The Impact of Weather and Climate Change on Global Food Supply

K. A.
National College of Ireland
Dublin, Ireland
X18xxxxxx

Billy Hanan
National College of Ireland
Dublin, Ireland
X18179797

N. L.
National College of Ireland
Dublin, Ireland
X18xxxxxx

A. G.
National College of Ireland
Dublin, Ireland
X18xxxxxx

ABSTRACT

Rice is in the top three crops consumed in the world and as a result, its production is crucial for global food supplies. In this paper, we investigate the factors that affect rice yield. In particular, we are interested in the relationship between rice yield and climate. Using annual data from 105 countries covering 1991-2015, we cluster these countries into three climatic categories and then build linear regression models to predict the rice yield.
We discover in all cases that fertiliser usage and GDP per capita are positively correlated with yield. However, in countries not having a temperate climate, the temperature has a strong negative correlation with yield. As such, our models suggest that if global warming continues, the vast majority of rice producing countries may struggle to meet their food requirements unless offset by increasing investment in their agricultural infrastructure.

I. *INTRODUCTION*

There are fifteen crops that provide nearly 90% of the world's plant-based food energy intake, with rice, maize, and wheat comprising nearly two thirds of the global diet. Of these three, rice is a primary source of nutrition for almost half of humanity with it being the third highest consumed globally. The ideal conditions for rice growth are well known and established to be 16ºC - 27ºC temperature with 100 cm – 200 cm average rainfall.

In 2017, NATO issued a warning to the G7 that climate change would trigger food shortages; in its 2018 report, the UN's Food and Agriculture Organisation ranked climate change as the leading cause of global hunger. Food security is a major concern for world leaders and disruptions to rice supply can lead to both economic troubles and social unrest [1] [2]

Figure 1. Rice Consumption and Poverty

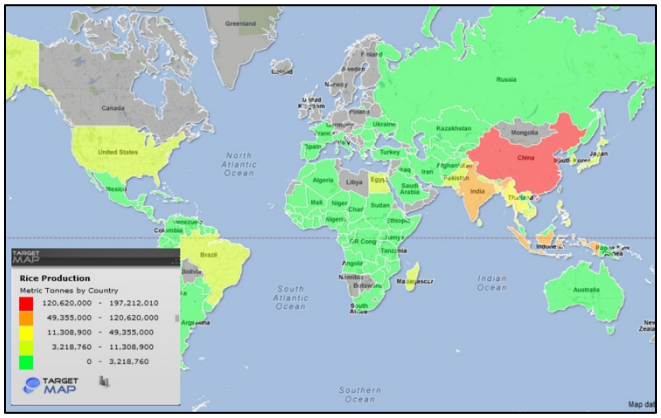

Figure 2. Rice Production by Country

In this paper we investigate the relationship between rice yield and climate in an attempt to understand how climate change may impact rice supply in the future years. We do this using annual agricultural and weather data relating to 105 countries covering the years 1991-2015. We begin our analysis by first obtaining a climate profile for each country, obtained simply by averaging the quarterly temperature and precipitation values across the 25 years. The countries are then clustered on this summary weather data and linear regression models built on each cluster.

In section II we provide a brief review of the related literature. In section III, we provide details of the data

sources and describe the data mining techniques applied to the processed data. We report and discuss the results of this work in section IV, before offering some final conclusions and suggestions of future work in section IV.

II. LITERATURE REVIEW

Due to its importance as a staple food in much of the developing world, there has been much academic attention devoted to the effects of weather and climate on rice growth. Wang and Hijmans [3] note in their research on China that increased temperatures in Southern China resulted in lower rice yields, whereas higher temperatures in Northern China resulted in higher rice yields. In the findings of Shrestha et al [1], they found that in central Vietnam, lower winter temperatures had an adverse effect on rice production and higher summer temperatures had a favourable impact. [3]
In a similar study Caruso et al [2] discuss rice plants in Indonesia and how they are disadvantaged due to Indonesia's high air temperatures which reduces photosynthesis and increases the respiration rate of the plants. These factors combined result in a shorter cultivation period and together with global warming has led to decreasing production levels of rice in Indonesia. Yang et al [4] suggest that moderate alternative wetting and drying (AWD) techniques of rice lands can improve rice quality and does not inhibit photosynthesis [2] . These research articles highlight the importance of temperature in rice production.

Li et al [5] discuss the geospatial variation of rice production across China in similar research to Wang and Hijmans [3]. They note that temperature and rainfall are factors that can influence rice production. However, they note that rainfall sparsity doesn't impact China's rice production as much as other countries as they have heavily irrigated lands. The impacts of climate change on rainfall is forcing rice farmers to schedule their irrigation patterns to optimise their results as shown in research carried out in Iran by Amiri et al [6] and Saliu et al [7] have similar findings when looking at Nigeria's rice crop and its vulnerability to heavy rainfall and the effects of climate change. Additionally, they find that excess rainfall can have adverse impacts on the rice crop, this will be something that will be considered when carrying out this research. In their research they create 'vegetation groupings' to run their analysis against which is something inspired our clustering approach in this project.

Food pricing for rice crops has been shown to have a major impact on society in areas where rice is a major economic crop [7]. Caruso et al [2] look at the impact that increased food prices and food security concerns have on Indonesia and how it leads to violence in times of uncertainty. In contrast Heady et al [8] looks at the concept that increasing global food prices benefits less advantaged countries. This is proposed as the majority of workers in these countries are employed in the farming industry so increased global food prices increase their income. The increased demand and production of rice can therefore be said to benefit the people of disadvantaged countries in Heady et al [8] opinion.

Similarly, the work of Davis et al [9] discusses how Sri Lanka has made a move toward self-sufficiency of food production with regard to rice production due to the uncertainty of global food trade. It is discussed how the projected population growth and increasing temperatures in Sri Lanka will bring with them increased food price uncertainty on one hand but also economic development through increased production and self-reliance through rice production. While this technique might be plausible for Sri Lanka Oort et al [10] discuss how many African countries do not have the capability to be self-sufficient to meet the rice consumption levels they need. An example of this is the fact that since 2000 they have increased their harvested rice area by 32%, the only plausible option would be to increase irrigation the areas to allow the African countries examined to be self-sufficient. This research project will aim to build on the work discussed to investigate if increasing climate change impacts rice production and in turn impacts rice pricing.

