Displaying the Correlation Between Aging Farmers and Farm Productivity

HDSI Agri Datathon 2024
Team Boom!
6 October 2024

Josh Zhang
Harvard University
joshzhang@college.harvard.edu

Chris Li
Harvard University
chris_li@college.harvard.edu

Joshua Zhang
Harvard University
joshua_zhang@college.harvard.edu

Abstract

This project seeks to analyze the aging farmer population in the United States and its economic impact on agricultural productivity on both state and county levels. Between the years 1997 to 2022, the average age of farmers has trended towards an increase at both the state and county level, correlating with a decrease of agricultural productivity and sales for key crops in select regions. Using NASS demographic and agricultural production and sales data, the study conducts multiple regressions to generate choropleth maps that visualize the economic impact of America's aging farmer population. The findings from the study aim to isolate key crops that may play a statistically significant role in reducing agricultural productivity as the average farmer age increases

1. Introduction

As the United States continues to grapple with the changing effects of climate change and the warming environment, food security serves as a main topic of concern when considering how daily human life may be impacted by such changes. While the demand for food remains high in the midst of an ever-increasing population, farmers serve as the backbone for supplying US food resources.

However, from the years 1997 onwards, the average age of farmers has increased, reaching an average of 58.1 years by 2022. As the average farmer age continues to grow, this may pose multiple challenges to the production capacity of farms among select counties in the United States. Thus, we seek to answer the following questions: How has the average age of farmers evolved by county and state? How does the average age of farmers correlate with the economic scale of farms? In counties where agricultural production has declined, what trends may exist regarding the ages of farmers in these locations? How may adjusting for inflation reveal affect the relationship between farmer age and farm production?

To analyze this impact, we generated multiple choropleth maps to visualize aging trends among farmers on both a county level. We then modeled the correlation between farmer age and agricultural productivity sourced from NASS sales data, using multiple linear regression to map the relationship between the two predictors. Lastly, the monetary sales data is adjusted to current inflation rates, giving a more accurate representation of the relationship between the sales data and the aging farmer population.

2. Data and Methodology

2.1. Data Cleaning

To answer the questions above, the team used the following data files:
- *prompt2_demos_landtotals_county.csv*
- *sales_data_county.csv*
- *state_level_2002_2007_2012_2017_2022*
- *inflation_GDP_price_index.csv*
- *prompt2_prompt3_sales.csv*

The first step of **data cleaning** was to get rid of unnecessary columns and rows in the dataset. For all respective datasets, **(D)** and **(Z)** values were all converted to **NaN**s.

Note: To view the trends for the average age of all farmers between 1997 and 2022 on a county level, the original dataframe from *prompt2_demos_landtotals_county.csv* was grouped into average ages by year, which was stored in an array of dataframes for years 1997, 2002, 2007, 2012, 2017, and 2022. For each year's dataframe, we concatenated each state FIP code with its corresponding county, where the resulting dataframe was processed to produce a choropleth map containing all average farmer ages by county for that specific year.

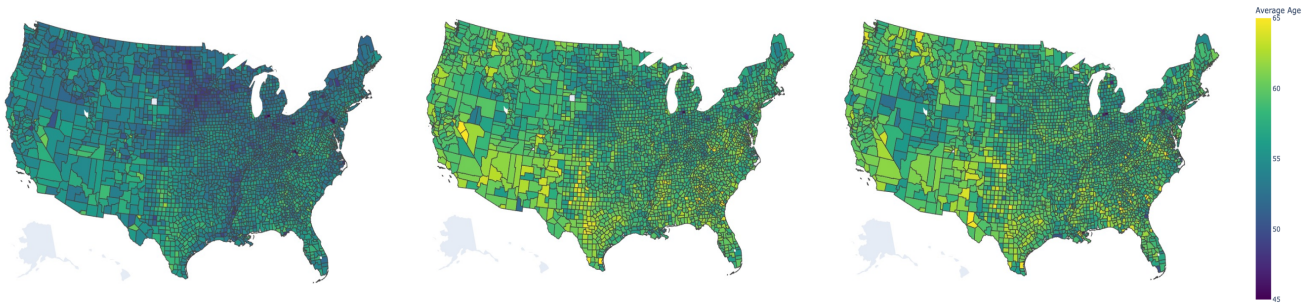


Figure 1: Average age of farmers in the United States between years 1997 (left), 2017 (middle), and 2022 (right). Each observation in the map is an individual county.

2.1.1 Demographic Data By County Cleaning Process

For the county-wide demographic data contained in *prompt2_demos_landtotals_county.csv*, we filtered all age data to only include the average age of farmers across all age ranges, as the original data was segmented into multiple age brackets. Because the label for farmers altered from "Principal Operators" to "Producers" from 2017 onwards. Thus, average age of farmers were joined to produce a single column.

