Introduction: Agriculture and the 2008 financial crisis

By Matthew De La Roca

Most times, the narrative we get of the 2008 financial crisis is that of the securities market, one in which risky derivatives and a housing bubble caused the crash of the global economy. This project is motivated by the desire to explore this event's effect on another important market, the agricultural market. The agricultural market, producing the world's entire supply of food is not to be overlooked due to its importance for survival and basic human needs. In this analysis we will take yearly production data from the FAO, weather data from Berkely Earth, and GNI measurements from the World Bank to answer this question. Our data spans from 1990 to 2013 and almost all countries in the world. We will also explore whether poor and rich countries were variably affected by the crisis and other factors. Primarily this is an inference question, so less flexible methods like OLS will be used. In the end this exploration will also attempt a more flexible model, in order to examine limitations of prediction models with agricultural data.

Introduction to the Data

This data includes 2485 rows of data and 17 variables. Not all variables were included in the model, as some were used just for feature transformations. The variables include:

Total_Yield: Our variable of interest, this measures the total yield in kilograms of all crops produced in a country in that calendar year. Possible inconsistencies could be introduced from the primary crop types grown in different climates (e.g bananas might have lower yield than potatoes per hectare)

Area: The country name and the area from which the data was collected. Almost all countries with complete data are included. This variable was used to group the other variables by, but not used in the model

Year: The Year the data was collected, the data here ranges from 1990-2013. Due to the fact that Berkely Earth's rainfall data only goes up to 2013, and also since we are only interested in pre and post 2008 and the immediate several years after. This variable was not used in the model due to collinearity issues, but used to create the **post_2008** variable.

Ha-cropland: Hectares of cropland or about 100 acres/ 10000 square meters of cropland. Total hectares per country per year measured. Not included in model to avoid multicollinearity, used to normalize other input variables to be per/hectare.

kg_nitrogen_ha: kilograms of nitrogen fertilizer per hectare. Nitrogen is an essential plant nutrient. Most pure nitrogen fertilizer is created through the synthetic process, called the haber-bosch process. Natural nitrogen from nitrogen fixing plants (peas) and bacteria in the soil, can affect how much synthetic nitrogen is needed. Organic nitrogen can also be received from manure, a separate variable.

kg_phosphate_ha: kilograms of phosphate per hectare. Another essential nutrient for plants. Phosphate is often gotten from bird poop and bat guano, but synthetic phosphate can also be created.

kg_potash_ha: kilograms of potash per hectare. Potash is a potassium rich salt used as a fertilizer. Potassium is another important plant nutrient

kg_manure_ha: kilograms of manure per hectare. Manure is usually rich in nitrogen and can received from a variety of animals, including buffalo, cows, llamas and camels. Manure is generally favored by small plot or more organic farmers of haber-bosch fixed nitrogen.

kg_Fungicides_Bactericides_ha: kilograms of fungicide and bactericides per hectare. Fungicides and bactericides can be especially important in wet tropical areas where fungus can destroy crops.

kg_Herbicides_ha: kilograms of herbicides per hectare. Herbicides are essentially weed killers used on a larger scale. Farmers do not want non-profitable plants using up soil nutrients and root space.

kg_Insecticdes_ha: kilograms of insecticides per hectare. Insecticides are used against a wide variety of insects. But common pests include beatles , locusts, aphids, scales and more. Can have negative long-term effects as many insect eating predators can be negatively affected, and insects can gain resistance to pesticides.

AverageTemperature: The average yearly temperature per country per year. This data is from Berkely earth.

GNI : The gross national income, includes gdp and income from overseas sources, divided by population. This variable was not used in the model, but used to create the **is_wealthy** dummy.

Is_wealthy: A dummy variable created as to whether the country's GNI crosses the world bank threshold to be considered "high income".

See: https://www.worldbank.org/en/country/mic/overview#1

Post_2008: used to create our counterfactual, is a dummy variable measuring whether the data point was collected pre or post 2008.

Rain_stedev: the average yearly rainfall in the country subtracted by the global average of 990 ml and divided by standard deviation, to see how many standard deviations a countries rainfall was away from the mean. This transformation was to reduce collinearity of the relatively static rainfall data with the constant.

Log_Yield: A log transformation of the Total Yield variable, to make the coefficients easier to interpret and to control outliers. Used in the model in place of Total_Yield.

A brief overview of summary statistics for the data

```
Column Means

Total_Yield                      1.142103e+12
ha_cropland                      9.068766e+06
kg_nitrogen_ha                   6.158812e+01
kg_phosphate_ha                  2.451669e+01
kg_potash_ha                     2.278187e+01
kg_manure_ha                     3.382582e+01
kg_Fungicides_Bactericides_ha    1.032345e+00
kg_Herbicides_ha                 1.096254e+00
kg_Insecticides_ha               6.133031e-01
GNI                              9.910109e+03
AverageTemperature               1.843273e+01
rain_stdev                       2.309593e-01
Log_Yield                        2.611710e+01
dtype: float64

Column st.devs
```

Observations on the means of each column:

Yield: the mean is set at about 1.142 e+12 meaning about 1,142,000,000,000 kg total per country. Or in **Log_Yield** is about 26.117.

Pesticides: The pesticide with the highest mean is herbicide with about 1.096 kg being used per hectare, and the least used pesticide being insecticides with about .613 kg being used per hectare.

Fertilizers: The fertilizer with the highest mean usage is nitrogen with about 61.58 kg being used per hectare with manure being the second most used at 33.82 kg per hectare. The least used fertilizer on average is potash at 22.7 kg per hectare

Weather: Rain_stdev on average deviates .2309 deviations from the mean , probably due to some countries and subregions being omitted. AverageTemperature is at about 18.4 degrees Celsius.