The prediction of rice yield has been examined in many studies. The impact of climate change on agriculture means that accurate prediction of future crop yield is important for low-lying countries like Bangladesh. In Hossain et al [11] they implement the WPSRY (Weather-based Prediction System for Rice Yield) approach to forecast rice yield. This involves estimating rice yields using Support Vector Regression (SVR) that uses as inputs predicted weather parameters from a neural network in combination with current agricultural data. The ultimate simulation produced demonstrates that the WPSRY approach achieves promising prediction accuracy. In a similar study Ramesh and Vishnu Vardhan [12] discuss how the agrarian sector in India is facing a rigorous problem to maximize the crop productivity. More than 60 percent of the crop still depends on monsoon rainfall. Recent developments in Information Technology for agriculture field has become an interesting research area to predict the crop yield. The problem of yield prediction is a major problem that remains to be solved based on available data from their findings.
Data Mining techniques are needed to accurately predict rice yield and its influences, and this research will be built on in within this project bearing in mind the research examined.

III. METHODOLOGY

The overall process that we follow can be summarized in Figure 3: This is the KDD approach of data mining.

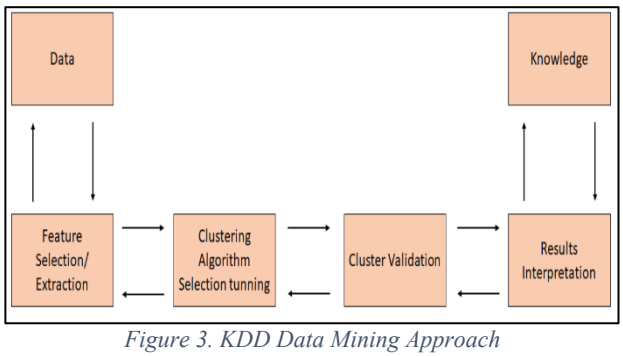

Figure 3. KDD Data Mining Approach

Though the above diagram refers specifically to clustering, it is applicable also to supervised learning techniques. We shall outline each step in the process below.

A. Data Collection and Feature Selection

The time period chosen to do this research project on was 1991- 2015 To get the most accurate and reliable data for this project, data from a range of online resources was examined such as timeanddate.com, Wunderground, foodsecurityportal.com and data.un.org (Links given in appendix B). After examining the publicly available data, rice yield, fertilizer and land data was collected from the WRS (World Rice Statistics) online query engine. A comparison between rice yield at the start of the research period (1991) and the end of the research period (2015) is shown geographically in Figure 4 and Figure 5 respectively. The trend shows an increase in rice yield production in the time period. The rainfall and temperature data were collected from the Climate Change Knowledge portal, this portal is a legitimate source for key indicators which impact past, current and future climate change vulnerabilities. GDP per Capita data was also retrieved from the official UN data repository and to help with the merging of data, ISO country codes from the International Organisation of Standardization website were
obtained.

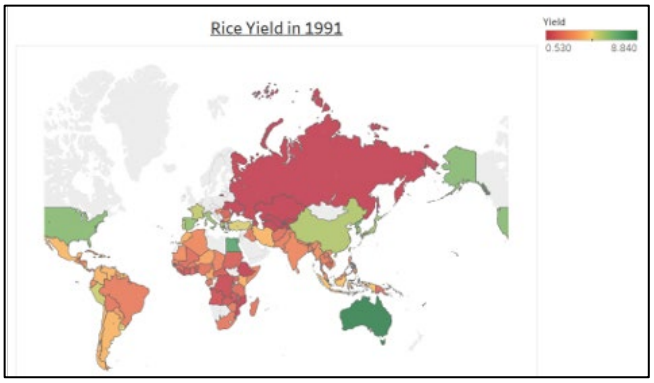

Figure 4. Raw Rice Yield Data in 1991

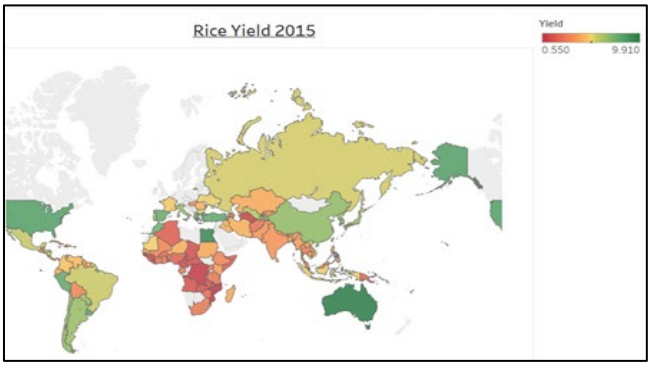

Figure 5. Raw Rice Yield Data in 2015

B. Data Cleansing

Although the data quality is reasonably sound, there were a few issues. Namely,

1. Missing values in fields like Yield for some of the years for some of the countries.
2. Temperature for Russia was around 30 degree for June and July for years 2001 and 2004. Russia never has that much mean temperature in any part of the year.
3. Approximately 20 countries have more than 70% of temperature and rainfall data missing.
4. For some countries we had no yield data.

1) Data Quality strategy:

a. For case 1 above, we imputed mean value of yield for that country over the period of 25 years.
b. For case 2, we imputed the mean monthly temperature value of the country. We then replaced any anomalous values with the mean Jun and July temperature.
c. For case 3 and case 4, we have removed the records. Out of 129 countries extracted from source, we were left with 105 countries with full data covering the years 1991-2015.

We used mainly the 'dplyr' package for handling data quality issues.

