Organic Farming in Ireland and Europe: Uncovering Trends and Drivers

Author: M. Vitucci
e-mail: sba24277@student.cct.ie
Student ID: 24277

Word Count (excluding titles, tables, caption and appendix): ~4000 words

Abstract
This report explores the adoption and growth of organic farming across Europe, using Ireland as a focal point for comparison. By integrating agricultural indicators, CAP funding data, and public sentiment analysis, the research investigates how financial support and societal attitudes influence organic farming practices. EDA, Hypothesis testing, Sentiment Analysis and Clustering are utilized to identify patterns and disparities among EU countries. Findings highlight Ireland's modest progress compared to leaders like Austria, which benefits from targeted CAP policies. Despite growing positive sentiment, financial incentives and policy reforms remain the primary drivers of organic farming adoption. The recommendations focus on enhancing Ireland's policies and expanding CAP support to align with the best practices of top-performing countries.

Keywords
GUI, NLP, Feature Extraction, Sentiment analysis, Hypothesis Testing, Linear Regression, Topic Modeling, Hierarchical Clustering, K-Means, Principal Component Analysis, VADER, Data manipulation, Optimization

1. Introduction
Organic farming is vital to sustainable agriculture, meeting growing consumer demand for eco-friendly practices. Using Ireland as a baseline, this report explores the adoption and growth of organic farming in Europe through data analysis techniques, such as exploratory data analysis (EDA), hypothesis testing, sentiment analysis, and clustering. Key questions include:

- How does CAP funding impact organic farming in Ireland compared to other countries?
- Does Ireland's growth align with trends in countries receiving higher CAP support?
- How do public sentiment and policy shape organic farming in Ireland and Europe?

Statistical methods and visualizations help uncover how financial support, farming growth, and societal attitudes are connected, offering practical insights to shape policies and drive sustainable agriculture forward.

2. Materials & Methods

2.1 Software and Libraries

2.1.1 Libraries and Tools
Key libraries included Pandas and NumPy for data manipulation, PRAW for Reddit scraping, Deep Translator for keyword translation, and Gensim for topic modeling to extract themes in textual data. Text pre-processing used NLTK for tokenization and lemmatization, while Scikit-learn enabled feature scaling and clustering. Challenges involved aligning different data formats, imputing missing values, and managing linguistic diversity during sentiment analysis. Integration relied on temporal and spatial keys for merging datasets, standardizing units, and engineering metrics like annual growth rates and sentiment scores to ensure consistency and relevance.

2.1.2 Aggregation and Transformation Methods

Aggregation and transformation methods refined the analysis for trends and regional comparisons. Grouping and pivoting created structured summaries, enabling clear comparisons across dimensions. Melting reshaped datasets for compatibility with analytical and visualization tools, aiding multidimensional trend analysis. Medians reduced outlier impacts in skewed data, while means captured averages in sentiment scores and growth rates. Summation provided cumulative insights into contributions. Winsorization minimized the influence of extreme values, and robust scaling normalized distributions for consistency.

2.2 Data Acquisition Process

The data for this project was obtained from two key sources, each providing unique perspectives and posing specific challenges. The open-access licensing of the Agricultural Data Portal allowed for smooth data acquisition, enabling the focus to remain on analysis. In contrast, Reddit's restrictive licensing highlighted the complexities of working with proprietary data, such as access limitations and the need for explicit permissions. This contrast underscores the importance of understanding licensing terms in research to ensure legal and ethical compliance while addressing potential challenges in accessing data.

2.2.1 Agricultural Data Portal

3. CMEF Indicators dataset, accessed through the European Commission (European Commission, n.d.a), aggregates data from various sources, including CATS, DOE, Eurostat. For dataset and domain info, refer to References

Bird, S., Klein, E., and Loper, E., 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3, 993–1022.

Blomer, J., Lammersen, C., Schmidt, M., and Sohler, C. (2016) *Theoretical analysis of the k-means algorithm—A survey*, Algorithm Engineering, 9220, pp. 81–116. doi: 10.1007/978-3-319-49487-6_3.

European Commission (n.d.a) *AgriData Portal: CAP Indicators*. Available at: https://agridata.ec.europa.eu/extensions/DataPortal/cap_indicators.html (Accessed: 23 November 2024).

European Commission (n.d.b) *Organic Production Sources.* Available at: https://agridata.ec.europa.eu/Qlik_Downloads/Organic-Production-sources.htm (Accessed: 23 November 2024).

Fawcett, T., 2006. *An Introduction to ROC Analysis*. Pattern Recognition Letters, 27(8), pp. 861–874.

García-Laencina, P.J., Sancho-Gómez, J.L. and Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: A review. Neural Computing and Applications, 19(2), pp. 263–282.

Guyon, I. and Elisseeff, A., 2003. *An introduction to variable and feature selection*. Journal of Machine Learning Research, 3, pp. 1157–1182.

Han, J., Kamber, M. and Pei, J., 2011. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann.

Hox, J.J., Moerbeek, M. and van de Schoot, R., 2017. *Multilevel Analysis: Techniques and Applications*. 3rd ed. Routledge.

Hutto, C.J. and Gilbert, E., 2014. VADER: *A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14).

Jolliffe, I.T. and Cadima, J., 2016. *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202.

Kim, Jae-Dong & Hwang, Ji-Hwan & Doh, Hyoung-Ho. (2023). *A Predictive Model with Data Scaling Methodologies for Forecasting Spare Parts Demand in Military Logistics*. Defence Science Journal. 73. 666-674. 10.14429/dsj.73.19129.

Medhat, W., Hassan, A., and Korashy, H., 2014. *Sentiment analysis algorithms and applications: A survey*. Ain Shams Engineering Journal, 5(4), pp. 1093–1113.

Mukaka, Mavuto. (2012). *Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research.* Malawi medical journal : the journal of Medical Association of Malawi. 24. 69-71.

Murphy, K.P., 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Murtagh, F. and Contreras, P., 2012. *Algorithms for hierarchical clustering: An overview.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp. 86–97.

Newbold, P., Carlson, W.L., and Thorne, B.M., 2019. *Statistics for Business and Economics*. 9th ed. Pearson.

Reddit Inc. (n.d.) *Developer terms*. Available at: https://redditinc.com/policies/developer-terms (Accessed: 23 December 2024).

Röder, M., Both, A. and Hinneburg, A., 2015. *Exploring the Space of Topic Coherence Measures*. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15), pp. 399–408.

Rousseeuw, P.J., 1987. Silhouettes: *A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, pp. 53–65.

Saito, T. and Rehmsmeier, M., 2015. *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. PLOS ONE, 10(3), p.e0118432.

Tufte, E.R. (2007). *The visual display of quantitative information.* 2nd ed., 5th printing. Cheshire, CT: Graphics Press.

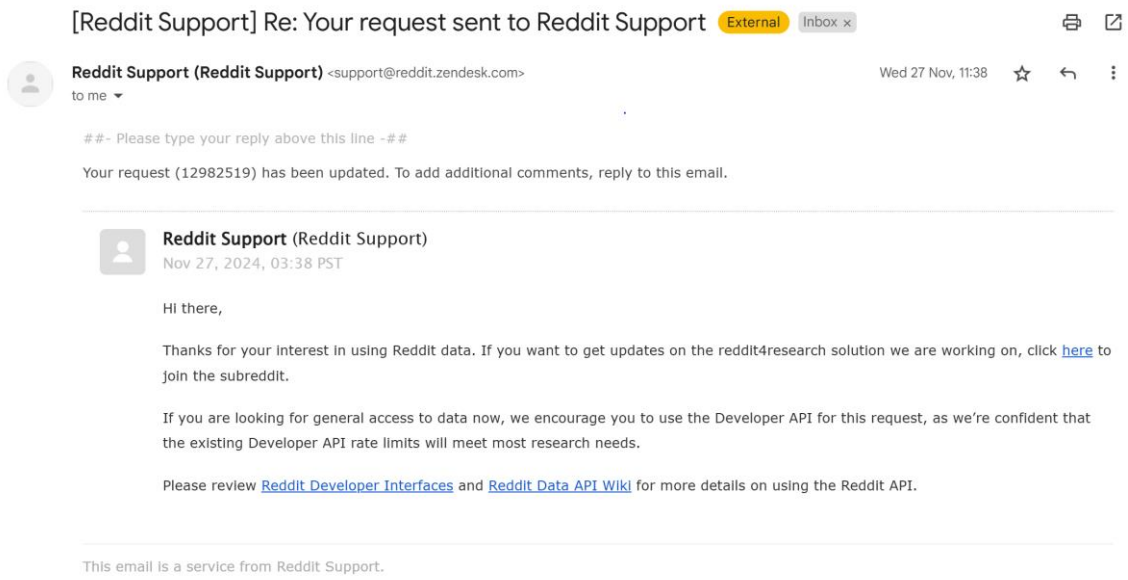
Yang, L. and Chiang, J.A., 2020. *Use case and performance analysis for missing data imputation methods in big data analytics*. Proceedings of the 2020 International Conference on Computing and Data Engineering (ICCDE '20), pp.107-111. DOI: 10.1145/3379247.3379270.

Wooldridge, J.M., 2013. *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning.

Appendix A. CMEF Dataset. The dataset provides detailed metrics on CAP funding allocation and organic farming adoption, essential for quantitative analysis. The dataset is openly available under public access licensing, explicitly permitting use for analysis and research without the need for formal permissions. The licensing facilitated smooth integration into the project and effectively eliminated the legal complexities typically associated with data acquisition.

3.1.1 Reddit

Explored as a potential source for public sentiment on organic farming. Data from Reddit was accessed via its Developer API using PRAW (Python Reddit API Wrapper), adhering to Reddit's Developer API terms (Reddit Inc., n.d.). Explicit authorization was sought and granted to ensure compliance with updated licensing policies (Figure 1). Reddit's licensing places limitations on the use, sharing, and storage of data, necessitating strict compliance with its terms to prevent violations. This involved adhering to API rate limits, anonymizing data to safeguard user identities, and refraining from any commercial use without explicit authorization.


[Reddit Support] Re: Your request sent to Reddit Support External Inbox ×

Reddit Support (Reddit Support) <support@reddit.zendesk.com> Wed 27 Nov, 11:38

to me

##- Please type your reply above this line -##

Your request (12982519) has been updated. To add additional comments, reply to this email.

Reddit Support (Reddit Support)
Nov 27, 2024, 03:38 PST

Hi there,

Thanks for your interest in using Reddit data. If you want to get updates on the reddit4research solution we are working on, click here to join the subreddit.

If you are looking for general access to data now, we encourage you to use the Developer API for this request, as we're confident that the existing Developer API rate limits will meet most research needs.

Please review Reddit Developer Interfaces and Reddit Data API Wiki for more details on using the Reddit API.

This email is a service from Reddit Support.

Figure 1: Authorization email from Reddit Support

3.2 Dashboard Design

A Streamlit dashboard was created to assist modern farmers by incorporating machine learning insights and dynamic visualizations (Figure 2 and Figure 3). Designed using Tufte's principles (Tufte, 2007), the dashboard focuses on clarity, precision, and accessibility, enabling users to explore organic farming indicators, CAP support, and clustering results. The design decisions are detailed in Table 1.