GNI: the world GNI is averaged at 9,910 dollars per capita.

```
Column st.devs

Total_Yield                        2.864433e+12
ha_cropland                        2.224036e+07
kg_nitrogen_ha                     8.296361e+01
kg_phosphate_ha                    5.236472e+01
kg_potash_ha                       3.702434e+01
kg_manure_ha                       5.733764e+01
kg_Fungicides_Bactericides_ha      2.208673e+00
kg_Herbicides_ha                   1.915097e+00
kg_Insecticides_ha                 1.128986e+00
GNI                                1.511899e+04
AverageTemperature                 8.647245e+00
rain_stdev                         9.785126e-01
Log_Yield                          2.074273e+00
dtype: float64

Column variances

Total_Yield                        8.204979e+24
ha_cropland                        4.946335e+14
kg_nitrogen_ha                     6.882960e+03
kg_phosphate_ha                    2.742064e+03
kg_potash_ha                       1.370802e+03
kg_manure_ha                       3.287605e+03
kg_Fungicides_Bactericides_ha      4.878235e+00
kg_Herbicides_ha                   3.667596e+00
kg_Insecticides_ha                 1.274609e+00
GNI                                2.285838e+08
AverageTemperature                 7.477484e+01
rain_stdev                         9.574869e-01
Log_Yield                          4.302609e+00
dtype: float64
```

Observations on the stdev/ var of each column:

Yield: the stdev is set at abou 2.864e+12 meaning about 2,864,000,000,000 kg total per country. Or in **Log_Yield** is about 2.074, which is interesting meaning that the data might have strong skew or outliers that Logistic regression corrects for. This suggests a non-normal distribution of Yield

Pesticides: The pesticide with the highest variances/ stdev is Fungicides_bactericides with a st.dev of about 2.20 kg per hectare, this makes sense as fungal crop damage is highly dependent on climate. The least variable pesticide is insecticides with st.dev being about 1.12 kg per hectare. This suggests a non-normal distribution of points

Fertilizers: The fertilizer with the highest variance/stdev is nitrogen with about 82.96 kg per hectare, suggesting a non-normal distribution with high levels of nitrogen being used by outliers. Manure had the lowest st.dev but it still suggests a non normal distribution as the st.dev Is 57.33 kg per hectare

Weather: Rain is already expressed in standard deviations, so it is not too useful to observe. AverageTemperature st.dev is at about 8.6 degrees Celsius.

GNI: the world GNI stdev is at 15,118 dollars per capita, suggesting outliers and a non-normal distribution.

```
Column medians

Total_Yield                         2.756873e+11
ha_cropland                         2.424000e+06
kg_nitrogen_ha                      4.124075e+01
kg_phosphate_ha                     1.416040e+01
kg_potash_ha                        7.947020e+00
kg_manure_ha                        1.680349e+01
kg_Fungicides_Bactericides_ha       2.481279e-01
kg_Herbicides_ha                    4.456067e-01
kg_Insecticides_ha                  1.778070e-01
GNI                                 3.170000e+03
AverageTemperature                  2.144217e+01
rain_stdev                          4.497796e-02
Log_Yield                           2.634253e+01
dtype: float64
```

Observations on the stdev/ var of each column:

Yield: the median is set at about 2.864e+11 meaning about 286,400,000,000 kg total per country. Or in **Log_Yield** is about 26.34,

Pesticides: The pesticide with the highest median is Herbicides with a median of about .445 kg per hectare, suggesting the mean is skewed by high outliers. The lowest median pesticide is insecticides with a median of about .177 kg per hectare, this again suggests heavy oultiers.

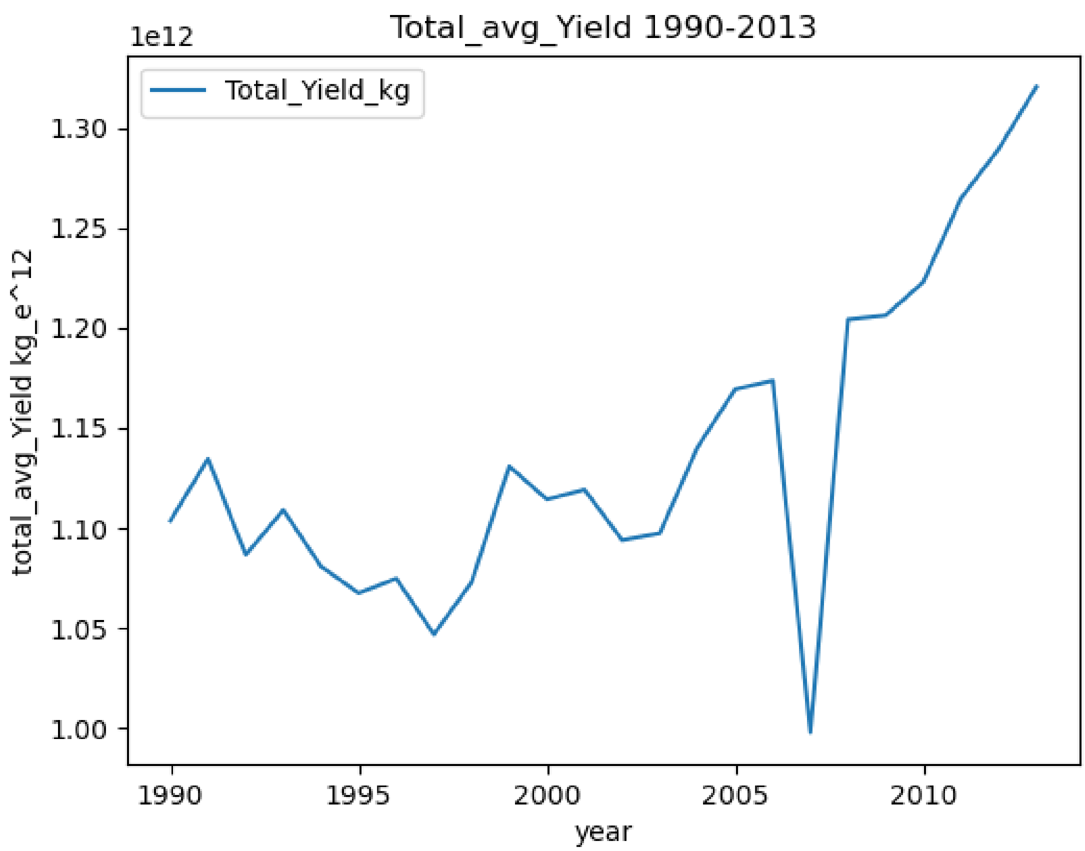
Fertilizers: The fertilizer with the highest median is nitrogen with about 41.2 kg per hectare. Potash has the lowest median 7.94 kg per hectare

Weather: Rain median deviation is 0.04497, so pretty small. AverageTemperature median is at about 21.44 degrees Celsius.

