Predicting crop yield in the United States using environmental indicators

Name	Student ID
Nicholas Ng Zhi Yong	1006021
Kwa Yu Liang	1006176
Chin Wei Ming	1006264
Dorishetti Kaushik Varma	1006012
S Hamsaraj	1005943

Abstract

Food security is a huge issue that affects 1 in 4 Americans households. This is made worse by disruptions in food supply chains due to the Ukraine conflict and the Covid-19 virus. As such, there is an increased importance to be self-sufficient through domestic food production. However, crop yields fluctuate from year to year, which might make it hard for policy makers to come out with a strategy to ensure that there is enough food available within the country. Therefore, if we can predict crop yield for the year, these policy makers can use this prediction as a guide to how they'll approach the problem, such as how much food to stockpile, or the amount of support to provide farmers to produce enough output.

For our prediction, we have chosen to use a multiple linear regression model, with the equation $y = B_1 * x_1 + B_2 * x_2 + \cdots + B_n * x_n$, where B_n values are constants.

Dependent variable

The dependent variable we have chosen is crop yield in kg per hectare. We chose this as it might indicate the amount of food that is available within the country. Additionally, there is sufficient data (more than 30) to ensure that our model might work well.

Independent variables

We chose environmental variables as these are natural factors that are harder to control. Therefore, the predicted yield will be an indication of the baseline amount of food available within the country. With this information, it will then be easier for policy makers to decide how and how much to top up to this amount.

So how do we decide which independent variables we should include in our model?

Firstly, the independent variables must be an environmental in nature, meaning that we do not include independent variables that involve manmade processes such as machinery or policies. Secondly, like the dependent variable, there must be sufficient data (more than 30), for the model to work well. Additionally, the data must also be measured annually, as the independent variable is measured annually too. Lastly, the independent variable must be justified to influence crop yield, with research done to support the justification.

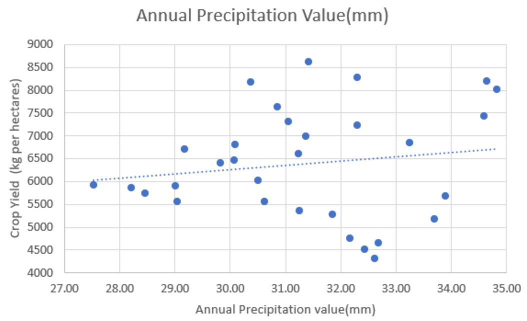
With this in mind, we identified 8 environmental variables:

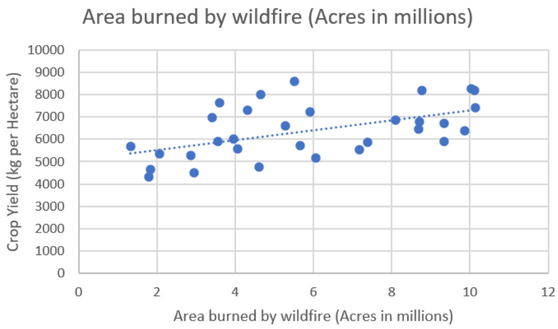
Response Variable	Predictor variables	Units	Justification
Crop Yield (kg per Hectare)	Annual Precipitation Value[i]	mm	Availability of water is higher; hence more crops will grow with access to the water that precipitated
	Average Annual Temperature[ii]	F (Fahrenheit)	Crops in general cannot grow under lower temperatures, affecting the yield
	Annual Average Droughts[iii]	Palmer Drought Severity Index	Droughts bring destruction to crops, lowering the yield
	Hurricanes[iv]	No. of hurricanes that reach the US	Hurricanes bring destruction to crops, lowering the yield
	Methane[v]	Million metric tons of CO_2 equivalent	diffusion of atmospheric methane into the soil is inhibited, reducing bacterial uptake in soil