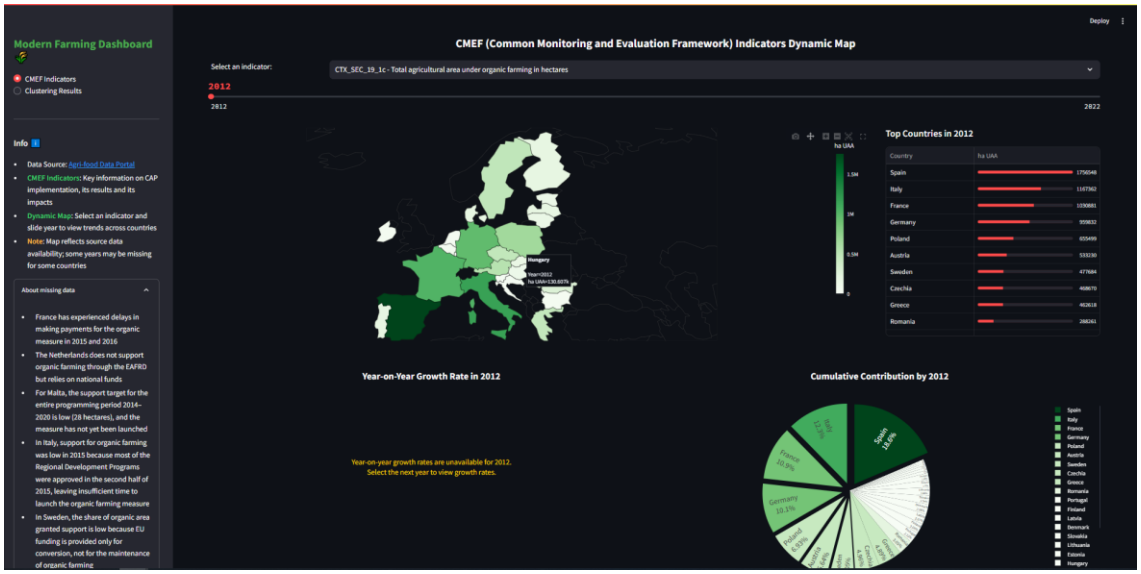


Figure 2: Modern Farming Dashboard: CMEF Indicators tab

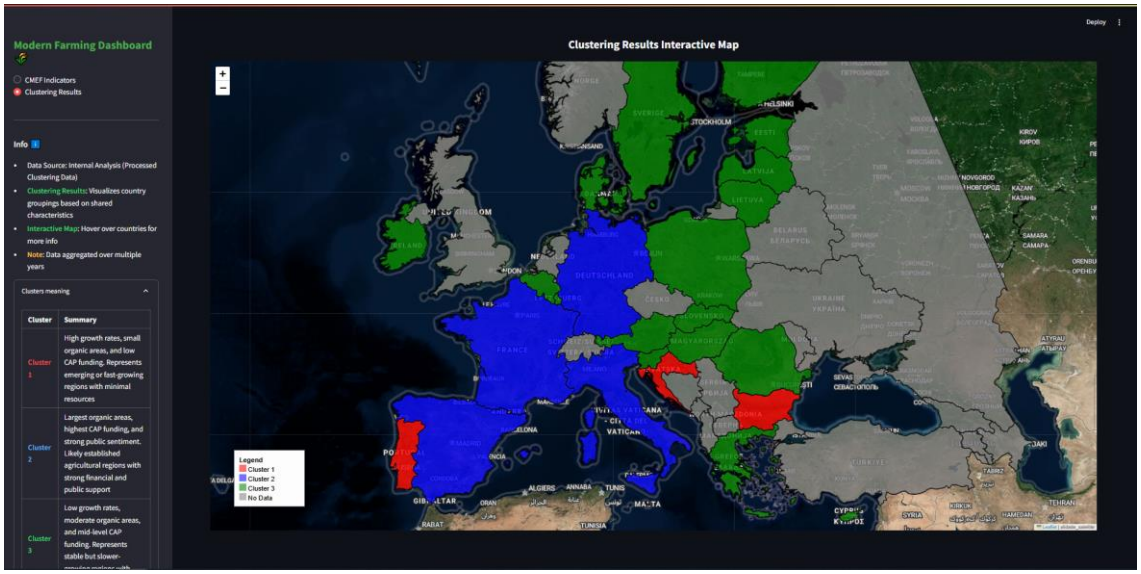


Figure 3: Modern Farming Dashboard: Clustering Results tab

Table 1: Dashboard Features and Design

Feature	Purpose	Design Choices	Alignment with Tufte's Principles
Dynamic Choropleth Maps	Highlight geographic trends in agricultural indicators	"Natural earth" projection, green color scales, interactive tooltips for country-level data	Ensures proportional data representation, clear scaling, and avoids unnecessary distractions
Top Countries Table	Rank countries by contribution to agricultural indicators	Ranked table with progress bars for quick identification of top contributors	Focused, data-driven view emphasizing ranking and minimizing non-essential elements

Feature	Purpose	Design Choices	Alignment with Tufte's Principles
Year-on-Year Growth Bar Charts	Show annual growth rates for agricultural indicators	Color-coded bars for positive/negative growth, interactive exploration	Encodes trends visually, avoiding redundant or distracting elements
Cumulative Contribution Pie Charts	Illustrate proportional contributions to agricultural indicators	Labels and percentages inside segments; annotations for large contributors	Accurate proportions with clear annotations, avoiding non-essential elements
Clustering Results Map	Visualize groups of countries by shared characteristics	Distinct cluster colors, tooltips for key metrics, and clear legend	Highlights groupings with intuitive labeling, avoiding extraneous elements
User-Centric Layout	Organize navigation and data source information	Sidebar with tools, context, and notes about missing data	A clean and simple layout tied to the data, making it easier to explore without feeling overwhelmed
Interactive Year Slider	Enable dynamic exploration of temporal data trends	Intuitive slider for year selection, responsive updates to visualizations	Enhances interactivity while keeping focus on relevant data, avoiding redundant controls
Accessibility and Responsiveness	Ensure usability across devices	Custom CSS styling to achieve a clear visual structure and uniform formatting	Emphasizes accessibility and clarity by removing distractions and concentrating on clear, effective communication
Multi-Device Support	Guarantee compatibility across different screen sizes and devices	Responsive design adjusts layout and visualizations for desktops, tablets, and mobile devices	Eliminates clutter and maintains a clear focus on data presentation

3.3 Optimization Strategies

3.3.1 Dashboard

The dashboard was designed to optimize resource use while ensuring a responsive user interface. Key optimization strategies summarized in Table 2, ensured optimal performance while maintaining a balance between system resources and user experience.

Table 2: Dashboard Optimization Strategies

Optimization Strategy	Details	Impact
Caching for Data Loading	Used @st.cache_data for loading and processing datasets like CMEF indicators and clustering data	Reduced redundant computations, lowering memory usage and processing time
Resource Optimization for Maps	Cached large GeoJSON files and map templates using @st.cache_resource	Conserved CPU and memory by avoiding repeated loading and rendering

Dynamic Filtering	Data subsets are filtered dynamically in response to user selections (e.g., year, indicator)	Reduced memory usage by focusing solely on processing relevant data
Pre-sorting and Grouping	Sorting and grouping were performed in advance as part of data preprocessing	Lower computational requirements during user interactions

These optimizations do involve certain trade-offs, such as increased memory usage due to caching large datasets and potential computational overhead from dynamic visualizations and interactivity on client devices. Still, they ensure smooth performance and scalability, helping the dashboard deliver practical insights for modern farming practices.

3.3.2 Scraping

The Reddit Scraper Class efficiently collects and processes Reddit data, using a modular structure to manage tasks like scraping, translating, and saving data. Table 3 outlines key optimization strategies that improved performance while ensuring reliability and scalability.

Table 3: Optimization Strategies in Reddit Scraper

Optimization Strategy	**Details**	**Impact**
Language Mapping and Translation	Maps countries to native languages and translates keywords, titles, and comments into English	Ensures query relevance, standardizes content for cross-language analysis, and avoids unnecessary translations
Data Caching and Incremental Saves	Loads existing data from a CSV file to avoid reprocessing and saves results incrementally	Prevents redundant computations, ensures scraping efficiency, and avoids data loss during interruptions
Dynamic Query Building	Combines translated and English keywords dynamically	Optimizes query accuracy while expanding coverage
Error Handling and Retries	Implements retry mechanisms for API errors and skips invalid posts or comments	Improves scraper stability and reduces interruptions during operation
Filtered Processing	Filters posts and comments by year, subreddit, and query dynamically	Minimizes memory usage and processing time by focusing only on relevant data
Avoiding Duplicate Processing	Monitors existing post IDs and queries to avoid duplicate processing	Saves resources and ensures scraping efficiency

However, these approaches come with compromises: caching increases memory usage as the dataset grows, translation of non-English posts and comments adds processing time, incremental saves introduce frequent file I/O operations, and retry mechanisms extend execution when handling API errors. These optimizations were designed to balance efficiency with the reliability needed for cross-language data analysis in this project.

4. Results

4.1 Exploratory Data Analysis (EDA)

4.1.1 Descriptive statistics

The CMEF dataset (2012-2022, 27 countries) shows substantial variability across indicators. Organic farming area and its share demonstrate right-skewed distributions, with means exceeding medians. Public expenditure indicators reveal considerable funding differences across countries, while CAP-supported areas highlight significant regional disparities. Detailed statistics in Table 4.

Table 4: Summary Statistics of CMEF Indicators

	count	mean	std	min	25%	50%	75%	max
CTX_SEC_19_1c	279	452400	595229	7	76538	225235	530991	2.77567 e+06
CTX_SEC_19_2	268	8.17392	5.66803	0.06	3.72	7.06	10.795	25.69
OIR_01a_2.11	212	7.19843 e+07	9.76171e +07	0	1.69431 e+07	3.46207 e+07	7.86491 e+07	5.26719 e+08
OIR_01b_2.11	212	4.79192 e+07	6.16623e +07	0	9.75609 e+06	2.56259 e+07	5.22362 e+07	3.03951 e+08
OIR_06_1.2	201	354290	393344	5.59	83161.3	205553	458044	1.81451 e+06

The distributions plots in Figure 4 shows significant right-skewness and variability in CMEF indicators, with most countries having small values for organic farming area, public expenditure, and CAP-supported areas. A few outliers drive long tails, while organic farming share is more balanced but still slightly skewed. These patterns highlight disparities, with most countries at lower levels and a few regions dominating high values.

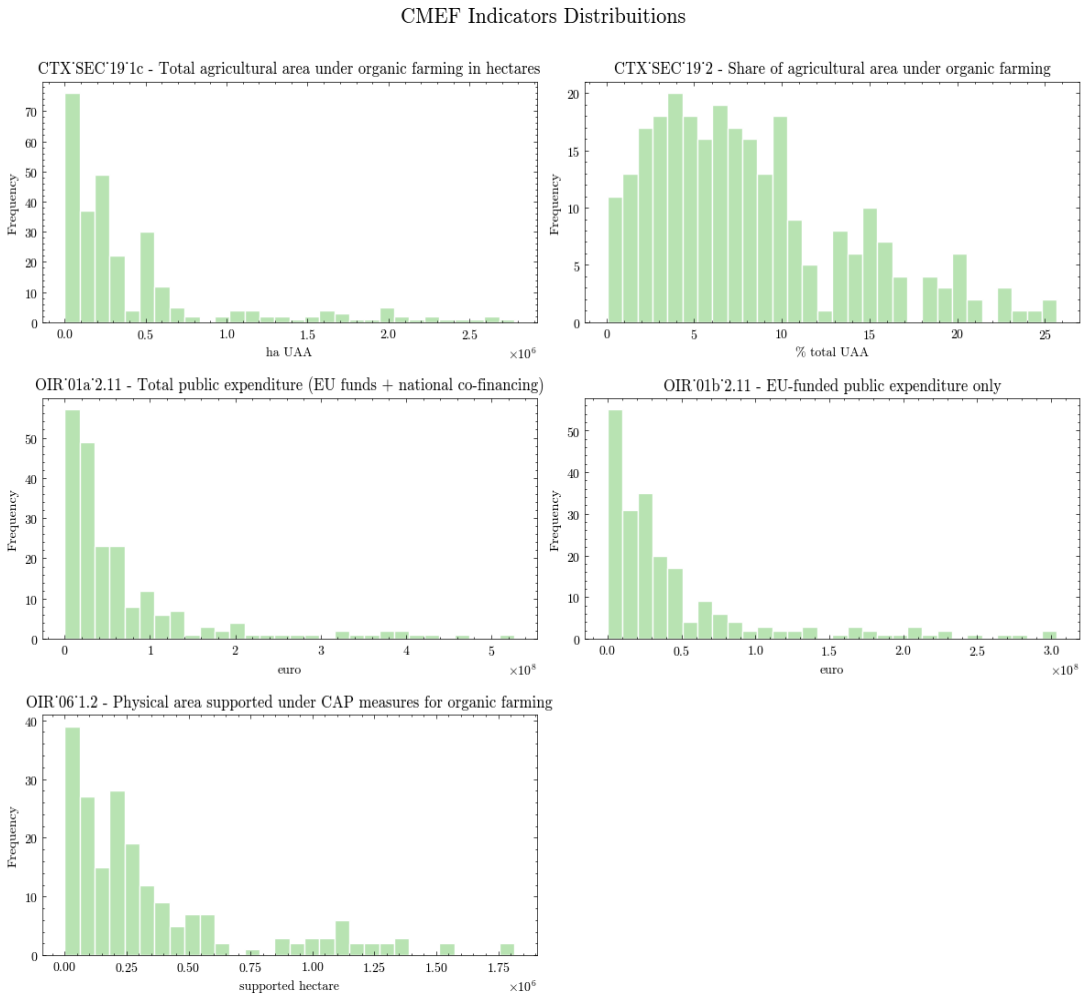


Figure 4: Distributions of CMEF Indicators

The trends in Figure 5 highlights steady growth in organic farming and policy support, with increases in total organic area and its agricultural share, especially post-2014. Public expenditure and CAP-supported areas grew notably from 2015, reflecting CAP reforms and expanded financial support.

