Analysis on Crop Diversification, Agricultural Land, and Production Metrics

HDSI Agri Datathon 2024

Team Crop Pop

Samuel Damon	Frank O	Sai Akhil Rayapudi	Harish Narava
Umass Boston	Student	NEU	NEU

Abstract

This study examines the interplay between crop diversification, agricultural land use, and production metrics across the United States, drawing on historical data from the Census of Agriculture. Since 1974, farms have been defined by a sales threshold, ensuring comprehensive coverage of various crop types and regions. By focusing on key agricultural areas such as the Midwest, Southeast, and West, we explore how shifts in crop production have correlated with changes in farmland size, land value, and productivity. Additionally, we assess the role of external factors, including urban expansion, in influencing crop distribution and land use patterns. This analysis aims to provide insights into the evolving dynamics of U.S. agriculture and contribute to strategies for sustainable land management and agricultural development.

1. Introduction

The U.S. agricultural sector has experienced significant changes in crop diversification, land use, and production patterns over recent decades. These shifts have been influenced by a variety of factors, including market demands, technological advancements, and external pressures such as urban expansion. Understanding the dynamics of crop diversification is crucial for addressing key issues in agricultural sustainability and economic efficiency. This project seeks to analyze how crop diversification has evolved over time across various regions in the contiguous United States, focusing on the relationship between changing crop patterns, farm sizes, and land values.

By examining historical trends in crop production, we aim to uncover patterns where certain crops have become increasingly dominant while others have declined, reflecting shifts in farming practices and regional agricultural preferences. Additionally, we explore how changes in farm size, whether measured by acreage or revenue, correlate with these diversification patterns. This analysis will provide insights into the underlying economic and environmental factors driving these changes, contributing to future land use strategies and agricultural policies.

2. Data and Methods

The analysis relies on three primary datasets, “land_use_farm_ops.csv”, “Prompt3_wide_3.0.csv”, and “sales_data_county.csv” which provide a detailed view of land use, management practices, and crop production across different regions of the United States. These datasets capture metrics related to cropland, pastureland, woodland, farm operation details, harvested and irrigated land, and the number of farming operations by crop type.

2.1. Data Preprocessing

a) Missing Data Handling: The datasets were first examined for missing values, and any gaps were handled through either imputation or exclusion, depending on their significance.

b) Data Cleaning: Special characters and formatting inconsistencies were resolved, particularly for county names, which were standardized using FIPS

codes. Inflation adjustment was applied to economic data to account for changing dollar values over time.

c) Feature Engineering: New variables were derived from the original datasets, including farm revenue per acre and diversification indices, which quantify the variety of crops grown in each county or region.

2.2. Methodologies

a) Exploratory Data Analysis (EDA): Descriptive statistics and visualizations were used to identify historical trends and spatial patterns in crop diversification and farm size across the U.S. Choropleth maps were created using Python Plotly's choropleth function to visualize diversification trends at the county level.

b) Trend Analysis: Historical shifts in crop diversification were analyzed through time-series models to assess how crop production patterns have changed over time in different regions.

c) Correlation Analysis: The relationships between farm size (acreage and revenue), crop diversity, and land value were explored using correlation coefficients and regression models.

2.3. Tools Used

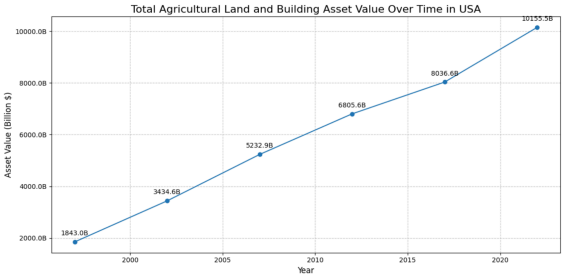
a) Python Libraries: Key libraries used for analysis include pandas for data manipulation and Plotly for visualization.

b) Choropleth Mapping: Python Plotly was employed to create choropleth maps that visualize crop diversification across counties, helping to highlight spatial patterns and trends.

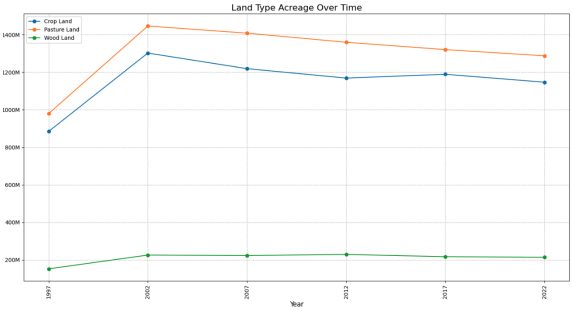
3. Results

We start at the country level and drill down to the state level.

