Soybean Yield Forecasting: A Machine Learning Approach for Iowa and Illinois, from 1982 to present.

Master Thesis Proposal

Presented to Tilburg University

Erion Mediu, 2066763, E.mediu@tilburguniversity.edu

1 Project Definition, Motivation & Relevance

Agricultural endeavors have a critical role in the United States economy. Today agriculture-based employment accounts for 21 million jobs in the U.S. and approximately 5.5 percent of the country's GDP based on, Economic Research Service USDA (2023a). Within the agriculture-based industry soybean production takes a crucial role, more than 80 percent of soybeans are cultivated in the upper Midwest. The United States counted Illinois, Iowa, and Minnesota as their leading soybean-producing states, Economic Research Service USDA (2023a). According to Food and Agriculture Organization of the United Nations (2023), soybeans contribute significantly to meeting the dietary needs of billions of people worldwide, emphasizing the importance of reliable supply chains sustained by accurate yield forecasts that aid mitigations of supply shocks. Furthermore, the U.S. soybean market has a significant impact on global trade dynamics. Accurate yield predictions can affect international trade agreements and partnerships, as well as the economic stability of countries that rely on U.S. soybeans for food security and industrial uses. Economic Research Service USDA (2023b) reports that U.S. soybeans and soy products are critical components of global trade, influencing trade balances and diplomatic relations. Thus soybean yield predictions can directly contribute to better agricultural practices, economic stability, and food security. With precise forecasting, farmers can make informed decisions that optimize production and reduce waste, contributing to sustainable agricultural practices while having a positive societal impact.

A comprehensive review of numerous studies highlights a fundamental truth, the choice and architecture of data sources significantly influence the effectiveness of machine learning-based yield prediction methods Mkhabela, Bullock, Raj, Wang, and Yang (2011); Panda, Ames, and Panigrahi (2010). This comes as a natural consequence since yield is influenced by many aspects like the seeds, climate, soil, topography, fertilization, and management Li et al. (2023). Such aforementioned studies have used different methods to extract various yield-related predictions, yet this research aims through the integration of both field-based samples and weather observations, to fill the knowledge gap in applying Machine learning techniques to soybean yield forecasting and create a new framework for predicting yields at a county level. By not limiting to weather-based variables or weather indices, the study contributes to the broader field of ecological predictive modeling. While Hamed, Van Loon, Aerts, and Coumou (2021) highlighted the importance of weather periods and agronomic factors.

While prior research has utilized weather data as well as field observations, comprehensive research merging both is not present. Furthermore, the focused spatial location adds to the general precision agriculture idea. By leveraging such data diversity and multidimensional feature engineering, this project aims to transcend traditional methodologies, providing a more nuanced understanding of how various factors contribute to yield outcomes.

In essence, this research has a specified approach to both feature engineering and validation methods tailored to a unique temporal and spatial domain of soybean cultivation. Through training and testing a set of models such as Random Forest, XGBoost, SVM, and Linear Regression It seeks within the sphere of soybean yield forecasting to create a framework for yield prediction and to decipher the intricate link of variables that define agricultural productivity. Through this endeavor, the project aspires to fill a critical knowledge gap, offering a new perspective and framework that can significantly impact both the scientific community and agricultural practices at large.

2 Literature Review

Agricultural activity is considered the cornerstone of numerous economies, and the integration of machine learning techniques has substantially improved the precision and efficiency of yield predictions in agriculture. This is particularly evident in research that has focused on essential crops such as corn and soybeans over recent years. However, the diverse nature of crops and the spatial-temporal aspects of agricultural yield estimation have given birth to a variety of methodological approaches which we will delve deeper into in this section. Nonetheless, the majority of research in the field indicates that models integrating multiple data types such as satellite imagery, climatic information, soil characteristics, and or farm management practices, consistently tend to outperform models relying on a singular data source Li et al. (2023).