Script Snippet:

```
colSums(is.na(crop3))
crop3 <- crop3 %>% group_by(ISO) %>%
mutate(FERT=ifelse(is.na(FERT),mean(FERT,na.rm=TR
UE),FERT))
crop3 <- crop3 %>% group_by(ISO) %>%
mutate(HARV_AREA=ifelse(is.na(HARV_AREA),mean(HAR
V_AREA,na.rm=TRUE),HARV_AREA))
crop3 <- crop3 %>% group_by(ISO) %>%
mutate(LAND_AREA=ifelse(is.na(LAND_AREA),mean(LAN
D_AREA,na.rm=TRUE),LAND_AREA))
crop3 <- crop3 %>% group_by(ISO) %>%
mutate(YIELD=ifelse(is.na(YIELD),mean(YIELD,na.rm
=TRUE),YIELD))
crop3 <- crop3[!is.na(crop3$FERT), ]
colSums(is.na(crop3))
```

2) Feature Engineering/Aggregation:

a. From the raw daily weather source data, we calculated average monthly temperature and rainfall for all 105 countries from 1991-2015. We subsequently calculated quarterly (Jan-Mar, etc.) averages from this monthly data to obtain mean quarterly temperature and quarterly rainfall data.
b. We converted the units used in the variable Irrigated Land from per Sq Km to per hectares.

3) Merging

Finally, ISO Country codes were added to data files that only had a name to identify the country. Data files were subsequently merged using the ISO code as the key.

C. Model Training: Unsupervised Algorithms

Two approaches to clustering, both using k-means, were applied in this project. In the first approach, countries were clustered on the variables Total Land Area, Irrigable Land Area, Fertilizer Usage and the quarterly temperatures and precipitation values. The resulting clusters were then

analysed in order to discern any significant differences in their rice yield output.

In the second approach, countries were clustered based solely on their monthly temperature and rainfall variables. To do this, we first calculated the mean of every weather variable over the 25 years for each country. This essentially summarizes the seasonal climate of every country.

Country	Q1 □C	Q2 □C	Q3 □C	Q4 □C	Q1 mm	Q2 mm	Q3 mm	Q4 mm
China	-3.60	13.28	17.74	0.67	17.9	65.88	87.84	20.41
India	20.46	29.24	26.85	21.64	14.08	79.92	216.32	36.0

Figure 6. Example of Quarterly Averages

We show in Figure 6 for illustrative purposes examples of averages of the quarterly temperature and precipitation (rather than the monthly values that were used in the clustering). As both these quantities are on different scales, they were separately rescaled before clustering. That is, all the temperature values were pooled together, and their median and interquartile range calculated. Likewise, for rainfall. The temperature and precipitation values were then standardised using these statistics.

Clustering in both the approaches above was then performed by executing the k-means algorithm with the squared Euclidean distance metric. Note that based on the gap statistic, the value k=3 was chosen in both cases. Please see the results section for an analysis of the clusters obtained.

D. Modelling

Linear regression was used to investigate the relationship of the annual rice yield of a country with the other variables in our data set. The independent variables were checked for correlations. Significant correlations (|r|>0.7) were found between the monthly temperatures and, to a lesser extent, monthly precipitations. As such, the monthly weather variables were discarded for the purposes of the linear regression analysis and the quarterly variables used instead. The following set of 10 independent variables were thus used for modelling purposes:

Year (YEAR), harvested area (HARV_AREA), fertiliser usage (FERT), GDP per capita (GDP), mean Q1 temperature (AVG_Q1_TEMP), mean Q3 temperature (AVG_Q3_TEMP), mean Q1 rainfall (AVG_Q1_RF), mean Q2 rainfall (AVG_Q2_RF), mean q3 rainfall (AVG_Q3_RF) and mean Q4 rainfall (AVG_Q4_RF)

Note that the average Q2 and Q4 temperatures were not used because they still had very strong (|r|>0.8) correlations to their neighbouring quarters.

Linear regression was initially performed on the entire data set of 105 countries. With each country having 25 data points (one for each year), 2625 data points were thus used in this regression. Note that we employed a best subsets regression strategy. That is, every possible combination of independent variables was fitted to the available data and the model producing the highest adjusted R^2 value was chosen as the best model.

The same approach was also implemented on the three clusters of countries. As such, you will find the details of four models in the results section. Standardised regression coefficients are also quoted there to aid in the interpretation of results.

IV. Results and Discussion

A. Cluster Analysis

105 countries were clustered into three groups. The resultant clusters are shown in the global map of Fig. 10. The primary variable of interest is the rice yield and in Fig. 7c, we can see that the mean rice yield in each grouping differs markedly. We shall henceforth refer to these three clusters as the high, medium and low yielding clusters.

In Fig. 7a and 7b, we see that countries in the high yield cluster (cluster 1) have an average Q1 and Q4 temperatures in the range of 5 to 8-degree Celsius and Q3 and Q4 average temperature in range of 15 to 20 degrees and average quarterly rainfall between 50 to 60 cm. That is, they have relatively low winter temperatures and low rainfall all year round.