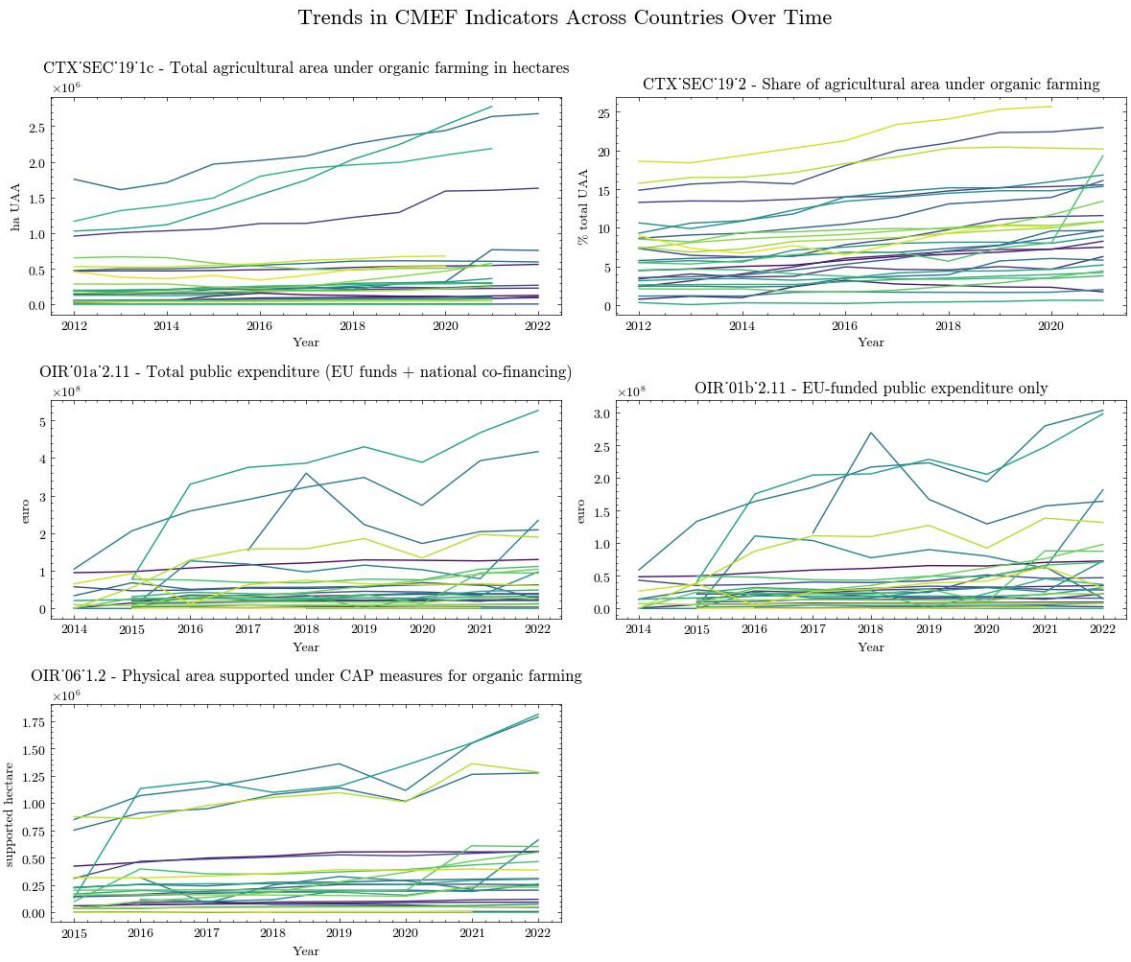


Figure 5: Trends in CMEF Indicators Across Countries Over Time

4.1.2 Data Cleaning

The data cleaning process involved renaming columns for consistency, filtering out non-EU entities, and mapping country names to their ISO codes for map representation.

4.1.3 Missing Values

The flag column was removed due to the high proportion of missing values, with 91.8% of the entries being absent (see Figure 6).

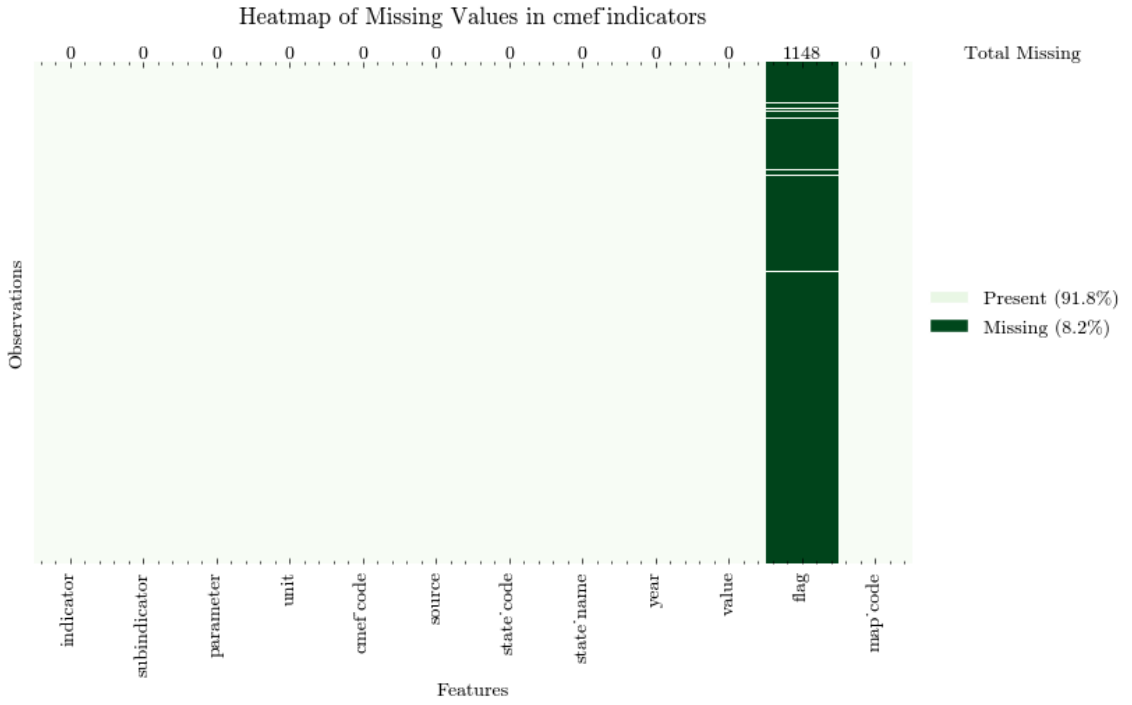


Figure 6: Heatmap of Missing Values in CMEF Indicators

4.1.4 CMEF Indicators Missing Countries / Years

5. The heatmaps in Figure 7, Figure 8, Figure 9 highlight data gaps in CMEF indicators, particularly CAP funding, absent before 2014 due to their introduction with that year's CAP reform. Additional details in References

Bird, S., Klein, E., and Loper, E., 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3, 993–1022.

Blomer, J., Lammersen, C., Schmidt, M., and Sohler, C. (2016) *Theoretical analysis of the k-means algorithm—A survey*, Algorithm Engineering, 9220, pp. 81–116. doi: 10.1007/978-3-319-49487-6_3.

European Commission (n.d.a) *AgriData Portal: CAP Indicators*. Available at: https://agridata.ec.europa.eu/extensions/DataPortal/cap_indicators.html (Accessed: 23 November 2024).

European Commission (n.d.b) *Organic Production Sources.* Available at: https://agridata.ec.europa.eu/Qlik_Downloads/Organic-Production-sources.htm (Accessed: 23 November 2024).

Fawcett, T., 2006. *An Introduction to ROC Analysis*. Pattern Recognition Letters, 27(8), pp. 861–874.

García-Laencina, P.J., Sancho-Gómez, J.L. and Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: A review. Neural Computing and Applications, 19(2), pp. 263–282.

Guyon, I. and Elisseeff, A., 2003. *An introduction to variable and feature selection*. Journal of Machine Learning Research, 3, pp. 1157–1182.

Han, J., Kamber, M. and Pei, J., 2011. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann.

Hox, J.J., Moerbeek, M. and van de Schoot, R., 2017. *Multilevel Analysis: Techniques and Applications*. 3rd ed. Routledge.

Hutto, C.J. and Gilbert, E., 2014. VADER: *A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14).

Jolliffe, I.T. and Cadima, J., 2016. *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202.

Kim, Jae-Dong & Hwang, Ji-Hwan & Doh, Hyoung-Ho. (2023). *A Predictive Model with Data Scaling Methodologies for Forecasting Spare Parts Demand in Military Logistics*. Defence Science Journal. 73. 666-674. 10.14429/dsj.73.19129.

Medhat, W., Hassan, A., and Korashy, H., 2014. *Sentiment analysis algorithms and applications: A survey*. Ain Shams Engineering Journal, 5(4), pp. 1093–1113.

Mukaka, Mavuto. (2012). *Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research.* Malawi medical journal : the journal of Medical Association of Malawi. 24. 69-71.

Murphy, K.P., 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Murtagh, F. and Contreras, P., 2012. *Algorithms for hierarchical clustering: An overview.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp. 86–97.

Newbold, P., Carlson, W.L., and Thorne, B.M., 2019. *Statistics for Business and Economics*. 9th ed. Pearson.

Reddit Inc. (n.d.) *Developer terms*. Available at: https://redditinc.com/policies/developer-terms (Accessed: 23 December 2024).

Röder, M., Both, A. and Hinneburg, A., 2015. *Exploring the Space of Topic Coherence Measures*. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15), pp. 399–408.

Rousseeuw, P.J., 1987. Silhouettes: *A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, pp. 53–65.

Saito, T. and Rehmsmeier, M., 2015. *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. PLOS ONE, 10(3), p.e0118432.

Tufte, E.R. (2007). *The visual display of quantitative information.* 2nd ed., 5th printing. Cheshire, CT: Graphics Press.

Yang, L. and Chiang, J.A., 2020. *Use case and performance analysis for missing data imputation methods in big data analytics*. Proceedings of the 2020 International Conference on Computing and Data Engineering (ICCDE '20), pp.107-111. DOI: 10.1145/3379247.3379270.

Wooldridge, J.M., 2013. *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning.

Appendix A. CMEF Dataset.

