Pesticide's Relation to Crop Yield in the US

Andrew Steininger
Sophomore Undergraduate – CSCI-B 365 – Professor Ye
Luddy School of Informatics, Computing, and Engineering
Bloomington, Indiana, United States
asteini@iu.edu

***Abstract* — Over the last several years, there has been a growing interest in pesticides and their effects on crops, the people who apply pesticides, and those who consume food that grows with pesticides. This is an exploration into the frequency of pesticide use over the mid-to-late 20th century into the early 21st century and the trends in crop yields that may result from it. The future allocation of pesticide resources is a crucial decision to make due to the side effects that pesticides have on the outside environment, so deciding where pesticides work the best is very important. The objective of this project is to isolate the frequency of certain pesticides and determine what may have the best effect on the yield of certain crops across the country.**

***Keywords* — Agriculture, Pesticides, Crop Yields, Farming, Climate, Ag Optimization, Yield Prediction, Pesticide Effectivity, Linear Regression, Logistical Regression**

I. Introduction

"Agriculture is our wisest pursuit, because it will in the end contribute most to real wealth, good morals, and happiness." — Thomas Jefferson, Third President of the United States of America. President Jefferson's words ring true even as we enter a new age of agriculture. Through artificial intelligence, machine learning, and other data analysis techniques, along with advancements in chemistry and biology, we are learning new things every day about what we can grow and how to be more efficient with increasingly shrinking farmland and a growing population.

Some of the biggest innovations in the industry are through the chemistry of pesticides that are used on the nation's best-yielding crops. Pesticides have a history steeped in controversy, but one thing that we do know is that they work. Crop yields grew dramatically in the mid-to-late 20th century because, from 1955, approximately 3% of all acreage planted with corn and soybean crops were treated with an herbicide; by 1985, that figure had increased to more than 95% [2]. Since that 1985 mark, pesticide usage has wavered, and questions about their side effects and consequences have made it easier for interest groups and lobbyists to turn legislation against some of these chemicals.

This project will find what pesticides are still (last 30 years) in use and at what frequency. In addition, it will investigate what trends, if any, are visible in US crop yields and whether they correlate to the increase in the use of certain herbicides or certain areas where those herbicides may have more prevalence.

II. Methods

A. Data Acquisition and Cleaning

There are two main datasets for the project. The first comes from Kaggle [3] and is a set with information about types of pesticides and locations within the 50 states. It also provides estimates for the total tonnage of each pesticide used for each year from 1994-2024. This will be the most important value from the pesticide set when compared with the second data set. The second set is from Our World in Data [1] and provides estimates for yields of all sorts of crops from all over the world, including corn, beans, cotton, and wheat, which, because of their prominence in the US, will be the main focus of the paper. There are also two supplemental datasets [6] [7] that provide temperature averages and rainfall averages for the regression model. These were not read into the project notebook but directly coded in due to the fact that they were not in CSV or text format and were only the years between 1992-2019, so the data was not large.

B. Data Preprocessing

For the dataset from Our World in Data, which has data from the whole world, there will need to be a new set with just the U.S. data. For this, we will subset the data so that the new data can be sorted by crop type and year and be treated as its own dataset. The other dataset, about pesticides, will have a lot of preprocessing to do as well. Some null values will need to be taken care of along with the state and county FIPS codes that will need to be sorted to make sense of the regionality of the data.

C. Creating Trend Lines and Regression Models

The regression model and trend lines will be the main methods for transforming the data into readable trends and comparable trends. Using sk-learn and numpy, there will be a bunch of regression lines for many variables, but especially the types of pesticides and crop types. Then, using simple linear regression, we will look at the trends of the crop yield data and the explanatory variables that lead the model to reach the given conclusion.

D. Predicting Future Performance

Once the trend lines are created and the variables that are most closely related are chosen, then the regression patterns will move the trend lines into the foreseeable future so that it can determine what resources are best used, how they should be applied, and foreshadow where they should be used the most. This is important because it will help set the trend of the country's agricultural market for the incoming crisis of shrinking farmland and growing population.