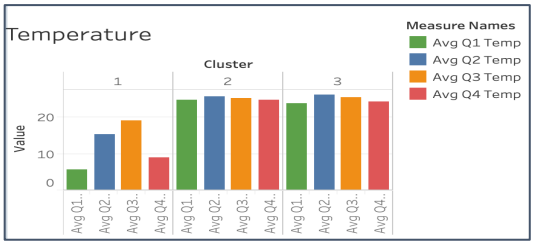

Figure 7a. Average Temperature

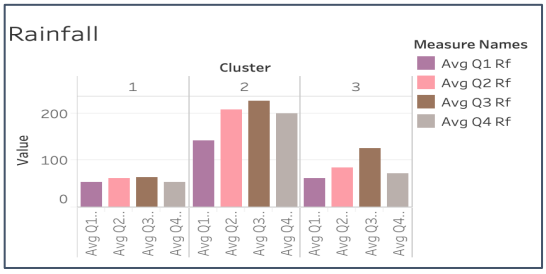

Figure 7b. Average Temperature

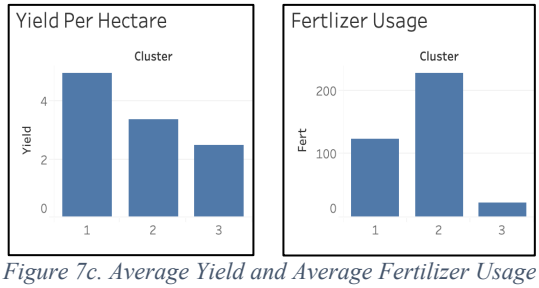

Figure 7c. Average Yield and Average Fertilizer Usage

The low and medium yield clusters (clusters 2 and 3 respectively) have very similar temperature patterns (and have much warmer climates than the high yield countries of cluster 1). The quarterly average rainfall differs markedly between the two clusters, however. Rainfall in the low yield cluster 2 lies between 130 to 200 cm per quarter while it is between 50 to 120 cm in the medium yield cluster - 3 countries. It should also be noted that fertilizer usage in cluster 2 countries is much higher than cluster 3 countries.

As a first approximation, the results suggest that climate and seasonal weather patterns have an impact on rice yield. The colder and low rainfall countries produce high yields while the hottest and wettest produce the lowest yields.

However, to arrive at more definitive conclusions, regression analysis was applied to each cluster group.

B. Regression Results

The table below summarizes the results of the regression analysis performed on our four sets of data. Standardised regression coefficients are quoted and all except those asterisked are statistically significant ($p<0.05$). The accuracy of the fitted model is also given via the root mean square error (RMSE) of the rice yield. This was calculated using 10-fold cross validation.

	All Countries	High Yield Countries	Medium Yield Countries	Low Yield Countries
YEAR	0.096	0.130	0.047*	0.187
HARV_AREA	0.093	0.178	-	-
FERT	0.234	0.283	0.485	0.065
GDP	0.387	0.357	0.381	0.183
AVG_Q1_TEMP	-0.072	-	-0.07	-0.360
AVG_Q3_TEMP	-0.131	-	-0.25	0.035
AVG_Q1_RF	-0.20	0.114	-0.299	-
AVG_Q2_RF	0.180	-0.076*	0.054*	0.240
AVG_Q3_RF	-0.228	-0.064*	-	-0.129*
AVG_Q4_RF	-0.075	0.223	-	-
Adjusted R^2	0.384	0.603	0.555	0.360
Yield RMSE	1.486	1.176	1.423	0.985

It should be noted that two countries, Brunei and Trinidad & Tobago were removed from the low yielding cluster before performing regression. As shown in the figure below, these countries have very high GDP per capita in comparison to the other countries in the cluster and as outliers, were subsequently skewing the regression results.

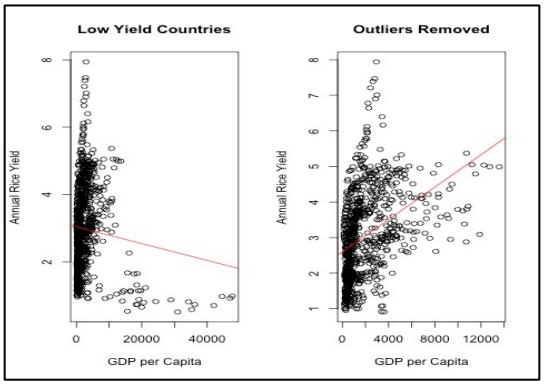

Figure 8. Outlier explained

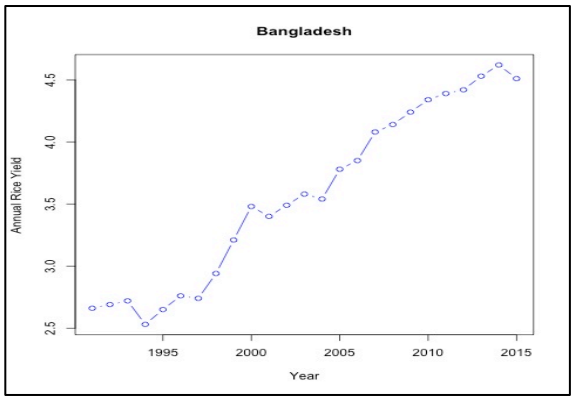

Figure 9. Yield Trend (Bangladesh)

In all models above, we note that the annual rice yield in a country is always positively related to time (YEAR), harvesting area, fertiliser usage and GDP per capita. This intuitively makes sense. Over time, a country gets progressively better at increasing their yield (See the figure 9 for Bangladesh above as an example). A greater harvesting area implies that a country has either specialised in the growth of rice or that the local climate is particularly efficient for its growth, again implying an improved yield. That increased fertiliser usage should produce larger yields is almost a truism. And finally, a larger GDP per capita suggests that the financial resources exist to buy machinery and advanced crop monitoring equipment, build required irrigation infrastructure and generally implement advanced crop management techniques. All will contribute to improved crop yield.

It is however the correlations with the weather variables that are of primary interest to us here. The model fitted to the entire dataset of 105 countries shows that rice yield is negatively correlated to all the quarterly variables other than Q2 rainfall. However, this one size fits all model is the most inaccurate of the four. The purpose of the cluster aggregation was to achieve greater accuracy and improved models. We can see for example that the RMSE of 1.486 for this model is reduced in the three cluster models to 1.176, 1.423 and 0.985. As such we will focus on these cluster models. Let us now consider each cluster separately.

The high yield cluster has one particular characteristic that distinguishes it from the other two. Namely, temperature is not a predictor in the model. This might be explained by the fact that the countries in this cluster have temperate summer climates with a median Q3 temperature of 20.15 Celsius and a maximum temperature of 28.06 degrees. That is, the temperatures lie almost entirely below the 27 degrees that is considered the ideal for rice growing (To be precise, only 5% lie above). Rainfall does have some impact on the rice yield. Oddly though, it is the rainfall in Q1 and Q4 that is significant. This is positively correlated with yield.