Hamed et al. (2021), by making use of a stepwise multilinear regression investigates the effects of combined hot–dry weather extremes on US soybean yields, pinpointing key climatic influencers and their synergistic impacts. Integrating weather data and satellite-derived soil moisture information, this research argues that approximately two-thirds of the annual variability in yields comes from weather-based features. It emphasizes the adverse consequences of high temperature and low soil moisture during the critical summer reproductive phase, where hot–dry conditions dictate diminishing yields. This underscores the necessity of incorporating weather-based features in forecasting models.

Joshi, Kazula, Coulter, Naeve, and Garcia y Garcia (2020) contrasts three main methods to approach the yield forecasting problem. Namely comparing three broad frameworks, weather-based frameworks, alternative yield estimation methods such as field observations/sampling, and crop simulation modeling. While field observations/sampling demands considerable manual labor and incurs high costs, and simulation modeling requires extensive soil and crop data, weather-based models stand out for their affordability and scalability. These models have been successfully deployed across a range of spatial and temporal scales for predicting yields in the short and long term. Nonetheless, despite their advantages, weather-based models encounter challenges related to modeling space-time data, including spatial and temporal correlation and collinearity among weather variables, Joshi et al. (2020). These complexities can hinder the modeling process and impact the precision of yield estimates.

Nonetheless within the weather-based frameworks, Joshi et al. (2020) employs and compares Step-wise multiple linear regression (MLR), generalized additive model (GAM), and support vector machine (SVM) model. These models correlate yield with readily available weather data, with SVM outperforming other models. While Johnson (2014); Li et al. (2023) make use of regression tree-based models and XGboost respectively, not employed on readily available weather data but on a series of agroclimatic indices. Indices, such as the Palmer drought severity index, Dai and for Atmospheric Research Staff (2023), and the SPEI index Vicente-Serrano and for Atmospheric Research Staff (2023) just to name a few serve as explanatory variables. This approach of using indices as an indirect approximation of weather-based impact highly increases the generalization and expansion capabilities of the deployed model since they mitigate several problems stemming from the spatiotemporal nature of real-time weather data Li et al. (2023). Yet the frequencies and real-time availability of these indices are not optimal for real-time predictions.

Meanwhile, Shahhosseini, Hu, and Archontoulis (2020), focusing primarily on the US Corn Belt, through the use of several ensemble models designed using blocked sequential procedure to generate out-of-bag predictions and "complete and partial in-season weather knowledge[1] " further supports the aforementioned idea of real-time implementation. Shahhosseini et al. (2020) bring two main innovations in corn prediction namely the block sequential procedure and feature selection procedure. The former allows for the preservation of temporal relationships and the latter categorizes yield-impacting factors into three groups: environmental, management-based (e.g., plant population, area harvested), and genome-based features. Excluding genome-based features, a three-stage process is employed to streamline data dimensionality. Initially, irrelevant features are pruned based on expert insights. A subsequent analysis using a random forest algorithm for permutation importance feature selection identifies the 80 most critical features. Finally, a filter-based selection eliminates highly correlated features to ensure the inclusion of variables with substantial predictive power. Echoing the findings of Hamed et al. (2021), weather-based features are identified as paramount.

Collectively, these studies highlight the dynamic and comprehensive nature of agricultural yield forecasting through machine learning. The body of research emphasizes the importance of accurate and timely predictions in agricultural management and policy formulation as well as the diversity in approaches and frameworks to such problems. As machine learning methodologies advance, the integration of diverse data inputs, including field observations and weather-based data, remains crucial for enhancing yield forecasting models for soybeans.

[1]Complete in-season weather knowledge suggests having a full set of data on weather conditions throughout the entire growing season. While partial indicates having incomplete or limited data during the season.

3 Main Research Question

As aforementioned accurately predicting yields is highly rewarding from a socio-environmental point of view. Thus the main research question that this research aims to answer is:

How do field observations and gridded weather observations enhance the accuracy of machine learning models (Random Forest, XGBoost, SVM, and Linear Regression) in predicting soybean yields in Iowa and Illinois from 1980 to 2022?