First we observe there is a clear trend that the value of the land has increased over time (Fig 1):

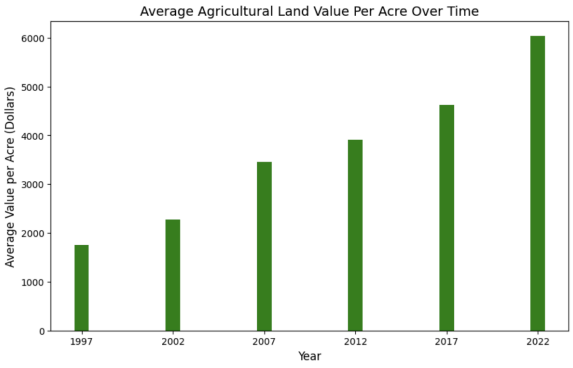


We compared that with the amount of land that is dedicated to agriculture. From 1997 till 2002 there was an increase, then constantly dropping for the next 20 years. In total we cannot see any significant increase over time (Fig 2):



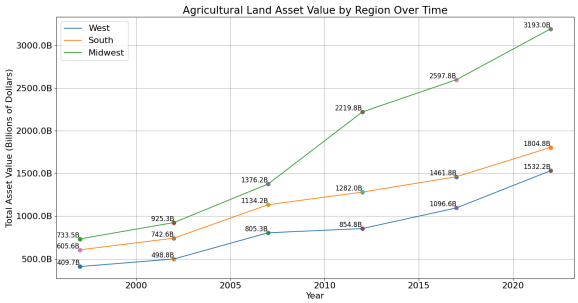
Hence, together with the previous result in Fig 1, it means that the value of the land is increasing. Further research, not covered in this study, could probably reveal the reason for that increase till 2002 and the later constant decrease for 20 years.

If we look into the price per acre the results look like this (Fig 3):



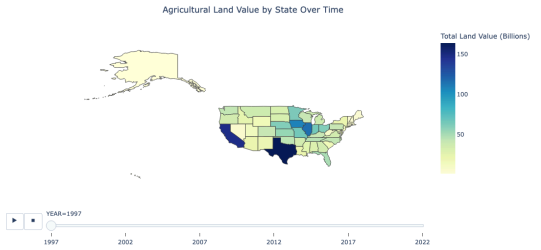
Which means that in the time of our observations (1997-2022) the price of an acre has increased from a bit less than $2000 till around $6000, a 30% increase in 25 years.

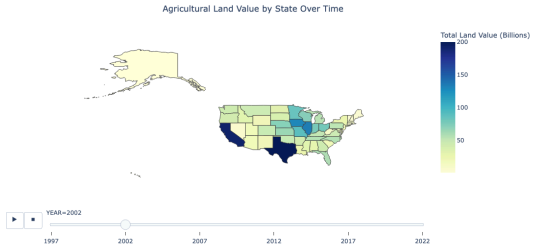
From the point of view of regions in terms of land value (Fig 4):

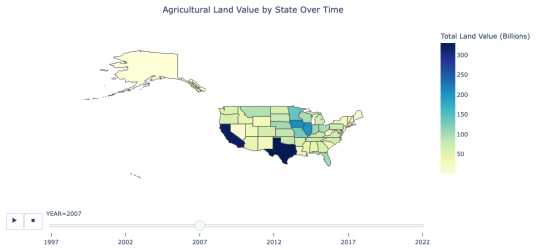


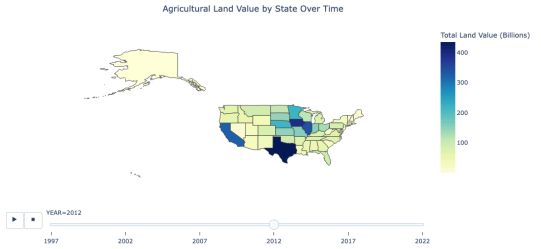
We can see that the midwest is leading this rank and double the South and West.

