Beyond Average Yield: A Stability-Based Analysis of Soybean Production in Argentina (1941-2024)

A Risk-Oriented Data Analysis Using Historical Production Data

Author: Valentino Lorenzati / Advanced Data Science Student

Tools: Python - SQL - Machine Learning - Tableau

1. Executive Summary

Agricultural production analysis is traditionally centered on average yield metrics, which often mask critical differences in production stability and risk. Two regions may exhibit similar long-term average yields while displaying fundamentally different temporal behaviors—one stable and predictable, the other highly volatile and unreliable. This limitation reduces the usefulness of average-based analyses for decision-making contexts where risk matters.

This project presents a stability-oriented analysis of soybean production in Argentina using historical data spanning from 1941 to 2024. Rather than focusing exclusively on production volume or mean yield, the analysis incorporates variability and consistency metrics to capture the long-term reliability of regional production systems.

An end-to-end analytical pipeline was implemented, combining agronomic data validation, relational data modeling, and risk-oriented feature engineering. Beyond standard yield averages, the analysis incorporates variability and consistency metrics to capture long-term production reliability.

Using these features, regions were grouped through K-Means clustering, with the optimal number of clusters selected based on established validation techniques. This process revealed three distinct production stability profiles, ranging from high stability and high yield to high volatility and elevated productive risk.

The results highlight a key insight: high production does not necessarily imply low risk. A strong correlation between harvested area and total production suggests that scale largely drives output, while stability varies significantly across regions. These differences become visible only when variability-based metrics are considered.

The analytical findings are integrated into interactive Tableau dashboards, allowing users to explore temporal trends, regional patterns, and stability profiles in a visual and intuitive manner. This combination of quantitative rigor and visual storytelling supports informed analysis and comparative assessment.

Overall, this project shows how a risk-aware analytical framework can enhance traditional performance assessments. Although focused on soybean production in Argentina, the approach is transferable to other domains where stability and uncertainty matter as much as averages.

2. Why Average Yield Is Not Enough

Agricultural performance is commonly evaluated using average yield indicators, which provide a simple and intuitive measure of productivity. While useful as a baseline, this approach assumes that long-term averages are sufficient to describe regional performance. In practice, however, averages often conceal meaningful differences in temporal behavior and production reliability.

Two regions may exhibit similar average yields over several decades while following very different production trajectories. One region may show stable and predictable yields year after year, while another may alternate between exceptionally high and critically low outputs. From a decision-making perspective, these two scenarios imply very different levels of risk, despite sharing the same average performance.

This limitation becomes especially relevant in contexts where uncertainty matters. Production planning, investment decisions, and long-term regional assessments require not only information about expected outcomes, but also about variability, volatility, and the likelihood of adverse events. An analysis based solely on averages fails to capture these dimensions.

To illustrate this issue, consider a simplified example: a region with moderate but consistent yields may offer greater long-term reliability than a region with higher average output driven by a small number of exceptional years. In such cases, focusing exclusively on averages can lead to overestimation of productive potential and underestimation of risk.

For this reason, this project adopts a stability-oriented analytical perspective, complementing traditional yield metrics with indicators that reflect temporal consistency and production risk. By explicitly measuring variability and instability, the analysis aims to provide a more realistic and informative representation of agricultural performance over time.

This shift in perspective—from performance alone to performance under uncertainty—forms the conceptual foundation of the entire analytical framework presented in this study.

3. Data and Analytical Pipeline Overview

This study is based on a historical dataset of soybean production in Argentina covering the period from 1941 to 2024, including information on planted area, harvested area, total production, and yield at regional levels. Given the temporal depth of the data, particular attention was placed on data consistency and logical validation prior to analysis.

3.1 Data Cleaning and Agronomic Validation

Raw agricultural data often contain inconsistencies that, if left unaddressed, can distort analytical results. In this project, the data cleaning process was guided by basic agronomic logic, rather than purely statistical criteria. Records where the harvested area exceeded the planted area were identified and removed, as such situations are not physically plausible. Additionally, yield values exceeding biologically realistic thresholds were filtered to avoid the influence of extreme outliers.

This validation step ensured that subsequent analyses were grounded in realistic production behavior, improving the reliability of all derived metrics.

3.2 Data Modeling and Persistence

After cleaning, the dataset was transformed from a flat file into a normalized relational database using SQLite. This modeling approach separates descriptive attributes (such as provinces and departments) from quantitative production measures, resulting in a clear distinction between dimension tables and a fact table.

