
3. Country, Production and three seasonal groups of climate variables - January to March, April to June, July to September - YM3
4. Country, Production and monthly aggregates of climate variables - YM4

Similarly, to predict Production quantity, we performed modelling techniques 1-3 on the following set of variables:

1. Country, Yield, and Area harvested - PM1
2. Country, Yield, Area harvested and annual aggregates of climate variables - PM2
3. Country, Yield, Area harvested and three seasonal groups of climate variables - January to March, April to June, July to September - PM3
4. Country, Yield, Area harvested and monthly aggregates of climate variables - PM4

Due to the large number of variables in the last variable configurations YM4 and PM4, we performed feature selection using Lasso Regression and Ridge Regression.

Based on preliminary runs, we found that the Random Forest performed best across all variable configurations and we performed Grid Search and Randomized Search Cross Validation to find the best model configuration. Here, we've defined the best model to be one that reduces the R^2 score and Mean Square Error. (Due to large numbers, the MSE is not shown in this report) .

We then used our best model to predict how Yield would change in different countries as Mean Temperature changes in 2046-2065 and 2081-2100 under different scenarios presented in Table

1 (IPCC, 2014). Since the temperature change is relative to mean temperature in 1986-2005, we created new data points for each scenario by aggregating the 1986-2005 values for each variable. We kept all variables at the 1986-2005 average, except for the 2m temperature variable to which we added the mean temperature change.

Scenario	2046-2065	2081-2100
RCP2.6	1	1
RCP4.5	1.4	1.8
RCP6.0	1.3	2.2
RCP8.5	2.0	3.7

Table 1: R^2 Score for all models

6 Results

In this section, we present our findings from the models introduced in Section 5: Modelling. Table 2 provides the R^2 Score of all the models considered. For predicting Yield, the Random Forest model using Country, Production, and annual aggregates of climate variables perform best. Interestingly, when predicting production, adding climate variables does not help much in terms of the R^2 score.

We also used the Random Forest model using Country, Production, and annual aggregates of climate variables to predict Yield change in different countries as Mean Temperature changes in 2046-2065 and 2081-2100 under different scenarios. Figure 1 shows our predictions. We decided not to predict Production for these scenarios as adding climate variables to our models did not show much improvement.

7 Discussion

Based on the models discussed in the previous sections, we find that a Random Forest performs better than Linear Regression and Support Vector Regression in predicting Sweetpotato yield. Interestingly, Support Vector Regression performs bad irrespective of the set of variables used, and even performs arbitrarily worse in some cases. This may imply that the model is unable to identify the trend of the underlying distribution. Also, the models suggest that annual climate variabls help predict yield better than seasonal or monthly variables.

We also found that when predicting Sweetpotato production quantity, adding climate variables does not help much. This may imply that either adding climate variables leads to overfitting or Sweetpotato production quantity is less affected by climate as compared to the country (soil type

Model	Predicting Yield				Predicting Production			
	YM1	YM2	YM3	YM4	PM1	PM2	PM3	PM4
Linear Regression	0.217	0.611	0.641	0.641	0.974	0.976	0.98	1
SVR - Linear	-0.067	-0.053	-0.045	-0.447	-0.079	-0.079	-0.797	-0.031
SVR - Polynomial	0.079	-0.054	-0.047	-0.067	-0.079	-0.797	-0.079	-0.067
SVR - RBF	-0.071	-0.071	-0.071	-0.071	-0.079	-0.797	-0.079	-0.071
Random Forest	0.941*	0.952*	0.927*	0.921	0.995*	0.994*	0.994*	0.996
Ridge Regression				0.641				0.999
Lasso Regression				0.641				0.999

* indicates the value is obtained after performing Grid Search Cross Validation.