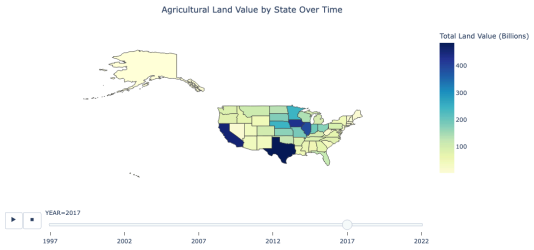
If we look into this by states, per year (Figs 5-11):

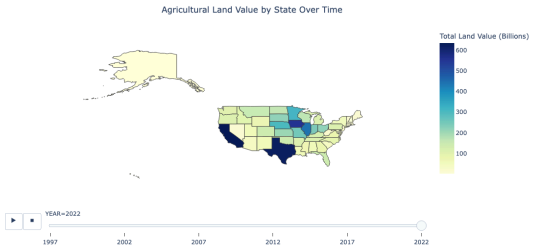






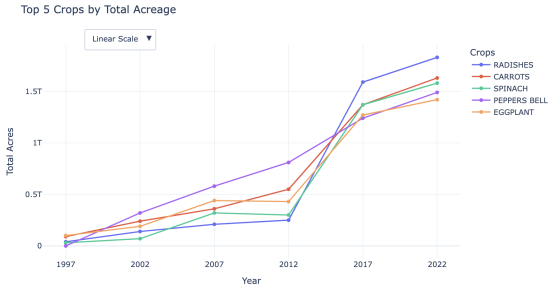






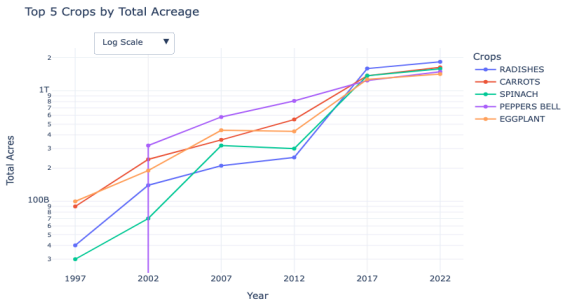
We can observe that Texas in the South has always been leading this rank. The value of land in California in the West and Iowa in the Midwest has significantly increased. The rest of the Midwest states also got an increase over the years but less noticeable compared with Iowa.

In terms of crops, the following (Fig 12):



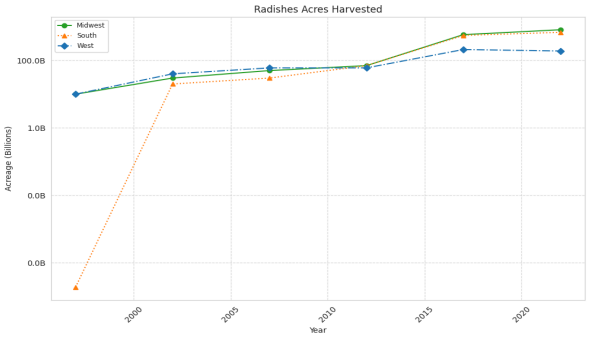
Showcases how the top 5 most crops based on total acres within the states had increased significantly their acreage in the observed time. We can see that radish has changed from the 5th position in the ranking till becoming 1st. This extreme increase happened in a period of 10 years, from 2012 till 2022.

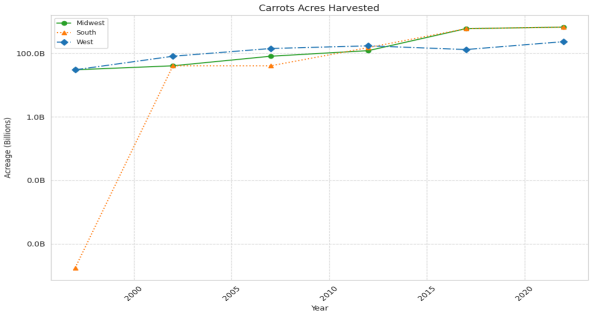
We also displayed the previous values in the log scale (Fig 13):