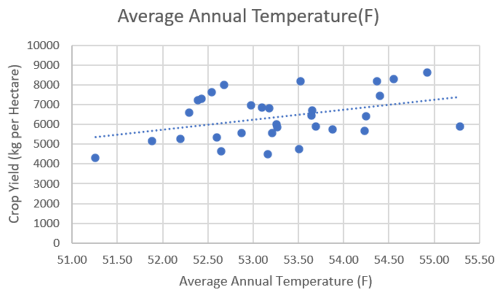
	Nitrous Oxide[vi]	Million metric tons of CO2 equivalent	critical ingredient in chlorophyll, needed for photosynthesis.
	Area burned by Wildfire[vii]	Acres in millions	Living things burnt will be absorbed by the soil, increasing soil fertility
	Carbon Dioxide[viii]	Million metric tons of CO2 equivalent	Crops require carbon dioxide to survive

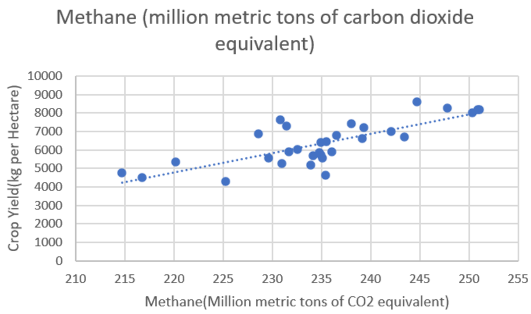
Table 1

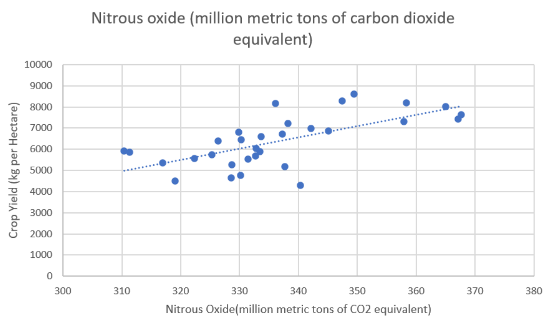
Next, we check if they are linear in relation to the dependent variable. This is to ensure that it will work well within the model, because in the equation we have defined above, x_n is to the power of 1, meaning it only affects y by a factor of B_n. We do this by plotting the individual independent variables against the dependent variable, then visually determining if the relationship can be modelled by a straight line. The plots are shown below:

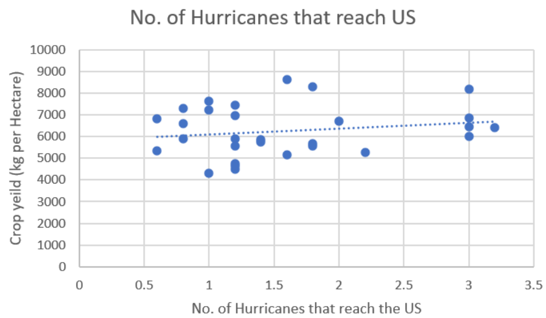


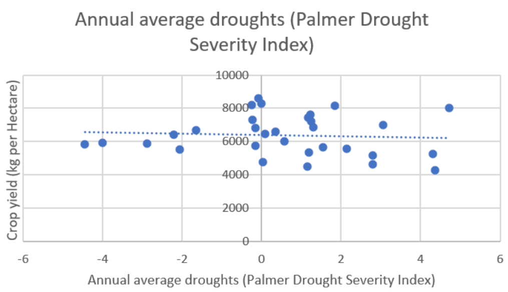


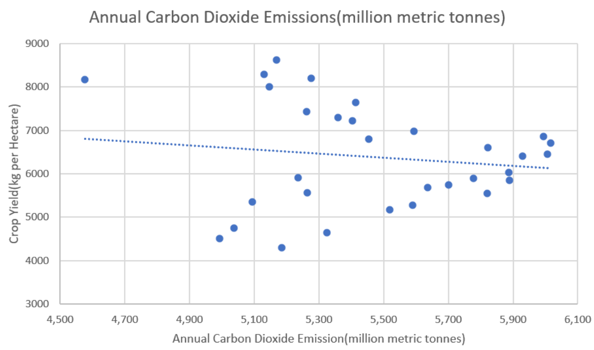












At the end of our analysis, we have decided to not use the following variables for these respective reasons.