GNI: the world GNI stdev is at 3,170 dollars per capita, suggesting wealthy countries are increasing the mean and st.dev

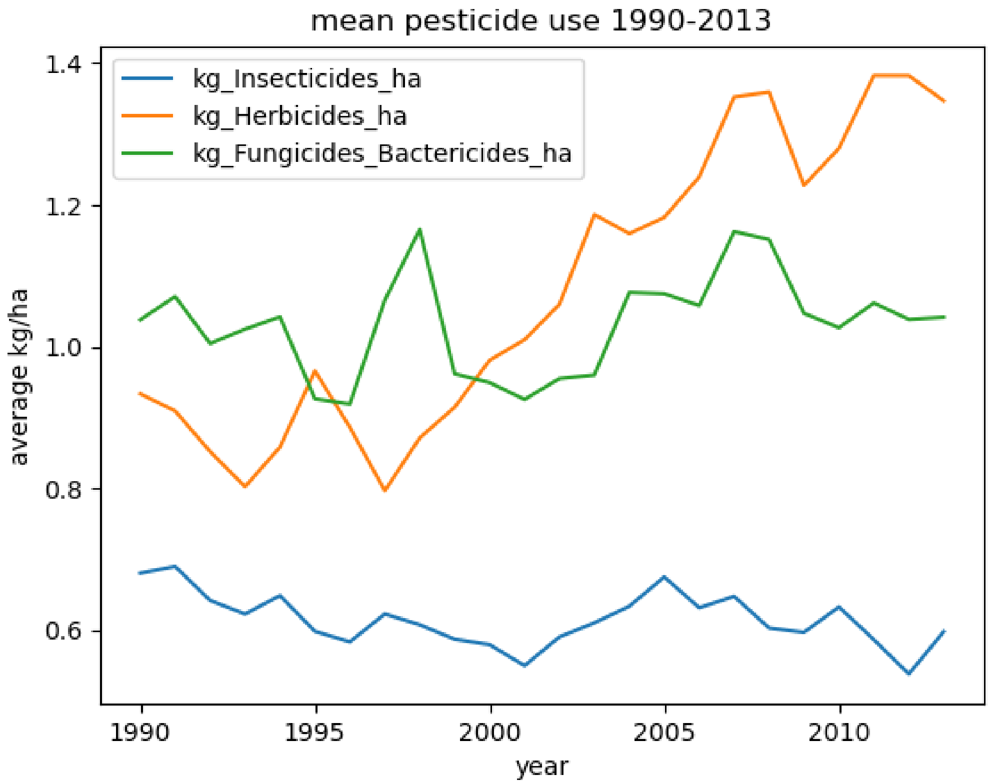
Overall impressions: The overall impression is that fertilizers are pretty heavily used, and that many of the variables seem to have non-normal distributions which is why the median is often very different from the mean and the st.dev is often larger than the median. This presents the presence of outliers and the large difference between wealthy and non-wealthy countries. Some of the large variance may also be caused by the large difference in inputs between industrial and non-industrial agriculture.

Some Charts to Explore the Data

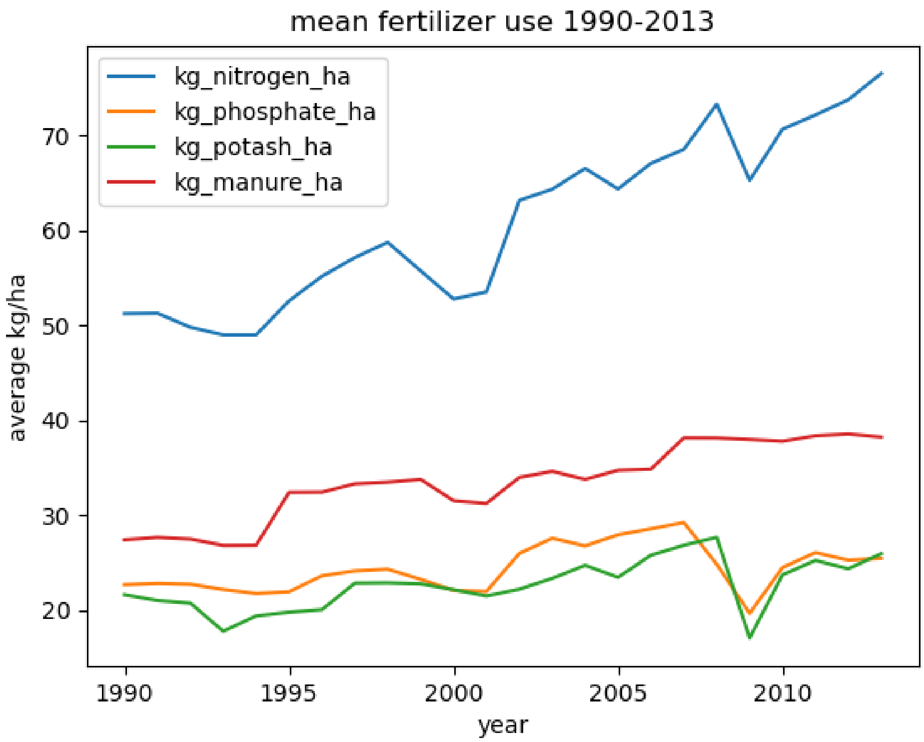


Interpretation: This chart shows the avg total yield for countries between 1990-203, as we can see there is a sharp dip around 2007/2008, which quickly recovers and rises. Some other possible minor dips are between 2000 and 2005 as well as during the late 1990s. This may be due to weather/demand shocks. Around those times the Asian financial crisis, the collapse of the USSR and Yugoslavia, and the early 2000s recession may have caused some of the turmoil.