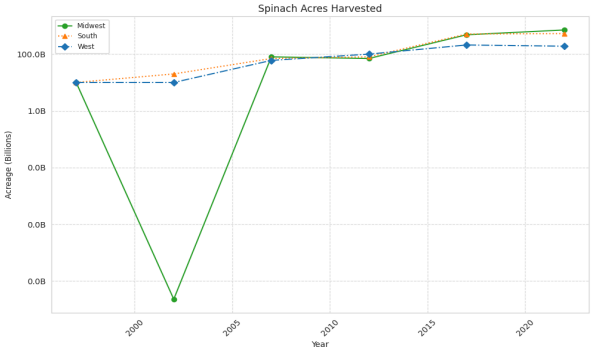


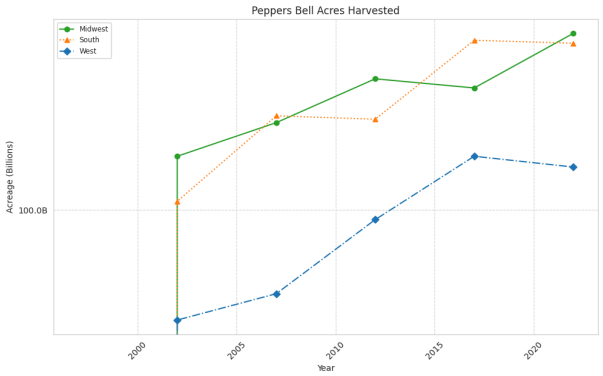
With this, we could detail that there is no data available for peppers bell till year 2002. A further investigation would likely give us insights of the reason behind that.

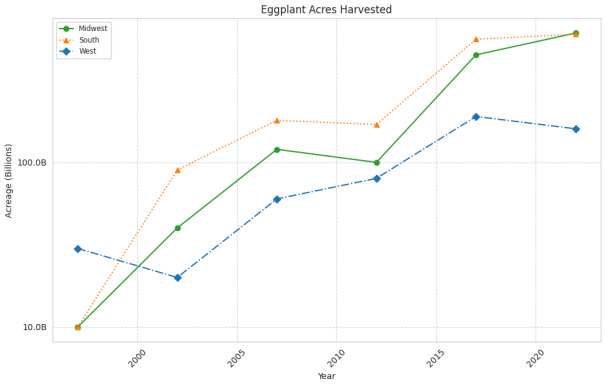
Unfortunately, there is no data in the sales dataset to explore a bit further all of this. We looked it at the region level and this is the breakout as the top 5 crops acreage by region over time in log scale (Fig 14-18):







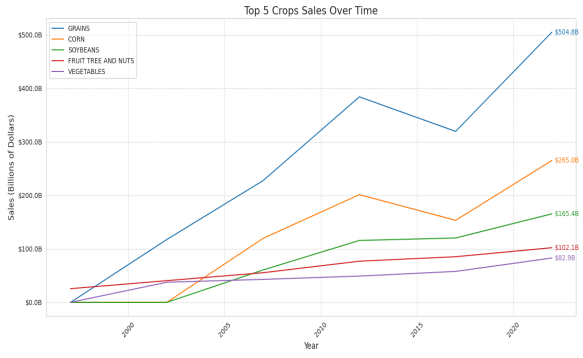




Interestingly, Spinach had a hard dip in the Midwest. Radish and Carrot had rapid growth in the South. A further investigation could potentially give us insights of the reason behind that.

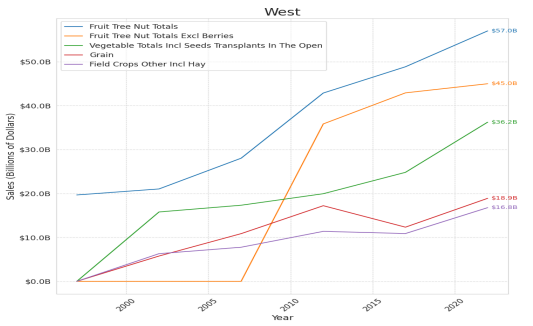
However, there is no data in the sales dataset to explore this any further.

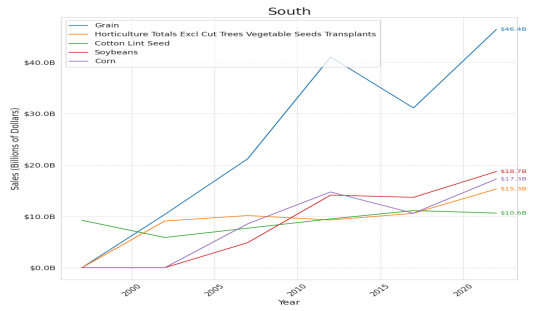
In the top 5 total sales per year excluding animal-based products (Fig 19):

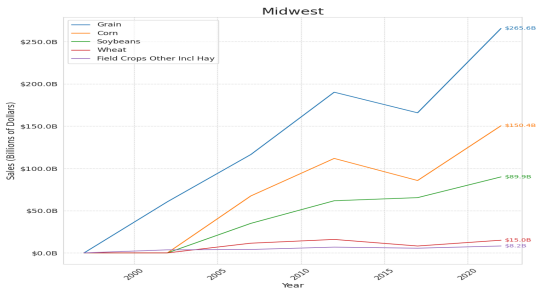


We can see that the best performance product is Grains, followed by Corn.