Annual Carbon Dioxide Emissions

The relationship between carbon dioxide and crop yield is not linear in nature.

No. of Hurricanes that reach the US

Hurricanes should decrease the crop yield due to its destructive nature - Crops are snapped or uprooted and food crops are flooded or washed away. However, we notice the opposite trend in the graph above. Additionally, its R^2 value is way too low at 0.7% for it to be useful.

Annual Average droughts

Its R^2 value is too low at 0.3% for it to be useful.

The rest of the variables appear linear in nature, with R^2 values above 10%, hence we narrowed down the independent variables to the remaining 5 variables: Methane, Nitrous Oxide, Average annual temperature, Annual precipitation, and Area burned by wildfires.

Metric to measure accuracy of the model

Chosen Metric: adjusted R^2 value

Now that we have determined several independent variables that are linearly related to the dependent variable, how do we decide which independent variables to include within the model, and which to exclude?

To do so, we need a way to measure the accuracy of the model. This will enable us to know which combination of variables that are included within the model produces the most accurate model.

For a simple linear regression, R^2 is a good indicator of the accuracy of the model, where $R^2 = 1 - \frac{residuals\ sum\ squares}{total\ sum\ squres}$, because it can measure the proportion of variance for a dependent variable that is explained by an independent variable. However, it does not work well for our model: a multiple linear regression. This is because R^2 will always increase as more predictors are added to the model, regardless of the quality of the predictor, since adding more predictors will increase the dimensions of the model, enabling the regression line to fit the points more closely.

However, this is not necessarily a good thing, because adding too many predictors may cause overfitting. Overfitting is when a model fits a training dataset too closely that it learns the "noise" or irrelevant data from the training dataset, making it too specific to the training dataset. This decreases the predictive ability of the model as it is now unable to generalise new data and output a useful value.

Hence, we need to find a way to compare models with different numbers of predictors. This is where adjusted R^2 comes in handy. $Adjusted\ R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$, where R^2 is defined above, N is total sample size, and p is number of independent variables. This imposes a penalty on models that uses more predictors in such a way that allows us to compare them all on the same level regardless of the number of predictors that the model utilises.

Selecting a model

We started off by including all the predictors within the model, and this is the results we obtained:

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.9037
R Square	0.816673
Adjusted R Square	0.780008
Standard Error	563.3469
Observations	31

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	35343861.9	7068772	22.27369159	1.78985E-08
Residual	25	7933992.83	317359.7		
Total	30	43277854.7			

	Coefficient:	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-33163.3	8104.79373	-4.09181	0.00039125	-49855.3907	-16471.1	-49855.4	-16471.1
Methane	55.84499	15.2426776	3.663726	0.001168558	24.45210803	87.23787	24.45211	87.23787
Nitrous oxide	46.21833	9.57940648	4.824759	5.86764E-05	26.48916915	65.94748	26.48917	65.94748
Annual Precipitation Value (mm)	-90.5422	70.0604292	-1.29234	0.20805605	-234.83437	53.74994	-234.834	53.74994
Average Annual Temperature (F)	251.7308	140.280967	1.794476	0.084840014	-37.1832511	540.6449	-37.1833	540.6449
Area burned by wildfire (Acres in n	48.36963	49.0205324	0.986722	0.333233166	-52.5900429	149.3293	-52.59	149.3293

Figure 1

As shown in Figure 1, the adjusted R^2 value for this model is pretty good, however, how might we find the best possible combination of predictors that achieves the highest R^2? This can be done through enumerating through all possible combinations of predictors. However, there are $6C1 + 6C2 + \cdots + 6C6 = 63$ possible combinations, which is too much work to iterate through. So how might we reduce the number of combinations to look through?

One way to do so is to estimate the quality of a predictor and only include the top n predictors for a model with n predictors. We decided to use P-value as it can indicate how statistically significant each predictor is. The lower the P-value, the easier it is to reject the null hypothesis, which is that the predictor has no effect on the output (its coefficient equals to 0). So, when testing a model with n predictors, we will select n predictors with the lowest P-values for the model. To run a quick test on whether this method holds, we ran the regression for 4 variables.