The medium and low yield countries suffer from much warmer and wetter climates. For the medium yield cluster, we find that Q1 rainfall and Q3 temperature are strong predictors, both being negatively correlated to rice yield. In the low yield cluster, it is Q1 temperature and Q2 rainfall that are dominant, the former being negatively correlated and the latter positively.

It is understandable that temperature in both these clusters are negatively correlated to yield. The temperatures in these regions are often quite close to the critical upper temperature of 27 degrees. In fact, we find that 25% of Q3 temperatures and 20% of Q1 temperatures in the medium and low yielding clusters respectively lie above this threshold. If the general trend of temperatures were to increase, yields may be impacted.

C. Geographical Mapping of Clusters

The location of the countries in each of our clusters is exhibited in Figure 10 below. Figure 8 Looking at the world map, we can see that countries above tropical lines are in the high yield cluster and countries at equatorial region are more in the medium yield cluster 2. This is quite imperative that weather pattern has an impact on the paddy crop yield.

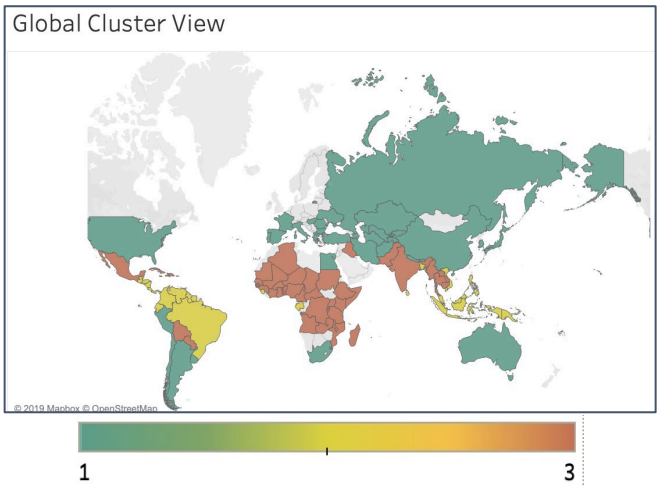

Figure 8. Geographical Distribution of Cluster Groupings

The notable lower average temperatures experienced by Group 1 compared to Group 2 and 3 appears to imply a correlation between increased rice yield and lower temperatures. The results of the clustering imply that Rice producers that are located farther away from the equator have a higher rice yield than those located in closer proximity to it (i.e. group 2 and 3).

D. Potential implications of Results

The strong positive correlation of Q2 rainfall in the low yielding countries may prove problematical in the future if the rainfall were to decrease in these regions. If that is the case, irrigation schemes will need to be set up as suggested in the works of Saliu et al [7] in Africa and Amiri et al [6] in Iran.

Our results suggest that global warming will have minimal impact on the high yielding cluster countries due to their more temperate climates. As such, we are likely in the future to see a migration of rice growing from hotter more tropical climes to areas away from the tropical zones. In fact, this migration is already taking place as detailed in the works of Wang and Hijmans [3] and Li et al [5] who examine the migration of rice paddy fields from southern to northern China. The potential impact of increased rice crop yield through understanding impacting factors will have a positive impact for food security also. This will potentially help to decrease conflicts in areas with food security concerns and will help local economies. This is in agreement with the works of Caruso et al [2] and our findings supports the idea of countries becoming more self -sufficient when producing rice, which has also built on the works of Davis et al [9] and Heady et al [8]. As a result of our findings we can propose methods for governments and NGOs to better prepare for incidents of climate related food scarcity.

V. CONCLUSIONS AND FUTURE WORK

The primary objective of this project was to identify the correlations that exist between weather and rice yield on a global scale. In particular, our interest was driven by the impact of climate change on future rice production. Assuming that both global temperatures and precipitation levels are to continue to rise in the coming years, our results suggest that the rich developed world, that constitute much of the high yielding cluster of countries, are relatively shielded from this trend. The temperatures and rainfall in these countries are currently temperate and can thus likely sustain their rice production.

The picture for the other two clusters of countries is however not so benign. In both clusters, the yield has a strong negative correlation to temperature, which is a result in agreement with Carruso [2] and Wang and Hijmans [3]. Higher temperatures in the future due to global warming will thus impact negatively on the efficiency of rice production in these regions.

However, in the case of the countries in the low yielding cluster, this may be offset by the strong positive correlation of yield with Q2 rainfall. Countries in the medium yielding cluster have no such counteracting force however. Increasing temperatures and increasing rainfall in these regions will likely prove detrimental to their rice yields.

All of the above analysis is of course contingent on the assumption that both temperatures and rainfall will increase in the future. This is too much of a simplistic generalisation of course. Some areas may experience reduced rainfalls, others increased precipitation.

In future work a more thorough analysis would need to be done to identify precisely the types of climatic changes that each region is forecast to experience. This would allow us to provide a more bespoke prediction for each country or region. Due to time constraints, we were also not able to

build alternative nonlinear models using, for example, regression trees or support vector regression. These may provide superior results.

Finally, as with all such projects, the quality of the data is paramount. The granularity of the data used here is restrictive. China or India for example are huge land masses. As such, we are losing a lot of information by quoting only a single average temperature for regions of that size. Ideally, you would want weather data at a more local level and specifically in rice-growing areas.

We have also used only temperature and rainfall in our study here. But it is known that humidity, sun exposure and wind speed are also predictive of yield. Additionally, some features derived from the available weather data could be engineered. For example, using daily data, one could construct input variables such as the maximum number of consecutive days without rainfall (to capture drought conditions) in a month or an indicator (0/1) to signify typhoon conditions in any given month. Data for other relevant model input variables could also be collected. For example, soil type, vegetation indices acquired from satellite data and technology adoption.

In conclusion, we have shown that using limited data and basic techniques such as clustering and linear regression, it is possible to gain some insight into the relationship between climate and rice production. The models built suggest that if temperature and rainfall are to increase in the future as a result of global warming, a significant proportion of the developing world may struggle to meet its rice production needs unless this is offset by increased wealth and technological advancements.