To further dissect the main objective to design a new soybean yield prediction framework at the county level to achieve high-accuracy prediction by combining diverse data input, multidimensional feature engineering, and a set of models, two sub-research questions are formed:

- How can a soybean yield prediction framework be developed at the county level, utilizing a combination of diverse data sources and machine learning models for enhanced accuracy?
- What role do multidimensional features play in improving the predictive accuracy of soybean yield models, and how can feature engineering be optimized?

4 Methodology and Evaluation

As aforementioned the main contribution that this thesis provides to the literature is linked to the diversity of data involved, spatial location, and feature engineering. Nonetheless, two components remain of crucial importance, the first being that by nature soybeans are only produced once a year thus field yield is available yearly. This immediately restricts the number of observations available per state. To overcome this it is detrimental to take a look at the county level this highly increases the information available in the dataset since each state contains a high number of counties Schwalbert et al. (2020). Furthermore, this may contribute to the increased predictive power of field observations.

A second aspect is linked to weather-based features. To have an overarching analysis there is a need to spatially link the area of interest with the data coming from field observations. By nature, the weather is spatially and temporally correlated thus it varies per geographic area. To overcome this we create boundaries enveloping each county in our study area and then retrieve gridded data of variables of interest daily within the county boundary. A high-level map of the ongoing processes is shown in §A.3.

After arranging the data several algorithms will be implemented, a regression model serving as the base model a random forest, XGBoost, and SVM will be in the main interest of comparing. As previously argued these are the most prominent models in the literature with XGBoost performing best when indices are used as approximators of weather data Li et al. (2023), and SVM being the best performer when using grid-based weather data Joshi et al. (2020).

Furthermore, the challenge outlined in this thesis is framed as a regression problem, aiming to predict soybean yield. This objective lays the groundwork for ”in-season yield forecasting” or ”dynamic yield prediction”. Given the retrospective approach of the current application, specific evaluation methods are necessitated. There seems to be a consensus across literature about RMSE as the main evaluation metric, Joshi et al. (2020); Li et al. (2023); Shahhosseini et al. (2020). It is particularly useful in yield prediction contexts because it directly relates to the magnitude of prediction errors, giving a sense of how much deviation there is from the actual yields, while also penalizing larger errors more than smaller ones, Joaquin Bogado (2022). Such consensus is lacking regarding cross-validation methods. Yet for this study, Leave-One-Out Cross-Validation (LOOCV) has been selected since understanding how slight variations in data can impact model predictions is considered important. This presents the necessity to de-trend yield, and de-trending will be done based on the pseudo-code presented in §A.1. De-trending is an essential step that enables models like random forests to extrapolate Shahhosseini et al. (2020).

To conclude, a linear regression model will be employed as a baseline for comparison. Furthermore, to determine the impact of data source diversity on the predictive accuracy of different machine learning models for soybean yield forecasts, the study will employ a feature ablation approach. This method aims to systematically assess the contribution of each data source to the overall predictive power, providing insights into whether the variety of data sources enhances model performance.

5 Dataset Description

The primary datasets utilized in this research include ORNL DAAC (2023); Physical Sciences Laboratory, NOAA (2023); USDA NASS (2023) and United States Census Bureau (2023), each varying in size, format, as well as accessibility. Field observations are sourced from USDA NASS (2023) via an API, enabling the download of both state and county-level observations. The variables of interest, detailed in the appendix have yearly, weekly, and daily frequencies, allowing for the creation of multi-level identifiers within the dataset through a combination of county, state, date, and week.

Weather-based data, derived from sources like ORNL DAAC (2023), is presented in NCDF4 format and daily frequencies, distinguishing it from the aforementioned datasets. This format facilitates both point and grid extraction of data. Data is extracted on 1X1 km grids, as illustrated in A.3, and undergoes cropping and adaptation to fit county borders. This cropping process is supported by the US Census Bureau data, which outlines the geometric shapes and borders of each US county.