Table 2: R^2 Score for all models

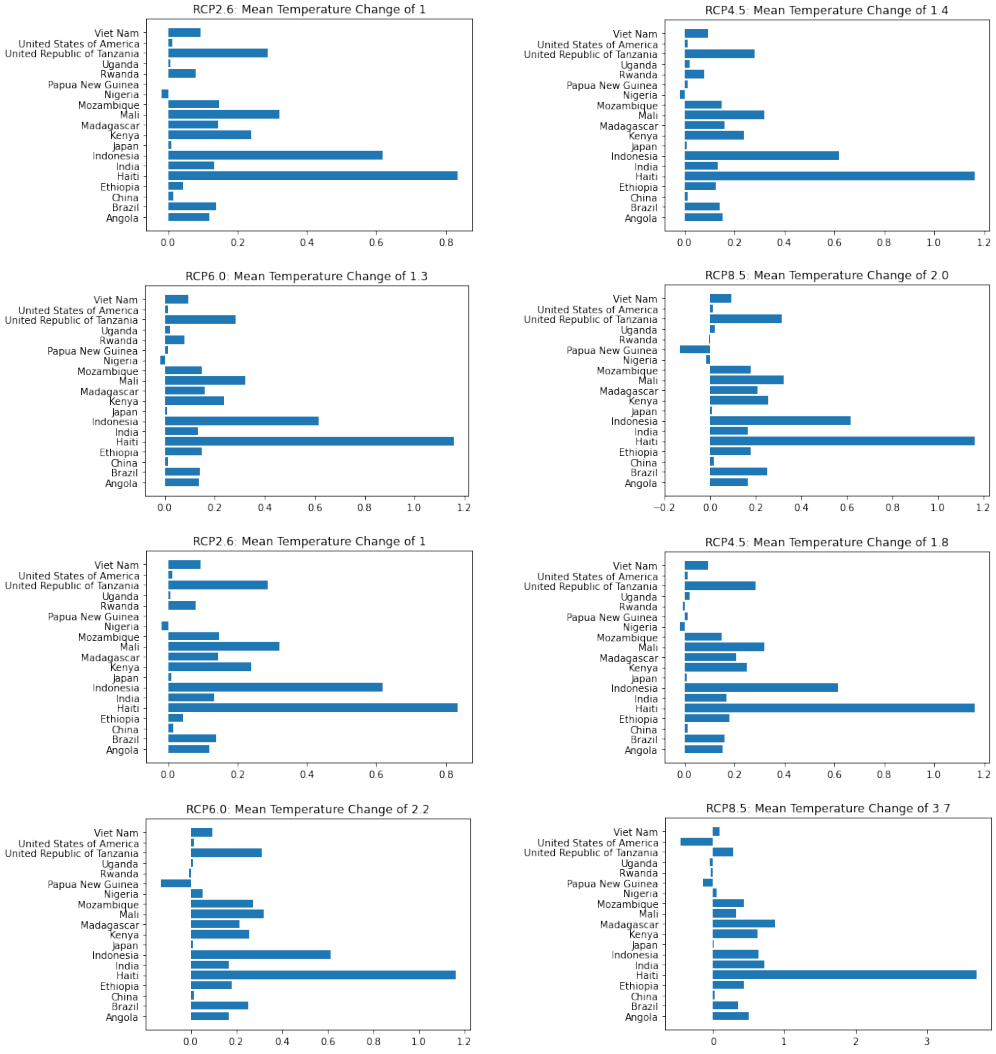


Figure 1: Yield Change for different scenarios in a) 2046-2065 and b) 2081-2100

and technological advancements etc), yield and area harvested or that other climate variables that we have not considered in our analysis exist.

Another interesting finding was that change in mean temperature does not affect change in yield for every country or for every time period in the same way. Some countries are predicted to have no change or decrease in yield (such as Nigeria, Papua New Guinea, United States of America in some time ranges) while some countries always show an increase. This may indicate how different countries will be affected by the same change in global mean temperature. The general increase in Sweetpotato yield suggests the crop's suitability to changing climates and may suggest its increasing relevance in these communities. Interestingly, we predicted an increase in yield for some of the most at-risk and vulnerable countries such as Haiti, Mali, and Madagascar (United Nations, n.d.).