If we look at this from the region level in the top 5 crops sales over time by region graphs (Fig 20-22)

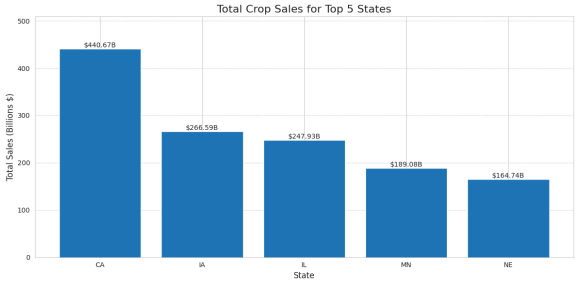






The top performance product is Grain in the South and Midwest. Interestingly, Fruit Tree is the top performance product for the West.

And if we go further into the State level (Fig 23):



This tells us that the top three states California ($440.67B), Iowa ($226.59B) and Illinois ($247.93B) are the most profitable to take into account when looking at the top selling non animal products.

4. Conclusion

This analysis of crop diversification, agricultural land use, and production patterns across the United States reveals significant trends and correlations that highlight the evolving nature of the agricultural sector. Our findings show that certain crops have become more dominant in specific regions, while others have diminished, reflecting both market dynamics and environmental changes. The correlation between shifts in farm size—measured by both acreage and revenue—and changes in crop diversity indicates that economic pressures and land value play a critical role in shaping farming practices.

While this project provides a comprehensive overview of historical trends, there are limitations. The analysis could be enhanced by incorporating more granular data on climate change impacts, regional water usage, and crop-specific factors that were not fully explored in this study. Future research could also investigate the long-term effects of urbanization on rural land values and crop distribution, offering a more complete picture of the factors shaping the agricultural landscape. This research could also be extended by adding some predictive models related to future crop diversification trends and whether that would be influenced by external factors such as urban expansion and economic fluctuations.

The insights gained from this analysis have important implications for policymakers and farmers, as they navigate the complexities of land management and sustainable agricultural practices in the face of shifting market conditions. Our findings contribute valuable knowledge to the field of agricultural economics and land management, helping to inform strategies for sustainable growth and efficient resource use in the future.

Please see this link to our Submission Video
Please see this link to our Google Colab Notebook1 and Google Colab Notebook2

References

US MAP DIVISION (worldatlas)

APPENDIX

EDA

We are working on the land_use_farm_ops.csv. We are identifying all the similar/common data and plotting them in the graph.

The data is 18775 rows and 75 columns

We found at least two columns that has many NAs:

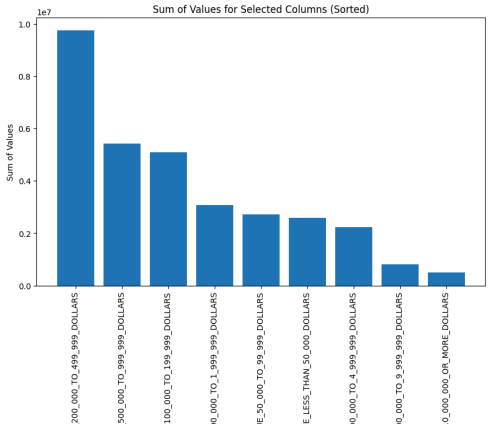
1. FARM_OPERATIONS_AREA_OPERATED_MEASURED_IN_PCT_OF_TOTAL_LAND 3129 non-null with a total of 15646 NAs

And

2. AG_LAND_INCL_BUILDINGS_OPERATIONS_WITH_ASSET_VALUE_WHERE_VALUE_10_000_000_OR_MORE_DOLLARS 11254 non-null with a total of 7521 NAs

This first one, FARM_OPERATIONS_AREA_OPERATED_MEASURED_IN_PCT_OF_TOTAL_LAND, we could drop it, as it doesn't seem to be adding much value.

For the second one, AG_LAND_INCL_BUILDINGS_OPERATIONS_WITH_ASSET_VALUE_WHERE_VALUE_10_000_000_OR_MORE_DOLLARS is part of a division in bins



We can observe that most of the data is in the range of 200000 to 499999 of total value.

In this subset of bins and for the rest of the division bins we found in the data, we decided that the value comes in adding them all, instead of having differentiated in bins so we proceeded to aggregate all the subsets bins together.

Although many additional steps were taken in the EDA process, time constraints limited how much we could include in this document.