REFERENCES

[1] S. D. P. a. B. T. Shrestha, "Adaptation strategies for rice cultivation under climate change in Central Vietnam," *Mitigation and Adaptation Strategies for Global Change,* vol. 21, no. 1, pp. 15-37, 2016.

[2] R. Caruso, I. Petrarca and R. Ricciuti, "Climate change, rice crops, and violence: Evidence from Indonesia," *Journal of Peace Research,* vol. 53, no. 1, pp. 66-83, 2016.

[3] H. a. H. R. Wang, "Climate change and geographic shifts in rice production in China," *Environmental Research Communications,* vol. 1, no. 1, pp. 8-11, 2019.

[4] Q. Z. j. z. Jianchang Wang, "Moderate Wetting and Drying Increases rice yield and reduces water use, grain arsenic level, and methane emission," *Science Direct,* pp. 151-158, 2016.

[5] Z. L. Z. A. W. Y. O. W. W. T. H. e. a. Li, " Chinese Rice production area adaptions to climate changes, 1949-2010.," *Environ Sci Technol.,* vol. 49, pp. 2032-2037, 2015.

[6] H. H. ,. B. S. E. Mohammad Javad Amiri, "Optimisation of deficit-irrigation under variable seasonal rainfall and planning scenarios for rice in a semi-arid region of Iran," *Internatinal Journal of Hydrology Science and Technology,* vol. 6, no. 4, pp. 331-343, 2016.

[7] J. N. E. T. M. Y. A. T. M. S. O. B. Saliu Akinlabi Tiamiyu, "Rainfall Variability and Its Effect on Yield of Rice in Nigeria," *International Letters of Natural Sciences ,* vol. 49, pp. 63-68 , 2015.

[8] D. Headey, "Food Prices and Poverty," The World Bank, 2016.

[9] J. G. a. T. K.F. Davis, "Sustaining food self-sufficiency of a nation: The case of Sri Lankan rice production and related water and fertilizer demands.," *Ambio 45,* vol. 3, pp. 302-312, 2016.

[10] b. K. S. A. T. E. A.-A. L. V. B. J. v. W. H. d. G. M. v. I. K. C. M. W. P.A.J van Oorta, "Assessment of rice self-sufficiency in 2025 in eight African countries," *Global Food Security,* vol. 5, p. 39=49, 2015.

[11] M. &. N. U. M. &. A. H. M. &. J. Y. M. Hossain, "Predicting rice yield for Bangladesh by exploiting weather conditions," ICTC, 2017.

[12] D. Ramesh and V. Vardhan, "ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES," 2016.

[13] "World Rice Statistics," July 2019. [Online]. Available: http://ricestat.irri.org:8080/wrsv3/entrypoint.htm. [Accessed July 2019].

[14] "Climate Knowlegde Change Portal," July 2019. [Online]. Available: https://climateknowledgeportal.worldbank.org/. [Accessed July 2019].

[15] "International Organisation of Standardisation," July 2019. [Online]. Available: https://www.iso.org/obp/ui/#search. [Accessed July 2019].

Appendix:

A. R Code:

We executed the code in R using R-studio on local machine. Snippet of the code is attached here with:

```
setwd('/Users/kulbhushanarora/Documents/Cropdata_2
/')
crop_data1 <-
read.csv('/Users/kulbhushanarora/Documents/Cropdat
a_2/Final_Temp_RF_V1.csv')
crop_data2 <-
read.csv('/Users/kulbhushanarora/Documents/Cropdat
a_2/Final_yield_v2.csv')
crop_data3 <-
read.csv('/Users/kulbhushanarora/Documents/Cropdat
a_2/fert.csv')
```

```
# Merged Weather data with yield and ferilizer
dataset based on Country and Year

dfc <- merge(x=crop_data2, y=crop_data1, by="Key",
all.x=TRUE)
dfc1 <- merge(x=dfc, y=crop_data3, by="Key",
all.x=TRUE)
cropdata <- data.frame(dfc1$Year.x    ,dfc1$ISO.x
,dfc1$Country.x       ,dfc1$LAND_AREA
       ,dfc1$ARABLE_LAND,     dfc1$IRGTD_LAND
       ,dfc1$HARV_AREA,dfc1$Fert     ,dfc1$YIELD
              ,dfc1$AVG_TEMP_JAN
       ,dfc1$AVG_TEMP_FEB     ,dfc1$AVG_TEMP_MAR
       ,dfc1$AVG_TEMP_APR     ,dfc1$AVG_TEMP_MAY
       ,dfc1$AVG_TEMP_JUN     ,dfc1$AVG_TEMP_JUL
,dfc1$AVG_TEMP_AUG     ,dfc1$AVG_TEMP_SEP
       ,dfc1$AVG_TEMP_OCT     ,dfc1$AVG_TEMP_NOV
       ,dfc1$AVG_TEMP_DEC     ,dfc1$AVG_RF_JAN
              ,dfc1$AVG_RF_FEB
       ,dfc1$AVG_RF_MAR       ,dfc1$AVG_RF_APR
       ,dfc1$AVG_RF_MAY       ,dfc1$AVG_RF_JUN
       ,dfc1$AVG_RF_JUL
       ,dfc1$AVG_RF_AUG       ,dfc1$AVG_RF_SEP
       ,dfc1$AVG_RF_OCT       ,dfc1$AVG_RF_NOV
       ,dfc1$AVG_RF_DEC                )
write.csv(dfc1,"Final_cropdata.csv")

# Applied Numeric Country code to each country and
realigned the variables.

crop <-
read.csv('/Users/kulbhushanarora/Documents/Cropdat
a_2/Final_cropdata_V2.csv')

# Removed data for year 2016 as temperature data
for 2016 is missing.
crop1 <- crop[crop$YEAR < 2016,]
crop2 <- select(crop1,SEQ_NO, YEAR    ,ISO
       ,COUNTRY_CDE    ,LAND_AREA     ,HARV_AREA
       ,FERT   ,YIELD  ,AVG_TEMP_JAN  ,AVG_TEMP_FEB
       ,AVG_TEMP_MAR   ,AVG_TEMP_APR  ,AVG_TEMP_MAY
       ,AVG_TEMP_JUN   ,AVG_TEMP_JUL  ,AVG_TEMP_AUG
       ,AVG_TEMP_SEP   ,AVG_TEMP_OCT  ,AVG_TEMP_NOV
       ,AVG_TEMP_DEC   ,AVG_RF_JAN    ,AVG_RF_FEB
       ,AVG_RF_MAR     ,AVG_RF_APR    ,AVG_RF_MAY
       ,AVG_RF_JUN     ,AVG_RF_JUL    ,AVG_RF_AUG
       ,AVG_RF_SEP     ,AVG_RF_OCT    ,AVG_RF_NOV
       ,AVG_RF_DEC)

# Check for NULL / missing values in the dataset.

colSums(is.na(crop2))

# Taking Backup
write.csv(crop2,"crop2.csv")

# Removed the following countries from analysis as
weather data is missing for these countries

crop3 <- crop2[crop2$ISO != "GMB", ]
crop3 <- crop3[crop3$ISO != "GUF", ]
crop3 <- crop3[crop3$ISO != "HKG", ]
crop3 <- crop3[crop3$ISO != "MAR", ]
crop3 <- crop3[crop3$ISO != "PKR", ]
crop3 <- crop3[crop3$ISO != "REU", ]
crop3 <- crop3[crop3$ISO != "TWN", ]
crop3 <- crop3[crop3$ISO != "ZWE", ]

# Taking Backup
colSums(is.na(crop3))

# Only Fields having missing values are LAND_AREA
and HARV_AREA and YIELD
# Impute Mean value for FERT, YIELD , LAND_AREA
and HARV_AREA

crop3 <- crop3 %>% group_by(ISO) %>%
mutate(FERT=ifelse(is.na(FERT),mean(FERT,na.rm=TRU
E),FERT))
crop3 <- crop3 %>% group_by(ISO) %>%
mutate(HARV_AREA=ifelse(is.na(HARV_AREA),mean(HARV
_AREA,na.rm=TRUE),HARV_AREA))
crop3 <- crop3 %>% group_by(ISO) %>%
mutate(LAND_AREA=ifelse(is.na(LAND_AREA),mean(LAND
_AREA,na.rm=TRUE),LAND_AREA))
crop3 <- crop3 %>% group_by(ISO) %>%
mutate(YIELD=ifelse(is.na(YIELD),mean(YIELD,na.rm=
TRUE),YIELD))

# Removed the records where Fertlizer value is
missing
crop3 <- crop3[!is.na(crop3$FERT), ]
colSums(is.na(crop3))

# taking backup0.. All missing values are handled

write.csv(crop3,"crop3.csv")
crp1 <-
read.csv('/Users/kulbhushanarora/Documents/Cropdat
a_2/crop3.csv')
head(crp1)

# Feature Engineering
# Adding Quarterly values for temperature and
Rainfall

crp1$AVG_Q1_TEMP <- (crp1$AVG_TEMP_JAN +
crp1$AVG_TEMP_FEB + crp1$AVG_TEMP_MAR)/3
crp1$AVG_Q2_TEMP <- (crp1$AVG_TEMP_APR +
crp1$AVG_TEMP_MAY + crp1$AVG_TEMP_JUN)/3
crp1$AVG_Q3_TEMP <- (crp1$AVG_TEMP_JUL +
crp1$AVG_TEMP_AUG + crp1$AVG_TEMP_SEP)/3
crp1$AVG_Q4_TEMP <- (crp1$AVG_TEMP_OCT +
crp1$AVG_TEMP_NOV + crp1$AVG_TEMP_DEC)/3
crp1$AVG_Q4_RF <- (crp1$AVG_RF_OCT +
crp1$AVG_RF_NOV + crp1$AVG_RF_DEC)/3
crp1$AVG_Q3_RF <- (crp1$AVG_RF_JUL +
crp1$AVG_RF_AUG + crp1$AVG_RF_SEP)/3
crp1$AVG_Q2_RF <- (crp1$AVG_RF_APR +
crp1$AVG_RF_MAY + crp1$AVG_RF_JUN)/3
crp1$AVG_Q1_RF <- (crp1$AVG_RF_JAN +
crp1$AVG_RF_FEB + crp1$AVG_RF_MAR)/3

# Dataset with only quarterly weather field.
Monthly fields are removed.
crp2 <- select(crp1,COUNTRY_CDE       ,YEAR
       ,LAND_AREA     ,HARV_AREA     ,FERT
       ,YIELD  ,AVG_Q1_TEMP ,AVG_Q2_TEMP
       ,AVG_Q3_TEMP   ,AVG_Q4_TEMP ,AVG_Q1_RF
       ,AVG_Q2_RF,AVG_Q3_RF,AVG_Q4_RF)