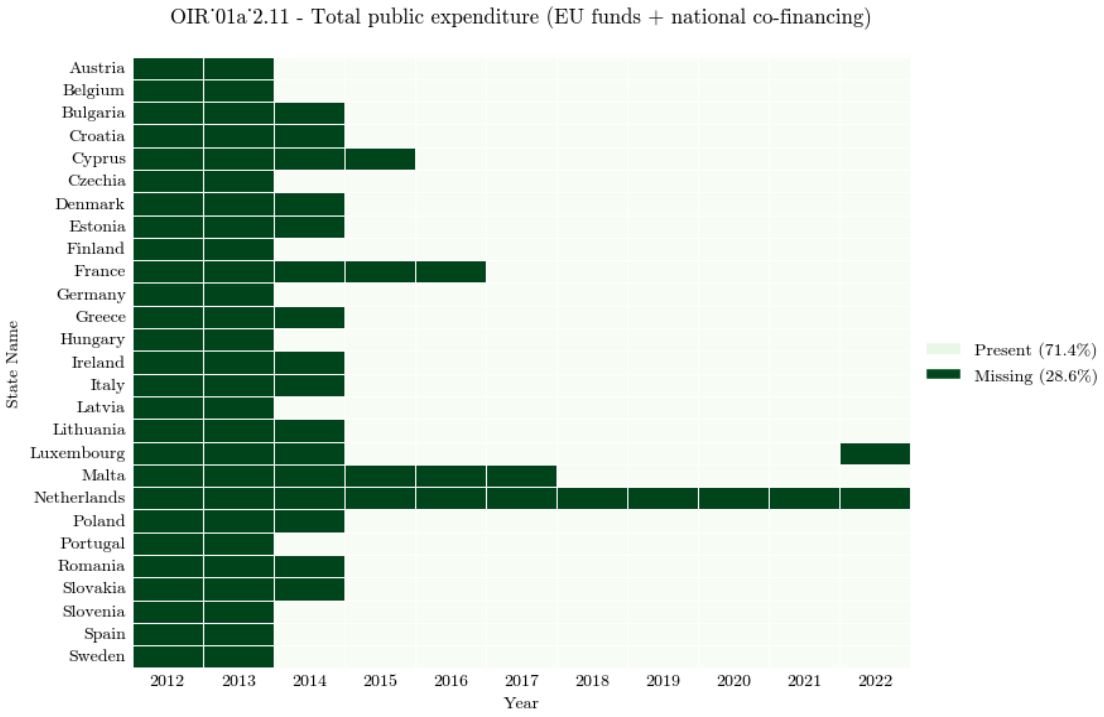


Figure 7: Missing Data in OIR'01a'2.11 - Total Public Expenditure

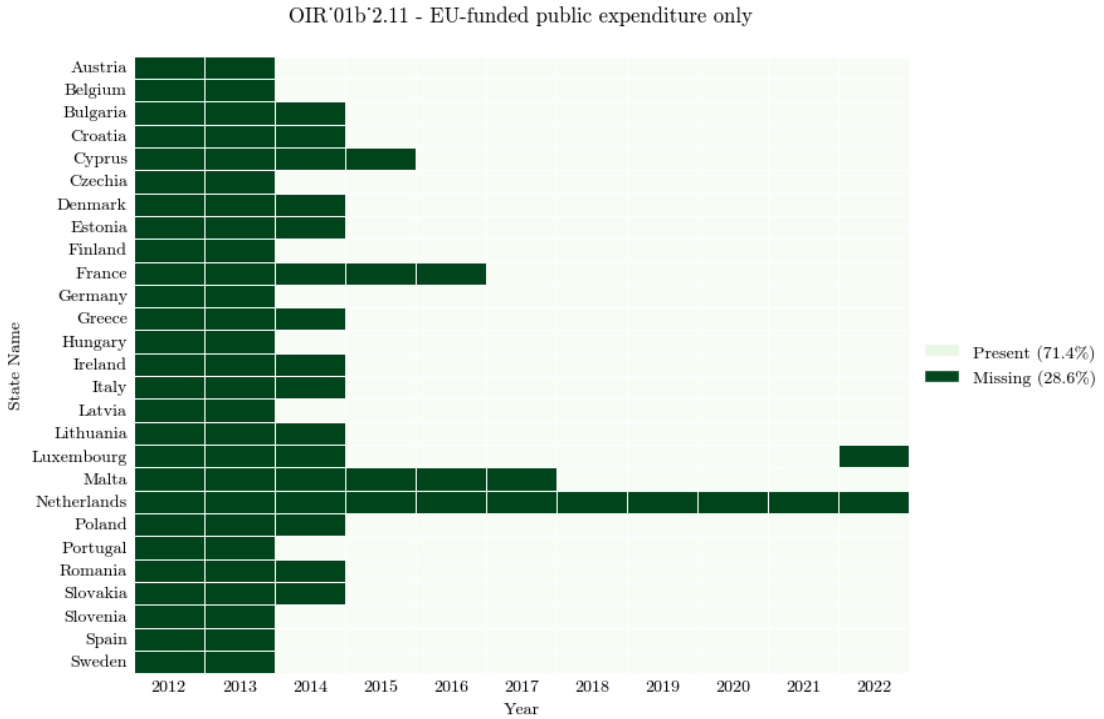


Figure 8: Missing Data in OIR'01b'2.11 - EU-Funded Public Expenditure

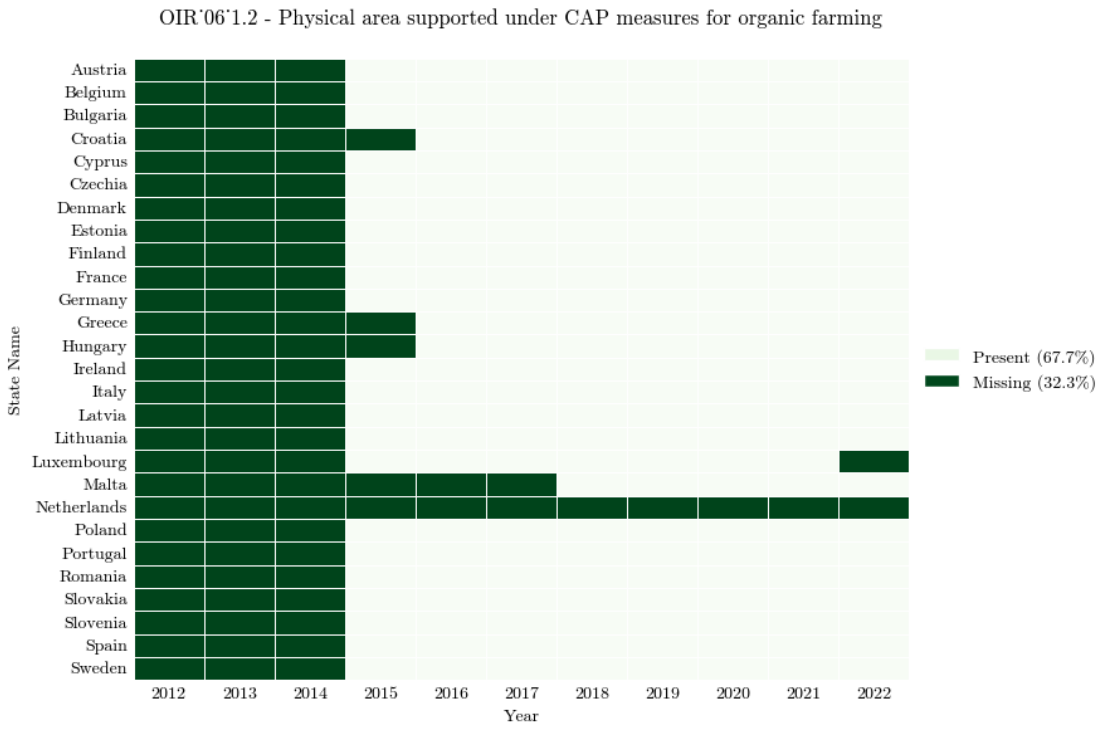


Figure 9: Missing Data in OIR'06'1.2 - Supported Area Under CAP Measures

Figure 10 and Figure 11 show how consistent data availability for organic farming area and share indicators, with recent gaps (2021–2022) likely due to reporting delays. Imputation will address these during ML pre-processing.

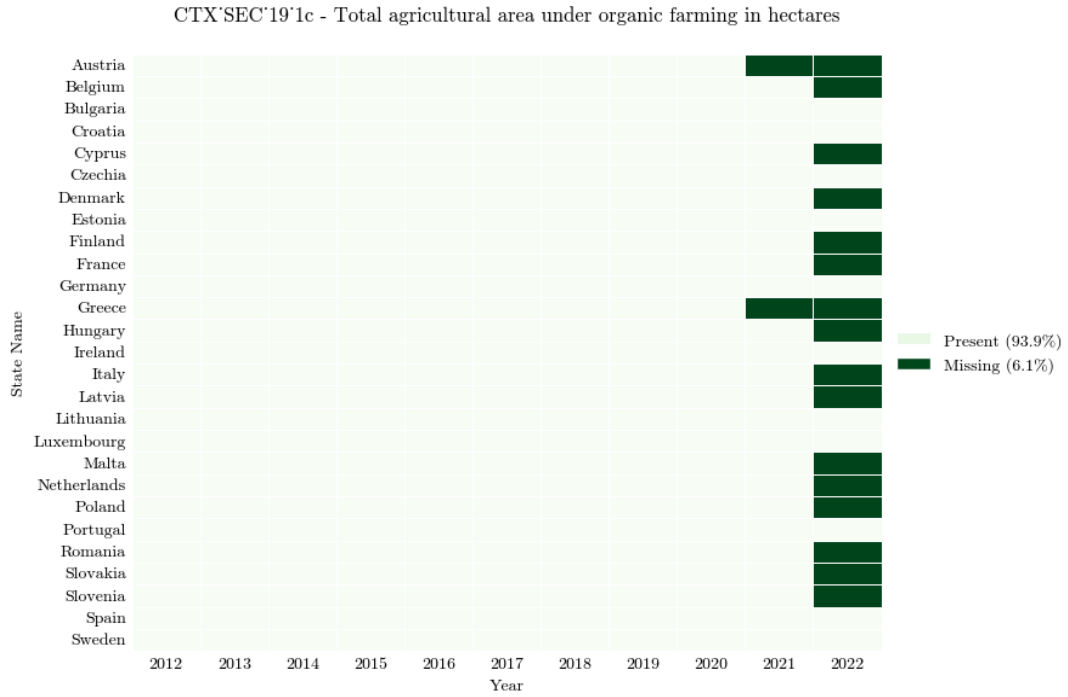


Figure 10: Missing Data in CTX'SEC'19'1c - Total Agricultural Area Under Organic Farming

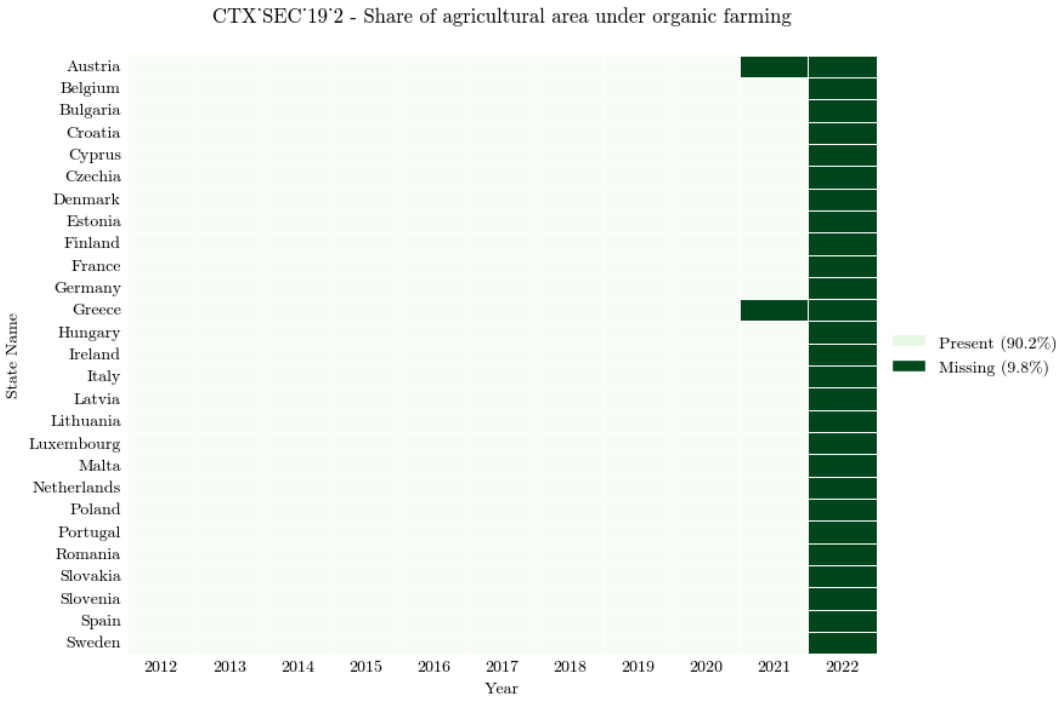


Figure 11: Missing Data in CTX'SEC'19'2 - Share of Agricultural Area Under Organic Farming

5.1.1 Outliers

Boxplots in Figure 12 highlight disparities and outliers in CMEF indicators. Organic farming areas and CAP-supported regions show wide variability, with outliers reflecting concentrated efforts. Public expenditure varies significantly, with some countries or years receiving exceptionally high amounts. Organic farming share, though less variable, shows uneven progress across regions.