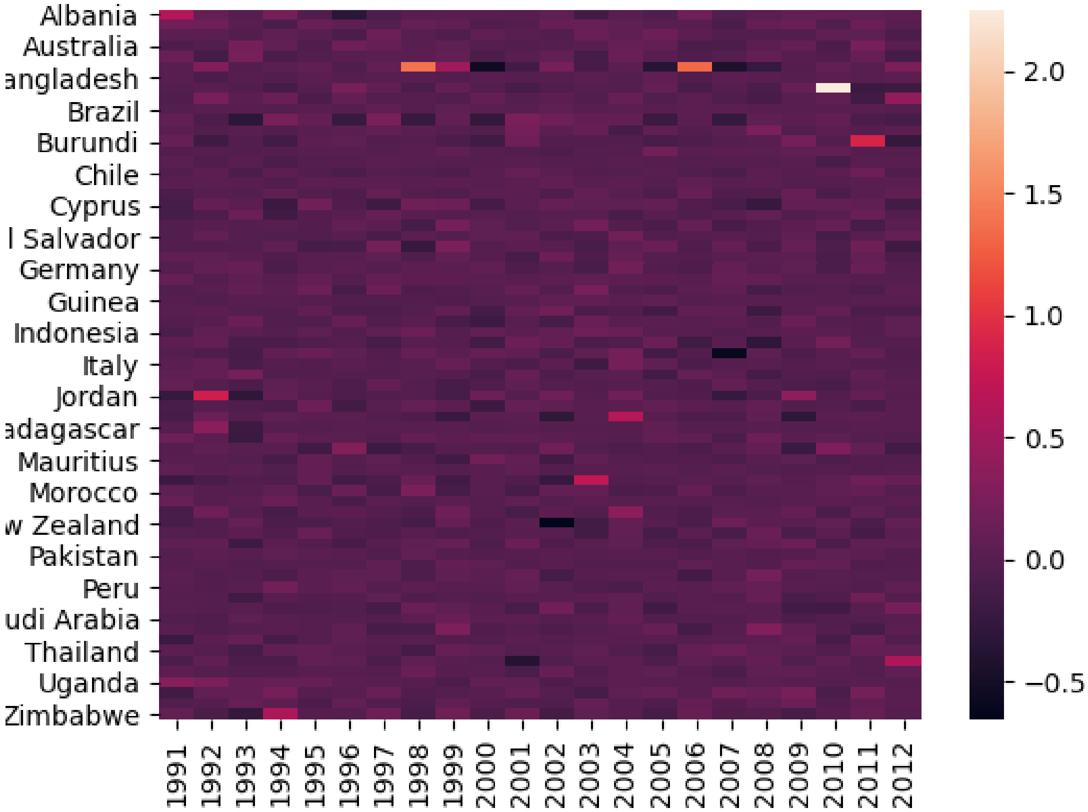
Interpretation: As wee can see in this chart, the agricultural yields of countries with a GNI > 12,000 , or high income countries is much more sensitive to shocks than the agricultural systems of poorer countries. This could due to a variety of reasons. For example a greater connectedness to the global financial markets (commodieties futures, insurance, loans),. It could be due to a higher variability in seasonal weather in the primarily wealthy and northern, Japan, Europe and North America. Or it could be due to the fact that small plot farmers in poor countries are less likely to let their fields lie fallow, or dump extra products due to price changes like large scale farmers. Another possible reason is farmers in wealthy countries witching to biofuel, is not counted as a agricultural product. As we can see the global decline in farming productivity for wealthy countries looks like it actually begins before 2008 in the early 200s , and quickly skyrockets post 2010. Poor countries themselves see a brief but sharp dip in productivity around 2008.



Interpretation: Global use of Insecticides and Fungicides have remained relatively flat, while Herbicide usage has rapidly increased since the 1990s. This could be an attempt to minimize competition for nitrogen in the soil, as well as a result of lower income countries increasing in wealth to middle income and gaining access pesticides. Again, we can see a dip in kg_per_ha around 2008, 2009, suggesting a decrease in intensity of usage, either due to increased price or an attempt by farmers to save on input costs during a demand shock. The rapid increase in herbicide. Regardless, we can see that the shocks to pesticide use occur around the same time as the supply shocks in agriculture seen in the first graph.