However, since we do not have accurate knowledge about future production, yield or area harvested and have assumed all other variables other than 2m temperature to be constant, these predictions may not be accurate. Moreover, the mean temperature after increase ranges from 285K (Japan, Scenario RCP2.6, 2045-2065) to 305K (Mali, Scenario8.5, 2081-2100) while in our training dataset, the values of 2m temperature ranges from 269K to 301K and the temperature range for each country is different (see Appendix E). Since we are extrapolating here using a Random Forest, our predictions may not be accurate for very high values. We are also not accounting for other changes in climate, such as precipitation and exposure to sunlight, and rising sea levels in our analysis which may affect Sweetpotato yield in the future.

8 Conclusion

In this paper, we tested various Machine Learning models to predict Sweetpotato yield and production quantity in the top 20 Sweetpotato producing regions based on open source agricultural and climate data. We found that a Random Forest model using agricultural variables and annual aggregates of climate variables performs the best when predicting yield while adding climate variables does not help in predicting production quantity. We used the yield prediction model we built to predict Sweetpotato yield in different temperature change scenarios. Through this research, we hope to highlight how Sweetpotato may be an important crop in the future, especially for low-income communities and developing countries.

Our dataset consisted of only 20 countries over a limited time period for one crops and thus was relatively small. While we are able to get a good model and reasonable predictions, this analysis should be carried out on more regions and with other crops. Due to the nature of climate dataset we used, we had to aggregate the climate variables temporally and spatially which may have affected our analysis. We also used a limited range of climate variables - adding other climate variables may help in improving these predictions.

References

[1] Boogaard, H., Schubert, J., De Wit, A., Lazebnik, J., Hutjes, R. & Van der Grijn, G. (2020). *Agrometeorological indicators from 1979 to present derived from reanalysis, version 1.0.* Copernicus Climate Change Service (C3S) Climate Data Store (CDS). 10.24381/cds.6c68c9bb

[2] FAO. (n.d.). *Crops and livestock products*. https://www.fao.org/faostat/en/#data/QCL

[3] FAO, IFAD, UNICEF, WFP & WHO. 2022. *The state of food security and nutrition in the world 2022. Repurposing food and agricultural policies to make healthy diets more affordable*. Rome, FAO. https://doi.org/10.4060/cc0639en

[4] IPCC. (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland

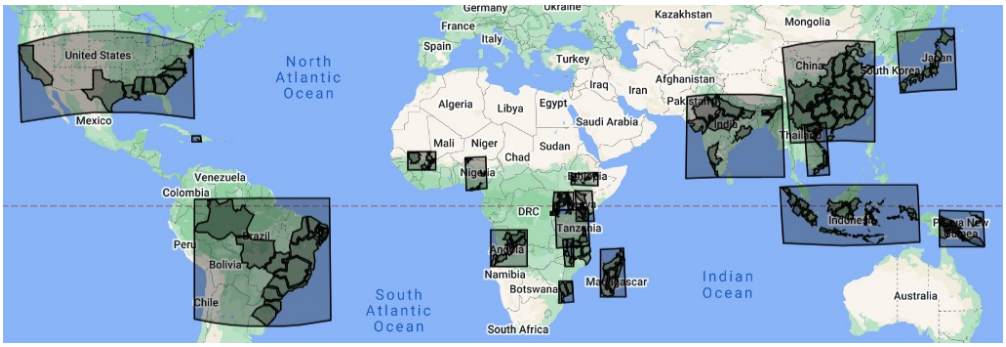
[5] Huaccho, L. & Hijmans, R.J. (2000). *A geo-referenced database of global sweetpotato distribution. Production systems and natural resource management department working paper no. 4*. International Potato Center. Lima, Peru.