III. Results

The data preprocessing was done in a fashion that led to the ability to create trend lines for the ~30 years' worth of data present for all variables. This was the starting point because there needed to be some kind of introduction to what was happening across the board before turning the data into a regression model.

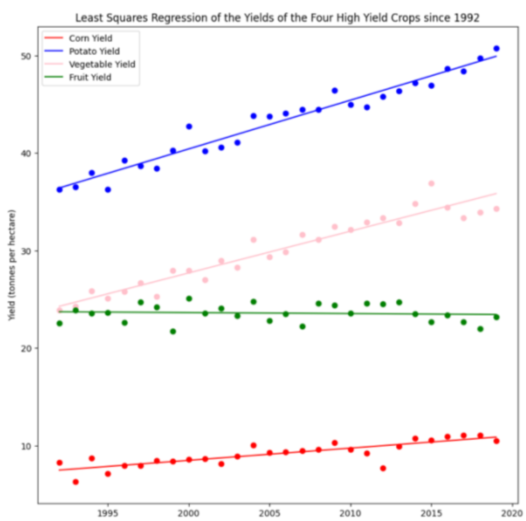

Fig. 1. The 30-year yield trend of the four highest-yielding crops of the eight that are being used in the study

Figure 1 shows that, on average, the yield (tons per hectare) trend of four of the eight crops that are being examined (corn, potatoes, fruit, vegetables) is going upward. Examining this and the other four crops (soybeans, cotton, wheat, tobacco), which are lower-yield crops for which the plot is very messy, we find that the upward trend generally holds for six of the eight being examined. The two that don't follow the trend are fruit and tobacco, which, when considered, makes sense as fewer resources have gone to bettering our fruit and tobacco yields in the U.S., as the country imports more fruit than before and the level of cigarette smoking has continued to drop dramatically, especially since 1992, when the data for this study starts.

For the regression model, the crop yields of all eight crops were averaged into one variable so that it could be properly compared with temperature, rainfall, and, of course, pesticide usage. This, and the upward trend across all of the yields, are displayed in Figure 2.

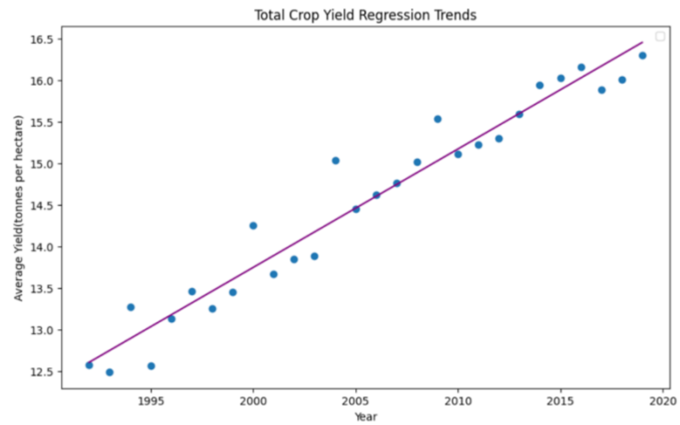

Fig. 2. The 30-year yield average trend of the eight crops that are being studied in this paper (corn, potatoes, fruit, vegetables, soybeans, cotton, wheat, tobacco)

Now, the study moved to finding the trends in specific pesticide use. For this, the nine highest-use pesticides, herbicides, or insecticides were selected. Their trends are displayed in Figure 3.

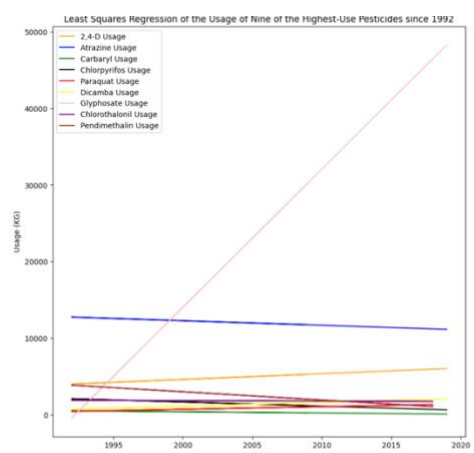

Fig. 3. The 30-year yield average trend of the nine pesticides that were selected to best represent the overall trend in usage (KG)