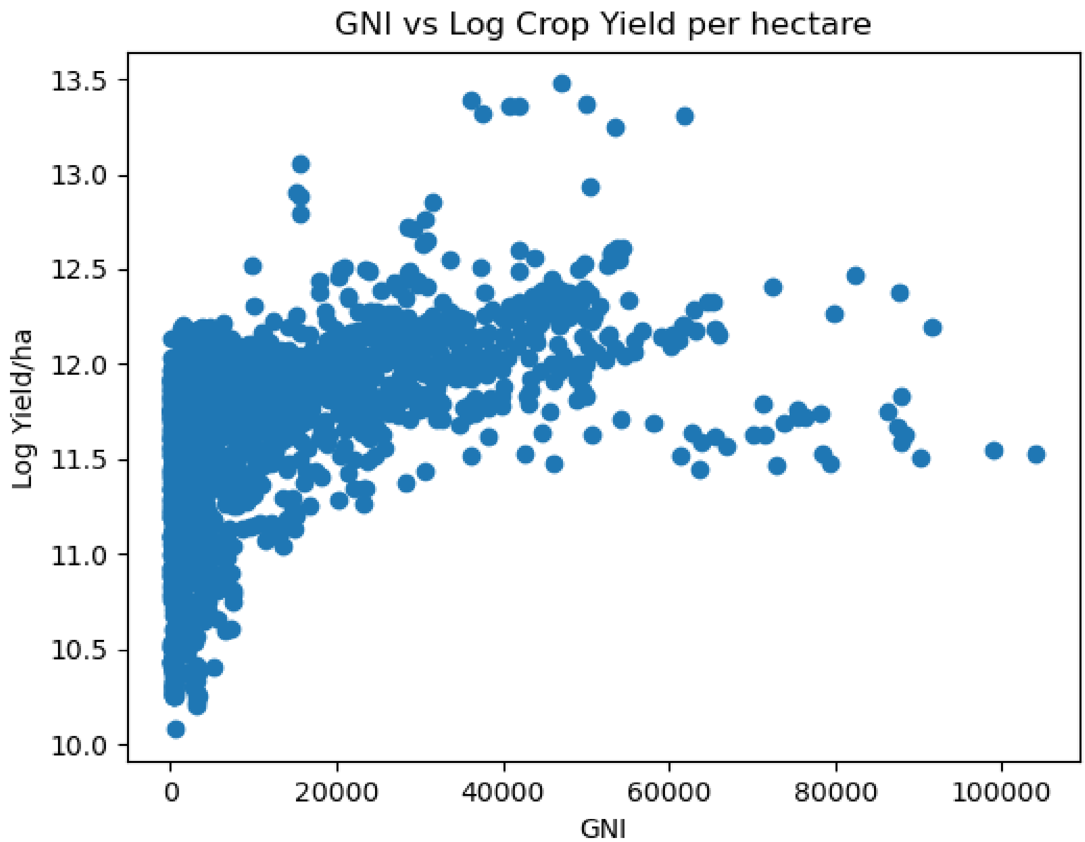


Interpretation: In this chart we see similar trends to previous figure. Fertilizer usage has increased gradually across the board, with nitrogen being the most popular and increasing the most clearly. All mineral and synthetic fertilizers saw a shock in usage around 2008. Manure was unaffected, due to its supply being a byproduct from animals, so essentially being very cheap or free to produce. Poor countries also tend to more heavily rely on natural fertilizers like manure, so may have also been less effected by global demand and financial shocks.



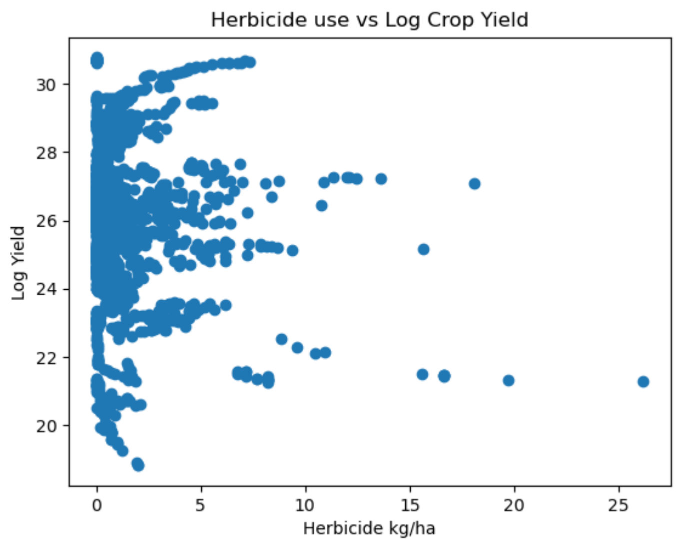
Interpretation: This chart shows the percent change year on year per country in crop production, it does not show any clear trend like the line graphs, as it only measures rate of changes from the previous year , not long term changes in yield . It is interesting to be able to see in the black, supply shocks in different countries, possibly representing droughts or other extreme events.

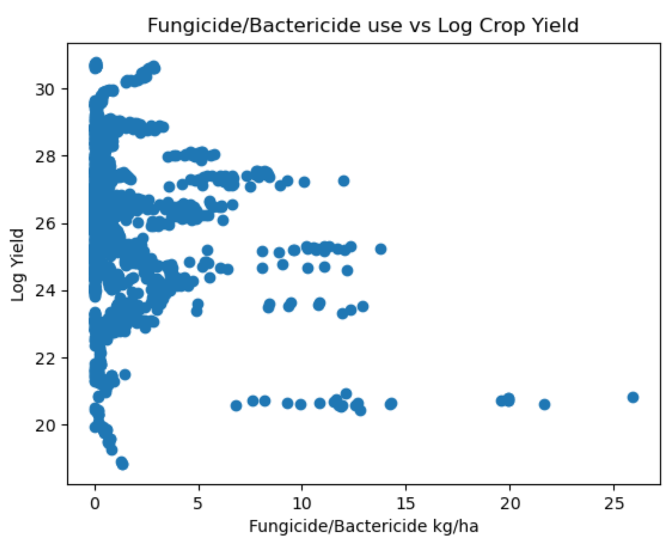
Some basic scatter plots:



Comments:

The GNI effect on crop yield per hectare shows some effects. As we see a higher GNI usually correlates with higher yield per hectare, I transformed Yield in Log Yield per hectare to normalize the graph, without the external effects of the other normalized data by hectare.





Comments:

These charts show a generally non-linear relationship between pesticide use and yield. This could be due to a range of factors, including soil quality. Areas with higher soil quality may use less fertilizer regardless of their wealth level, and thus there may be some omitted variable bias in these variables.