[6] Loebenstein, G. & Thottappilly, G. (2009). *The sweetpotato*. Springer Dordrecht. https://doi.org/10.1007/978-1-4020-9475-0

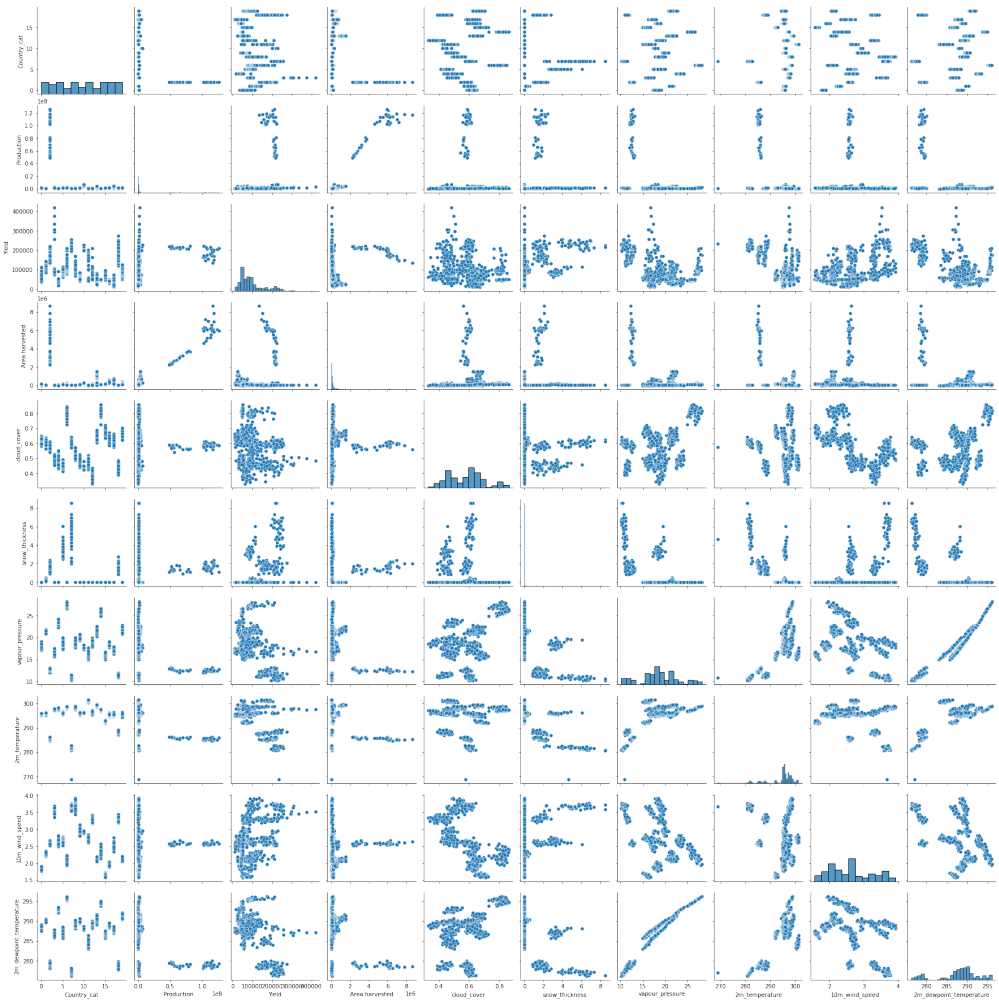
[7] Manners, R., & van Etten, J. (2018). Are agricultural researchers working on the right crops to enable food and nutrition security under future climates? Global Environmental Change, 53, 182–194. https://doi.org/10.1016/j.gloenvcha.2018.09.010

[8] United Nations (n.d.). *Profiles of LDCs | Office of the High Representative for the Least Developed Countries, Landlocked Developing Countries and Small Island Developing States*. https://www.un.org/ohrlls/content/profiles-ldcs