Beyond facilitating structured querying, this step demonstrates an emphasis on data integrity, scalability, and analytical reproducibility, aligning the project with real-world data engineering practices.

3.3 Feature Engineering for Stability Analysis

To move beyond traditional performance metrics, the analysis focused on constructing features that capture temporal stability and productive risk. Alongside average yield, additional indicators were computed to quantify variability, volatility, and consistency over time.

These features form the analytical foundation for the segmentation process, enabling regions to be compared not only by how much they produce, but by how reliably they sustain production across decades.

4. Stability Metrics and Risk Indicators

Traditional agricultural analyses often rely on average yield as the primary indicator of performance. While useful as a reference point, this metric alone provides limited insight into the reliability and consistency of production systems over time. To address this limitation, this project incorporates a set of complementary indicators designed to capture production stability and risk.

4.1 Average Yield as a Baseline Indicator

Average yield serves as a baseline measure of productive performance, summarizing long-term output relative to harvested area. It provides a useful point of comparison across regions but does not reflect how production behaves from year to year. Regions with identical average yields may experience very different levels of variability, making average yield an incomplete indicator when used in isolation.

For this reason, average yield is treated in this analysis as a reference metric, rather than a definitive measure of performance.

4.2 Coefficient of Variation: Measuring Relative Variability

The Coefficient of Variation (CV) is used to quantify yield variability relative to the mean. By expressing dispersion as a proportion of the average, the CV allows for meaningful comparisons between regions with different yield levels.

A higher CV indicates greater relative variability, suggesting lower production stability and increased exposure to risk. Conversely, a lower CV reflects more consistent yields over time. This metric helps distinguish regions where production is not only high, but also predictable.

4.3 Interannual Volatility: Capturing Year-to-Year Instability

While the CV summarizes overall variability, it does not explicitly account for temporal fluctuations between consecutive years. To address this, an interannual volatility metric was computed, measuring the magnitude of yield changes from one year to the next.

High interannual volatility signals abrupt production shifts, which can complicate planning and increase uncertainty. This indicator provides additional insight into the dynamic behavior of production systems, complementing long-term variability measures.

4.4 Frequency of Production Failures

To further characterize productive risk, the analysis incorporates a metric capturing the frequency of production failures, defined as years in which yields fall significantly below a region's historical norm. This indicator highlights regions that may experience acceptable average performance but suffer from recurrent low-output events.

From a risk perspective, the recurrence of such failures can be as relevant as overall variability, particularly for stakeholders concerned with resilience and long-term reliability.

4.5 A Multidimensional View of Production Stability

Together, these indicators form a multidimensional representation of production stability, combining level, variability, volatility, and failure frequency. This framework enables a more nuanced comparison of regions, revealing differences that remain hidden under average-based analyses.

By explicitly incorporating risk-related dimensions, the analysis shifts the focus from "how much is produced" to "how reliably production is sustained over time."

Illustrative Example

Two regions may share the same average yield over several decades. However, one region may display stable year-to-year production, while the other alternates between extreme highs and lows. Despite identical averages, the second region represents a higher productive risk, which becomes visible only when variability-based metrics are considered.

5. Unsupervised Clustering and Segmentation Approach

Once stability-related features were constructed, the analysis shifted from individual metrics to a segmentation perspective, aiming to identify groups of regions with similar long-term production behaviors. Given the exploratory nature of the problem and the absence of predefined labels, an unsupervised learning approach was selected.

5.1 Why Clustering?

Clustering enables the identification of latent structures within data by grouping observations based on shared characteristics. In this context, the goal was not to predict outcomes, but to discover patterns of productive stability across regions. This approach allows regions to be compared holistically, considering multiple dimensions of performance and risk simultaneously.

5.2 Algorithm Selection: K-Means

The K-Means algorithm was chosen due to its interpretability, computational efficiency, and suitability for numerical feature spaces. While simple in structure, K-Means provides clear and actionable segmentations when features are properly engineered and scaled.

Prior to clustering, all stability-related features were standardized to ensure that no single metric disproportionately influenced the results.

5.3 Selecting the Number of Clusters

Determining the appropriate number of clusters is a critical step in unsupervised analysis. In this project, the optimal number of clusters was selected by combining two widely used validation techniques: the Elbow Method, which evaluates within-cluster variance, and the Silhouette Score, which measures cluster separation and cohesion.