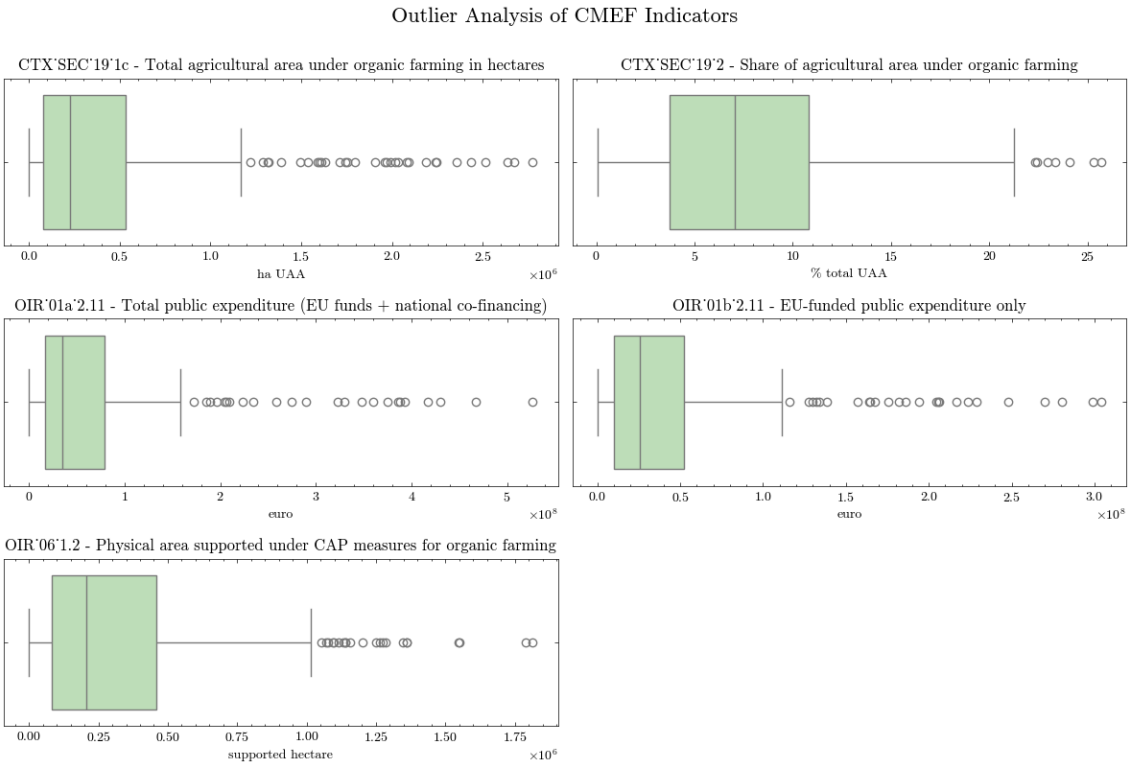


Figure 12: Outlier Analysis of CMEF Indicators

5.1.2 Organic Farming Area and Growth

To assess the dynamics of organic farming adoption over time, the annual growth rate is computed from organic farming area, using the standard percentage change formula (Wooldridge, 2013):

$$\text{Growth Rate (\%)} = \left(\frac{\text{Value}_{\text{current year}} - \text{Value}_{\text{previous year}}}{\text{Value}_{\text{previous year}}} \right) \times 100$$

Where:

- $\text{Value}_{\text{current year}}$: The value of total agricultural area for the current year.
- $\text{Value}_{\text{previous year}}$: The value of total agricultural area for the previous year.

Figure 13 compares average annual growth rates and organic farming shares across countries, highlighting long-term trends while reducing yearly fluctuations. Ireland shows steady growth like Finland and Belgium but lags in organic farming share compared to emerging adopters like Bulgaria and Romania. Leaders Austria and Sweden have high land shares, while Malta and Portugal display rapid growth despite smaller shares. These countries will serve as references for time series analysis.

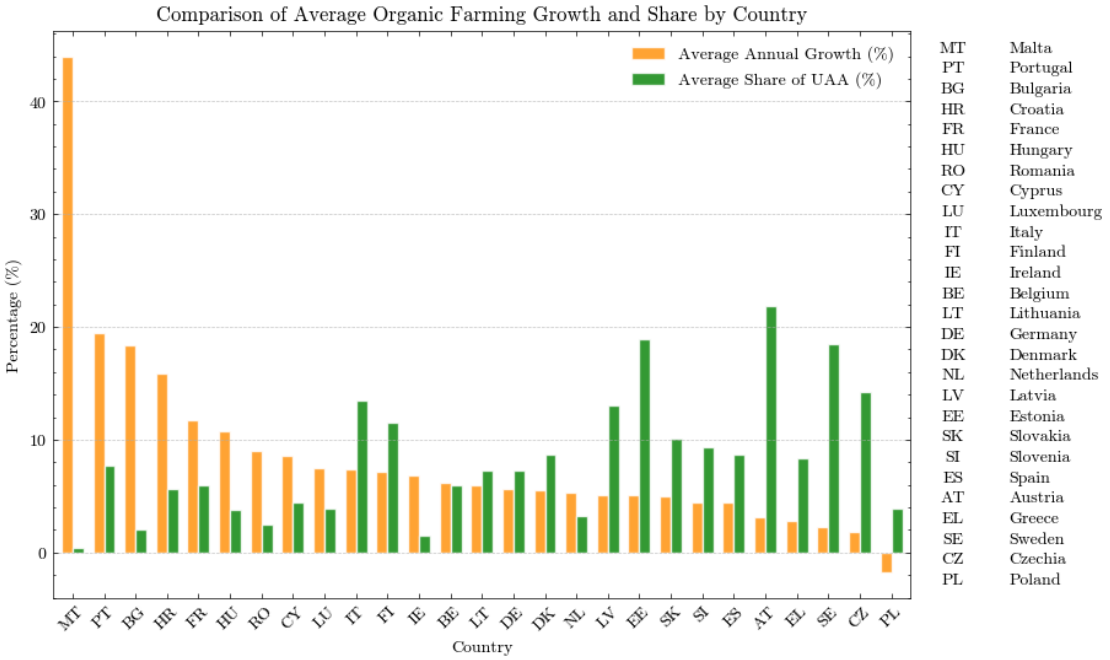


Figure 13: Comparison of Average Organic Farming Growth and Share by Country

Time-series plot in Figure 14 confirms Austria and Sweden's leadership in organic farming with consistently high areas and shares. Portugal shows rapid post-2020 growth, reflecting successful policies. Ireland displays steady but modest progress, while Malta and Bulgaria, despite rapid growth, remain in early adoption stages with low metrics.

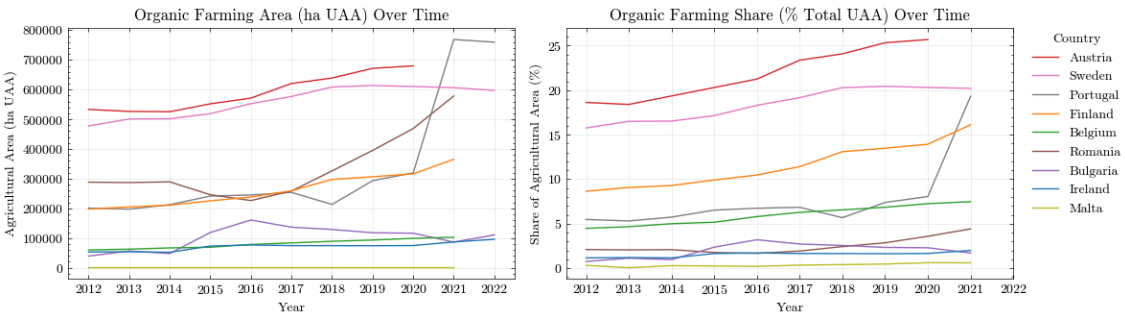


Figure 14: Trends in Organic Farming Area and Share Over Time

5.1.3 CAP Support for Organic Farming

The impact of CAP funding on the adoption of organic farming has been examined by analysing the relationship between financial support and organic farming growth across different countries.

Figure 15 shows CAP funding trends (2014–2022), with a steady increase peaking in 2018. Lines represent means, shaded areas 95% confidence intervals. The increasing gap between total funding and EU contributions underscores the significance of national co-financing, while the broader shading reflects variations in priorities or capabilities.

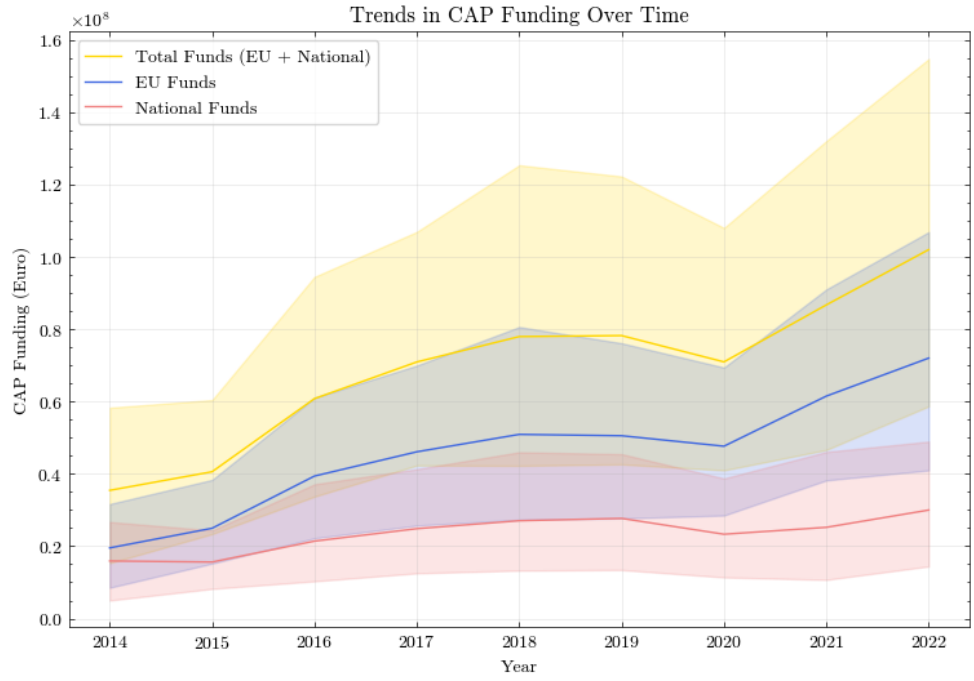


Figure 15: Trends in CAP Funding Over Time

Figure 16 shows growth in organic farming, with total and CAP-supported areas rising. Faster total area expansion suggests market demand and non-CAP factors drive growth, while CAP funding may not fully match adoption rates.

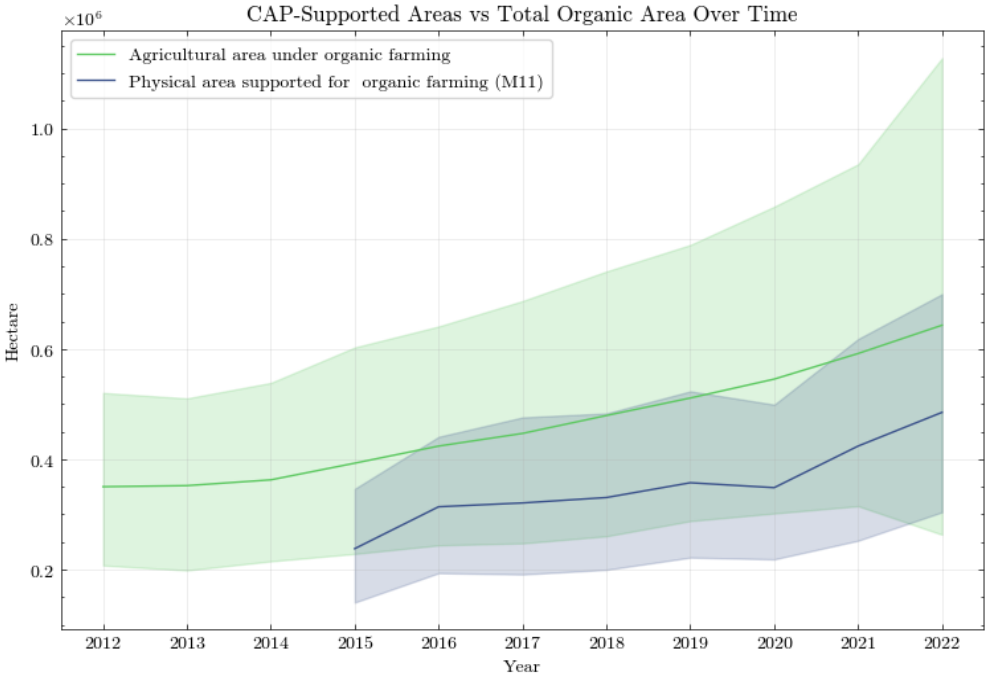


Figure 16: CAP-Supported Areas vs Total Organic Area Over Time

Figure 17 highlights CAP's role in organic farming growth. Portugal's post-2020 surge shows policy impact, Austria and Sweden align CAP support with high organic levels,

Romania and Bulgaria demonstrate CAP-driven adoption, whereas Ireland's slower progress highlights the need for more robust policies.

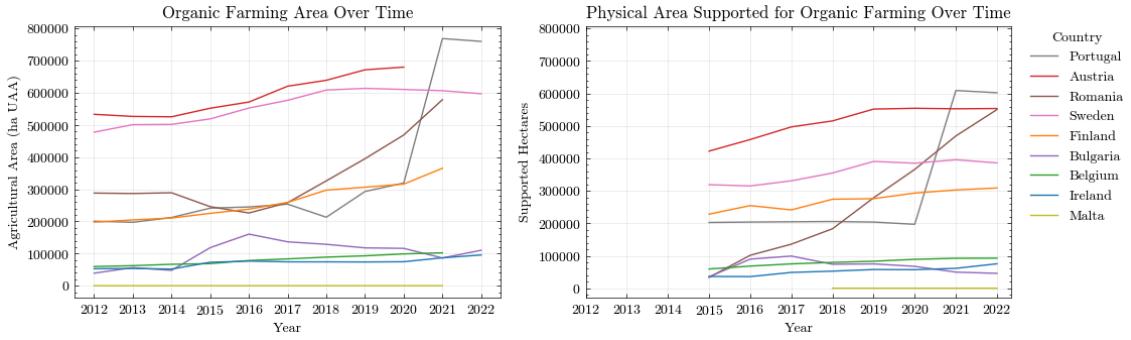


Figure 17: Organic Farming Area vs CAP-Supported Area Over Time

Figure 18 shows how CAP support helps drive organic farming growth by supporting established systems, boosting emerging sectors, scaling successful practices, and strengthening smaller regions.