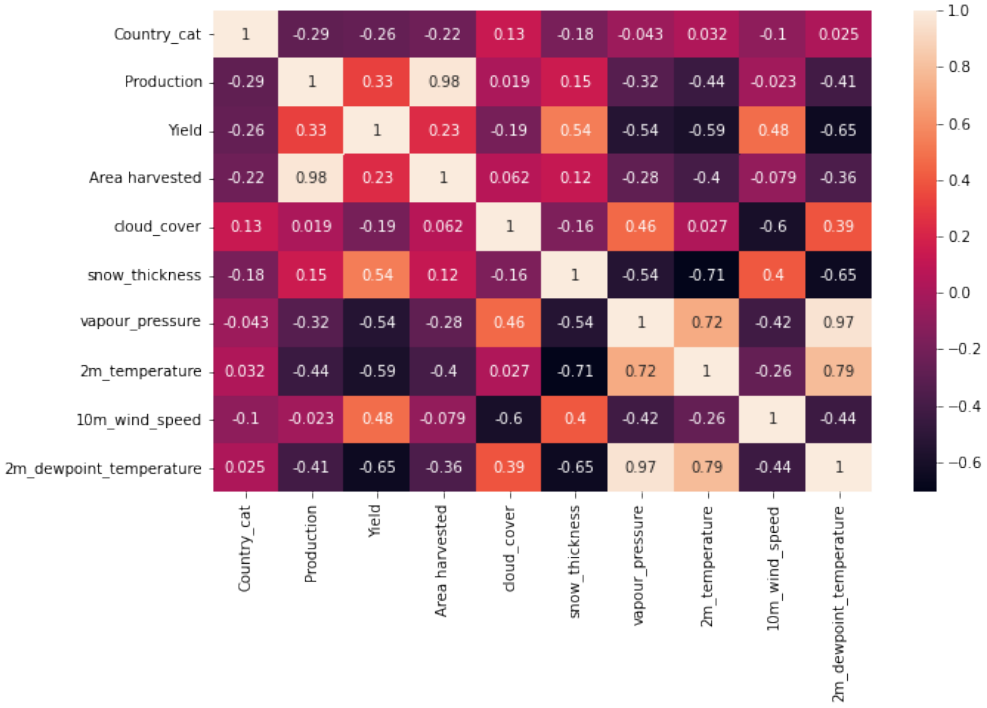
A Sweetpotato Production Zones in Top 20 Producing Countries



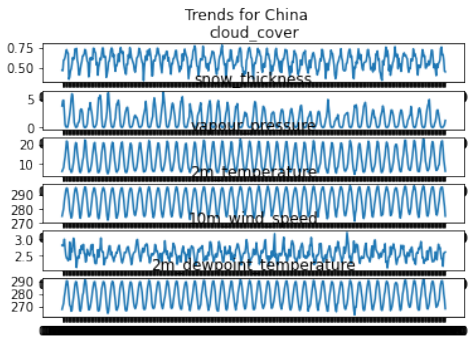
B Univariate and Bivariate Distributions of Variables, climate variables aggregated annually

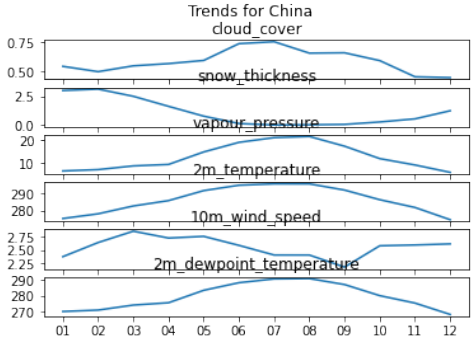


C Correlation between variables, climate variables aggregated annually



D Trend over a) all time and b) over a year for China for monthly aggregated variables





E 2m Temperature Ranges by Country