OLS difference in difference: introduction and interpretation

Some comments on variable selection and omission

For this analysis, the difference in difference technique was used in order to see the effects of the 2008 financial crisis and wealth levels on agricultural output. For this diff-in-diff , the dummy variables is_wealthy and post_2008 were created. They were interacted with eachother in order to see the combined effect of wealth and the 2008 financial crisis. The ha_cropland and GNI were eliminated due to redundancy or collinearity. Year was also not used due to collinearity. The average_rain_mm datapoint was changed from mm to st.dev into rain_stdev in order to eliminate multicollinearity with the intercept as the rainfall data was mostly static by country. Log Yield was used to allow for better interpretation of large numbers in the dataset and mitigation of outliers

The diff in diff table:

	Total_Yield	
is_wealthy	0.0	1.0
post_2008		
0	1.146303e+12	9.588017e+11
1	1.434921e+12	8.934289e+11

As we can see the in post 2008, we actually see marginally higher yields of 1.434921 e+12 compared to 1.146303 e+12 for non-wealthy countries. In contrast we can see that for wealthy countries the level of agricultural production actually fell from 9.588017 e+11 kg per country to 8.934289 e+11 kg per country on average. The difference being about 65 billion kgs of crops produced per country on average. In order to see if this change is statistically significant, let’s move on to the actual OLS dif-in-diff regression.