Subsequently, variables of interest are averaged for a specified area, transitioning, for instance, the minimum temperature of the county to its average minimum temperature. This method marks a progression from simple point extraction per county. However, it also highlights the potential for increased resolution since assuming a uniform temperature distribution across a geographic area may not always be accurate. Nonetheless, such an enhancement in resolution would necessitate finer data at the field level, thus significantly elevating computational complexity and bringing us full circle to the challenges inherent in managing large-scale agricultural data.

Finally, adjusting the temporal resolution across datasets is accomplished through the use of field observations of progress. This variable provides a per-county overview of the phases the plant is undergoing, allowing for the coordination of weather data within this specific temporal timeframe. According to the dataset, planting starts at the earliest in week 14 and ends at the latest in week 47. This method ensures that the analysis is finely tuned to the diverse stages of soybeans [2], enhancing the predictive accuracy of our model by focusing on the most relevant environmental conditions. Such format is also portrayed in §A.4.

6 Algorithms and Software

To implement the aforementioned models and data manipulation, a combination of software and packages is employed. Mainly VScode software is used for weather-based data and data manipulation, R-studio for field observations due to the presence of pre-determined API packages, and finally QGIS for mapping and coordinate systems. As for the algorithms, the " in-house " developed algorithms used to de-trend soy yield as well as for feature ablation are presented in §A.1 and §A.2 respectively. While main models of interest such as Random forest, SVM, and XGBoost remain well documented within the literature.

7 Milestones and Plan

Task name	Feb				Mar				Apr				May				Jun			
	w-1	w-2	w-3	w-4	w-1	w-2	w-3	w-4	w-1	w-2	w-3	w-4	w-1	w-2	w-3	w-4	w-1	w-2	w-3	w-4
Gathering Data: USDA NASS, DAYMET, PRSIM																				
Processing Data: Clean and align																				
Merging Data: Creating the data frame																				
Feature engineering																				
De-trending Yields																				
Testing different models																				
Concluding based on valuation metrics																				
Finalizing proposal																				
First submission																				
Second submission																				
Writing thesis																				

Table 1: Gantt Chart presenting the main tasks and timelines of the project. Tasks are color-coded to indicate their status: green for completed tasks, yellow for ongoing ones, and red for critical deadlines.

[2] The stages considered are: Planted, Emerged, Blooming, Setting pods, Coloring, Dropping Leaves, Harvested