2.1.2 Harvest Data Cleaning Process

After viewing all the columns in the sales data, we found that the useful information of this dataset primarily consists of the quantity of how much each crop was harvested with respect to the state, county, and year. Next, we focused on finding the quantity of acres harvested for any crop. Hence, we removed all the columns that did not have the keyword `'ACRES'`. From there, we removed columns with keywords `'_TO_'`, `'IRRIGATED'`, `'OPERATION'`, `'NON_BEARING'` and `'PRODUCTION'`, as its data was irrelevant for quantifying agricultural production. It is notable to also address that the data did have the quantity of operations (the quantity of harvesting jobs) too. However, we chose to remove this quantity as believe that there would be no efficient or realistic way to standardize values of operations for each crop.

Finally, we filtered out less significant crops that will not play a role in quantifying the overall production by proportioning each respective crop harvested over the total amount of crops harvested for each respective county by year, finding the mean for said proportions, and only keeping the columns where said column proportion was able to be greater than the mean at least once.

2.1.3 Sales Data Cleaning Process

Much of the analysis would require using the sales data, originally provided in nominal dollars. This would not reflect the changes of the purchasing power of the farmers' revenue from 1997 to 2022. Thus, using the inflation data, the proper calculations were done by the *nominal_to_2022real* function to convert all nominal dollars to to 2022 real dollars. Note, this was only done for the columns containing the keyword "DOLLARS", as other columns measured sales, that is quantity of units of crop sold. It would not make sense to perform such conversions on said sales data. A separate csv file was extracted after this, titled *real_dollar_sales_county.csv*.

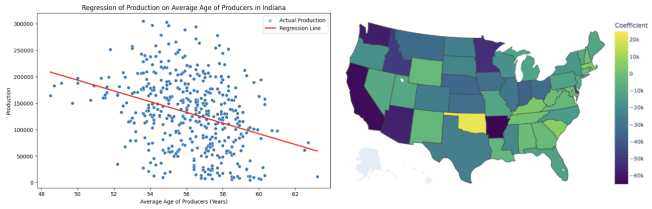
3. Results

3.1. Agricultural Production as a function of Spatiotemporal Trends In Farmer Age

On a state level, all states saw an increase in average age when only viewing the average between years 1997 and 2022, with Washington State having the largest increase in age of 6.56 years, while the average change of age is 4.00 years. Iowa has experienced the largest rate of change with an average rate of 0.65, with Utah the lowest at 0.13. Overall changes in age by state can be seen in Figure 1.

Goal: examine correlation of agricultural production, defined as the sum of total acres harvested (all crops), animals and animal products sold, average age over time as our predictor. Correlation was measured at the national, state, and county level. At the national level, a clear negative correlation is observed. This was done with a linear regression, using the ordinary least squares method (OLS).

At the state level, negative or weak negative trends were observed for all states except for Massachusetts and Oklahoma. In our Python notebook submission, scatter plots with linear regression, similar to the one below are plotted for every state. Such calculations were also done at the county level. More can be found in the Python notebook.

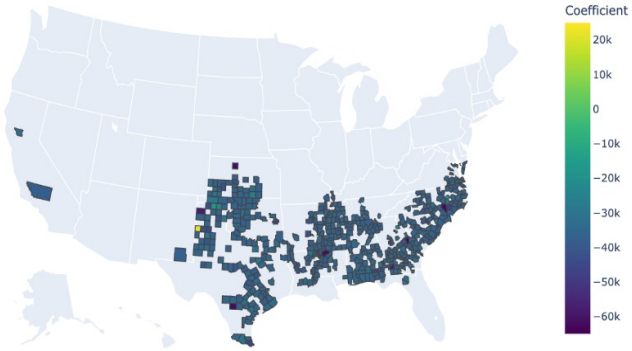


3.2. Agricultural Production Measured in Dollars as a Function of Average Age over Time

First, we quantified agricultural production measured in dollars also known as `'PRODUCTION'`. To identify the variables used to quantify this value, we examined a frequency map that counted the number of times a crop's harvest proportion exceeded the mean proportion for its respective county and year. We only included crops that were commonly harvested within America, defined as those that appeared over 5000 times in the frequency map, as all other crops appeared fewer than 2000 times. The crops utilized were corn, cotton, hay, soybeans, and wheat. After merging our datasets from the data cleaning section, we calculated the corresponding dollar quantity made for each respective crop to determine our variable. In addition, animal-related sales, including animal products, were considered in this variable. All variables were added together to make `PRODUCTION`. Rows that did not contain data for all of these variables or for `AGE_AVG_MEASURED_IN_YEARS_AND_PRODUCERS` were dropped. Rows missing data for certain years were also dropped, as we were concerned that the lack of degrees of freedom at the county level could lead to inconsistent linear regression models.