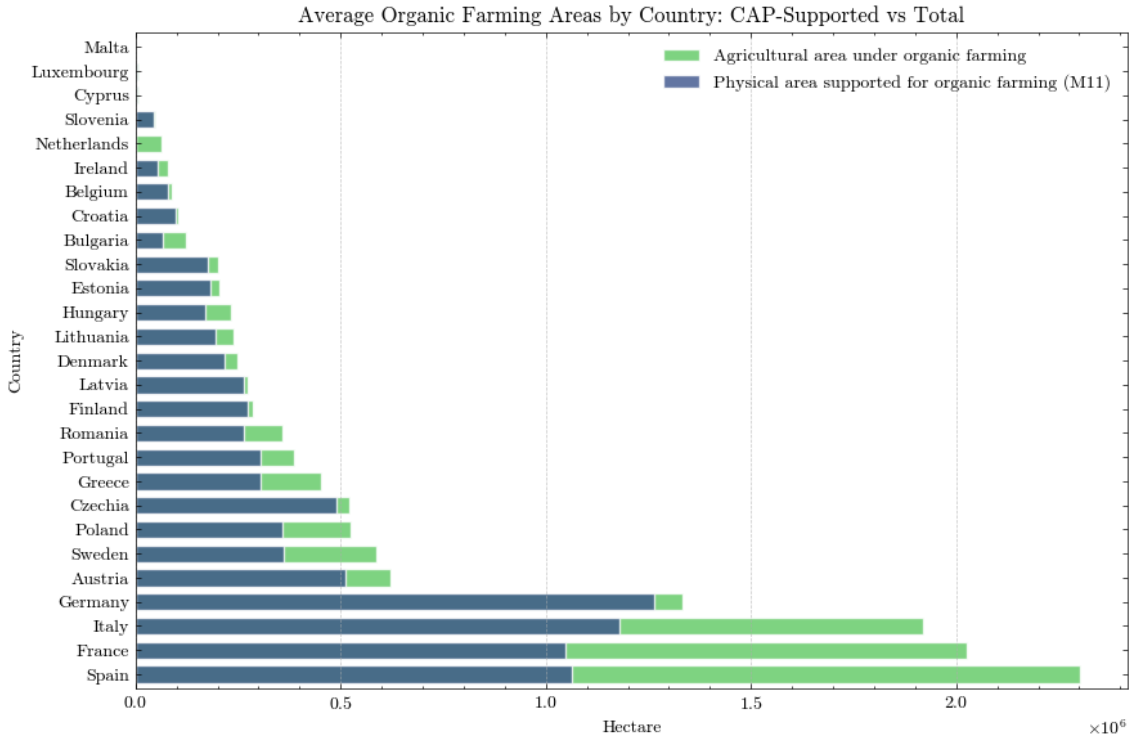


Figure 18: Average Organic Farming Areas by Country: CAP-Supported vs Total

Figure 19 reveals disparities in CAP funding per hectare across the EU. Ireland receives low support from both EU and national sources. Variations reflect land size, geographic challenges, and policy priorities, with some countries balancing EU and national funding, while others rely more on EU support.

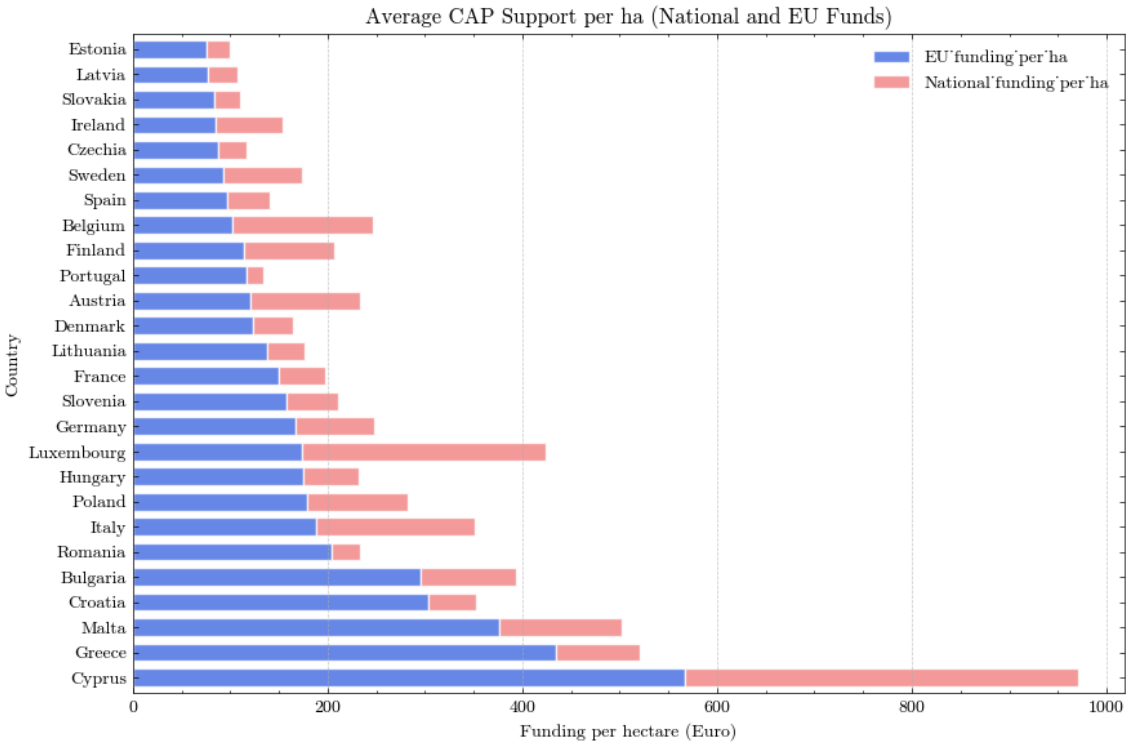


Figure 19: Average CAP Support per Hectare (National and EU Funds)

5.1.4 Confidence Intervals for Organic Farming Share

Exploratory Data Analysis (EDA) shows clear differences in the share of organic farming across Europe, leading to a closer look at this variable. Figure 20 presents an estimate of the population mean alongside confidence intervals to evaluate the uncertainty in country-level averages. The confidence intervals are calculated using the formula:

$$CI = \bar{x} \pm Z \cdot \frac{s}{\sqrt{n}}$$

Where $Z = 1.96$ corresponds to a 95% confidence level. These intervals provide a measure of the reliability of observed trends, incorporating the mean ($\bar{x}$), standard deviation (s) and sample size (n) to reflect both the variability and accuracy of the estimates (Newbold et al., 2019).

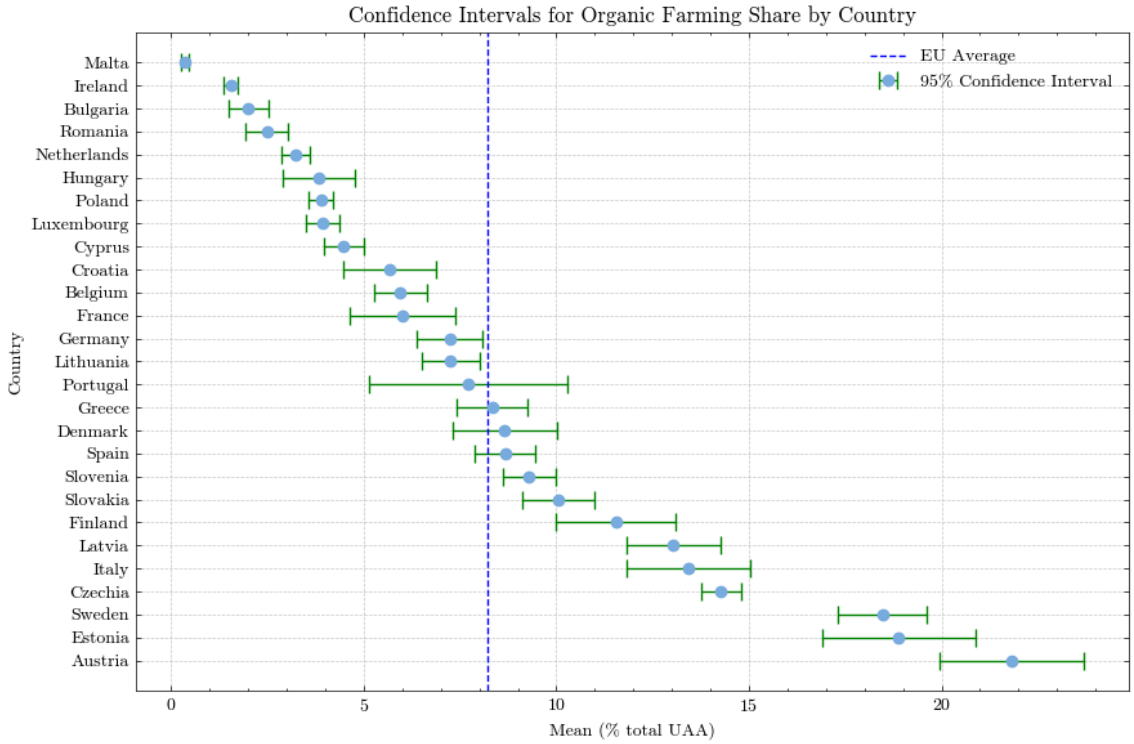


Figure 20: Confidence Intervals for Organic Farming Share by Country

5.2 Hypothesis Testing

Table 5 provides an overview of the tests utilized in this analysis, comparing the characteristics of parametric and non-parametric approaches (Newbold et al., 2019).

Table 5: Comparison of Parametric and Non-Parametric Tests

Feature	Parametric Tests	Non-Parametric Tests
Distribution	Assumes a normal distribution	Does not rely on assumptions about the underlying distribution
Data Type	Continuous (interval/ratio scale)	Ordinal, ranked, or continuous
Statistical Power	More powerful when assumptions are satisfied	Less powerful but remains robust when assumptions are violated
Examples of Tests	Paired T-test, Repeated Measures ANOVA	Wilcoxon Signed-Rank Test, Friedman Test
Typical Use Case	Normally distributed data (e.g., comparing means)	Skewed distributions, small samples, or cases with assumption violations (e.g., comparing medians or ranks)

5.2.1 Paired T-test

The Paired T-test compares organic farming share between Ireland and Bulgaria using matched yearly observations, as shown in Figure 21. Unlike an independent T-test, the paired T-test is suitable here as it accounts for matched pairs. The test statistic is calculated as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

Where:

- $\bar{d}$ = mean of the differences between paired observations.
- s_d = standard deviation of the differences.
- n = number of paired observations.

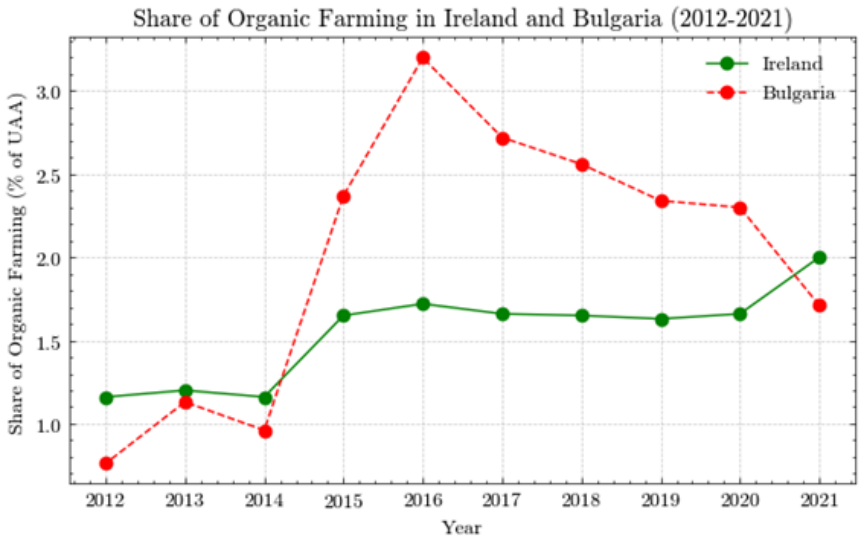


Figure 21: Share of Organic Farming in Ireland and Bulgaria (2012–2021)

Ireland and Bulgaria, both in early stages of organic farming adoption, are compared. Ireland shows steady growth, while Bulgaria displays fluctuations, as shown by yearly differences in Table 6. The assumptions and hypotheses for the test are outlined in Table 7, ensuring the validity of the results.