A Appendix

A.1 Pseudo code for de-trending yields

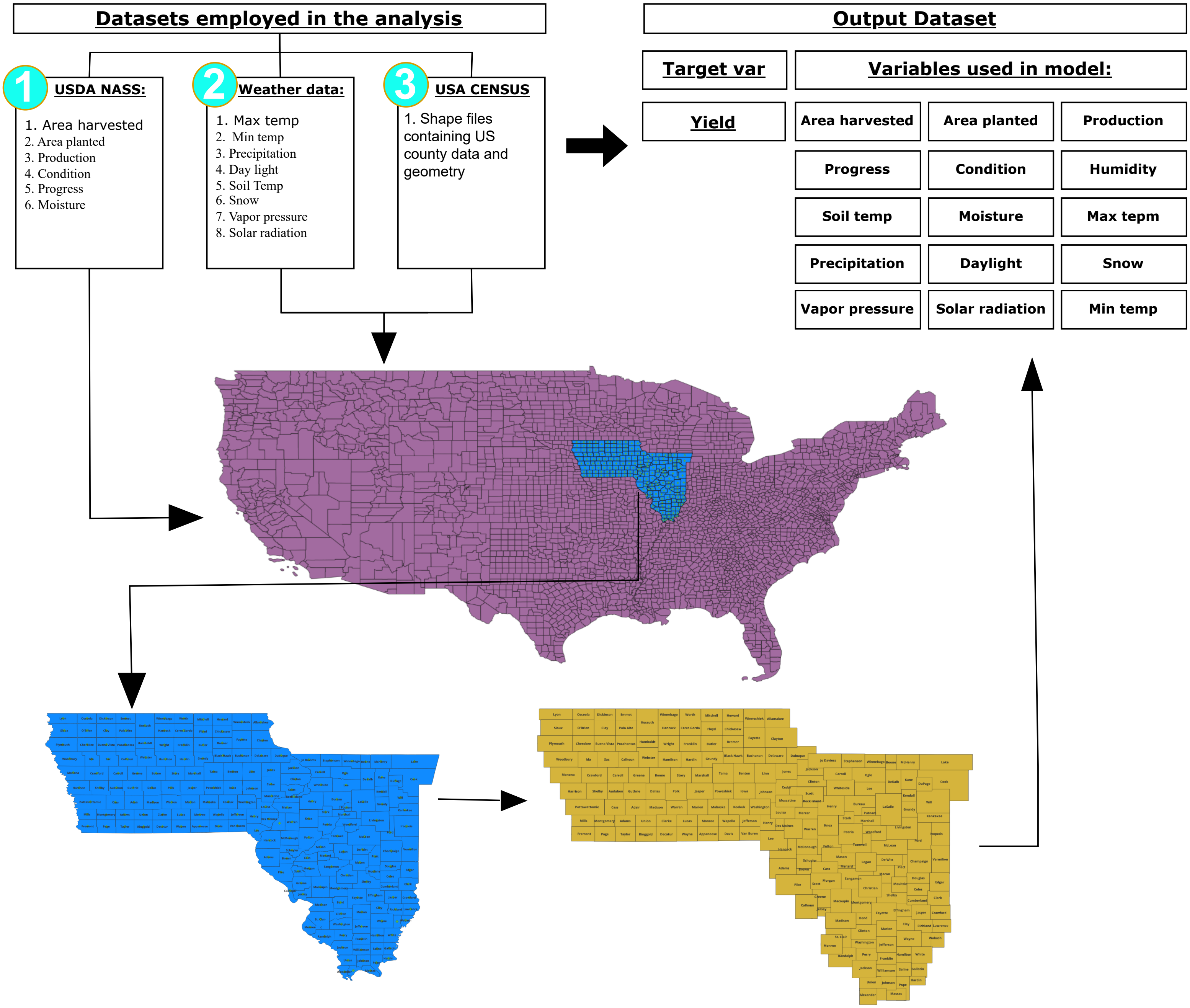
```
1. Initialize the window size (e.g., 10 years for the rolling window).
2. For each year in the dataset, do the following:
   a. Select the window of data centered around the current year,
      including 10 prior years.
   b. Perform a linear regression of yield on year within this window.
   c. Calculate the predicted yield for the current year using the
      regression model.
   d. Detrend the actual yield by subtracting the predicted yield
      from the actual yield for the current year.
3. The result is a series of detrended yields, where the linear
   trend due to time has been removed.
```

A.2 Pseudo code for feature ablation

```
1. Initialize baseline model performance with the full set of features.
   baseline_performance = evaluate_model_performance(model, full_features, target)
2. Initialize an empty dictionary to store feature importances.
   feature_importances = {}
3. For each feature in the dataset:
   a. Create a modified feature set by removing the current feature.
      modified_features = full_features - feature
   b. Evaluate the performance of the model with the modified feature set.
      modified_performance = evaluate_model_performance(model, modified_features, target)
   c. Calculate the change in performance due to the removal of the feature.
      performance_change = baseline_performance - modified_performance
   d. Store the performance change associated with the feature.
      feature_importances[feature] = performance_change
4. Rank features based on their importance.
   ranked_features = rank_features_by_importance(feature_importances)
```