3.2.1 County Level Trends

Next, we ran an Ordinary Least Squares (OLS) linear regression on `PRODUCTION`, with `AGE_AVG_MEASURED_IN_YEARS_AND_PRODUCERS` as the predictor, for each county. The null hypothesis was that the predictor has no effect on agricultural production (i.e., the coefficient is equal to 0), while the alternative hypothesis was that the predictor does have an effect. We ran a regression for each county and stored each county's p-value and predictor coefficient (slope) in two new columns: `REG_PVAL_COUNTY` and `REG_COEF_COUNTY`, respectively. Given that the threshold for statistical significance was set at the standard of 0.05, we excluded all counties where the p-value exceeded this threshold. From there as analyzing the choropleth as shown below, we found that the only remaining counties with statistically significant relationships between `AGE_AVG_MEASURED_IN_YEARS_AND_PRODUCERS` and `PRODUCTION` were located in America's South. Practically all counties in the South showed a significant inverse relationship between average age over time and agricultural production. Therefore, we have found that there is a strong trend in an inverse relationship between the two, however, this is only applicable to the American South.



3.3. Variations In Correlation Coefficient By Crop

Thus, we implemented the functionality to extract all of the crops that are the top produced crop by acres in at least one state. From there, we obtained the average correlation coefficient of all the states in which each particular crop is the top produced. For example, soy beans as a crop was extracted, because it is the top produced crop in 10 different states, like Arkansas for instance. Then, we averaged the correlation coefficients of all of these 10 states. This average correlation coefficient gives us an idea of how much average age in different states correlates with production levels for a certain crop. Pictured in the following bar graph, for all of the crops that are the top produced in at least one state, average age increasing correlates with acres produced decreasing, hence all of the negative values plotted. For some crops, like hay, the average coefficient is not very

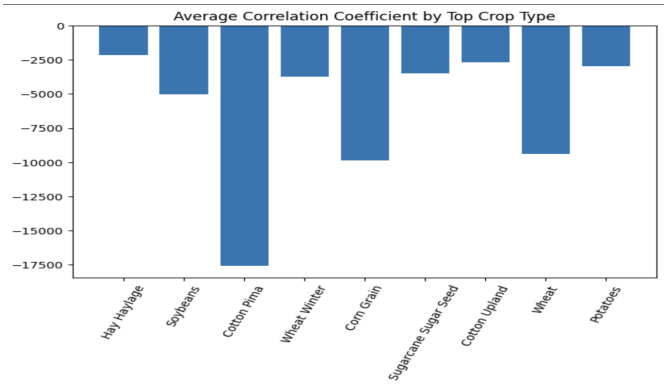


Figure 2: Average correlation coefficient for top produced crops in 10 states experiencing lowering agricultural productivity as a result of aging farmers.

steep, while for cotton, the slope is much more negative. We can see that for crops we know to be more labor intensive, like cotton, average age increasing seems to correlate with less production, as we defined in section 3.1. Note that the data is not completely cleaned these crops, as there are some crop labels that are subsets of other crops, and both were extracted, as perhaps different labels were used in data set, suggesting that correlation can vary between states. Note that the crops extracted are the top produced for different number of states, which means we are not guaranteed that for each correlation coefficient we are operating with the same sample size.

3.4. Effects of Inflation

When we only observe nominal dollars, slope (coefficient of the predictor) of linear regression at the national level, measuring agricultural production is $-1.176e + 04$ when working with nominal data, and $-1.392e + 07$ when working with real 2022 dollars. Thus, there is a sharper decrease, showing that actual purchasing power and economic strain on farmers is greater.

4. Conclusion

After extensive data cleaning and analysis of all five NASS datasets provided, we have provided a useful metric to define the monetary value of agricultural production that we can measure the correlation between the average age of farmers and its economic impact on farm production. From this metric, we implemented the negative relationship between agricultural productivity and average age to all US states, isolating key southern states that exhibit statistically significant decreases in productivity from a select few crops. While we have successfully isolated this correlation to specific states and select crops, there exists room for inquiry to investigate the drivers behind the selected crops, which may be due to the demand of physical labor required to raise and harvest the crop. With these findings, one may pivot attention to a select few states and crops to produce recommendations toward reducing the negative impact that aging farmers may have on agricultural productivity.

Please see this link to our Submission Video.
Please see this link to our Google Colab Notebook.

References

We only used the provided data as our sources. This included inflation data from 1997 to 2022, NASS data about agricultural sales by county.