SUMMARY OUTPUT	
ession Statistics	
Multiple R	0.896898
R Square	0.804426
Adjusted F	0.774337
Standard E	570.5609
Observatic	31

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	4	34813822.1	8703456	26.73546465	7.01685E-09
Residual	26	8464032.57	325539.7		
Total	30	43277854.7			

	Coefficient:	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-34943.8	8089.10215	-4.31986	0.0002022	-51571.2156	-18316.4	-51571.2	-18316.4
Methane	56.64114	15.4252553	3.671974	0.001093129	24.93407706	88.34821	24.93408	88.34821
Nitrous ox	39.10173	7.9388029	4.925394	4.09173E-05	22.78328895	55.42018	22.78329	55.42018
Wildfire	60.22012	48.7719063	1.23473	0.227974041	-40.0319645	160.4722	-40.032	160.4722
Average A	272.0318	141.183751	1.926792	0.06500156	-18.1755821	562.2391	-18.1756	562.2391

Figure 2 (next 4 lowest P-values)

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.899741
R Square	0.809534
Adjusted R Square	0.780231
Standard Error	563.061
Observations	31

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	35034874.1	8758719	27.62673983	5.00313E-09
Residual	26	8242980.63	317037.7		
Total	30	43277854.7			

	Coefficient:	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-37201	6992.33973	-5.32025	1.44609E-05	-51573.9716	-22828.1	-51574	-22828.1
Methane	61.36966	14.169911	4.330984	0.000196371	32.24298725	90.49633	32.24299	90.49633
Nitrous oxide	47.00919	9.54097297	4.927085	4.07351E-05	27.39743722	66.62094	27.39744	66.62094
Annual Precipitation	-103.474	68.7888387	-1.50422	0.144575935	-244.871159	37.92381	-244.871	37.92381
Average Annual Temperature (	311.0397	126.686987	2.455183	0.021083513	50.63091039	571.4486	50.63091	571.4486

Figure 3 (4 lowest P-values)

SUMMARY OUTPUT	
ession Statistics	
Multiple R	0.890482
R Square	0.792958
Adjusted F	0.769953
Standard E	576.0767
Observatic	31

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	3	34317518.1	11439172.7	34.46942627	2.23704E-09
Residual	27	8960336.62	331864.319		
Total	30	43277854.7			

	Coefficient:	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-40481.2	6797.18357	-5.9555841	2.37854E-06	-54427.8669	-26534.5	-54427.9	-26534.5
Methane	63.91545	14.3936821	4.4405211	0.000136856	34.38205303	93.44885	34.38205	93.44885
Nitrous Ox	38.81074	8.01201708	4.84406652	4.6312E-05	22.37144258	55.25004	22.37144	55.25004
Average A	352.2894	126.54249	2.78396151	0.009688687	92.64567919	611.9332	92.64568	611.9332

Figure 4 (next 3 lowest P-values)

SUMMARY OUTPUT	
ession Statistics	
Multiple R	0.890482
R Square	0.792958
Adjusted F	0.769953
Standard E	576.0767
Observatic	31

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	3	34317518.1	11439172.7	34.46942627	2.23704E-09
Residual	27	8960336.62	331864.319		
Total	30	43277854.7			

	Coefficient:	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-40481.2	6797.18357	-5.9555841	2.37854E-06	-54427.8669	-26534.5	-54427.9	-26534.5
Methane	63.91545	14.3936821	4.4405211	0.000136856	34.38205303	93.44885	34.38205	93.44885
Nitrous ox	38.81074	8.01201708	4.84406652	4.6312E-05	22.37144258	55.25004	22.37144	55.25004
Average A	352.2894	126.54249	2.78396151	0.009688687	92.64567919	611.9332	92.64568	611.9332

Figure 5 (3 lowest P-values)

Figure 3 shows the regression for the predictors with the 4 lowest P-values, while figure 2 shows the regression for the predictors the next 4 lowest P-values. The results show that the 4 lowest P-values has a higher adjusted R^2 value compared to the next best option. We ran the same test for a model with 3 variables, and the results were similar as shown in Figure 4 and 5.