A.3 High-level overview of data processing procedures

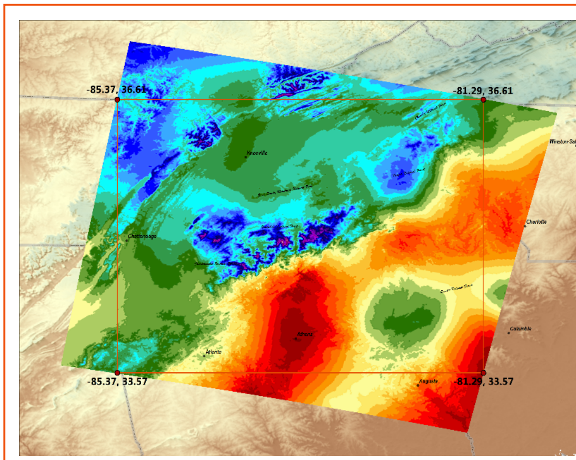
High level view of data and processing:



High level view of data processing steps:

1. Aggregate Agricultural Data: Aggregate agricultural data (such as area harvested, production, and yield) specifically excluding combined county records to ensure data accuracy and relevance to specific geographic locations.
2. Transform Data for Comprehensive Analysis: Agricultural data is converted from a long to a wide format, thus aligning various statistical categories across different dimensions such as state, year, and county.
3. Track Soybean Growth Stages: Fetch and categorize state-level soybean growth stages, aligning agricultural practices with growth milestones. Assumed to be uniformly distributed within a state thus uniformity across counties.
4. Synchronize Weather Data with Growth Periods: Align weather data collection to critical soybean growth periods (weeks 14 to 47) for environmental impact accuracy.
5. Monitor Soybean Plant Conditions: Process state-level plant condition data to gain insights into soybean health throughout the season.
6. Create a Unified Agricultural Dataset: Merge agricultural data with growth stage and condition data, capturing spatial and temporal dynamics.
7. Adjust Weather Data Collection to County Borders: Use GIS techniques to tailor weather data collection to specific county boundaries, ensuring relevant environmental conditions.
8. Automate Environmental Data Retrieval: Collect high-resolution weather data for key variables, tailored to study needs using spatial and temporal parameters.
9. Prepare Dataset for Predictive Modeling: Structure integrated dataset for machine learning analysis, predicting yield based on comprehensive variables.
10. Assume Environmental Uniformity for Analysis: Simplify spatial variability with uniform environmental assumptions across geographic areas, managing computational complexity.

Grid weather data example:



A.4 Final dataset format for one observation

Listing 1: JSON structure for county-level agricultural data

```
{
  "Iowa_county": {
    "2022": {
      "Yield": 57.2,
      "Area planted": 21312312,
      "Area Harvested": 19879866,
      "Production": 2208000,
      "Progress": {
        "Week 14": {
          {"PCT PLANTED": "50",
           "PCT_Emerged": "100"
            ...
          },
          "Condition": {
            "Exellent" : "45"
            "Good" : "80"
            ...},
          "Temperature_daily": [{"date": "2022-04-10", "value": 15.2}...],
          "Soil_temp": [...]
        },
        "Week 15": {
          {"PCT PLANTED": "80",
           "PCT_Emerged": "70"
            ...
          },
          "Condition": {
            "Exellent" : "45"
            "Good" : "80"
            ...},
          "Temperature_daily": [{"date": "2022-04-18", "value": 15.2}...],
          "Soil_temp": [...]
        },
        ....
        }
      }
    }
  }
}
```