head(crp2)

# taking backup
write.csv(crp2,"crp2.csv")

#key <- group_by(crp2, COUNTRY_CDE)
#crp21k <- summarize(key, LAND_AREA =
mean(LAND_AREA), HARV_AREA = mean(HARV_AREA), FERT
= mean(FERT), YIELD = mean(YIELD), AVG_Q1_TEMP =
mean(AVG_Q1_TEMP), AVG_Q2_TEMP =
mean(AVG_Q2_TEMP), AVG_Q3_TEMP =
mean(AVG_Q3_TEMP),AVG_Q4_TEMP = mean(AVG_Q4_TEMP),
AVG_Q1_RF = mean(AVG_Q1_RF), AVG_Q2_RF =
mean(AVG_Q2_RF), AVG_Q3_RF = mean(AVG_Q3_RF),
AVG_Q4_RF = mean(AVG_Q4_RF) )
#write.csv(crp21k,"crp21k.csv")

head(crp2)
crp21 <-
read.csv('/Users/kulbhushanarora/Documents/Cropdat
a_2/crp21.csv')
write.csv(crp21,"crp21.csv")
```

```
# checking the optimal value of K. Number of
clusters.
fviz_nbclust(crp21, kmeans, method = "wss")

# scale the variables
crpx <- crp21
AVG_Q4_TEMP_scal = scale(crpx$AVG_Q4_TEMP)
AVG_Q3_TEMP_scal = scale(crpx$AVG_Q3_TEMP)
AVG_Q2_TEMP_scal = scale(crpx$AVG_Q2_TEMP)
AVG_Q1_TEMP_scal = scale(crpx$AVG_Q1_TEMP)
AVG_Q4_RF_scal = scale(crpx$AVG_Q4_RF)
AVG_Q3_RF_scal = scale(crpx$AVG_Q3_RF)
AVG_Q2_RF_scal = scale(crpx$AVG_Q2_RF)
AVG_Q1_RF_scal = scale(crpx$AVG_Q1_RF)
YIELD_scal = scale(crpx$YIELD)
LAND_AREA_scal = scale(crpx$LAND_AREA)
HARV_AREA_scal = scale(crpx$HARV_AREA)
FERT_scal = scale(crpx$FERT)
YEARX <- crpx$YEAR
COUNTRYX <- crpx$COUNTRY_CDE
# datframe with scaled fields
crpX1 <- data.frame(YIELD_scal, LAND_AREA_scal,
FERT_scal, AVG_Q1_TEMP_scal, AVG_Q2_TEMP_scal,
AVG_Q3_TEMP_scal, AVG_Q4_TEMP_scal,AVG_Q1_RF_scal,
AVG_Q2_RF_scal, AVG_Q3_RF_scal,AVG_Q4_RF_scal)

kx <- kmeans(crpX1, centers = 3, nstart = 25)
fviz_cluster(kx, data = crpX1)
kx
k2 <- kmeans(crpX1, centers = 2, nstart = 25)
k3 <- kmeans(crpX1, centers = 3, nstart = 25)
k4 <- kmeans(crpX1, centers = 4, nstart = 25)
k5 <- kmeans(crpX1, centers = 5, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point", data =
crpX1) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data =
crpX1) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data =
crpX1) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data =
crpX1) + ggtitle("k = 5")

library(gridExtra)
grid.arrange(p1, p2, p3, p4, nrow = 2)

set.seed(123)

# function to compute total within-cluster sum of
square
wss <- function(k) {
  kmeans(crp21, k, nstart = 10 )$tot.withinss
}

# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15

# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")

set.seed(123)

# calculating the average values for three
clsuters
library(dplyr)
crp22 <- crp21 %>%
  mutate(Cluster = k3$cluster) %>%
  group_by(Cluster) %>%
  summarise_all("mean")

# taking backup
write.csv(crp22,"crp22.csv")

crp23 <- k3$centers
write.csv(crp23,"crp23.csv")

# with in clusters, predicting the yield value
based on input variables
m1 <- crp21[crp21$cluster ==1, ]
m2 <- crp21[crp21$cluster ==2, ]
m3 <- crp21[crp21$cluster ==3, ]


# Cluster 1 - Multiple Linear regression
m1_train <- m1[c(-3,-10,-32),]
Model <-lm(YIELD~LAND_AREA + FERT + AVG_Q1_TEMP +
AVG_Q2_TEMP + AVG_Q3_TEMP + AVG_Q4_TEMP +
AVG_Q1_RF +  AVG_Q2_RF +  AVG_Q3_RF + AVG_Q4_RF,
data = m1_train)
Model
m1_test <- m1[c(3,10,32,76,100),c(-6)]
m1_test
m1_test$YIELD <- predict(Model, newdata = m1_test)
summary(Model)

# Cluster 2 - Multiple Linear regression
m2_train <- m2[c(-3,-10,-32,-76,-100),]
m2_train <- m2[c(-3,-10,-32,-76,-100),]
Model <-lm(YIELD~LAND_AREA + FERT + AVG_Q1_TEMP +
AVG_Q2_TEMP + AVG_Q3_TEMP + AVG_Q4_TEMP +
AVG_Q1_RF +  AVG_Q2_RF +  AVG_Q3_RF + AVG_Q4_RF,
data = m2_train)
Model
m2_test <- m2[c(3,10,32,76,100),c(-6)]
m2_test
m2_test$YIELD <- predict(Model, newdata = m2_test)
summary(Model)

# Cluster 3 - Multiple Linear regression
m3_train <- m3[c(-3,-10,-32,-76,-100),]
Model <-lm(YIELD~LAND_AREA + AVG_Q1_TEMP +
AVG_Q2_TEMP + AVG_Q3_TEMP + AVG_Q4_TEMP +
AVG_Q1_RF +  AVG_Q2_RF +  AVG_Q3_RF + AVG_Q4_RF,
data = m3_train)
Model
m3_test <- m3[c(3,10,32,76,100),c(-5)]
m3_test
m3_test$YIELD <- predict(Model, newdata = m3_test)
summary(Model)

# Multiple linear regression on Complete dataset
of 105 countries
crp21_train <- crp21[c(-3,-10,-23, -45, -32,-76,-
100),]
Model <-lm(YIELD~LAND_AREA + FERT + AVG_Q1_TEMP +
AVG_Q2_TEMP + AVG_Q3_TEMP + AVG_Q4_TEMP +
AVG_Q1_RF +  AVG_Q2_RF +  AVG_Q3_RF + AVG_Q4_RF,
data = crp21_train)
Model
crp21_test <- crp21[c(3,10,23, 45, 32,76,100),c(-
5)]
crp21_test
crp21_test$YIELD <- predict(Model, newdata =
crp21_test)
summary(Model)
```

B:

Links to data sources:

World Rice Statistics

http://ricestat.irri.org:8080/wrsv3/entrypoint.htm

Climate Knowledge Portal

https://climateknowledgeportal.worldbank.org/download-data

U.N. GDP per Capita Data Source:

http://data.un.org/Data.aspx?q=GDP+per+capita&d=SNAAMA&f=grID%3a101%3bcurrID%3aUSD%3bpcFlag%3a1