We then ran the regression on 1-5 predictors with the lowest P-value predictors, with the results shown in the table below.

Number of predictors	Best Adjusted R^2	Combination of predictors

1	0.58*	Methane
2	0.71	Methane, Nitrous Oxide
3	0.77	Methane, Nitrous Oxide, Average annual temperature
4	0.78	Methane, Nitrous Oxide, Average annual temperature, Annual precipitation
5	0.78	Methane, Nitrous Oxide, Average annual temperature, Annual precipitation, Area burned by wildfires

For one predictor, the best adjusted R^2 came from methane even though it didn't have the lowest P-value. This shows that P-value is just an estimation on the quality of the predictor. However, it is still a good indicator if the difference in P-values is sufficiently large enough, such as swapping Area burned by wildfires with Annual precipitation value in Figure 2 and Figure 3. Whereas if the difference in P-values is too small, such as between methane and nitrous oxide for 1 predictor, using P-values might not be as accurate.

However, we notice adjusted R^2 only becomes sufficiently high from 3-5 predictors, and the n lowest P-value predictors are significantly lower than the next lowest option, hence we can safely conclude that the combination will yield the highest adjusted R^2.

Finally, from the table above, we have identified having 4 and 5 predictors yield the highest adjusted R^2 at 0.78. However, given a choice between either using greater or smaller number of predictors to produce a model with the same accuracy, it makes more sense to choose one with a lower number of predictors so that less data collection is required for the same level of accuracy. Therefore, our group decided to settle with the model $Crop\ yield = B_0 + B_1 * Methane +\ B_2 * Nitrous\ Oxide + B_3 * Average\ annual\ temperature +\ B_4 * Annual\ precipitation.$

Analysis of Regression

The result of the regression is shown in Figure 3. For the signs of the coefficients, methane and nitrous oxide is positive, which make sense because they are essential for plant growth so higher values might increase crop yield. The sign for annual temperature is positive, as plants might survive better when the weather is not too cold. Finally, the sign for annual precipitation is negative, which might be because high amounts of rainfall could potentially drown the crops and reduce its yield.

For the magnitudes of the coefficients, Average annual temperature is the highest possibly due to the high sensitivity of crops to fluctuations in temperature. Annual precipitation is second highest possibly due to the susceptibility of crops to drowning under high rainfall, however it might have a lesser impact compared to temperature because the effect can be mitigated through human intervention. Finally, methane and nitrous oxide is the lowest at around 50 as these factors might not be as essential to plant growth.

Area for improvement

Firstly, the model might have missed out other variables that affects crop yield. Thus, in order to improve our model, we can consider a greater number of factors to better predict the crop yield. Secondly, the model only accounts for environmental variables, hence we assumed that the other non-environmental variables were constant during the period, which is fictitious. Ideally, the environmental data we use should be from periods of time during which the non-environmental variables are constant. Lastly, we only used 31 data points for the regression, which is not a very large dataset, hence using a larger dataset in the future as more data collected might increase the accuracy of the model.

References

[i] https://www.epa.gov/climate-indicators/climate-change-indicators-drought

[ii] https://www.statista.com/statistics/500472/annual-average-temperature-in-the-us/

[iii] https://www.epa.gov/climate-indicators/climate-change-indicators-drought

[iv] https://www.epa.gov/climate-indicators/climate-change-indicators-tropical-cyclone-activity

[v] https://www.epa.gov/climate-indicators/climate-change-indicators-tropical-cyclone-activity

[vi] https://www.epa.gov/climate-indicators/climate-change-indicators-tropical-cyclone-activity

[vii] https://www.epa.gov/climate-indicators/climate-change-indicators-wildfires

[viii] https://www.statista.com/statistics/183943/us-carbon-dioxide-emissions-from-1999/