A.5 Soybean production density

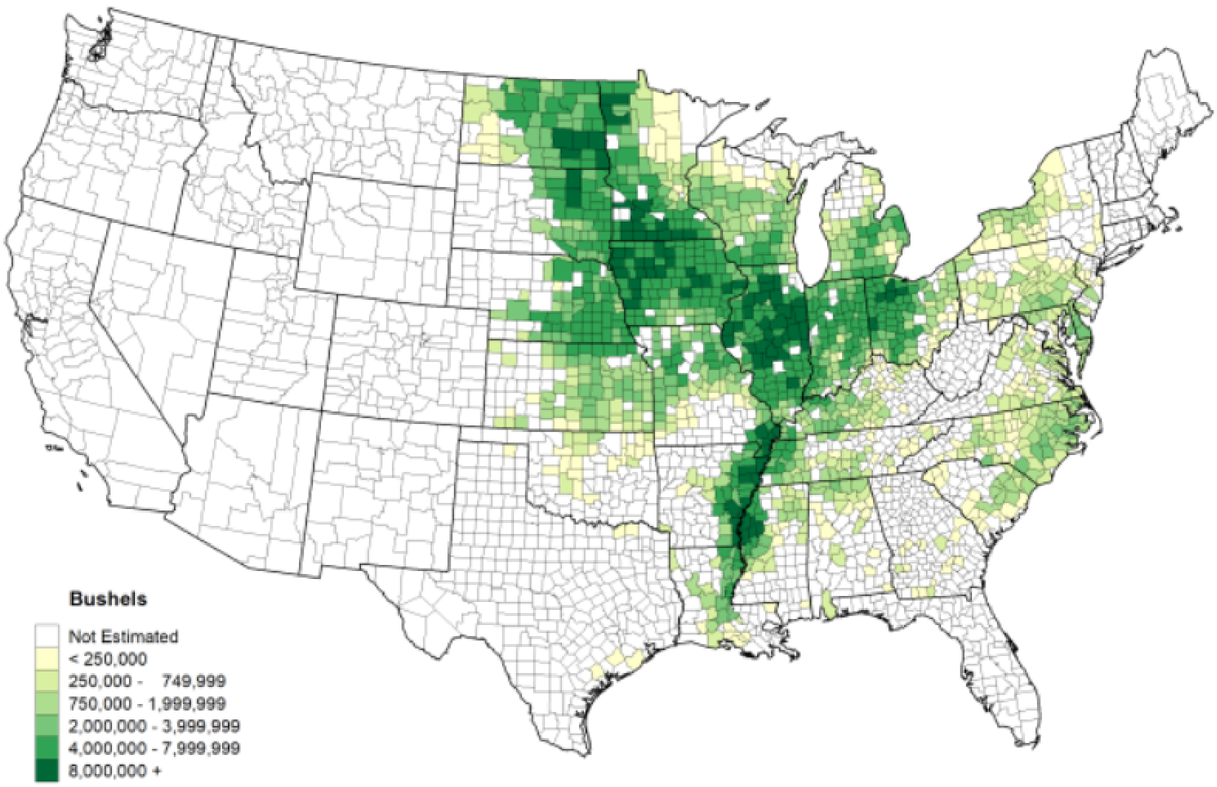


Figure 1: Soybean 2022 Production by county for selected states retrieved from: USDA NASS (2023).

References

Dai, A., & for Atmospheric Research Staff, N. C. (2023, August). *The climate data guide: Palmer drought severity index (pdsi).* Retrieved from https://climatedataguide.ucar.edu/climate-data/palmer-drought-severity-index-pdsi. Retrieved 2024-02-15, from `https://climatedataguide.ucar.edu/climate-data/palmer-drought-severity-index-pdsi` (Last modified 2023-08-19)

Economic Research Service USDA. (2023a). *Agriculture the food sectors and the economy.* Online. Retrieved from `https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/ag-and-food-sectors-and-the-economy/` (Accessed: 2024-02-15)

Economic Research Service USDA. (2023b). *Major factors affecting global soybean and products trade projections.* Online. Retrieved from `https://www.ers.usda.gov/amber-waves/2016/may/major-factors-affecting-global-soybean-and-products-trade-projections/` (Accessed: 2024-02-29)