Both methods indicated that three clusters provided a balanced trade-off between interpretability and explanatory power.

5.4 Interpreting the Clusters

The resulting clusters represent distinct production stability profiles, rather than simple performance rankings. Each cluster reflects a characteristic combination of yield level, variability, volatility, and failure frequency.

6. Results and Production Stability Profiles

The clustering analysis reveals three distinct production stability profiles, each characterized by a unique combination of yield level, variability, volatility, and failure frequency. These profiles provide a structured framework for understanding regional production behavior beyond traditional performance metrics.

6.1 Profile I: High Stability / High Yield

Regions in this profile combine consistently high yields with low variability and limited interannual volatility. Production in these areas remains relatively stable across decades, indicating strong resilience and predictable output.

From a risk perspective, this profile represents the lowest level of productive uncertainty, making these regions particularly suitable for long-term planning and investment. High average performance in this cluster is reinforced by stability rather than driven by isolated exceptional years.

6.2 Profile II: Medium Stability / Structurally Lower Yield

This profile is characterized by moderate variability and lower average yields when compared to other clusters. Production patterns are relatively stable, but output levels remain structurally constrained.

Regions in this group may not exhibit extreme risk, but their long-term productive potential is limited. This profile highlights areas where stability exists without high performance, suggesting the presence of structural constraints rather than volatility-driven risk.

6.3 Profile III: Low Stability / High Productive Risk

The third profile includes regions with high interannual volatility, elevated variability, and a higher frequency of production failures. While these regions may occasionally achieve high yields, their production patterns are inconsistent and less predictable over time.

Despite sometimes strong average performance, this profile represents the highest level of productive risk. The results underscore a critical insight: high output does not necessarily imply low risk, particularly when production is driven by a limited number of exceptional years.

7. Cross-Profile Interpretation and Decision Relevance

While the clustering results identify three distinct production stability profiles, their true value emerges when these profiles are compared and interpreted in relation to decision-making contexts.

A central finding of this analysis is that regions with similar average yields can exhibit fundamentally different risk and stability behaviors. This demonstrates the limitation of evaluating agricultural performance solely through productivity metrics, as average yield may mask underlying volatility, structural instability, or dependence on exceptional years.

From a strategic perspective, the High Stability / High Yield profile represents regions where performance is both strong and reliable. These areas are well suited for long-term planning, infrastructure investment, and policies focused on sustaining productivity rather than mitigating risk.

In contrast, regions classified under Medium Stability / Structurally Lower Yield show consistent but constrained performance. While these areas do not present elevated production risk, their long-term output potential appears limited, suggesting that productivity improvements would likely require structural interventions rather than risk management strategies.

The Low Stability / High Productive Risk profile highlights regions where average yields can be misleading. Despite occasional high output, production in these areas is driven by a small number of exceptional years and characterized by significant interannual volatility. For decision-makers, this implies higher uncertainty and the need for risk-aware planning, adaptive strategies, or contingency mechanisms.

Overall, this cross-profile interpretation emphasizes the importance of integrating stability, volatility, and failure frequency into agricultural assessments. By shifting the focus from average performance to production reliability, the proposed segmentation framework enables more informed decisions in planning, investment, and regional analysis.

8. Visualization and Analytical Integration

To ensure that analytical results are both interpretable and actionable, the final outputs of this project were integrated into an interactive visualization layer using Tableau. This step bridges the gap between computational analysis and decision-oriented exploration.