Table 6: Yearly Differences in Organic Farming Shares

year	Ireland	Bulgaria	Difference
2021	2	1.71	0.29
2020	1.66	2.3	-0.64
2019	1.63	2.34	-0.71
2018	1.65	2.56	-0.91
2017	1.66	2.72	-1.06
2016	1.72	3.2	-1.48
2015	1.65	2.37	-0.72
2014	1.16	0.96	0.2
2013	1.2	1.13	0.07
2012	1.16	0.76	0.4

Table 7: Assumptions and Hypotheses for the Paired T-Test

Category	Details
Hypotheses	

Category	Details
Null (H_0)	The mean difference in organic farming shares between Ireland and Bulgaria is zero.
Alternative (H_1)	The mean difference is not zero.
Assumptions	
Paired Observations	Yearly comparisons are paired between Ireland and Bulgaria.
Normality	Differences are approximately normally distributed (Shapiro-Wilk test).
Outliers	No extreme outliers in the differences.
Independence	Pairs are independent of each other.

The assumptions for the Paired T-test were checked as follows:

- Normality: The Shapiro-Wilk test result indicates that the differences appears to be normally distributed (P-value $\geq \alpha$):
  - α: 0.05
  - Statistic: 0.9147334
  - P-value: 0.3151123
- Outliers: No extreme outliers were detected in the differences (see Figure 22).

The Paired T-test results indicate:

- α: 0.05
- Statistic: -2.2239228
- P-value: 0.5322218

At a significance level of 0.05, the P-value is slightly above the threshold, so we fail to reject the null hypothesis (P-value $\geq \alpha$). This indicates no statistically significant difference in the mean share of organic farming between Ireland and Bulgaria across the observed years.

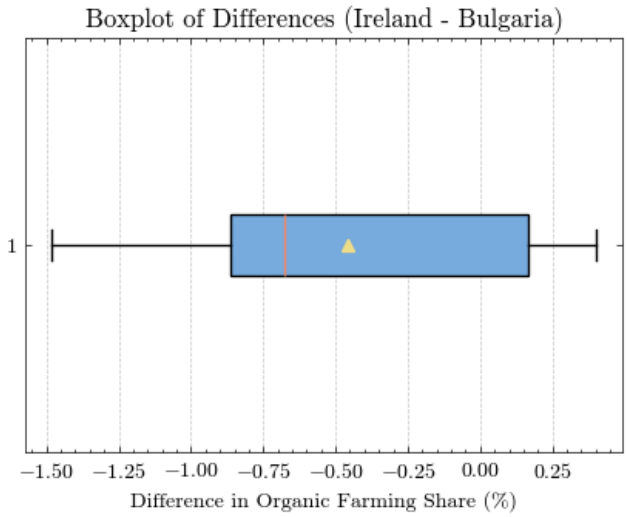


Figure 22: Boxplot of Differences in Organic Farming Share (Ireland - Bulgaria)

Figure 23 shows the T-distribution for the paired T-test. The T-statistic (-2.22) is within the critical range (±2.26), and the P-value (0.053) exceeds 0.05. Thus, we fail to reject the null hypothesis.

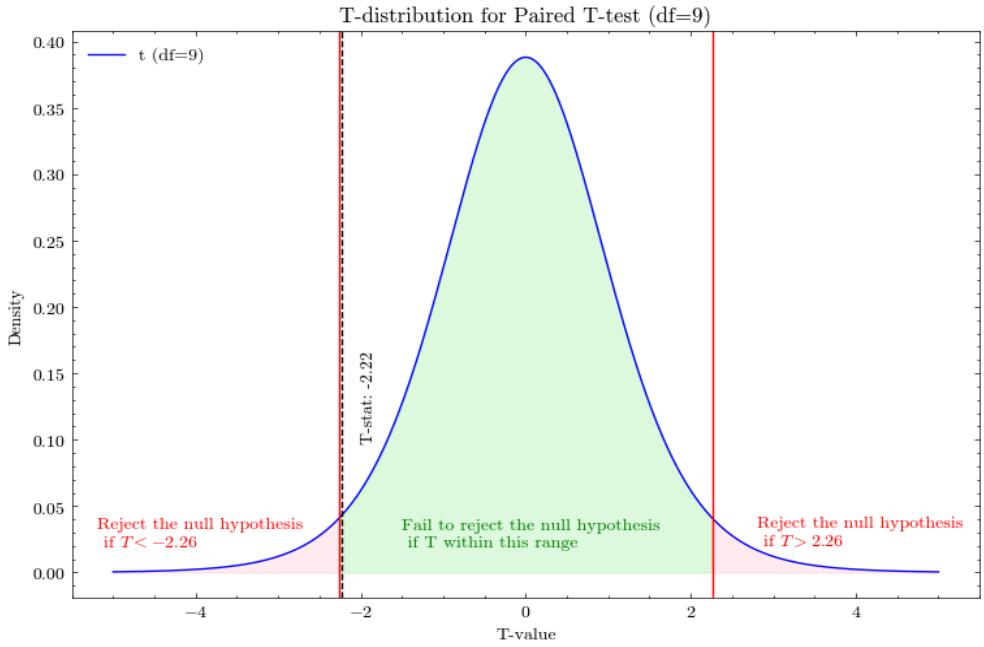


Figure 23: Distribution for Paired T-Test (df = 9)

5.2.2 Wilcoxon Signed-Rank test

The Wilcoxon Signed-Rank Test, a non-parametric alternative to the Paired T-test, is used when normality is questionable. Figure 24 shows boxplots of organic farming shares, with Bulgaria displaying greater range and variability than Ireland. The Wilcoxon test assesses whether the median of paired differences significantly deviates from zero by ranking the absolute differences. The test statistic is calculated as:

$$W = \sum R^+$$

Where:

- R^+ = sum of ranks for the positive differences between paired observations.

This test complements the Paired T-test by not assuming normality and focusing on ranks, making it less sensitive to variability or outliers. It is ideal for small samples or skewed distributions. The Wilcoxon Signed-Rank Test results indicate:

- α = 0.05
- Statistic: 10.0
- P-value: 0.084

We fail to reject the null hypothesis at α = 0.05 (P-value ≥ α). This means there is insufficient evidence to conclude that the medians of organic farming shares between Ireland and Bulgaria are different over the observed years.

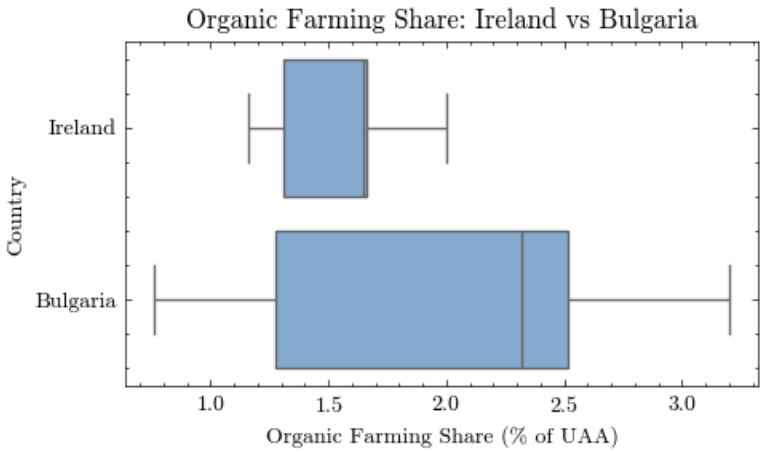


Figure 24: Organic Farming Share: Ireland vs. Bulgaria

Both the Paired T-test and Wilcoxon Signed-Rank Test confirm that, despite Bulgaria's fluctuations, the differences between the two countries are not statistically significant.

5.2.3 Repeated Measures ANOVA test

Repeated Measures ANOVA analyzes datasets with repeated measurements of the same subjects over time. Here, it compares organic farming shares across Ireland, Bulgaria, and Romania, focusing on early adopters. Assumptions and hypotheses are in Table 8.

Table 8: Assumptions and Hypotheses for Repeated Measures ANOVA

Category	**Details**
Hypotheses	
Null (H_0)	No differences in organic farming shares between countries or over time.
Alternative (H_1)	At least one group or time period has a different mean.
Assumptions	
Sphericity	Variances of differences between repeated measures (years) are equal (Mauchly's test).
Normality	Organic farming shares are normally distributed (Shapiro-Wilk test or Q-Q plots).
Outliers	No extreme values present (checked via boxplots or IQR).
Random Sampling	Data is collected randomly or is representative.
Independence	Groups (countries) are independent from one another.

Table 9 shows that Ireland and Romania's data violate the normality assumption ($p = 0.037$ for Ireland and $p = 0.032$ for Romania, Shapiro-Wilk Test). Although variances are homogeneous (Levene's Test: $p = 0.176$), the lack of normality renders the Repeated Measures ANOVA unsuitable. The yearly organic farming shares for the three countries are presented in Table 10.

Table 9: Assumption Testing Results

Country	Normality (Shapiro-Wilk Test)	Levene's Test for Homogeneity of Variances
Ireland	Statistic = 0.833, p = 0.037	Statistic = 1.852, p = 0.176
Bulgaria	Statistic = 0.833, p = 0.037	
Romania	Statistic = 0.829, p = 0.032	

Table 10: Yearly Organic Farming Shares for Ireland, Bulgaria, and Romania

year	Ireland	Bulgaria	Romania
2021	2	1.71	4.42
2020	1.66	2.3	3.59
2019	1.63	2.34	2.86
2018	1.65	2.56	2.43
2017	1.66	2.72	1.93
2016	1.72	3.2	1.67
2015	1.65	2.37	1.77
2014	1.16	0.96	2.09
2013	1.2	1.13	2.06
2012	1.16	0.76	2.1

Therefore, a non-parametric alternative like the Friedman Test, is more appropriate as it does not require normality and is suitable for repeated measures data.

5.2.4 Friedman test

The Friedman Test, a non-parametric alternative to repeated measures ANOVA, is used for repeated measures or matched groups when normality is violated. The test statistic is calculated as:

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)$$

Where:

- n = number of blocks (subjects or groups),
- k = number of treatments (conditions or time points),
- R_i = sum of ranks for the i-th treatment, and
- χ_F^2 = Friedman chi-square statistic.

This statistic is used to test if there are significant differences in the ranks across the treatments. The Friedman Test results indicate:

- $\alpha = 0.05$
- Statistic: 5.0
- P-value: 0.082

At a significance level of 0.05, the P-value is slightly above the threshold, so we fail to reject the null hypothesis (P-value $\geq \alpha$). This means the Friedman Test did not find statistically significant differences in the median organic farming shares among Ireland, Bulgaria, and Romania.

However, visual trends in Figure 25 shows practical differences: Ireland's steady growth, Bulgaria's fluctuations, and Romania's sharp post-2020 increase. As the test focuses on median ranks, it misses year-to-year change magnitudes or direction, highlighting the need for regression analysis to compare trend slopes.

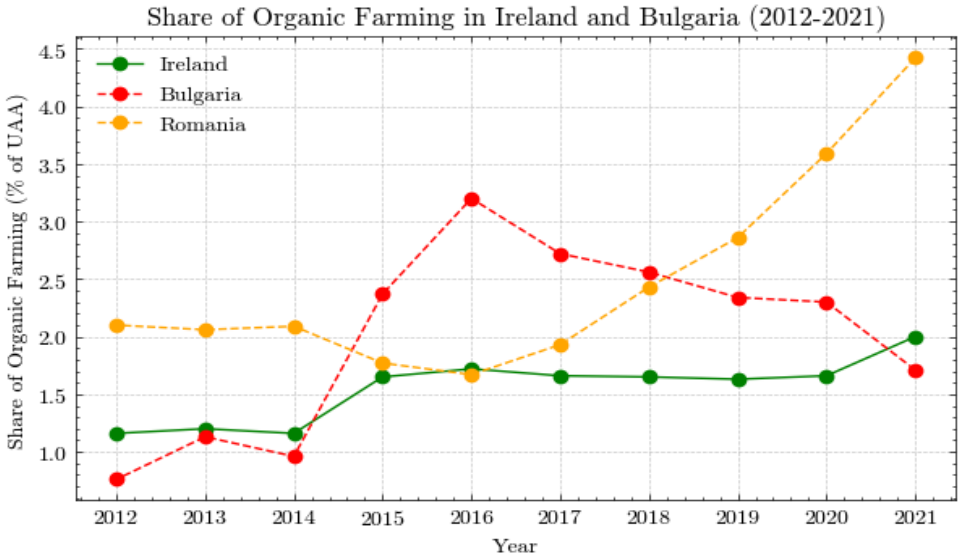


Figure 25: Trends in Organic Farming Share (% of UAA) in Ireland, Bulgaria, and Romania (2012–2021)

5.2.5 Ordinary Least Squares (OLS)

Regression analysis, shown in Figure 26 and summarized in Table 11, confirms the trends in organic farming shares.