Food and Agriculture Organization of the United Nations. (2023). *The State of Food and Agriculture 2023: Revealing the true cost of food to transform agrifood systems* (Tech. Rep.). Rome: Author. Retrieved from `https://doi.org/10.4060/cc7724en` doi: {10.4060/cc7724en}

Hamed, R., Van Loon, A. F., Aerts, J., & Coumou, D. (2021). Impacts of compound hot–dry extremes on us soybean yields. *Earth System Dynamics*, *12*(4), 1371–1391. Retrieved from `https://esd.copernicus.org/articles/12/1371/2021/` doi: 10.5194/esd-12-1371-2021

Joaquin Bogado, F. M. J. D., Mario Lassnig. (2022). Modelling large-scale scientific data transfers. *Computational Software for Big Science*, *6*, 12. Retrieved from `https://doi.org/10.1007/s41781-022-00084-4` doi: 10.1007/s41781-022-00084-4

Johnson, D. M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the united states. *Remote Sensing of Environment*, *141*, 116-128. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0034425713003957` doi: https://doi.org/10.1016/j.rse.2013.10.027

Joshi, V., Kazula, M., Coulter, J., Naeve, S., & Garcia y Garcia, A. (2020, 11). In-season weather data provide reliable yield estimates of maize and soybean in the us central corn belt. *International Journal of Biometeorology*, *65*. doi: 10.1007/s00484-020-02039-z

Li, Y., Zeng, H., Zhang, M., Wu, B., Zhao, Y., Yao, X., … Wu, F. (2023). A county-level soybean yield prediction framework coupled with xgboost and multidimensional feature engineering. *International Journal of Applied Earth Observation and Geoinformation*, *118*, 103269. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1569843223000912` doi: https://doi.org/10.1016/j.jag.2023.103269

Mkhabela, M., Bullock, P., Raj, S., Wang, S., & Yang, Y. (2011). Crop yield forecasting on the canadian prairies using modis ndvi data. *Agricultural and Forest Meteorology*, *151*(3), 385-393. Retrieved from https://www.sciencedirect.com/science/article/pii/S0168192310003138 doi: https://doi.org/10.1016/j.agrformet.2010.11.012

ORNL DAAC. (2023). *Daymet: Daily surface weather data on a 1-km grid for north america, version 4 r1.* Online. Oak Ridge National Laboratory Distributed Active Archive Center. Retrieved from https://doi.org/10.3334/ORNLDAAC/2129 doi: 10.3334/ORNLDAAC/2129

Panda, S., Ames, D., & Panigrahi, S. (2010, 03). Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing*, *2*. doi: 10.3390/rs2030673

Physical Sciences Laboratory, NOAA. (2023). *Gridded data sets.* Online. Retrieved from https://psl.noaa.gov/data/gridded/ (Accessed on 2024-02-04)

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern brazil. *Agricultural and Forest Meteorology*, *284*, 107886. Retrieved from https://www.sciencedirect.com/science/article/pii/S0168192319305027 doi: https://doi.org/10.1016/j.agrformet.2019.107886

Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020, July). Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, *11*. Retrieved from http://dx.doi.org/10.3389/fpls.2020.01120 doi: 10.3389/fpls.2020.01120

United States Census Bureau. (2023). *Counties (and equivalent) shapefiles.* Online. Retrieved from https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2023&layergroup=Counties+(and+equivalent) (Accessed on YYYY-MM-DD)

USDA NASS. (2023). *Usda national agricultural statistics service database.* Online. Retrieved from https://www.nass.usda.gov (Accessed: 2024-02-15)

Vicente-Serrano, S. M., & for Atmospheric Research Staff, N. C. (2023, September). *The climate data guide: Standardized precipitation evapotranspiration index (spei).* Retrieved from https://climatedataguide.ucar.edu/climate-data/standardized-precipitation-evapotranspiration-index-spei. Retrieved 2024-02-15, from https://climatedataguide.ucar.edu/climate-data/standardized-precipitation-evapotranspiration-index-spei (Last modified 2023-09-04)