As will be displayed later, the average pesticide use is climbing at a rate similar to that of the overall crop trend. This chart shows that while many of the older, more popular herbicides are trending downward, the pesticide glyphosate has exploded in usage over the last 30 years, dragging the overall trend upward. Much of this is due to the Environmental Protection Agency's (EPA) efforts to ban or limit the use of things they find harmful to the overall environment, as explained in a Sciendo study from 2009 [5]. These efforts, however, have yet to be targeted at glyphosate.

Using these two variables, along with average rainfall across the contiguous U.S., which shows a slight downward trend, and average temperature across the contiguous U.S., which shows a slight upward trend, there is a linear regression model to be made to predict crop yield.

This regression model examines the crop yield as the variable to be explained, with the other three doing the said explanation. The first output is a heatmap describing the correlation between all the data points, as shown in Figure 4.

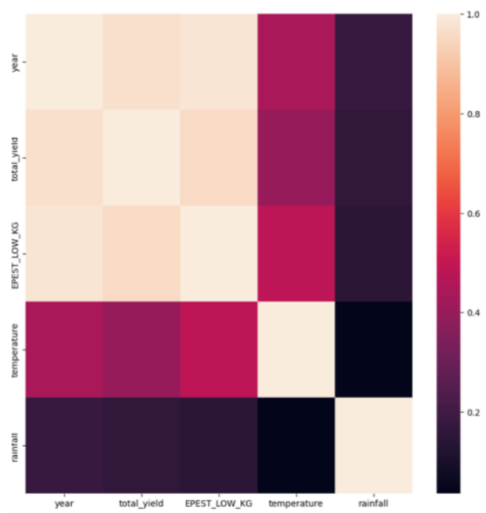

Fig. 4. The correlation heatmap between the variables in question with lighter colors being closer to 1.0 and darker colors being closer to 0.0 (no negative correlation between variables)

The heatmap displays a few things that are very interesting in terms of the relationships that are relevant. The first noticeable thing is the very high correlation between the average yield of the crops (total_yield variable) and the total pesticide usage by year (EPEST_LOW_KG variable). The correlation sits at 0.955, which is extraordinarily high when compared with the correlation between crop yield and temperature, which is 0.399, and between crop yield and rainfall, which is only 0.163.

The performance metrics indicate that this regression model does a good job of capturing the variance in the dataset with an r-squared score of 0.861 and a mean-squared error of 0.179 tons-per-hectare-squared. This is largely due to the strong correlation that the pesticide usage shares with the crop yield.

One other interesting thing that was done to help reach our goal of determining the most efficient use of pesticide resources was extrapolating the pesticide use data for the years 2020-2022 and creating crop-yield predictions to compare to the real data we have for those years. This is displayed in Figure 5.

year	predicted_yield	total_yield
2020	16.008512	16.397974
2021	16.199462	15.946505
2022	16.202097	15.828090

Fig. 5. Predicted yield vs. actual yield (2020 – 2022)

As is visible by the results in this table, the model does a good job of generally getting the yield value close. This was predicting the value just based upon extrapolated pesticide use, so it will not be perfect. The important thing is that we know we can do something similar to this, as will be talked about in the discussion.

The last thing we can do to help reach our goal of determining the most efficient use of pesticide resources is to determine where the need is most significant. For this study, in addition to the regression model that predicts crop yield, we can see a map that determines where this is most relevant and most needed, which is displayed in Figure 6.