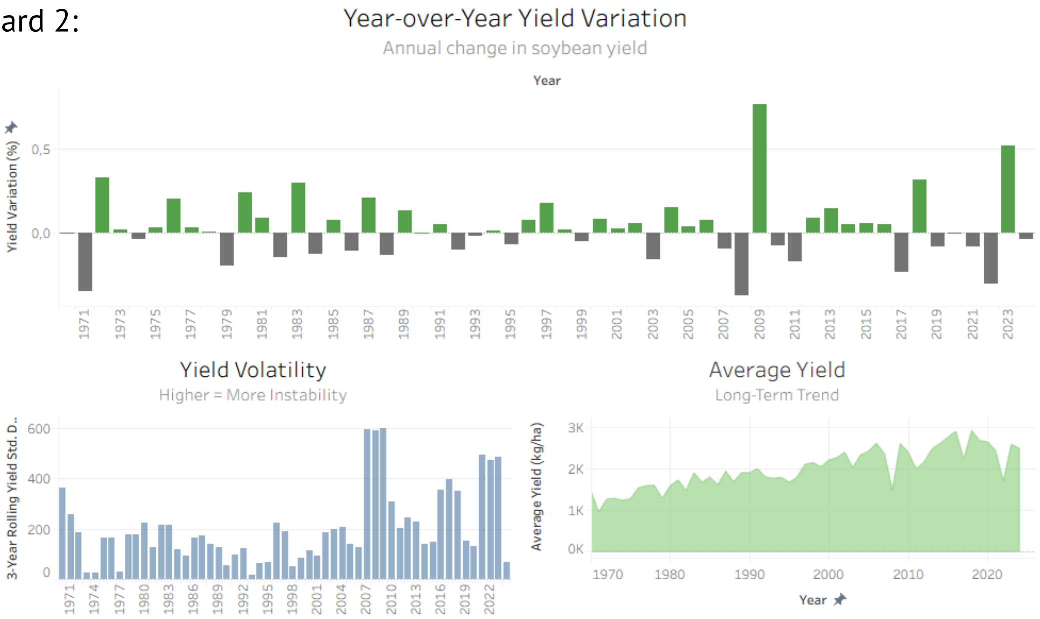
The clustering results generated in Python were exported as a structured dataset, enabling seamless integration with Tableau through a dedicated output file. This approach preserves the analytical logic developed during the modeling phase while allowing for flexible, visual exploration of results.

The dashboards are designed to support multiple analytical perspectives. Geographic visualizations display the spatial distribution of production stability profiles, revealing regional patterns that are not evident through tabular analysis alone. Temporal dashboards illustrate long-term trends and interannual variability, enabling users to assess both structural performance and year-to-year dynamics.

In addition, supporting dashboards focus on production determinants, highlighting the strong relationship between harvested area and total output, as well as the relative contribution of yield stability to overall production behavior. Together, these visual components reinforce key analytical findings while enabling intuitive interpretation.

By integrating Python-based machine learning, SQL-based data structuring, and interactive visualization, this project demonstrates an end-to-end analytical workflow. The visualization layer is not treated as a presentation artifact, but as an extension of the analytical process itself, enabling validation, exploration, and communication of results.

Dashboard 2:



9. Limitations and Future Work

While this study provides a structured and robust framework for analyzing soybean production stability, several limitations must be acknowledged to properly contextualize the results and identify opportunities for future improvement.

First, the analysis is based exclusively on historical production records, without incorporating exogenous variables such as climate conditions, soil characteristics, or technological adoption. Factors like precipitation patterns, temperature extremes, and soil fertility are known to significantly influence agricultural performance and could help explain part of the observed variability.

Second, the temporal aggregation of data at the regional level may mask intra-regional heterogeneity. Variations within departments or provinces—driven by localized environmental or management differences—are not captured in the current dataset.

Third, the clustering approach assumes static production behavior over time. While this allows for clear segmentation, it does not account for structural changes such as technological improvements, policy shifts, or changes in land use that may alter production dynamics across decades.

Future work could extend this framework by integrating climatic and environmental datasets, enabling the development of hybrid models that combine stability metrics with causal drivers. Additionally, time-aware clustering techniques or regime-shift detection methods could be applied to capture changes in production behavior over different historical periods.

Despite these limitations, the current approach provides a solid foundation for risk-aware agricultural analysis and demonstrates how production stability metrics can complement traditional productivity indicators.

10. Conclusions and Strategic Value

This project demonstrates that agricultural performance assessment can be significantly enhanced by incorporating production stability and risk metrics alongside traditional productivity indicators. By moving beyond average yield, the analysis provides a more complete and realistic understanding of long-term production behavior.

The proposed segmentation framework reveals that regions with similar output levels may differ substantially in terms of reliability, volatility, and exposure to production risk. This distinction is critical for informed decision-making, particularly in contexts where stability and predictability are as important as peak performance.

From a methodological perspective, the project illustrates an end-to-end analytical workflow, integrating data cleaning, relational database design, feature engineering, machine learning, and interactive visualization. Each stage of the pipeline contributes to analytical rigor and interpretability, reinforcing the coherence of the overall approach.

From a practical standpoint, the results support applications in regional planning, risk assessment, and strategic investment analysis. The framework can be adapted to other crops or geographic contexts, demonstrating its scalability and broader relevance.

Overall, this work highlights the value of risk-aware data science in agricultural analysis and showcases how technical methodologies can be translated into actionable insights through thoughtful interpretation and communication.