The regression table

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Log_Yield   R-squared:                       0.203
Model:                            OLS   Adj. R-squared:                  0.199
Method:                 Least Squares   F-statistic:                     52.45
Date:                Thu, 16 Dec 2021   Prob (F-statistic):          1.51e-112
Time:                        14:57:45   Log-Likelihood:                -5056.8
No. Observations:                2485   AIC:                         1.014e+04
Df Residuals:                    2472   BIC:                         1.022e+04
Df Model:                          12
Covariance Type:            nonrobust
=================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                       27.1292      0.116    233.246      0.000      26.901      27.357
is_wealthy                       0.2801      0.132      2.125      0.034       0.022       0.538
post_2008                        0.0955      0.102      0.940      0.347      -0.104       0.295
post_2008:is_wealthy            -0.6330      0.185     -3.413      0.001      -0.997      -0.269
rain_stdev                      -0.1226      0.044     -2.797      0.005      -0.208      -0.037
AverageTemperature              -0.0312      0.005     -5.908      0.000      -0.042      -0.021
kg_Insecticides_ha              -0.7266      0.049    -14.881      0.000      -0.822      -0.631
kg_Fungicides_Bactericides_ha    0.0264      0.021      1.272      0.204      -0.014       0.067
kg_Herbicides_ha                 0.2533      0.028      9.049      0.000       0.198       0.308
kg_potash_ha                    -0.0004      0.002     -0.260      0.795      -0.004       0.003
kg_nitrogen_ha                   0.0015      0.001      1.680      0.093      -0.000       0.003
kg_phosphate_ha                  0.0014      0.001      1.232      0.218      -0.001       0.004
kg_manure_ha                    -0.0124      0.001    -12.822      0.000      -0.014      -0.010
==============================================================================
Omnibus:                       26.404   Durbin-Watson:                   0.135
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               28.634
Skew:                          -0.213   Prob(JB):                     6.06e-07
Kurtosis:                       3.309   Cond. No.                         738.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
test score array
```

R^2 : The r^2 is .203 which means 20.3% of the variance in crop yield is explained by the above variables. This is to be expected given our limited variables, and the large scope of our model. As this model is primarily an inference model, a very high r^2 is not required for useful results.

Cond. No.: The condition number is a little high, but not dangerously so. This suggests some levels of multicollinearity but not excessive levels. This possibly due to fact that some inputs to farming are often compliments and are used simultaneously in most cases.

Significant Variables: Interpretations

Is_wealthy: is_wealthy is statistically significant at the 95% level with a p value of 0.034. Meaning that if the null hypothesis were true, we would expect to see this outcome 3.4% of the time. If a country is wealthy, we can expect around a .28 increase in log crop yield. This makes sense as wealthy countries often have higher access to inputs and markets, as well as use modern machinery and technology more often.

post_2008:is_wealthy: The interaction variable between the post_2008 and is_wealthy is highly statistically signiciant at the 95% level and even the 99% level with a p value of 0.001, meaning that if the null hypothesis were true, we would expect to see such a result 0.1 % of the time. If the data point is from a wealthy country and it was collected post 2008. We can expect to see a .63 decrease or 27.1219/(27.1219-0.6330) = around 3.5% decrease in log crop yield as opposed to a non-wealthy country post 2008 or a wealthy country pre-2008. This is important as it means wealthy countries had more of a supply shock than poor countries who were affected minimally. This could be due to a variety of reasons, which will be discussed in the overall interpretation of the model.

Rain_stdev: This variable is significant at the 95% and 99% levels with, a p value of 0.005 , meaning that if the null hypothesis were true, we would expect to see such a result 0.5 % of the time. The coefficient on this variable is actually negative meaning that more every st.dev over the average, the log crop yield goes down by .112 . This is interesting, as it means that very high rainfall actually is not-beneficial for crop yields. This could be due to wet climates promoting fungus and bacteria or due to erosion effects. High rain countries also see high levels of nutrients and fertilizers washed from the soil which could cause poor soil quality.

Average temperature: Average Temperature is highly significant at the 95% and 99% levels with a p value of .000 or practically zero. Meaning we would have an essentially 0% chance of seeing such a result if the null were true. Each 1 degree increase in average temperature actually decreases log farm yield by about 0.03. This may be due to drought prevalence or hot conditions paired with high rain creating humid conditions that increase spoilage and fungal/bacterial growth.

Kg_Insecticides_ha: kg_insecticides per hectare is highly significant at the 95% and 99% levels with a p value of .000 or practically zero. Meaning we would have an essentially 0% chance of seeing such a result if the null were true. For each 1kg increase per hectare we expect to see Log Yield decrease by .726. This result is interesting as we would expect a farm input to have beneficial effects on yield. This result could be due to some omitted variable bias with countries with high insect problems being more likely to use insecticides. It could also be due to high use of insecticides being correlated with overuse in general, which leads to insect resistance and the death of natural predators to insects

Kg_Herbicides_ha: kg_Herbicides per hectare is highly significant at the 95% and 99% levels with a p value of .000 or practically zero. Meaning we would have an essentially 0% chance of seeing such a result if the null were true. For every kg increase in herbicide use per kg we expect to see a .2533 increase in log Yield . This could be due to herbicides reducing competition for nutrients in the soil from other plants as well as the fact that resistance is not as big a problem as any weed not killed by herbicide can be uprooted, burned or killed manually.

Kg_manure_ha: The only fertilizer that was statistically significant was manure, being highly significant at the 95% and 99% levels with a p value of .000 or practically zero. Meaning we would have an essentially 0% chance of seeing such a result if the null were true. For every unit increase in manure usage per hectare we actually expect to see 0.0124 decrease in log Yield. This is interesting as we would expect a fertilizer to have a positive effect on yield. This could be due to a omitted variable bias from technology , as higher usage of manure could be correlated with low levels of technology or pure fertilizers. But that is spectulative. The standard for this variable is also pretty low at 0.001

Non-signiciant variables

Post-2008: Post 2008 is not statistically significant at the 95% level, with a p value of .347, meaning that we would expect to see such a result 34.7% of the time if the null were true. This means, that on a global scale, regardless of wealth level the 2008 financial crisis did not have a large effect on agricultural production

Kg_fungicides_bactericide: kg_fungicides_bacericides is not statistically significant at the 95% level, with a p value of .204, meaning that we would expect to see such a result 20.4% of the time if the null were true. It is unclear why this is the case, but it could be because it is used primarily in humid environments and the tropics, where small plot diversified farming might be common. Small plot diversified farming is often practices in countries like Ecuador, Peru, the Brazilian north, and many south Asian countries.

Kg_potash_ha: kg_potash_hectare is not statistically significant at the 95% level, with a p value of .795, meaning that we would expect to see such a result 79.5 % of the time if the null were true. This insignificance could be due to that it the least used fertilizer and might be used in certain specific soil health situations. The inclusion of soil data would have certainly helped the model, but soil data is often highly localized, so would only really be plausible with a more local agricultural data set. It could also be due to nonlinearities in the relationship

Kg_nitrogen_ha: kg_nitrogen_hectare is not statistically significant at the 95% level with a p value of 0.093 , meaning that we would expect to see such a result 9.3 % of the time if the null were true. . Nitrogen is the most used fertilizer in the world, so one would expect to have an significant effect. This variable might also be statistically insignificant due to the same reasons as other fertilizer data due to omitted variable bias with soil quality . It could also be due to non-linearities in the relationship

Kg_phosphate_ha: kg_phosphate_hectare is not statistically significant at the 95% level with a p value of 0.218 , meaning that we would expect to see such a result 21.8 % of the time if the null were true. This might also be insignificant due to omitted variable bias with soil quality. Meaning that poor soil areas are more likely to use heavy fertilizer, such that the effect of poor soil is incorrectly being combined with the fertilizer coefficients decreasing their magnitude. It could also be due to non-linearities in the dataset.

General discussion of the model results

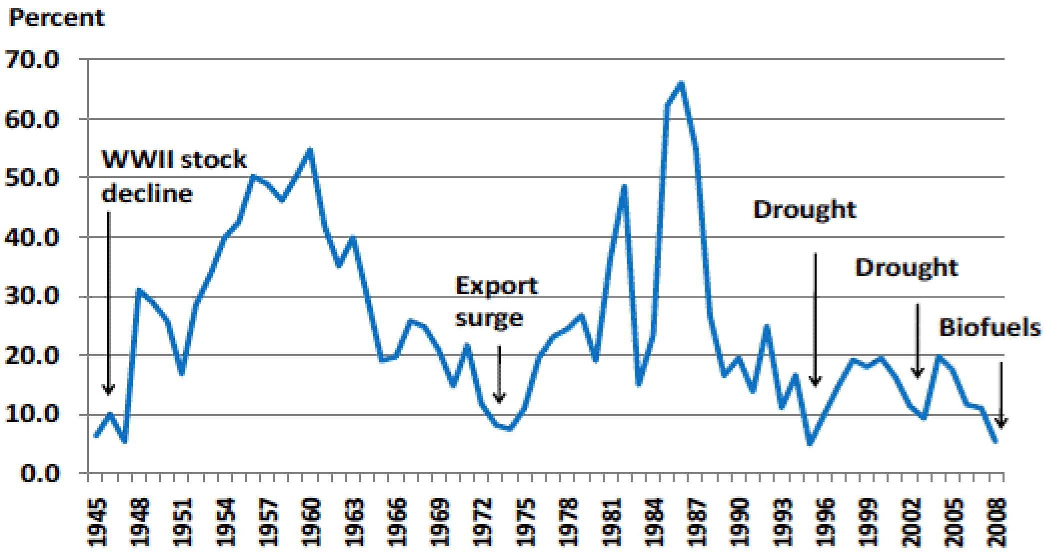
The OLS diff-in-diff shows the 2008 financial crisis seemed to have no major effect on global agriculture yields, but the affect of the 2008 financial crisis on wealthy countries vs non-wealthy was highly significant representing around 3.5% decrease in output for wealthy countries. This brings up several questions about the reliance of wealthy countries on the financial markets for agriculture, as well as the interconnectedness of wealthy markets with global trade. There could be several reasons for this decrease in productivity:

Hypothesis 1: Wealthy countries tend to have larger plot industrial farms, so farmers will often leave fields fallow and reduce complex inputs (as evidenced by the fertilizer and pesticide graphs). Another practice is to allow spoilage if a farmer cannot sell their produce. A smaller plot farmer in a poorer country is more likely to use all the space available with less input demands.

Hypothesis 2: The financial markets may have affected the dip in productivity. For example US farm insurance, subsidies and futures, all prop up farming and might have faltered with government spending cuts or changes in rates. The futures market though as shown by sources, actually skyrocketed post 2008.

An alternate interpretation: As shown by figures in earlier in the report, the decrease in yield for wealthy countries actually started to decrease a lot earlier than 2008. This was for a variety of reasons , according to an article by Professor David Harvey from University of Newcastle in the UK, there were several droughts in the USA leading up to 2008 and biofuels continued to press down the crop yield in 2008 as farmers switched from edible crops to biofuels . This process might have been replicated in other developed economies. The interesting thing, is that despite this decrease in yield, farmers actually started to do better in 2008 since biofuels caused a skyrocket in futures for corn and other grain crops. This rapid increase is US commodity prices is reported by the USDA that same year. So this might be an example of how edible crop yield can decrease, but farmers can financially improve their profits.

Graph of US corn stocks (Credit David Harvey)



The futures market may have also helped non-biofuel farmers as remembering the crop yield figure showing yields between rich and poor countries we can see a rapid rebound by rich countries in 2009 and beyond.

Sources for analysis:

USDA:

https://www.ers.usda.gov/amber-waves/2009/march/agricultural-commodity-price-spikes-in-the-1970s-and-1990s-valuable-lessons-for-today/

Article on the mid 2008 Food price Crisis by Professor David Harvey

https://www.staff.ncl.ac.uk/david.harvey/ACE2006/Principles/2008FoodPriceCrisis.html

b

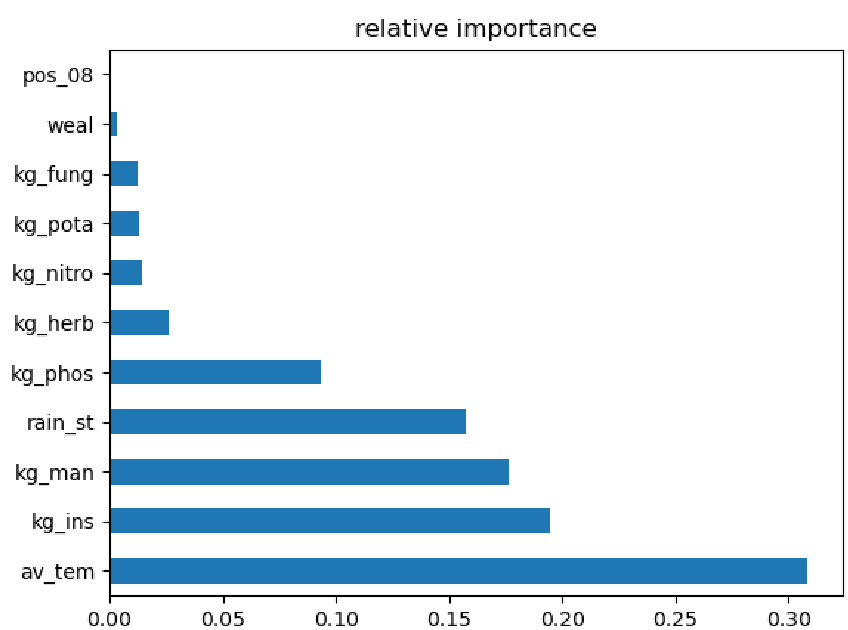
A small experiment in Random Forest:

Even though this very much a inference problem, I decided to see if more flexible model such as random forest could help with some of the non-linear relationships in the data as well as provide a small benchmark for prediction accuracy on this dataset. For the data, I used all the same variables as the OLS regression, but I also preprocessed with by normalizing the data between 0 and 1 which improved the accuracy by around .2%.

Results:

```
test score array
[-0.06638515 -0.04801939 -0.06154591 -0.06472422 -0.05219917]
Mean percentage error
-0.05857476807669276
```

The mean percentage error ended up being around 5.857 % which is surprisingly good for the non-local level of the data, and the lack of soil and other data. This could be due to the fact that many of the inputs are non-linear and predict yield much better in a more flexible model. The importances are as follows:



The importances show that av_tem is the most relevant, but that also phosphorous is very important for node purity in random forest when the dataset is iterated. The importances mostly align with significance results from the OLS, with the dummy variables being reduced in significance. The importances show that some of the fertilizers seen as insignificant by OLS might actually be important, but just have non-linear relationships. For example, a fertilizer might have a parabolic relationship with output, with a poor country with lack of inputs to use a low amount, the rich country with good soil to use some but not a lot, and a rich country with poor soil to use excessive amounts of fertilizer.

Some ideas on how to improve prediction of agricultural data:

According to this paper by Shahhoseini , Hu, Huber and Archontoulis, localized soil data is very important for making accurate predictions, as well as not only rainfall data but water table measurements. The paper suggests using hybrid models for the strongest results, especially for time series data. Other things potentially missing from this model are more detailed daily weather data and monthly harvest data. For future investigations into crop yield and inputs, a more localized dataset would be beneficial. For our purposes in this report, a OLS model with a broader dataset was suitable as we were trying to measure global effects of the 2008 recession across multiple countries, and cared less about the effects of certain hard inputs.

Link to Shhoseinei , Hu, Huber and Archontoulis paper:

https://www.nature.com/articles/s41598-020-80820-1

Ideas for future investigations:

Since we have shown possible effects of the financial markets on crop yield, it would be interesting to include more US financial data as well as just focus on US crop yields to see the localized effects of crisis such as the 2008 crisis, as well as the current financial turmoil caused by covid-19. Localized data would also give us the benefit of county level data and weather reports that would be useful for partitioning drought and extreme weather effects out from our data. In conclusions this investigation proved to be interesting and educational, and revealed that more attention should be paid to agricultural markets in the wake of a recession and not just financial, and services markets.

Total list of sources:

USDA report on commodities skyrocket in 2008:

https://www.ers.usda.gov/amber-waves/2009/march/agricultural-commodity-price-spikes-in-the-1970s-and-1990s-valuable-lessons-for-today/

The 2008 food price crisis:

https://www.staff.ncl.ac.uk/david.harvey/ACE2006/Principles/2008FoodPriceCrisis.html

Predicting agricultural data using hybrid models

https://www.nature.com/articles/s41598-020-80820-1

effect of agricultural commodities FAO

https://www.fao.org/3/am320e/am320e.pdf