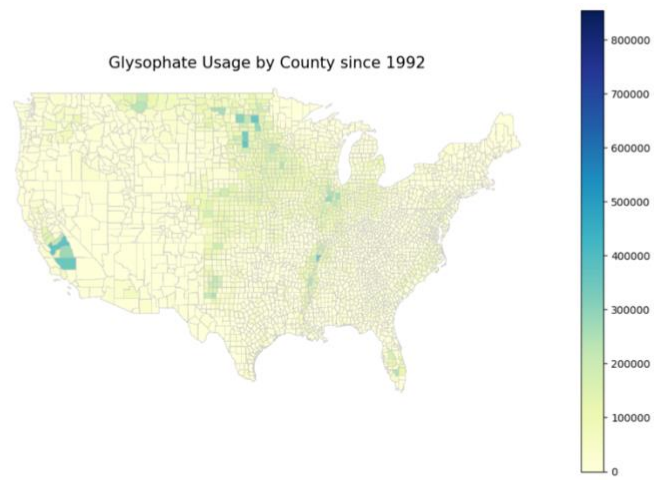

Fig. 6. Map showing the total glyphosate usage by U.S. County since 1992 with darker colors meaning more usage

The map shows three main areas where glyphosate usage is the highest: Tornado Alley (South Dakota to West Texas), the Corn Belt (North Dakota to Indiana), and central California. As will be mentioned in the discussion, this is where this study should be most relevant and further researched.

IV. Discussion

The results of this study demonstrate a strong correlation between crop yield and pesticide use since 1992. This suggests that pesticides play a large role in increasing agricultural productivity. This relationship is very strong, but it does not imply any sort of causation. There have been other significant advances in agriculture in the last 30 years that could explain some of these trends as well. These include genetic modification of seeds and improved irrigation systems, just to name a few.

When considering the original purpose of this study, the allocation of pesticide resources in the future, we also charted the trends and patterns of the location of glyphosate usage to better describe where more studies and research could be conducted on a local level.

This was important because this model had a very small sample size due to the pesticide data only being available for around 30 growing seasons, along with the lack of regionalized data for crop yields. For all intents and purposes, this model should be considered a pilot model until more regionalized data can be specified and more total data points are collected.

This model considered averages over a large swath of land, which is very different across it. Averaging the crop yields, temperature, and rainfall data may obscure crop-specific or location-specific trends that will be better discovered through local tests.

Another important piece of value that this study found was a method for government regulators to see what kind of effect capping pesticide use would have on crop yield. This is because, as was done for the years 2020-2022, we could extrapolate the pesticide use data and feed it into the model to determine what kind of crop yield would result from it. This extrapolation could easily be done with a "capped" pesticide tally to determine the crop yield.

Overall, this study is at a very basic level in ag data analysis. The field of agriculture still has a long way to go when it comes to having not only enough data, but also enough data for specific regions to help grow the number of data points for the models. There is a hope that this project will help continue the data revolution in agriculture and grow into a good basis for the future. These things must happen as we continually lose farmland and gain people in the world to feed. Agricultural data science has a long way to go, but this is the right start for the future.

References

[1] In. (2024). *Background and Approach to the Study*. Nih.gov; National Academies Press (US). https://ncbi.nlm.nih.gov/books/NBK236265/

[2] Konrad Banachewicz. (2019). *Pesticide Usage in the United States*. Kaggle.com. https://www.kaggle.com/datasets/konradb/pesticide-usage-in-the-united-states

[3] *2024 TIGER/Line® Shapefiles*. (2024). Census.gov. https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2024&layergroup=Counties+%28and+equivalent%29

[4] Tudi, M., Daniel Ruan, H., Wang, L., Lyu, J., Sadler, R., Connell, D., Chu, C., & Phung, D. T. (2021). Agriculture Development, Pesticide Application and Its Impact on the Environment. *International Journal of Environmental Research and Public Health*, *18*(3). https://doi.org/10.3390/ijerph18031112

[5] Aktar, W., Sengupta, D., & Chowdhury, A. (2009). Impact of Pesticides Use in Agriculture: Their Benefits and Hazards. *Interdisciplinary Toxicology*, *2*(1), 1–12. National Library of Medicine. https://doi.org/10.2478/v10102-009-0001-7

[6] National Weather Service. (2018). *National Weather Service*. Weather.gov; National Weather Service. https://www.weather.gov/

[7] *IBISWorld - Industry Market Research, Reports, and Statistics*. (2024). Ibisworld.com. https://www.ibisworld.com/us/bed/average-annual-precipitation/489/