Table 11: Regression Results for Organic Farming Trends

Country	Trend Description	P-value	R^2	Interpretation
Ireland	Steady growth trend is statistically significant	0.002	0.730	High R^2 indicating 73% of the variation in organic farming shares is explained by the model
Bulgaria	Positive but statistically insignificant growth	0.115	0.281	Low R^2 suggests a weak fit, with only 28% of the variation explained
Romania	The rapid growth trend is statistically significant	0.008	0.608	Model explains 60.8% of the variation in organic farming shares

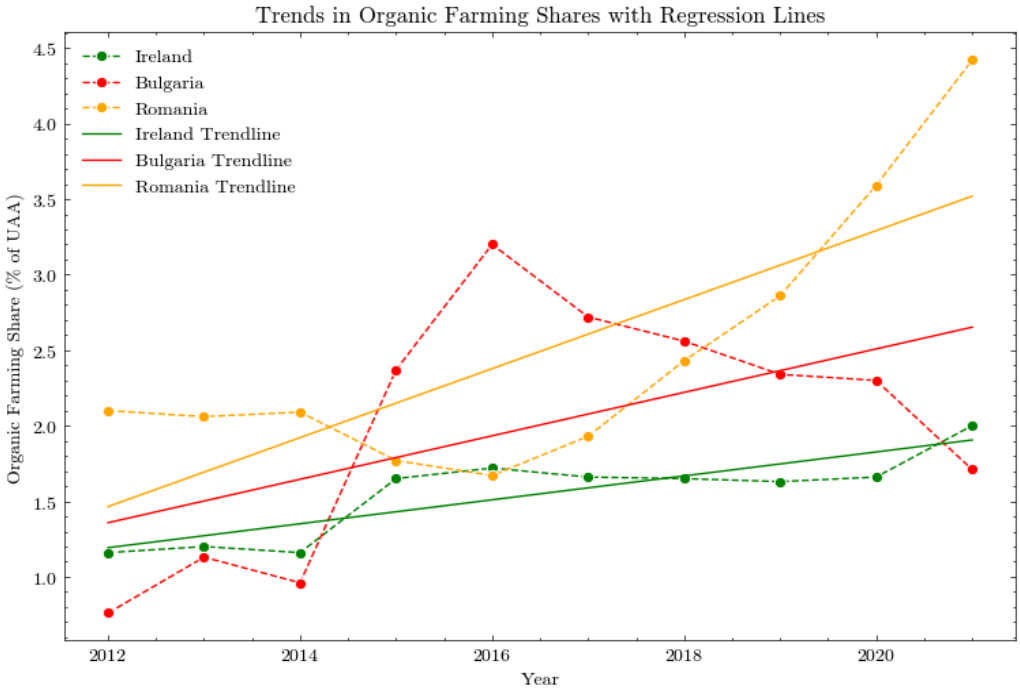


Figure 26: Trends in Organic Farming Shares with Regression Lines (2012–2021)

5.3 Sentiment Analysis

Reddit data was scraped using Python and PRAW, targeting country-specific subreddits (e.g., r/Austria, r/Bulgaria) from 2012 to 2022, focusing on the keyword "organic." Metadata like titles, bodies, scores, and comments were extracted and stored in a CSV. Due to time constraints, scraping was limited to 10 posts per keyword and 50 comments per post, yielding 31,416 entries. Refer to CA2_notebook-sentiment.ipynb for code regarding this section.

5.3.1 Feature Extraction

Table 12 outlines the pre-processing steps applied to clean the Reddit dataset. Lemmatization, preferred over stemming, preserves grammatical accuracy and contextual meaning by reducing words to their base form while considering part of speech, ensuring precise text representation for analysis (Bird et al., 2009).

Table 12: Sentiment Analysis pre-processing steps

Step	Description	Rationale
URL Removal	Identified and removed URLs (e.g., http://example.com) using regular expressions.	URLs do not contribute meaningful context to text analysis and could introduce noise
Special Characters & Numbers	Removed special characters, punctuation, and numeric values	Simplifies text and focuses on meaningful linguistic content, reducing noise
Lowercase Conversion	Converted all text to lowercase	Prevents issues caused by case-sensitive duplication (e.g., "Organic" vs. "organic")
Stopword Removal	Removed common stopwords using a combined list from Gensim and NLTK for broader coverage.	Eliminates non-informative words to improve the focus on meaningful terms during analysis

Step	Description	Rationale
Irrelevant Content Exclusion	Excluded entries marked as "delete" or "remove"	Avoids processing incomplete or irrelevant data that does not contribute to insights
Tokenization	Split text into individual words using NLTK's word_tokenize.	Breaks down text into manageable components for analysis and further processing
Lemmatization	Reduced words to their base forms using NLTK's WordNetLemmatizer (e.g., "running" → "run").	Groups words with similar meanings, reducing dimensionality and improving analysis accuracy
POS Tagging for Lemmatization	Performed POS tagging using NLTK's pos_tag	Ensures lemmatization considers grammatical roles, improving the reliability of results
Duplicate Removal	Dropped duplicate rows	Reduces bias and ensures each data point contributes uniquely to the analysis

The processed text was reviewed to ensure it was free of noise and irrelevant entries. Table 13 shows examples of raw text transformed into cleaned text. The Word Cloud in Figure 27 illustrates the effectiveness of pre-processing, showcasing the most frequent meaningful terms post-cleaning.

Table 13: Example of Raw and Cleaned Text

Country	Year	source	Text	Cleaned Text
Austria	2012	title	Animal rights organization/activists in Austria?	animal right organization activist austria
Austria	2015	title	Unable to secure a job for almost two years.	unable secure job year
Austria	2015	title	Drei charged us over 90 Euros incorrectly, and say that "maybe we will return" that money but "as a coupon".	drei charge euros incorrectly maybe return money coupon
Austria	2015	title	Hi Austria! Trying to identify some pics….	hi austria try identify pic
Austria	2015	title	Death threat via email is not considered a death threat - (digital Stone Age in Austrian justice)	death threat email consider death threat digital stone age austrian justice

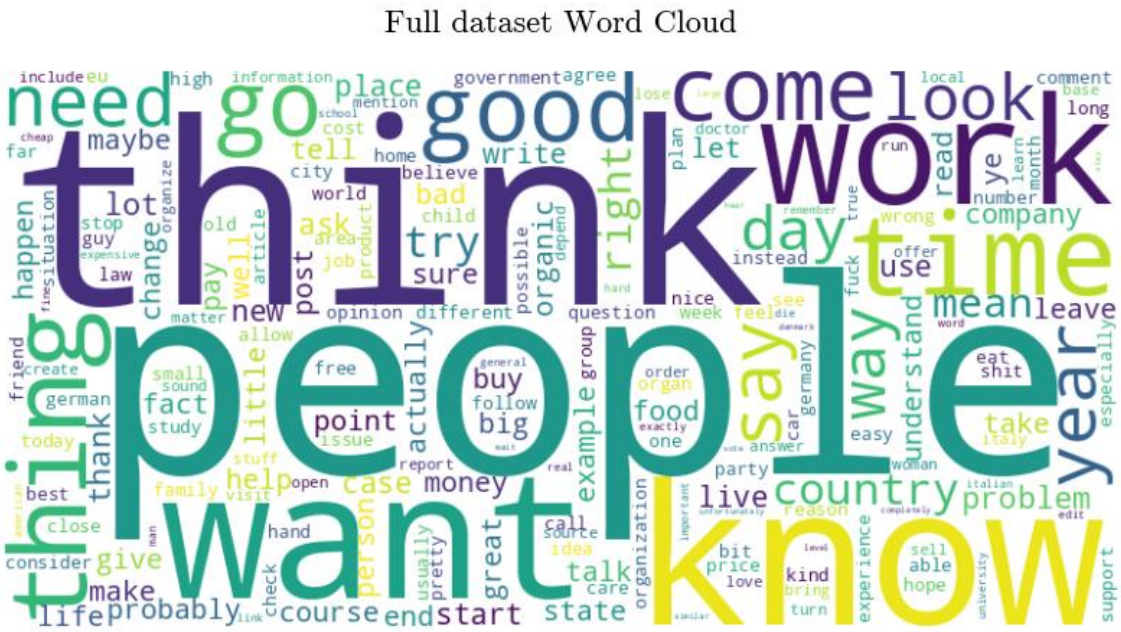


Figure 27: Word Cloud representing the most common words in the cleaned Reddit dataset

5.3.2 VADER

The sentiment labels for the model were derived using VADER Sentiment, a lexicon-based tool designed for analysing social media text, which classifies text as positive, negative, or neutral (Hutto and Gilbert, 2014). Figure 28 shows that positive sentiment dominates, while neutral and negative sentiments are less frequent but relatively balanced.

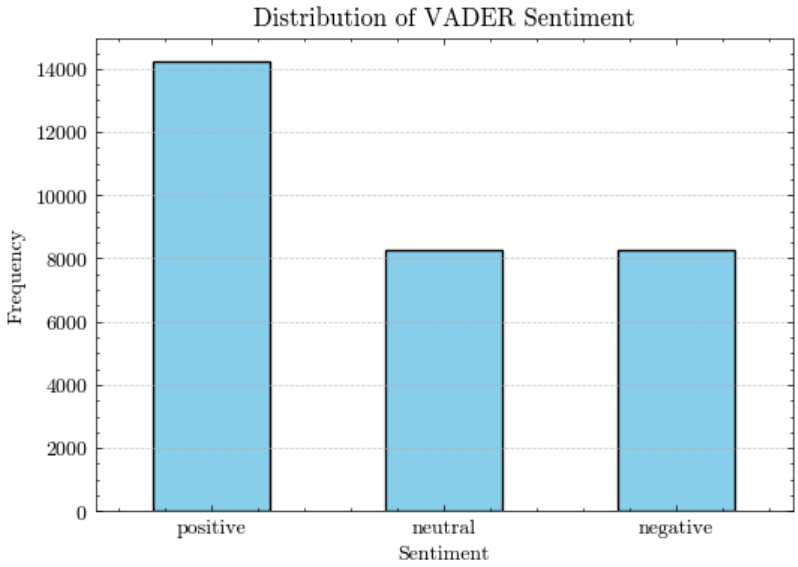


Figure 28: Distribution of VADER Sentiment Scores in the Reddit dataset

5.3.3 ML Model

Machine learning models improve upon VADER by adapting to domain-specific language, handling complex sentences, and providing nuanced sentiment analysis (Medhat et al., 2014). Multinomial Naive Bayes (NB) and Logistic Regression (LR) were used for sentiment classification. NB was chosen for simplicity and efficiency, while LR was selected for its ability to handle complex relationships with regularization (Murphy, 2012).

5.3.3.1 Multinomial Naive Bayes (NB)

NB achieved 66.6% test accuracy (Table 14) with a 66.4% CV mean score and low standard deviation, indicating stable performance ((Table 15).

Table 14: Classification Metrics for NB model

Metric	Train Set	Test Set
Accuracy	0.765	0.666
Precision (Weighted)	0.799	0.698
Recall (Weighted)	0.765	0.666
F1 Score (Weighted)	0.755	0.641

Table 15: Cross-Validation Metrics for NB Model

Metric	Value
Score Metric	accuracy
CV Scores Mean	0.664 (+/- 0.014)
CV Scores Std	0.007
CV Folds Score	0.676, 0.667, 0.661, 0.656, 0.662

Table 16 shows the model excelled in identifying positive sentiment but struggled with n eutral sentiment, with a recall of only 30%. The confusion matrix in Table 17 highlights frequent misclassification of neutral posts as positive, lowering performance for neutral sentiment prediction.

Table 16: Classification Report for NB Model (Test Set)

Class/Metric	precision	recall	f1-score	support
Negative	0.784677	0.589697	0.673356	1650
Neutral	0.752294	0.30559	0.434629	1610
Positive	0.61831	0.910159	0.736371	2894
accuracy	0.666071	0.666071	0.666071	0.666071
macro avg	0.718427	0.601815	0.614786	6154
weighted avg	0.697969	0.666071	0.640534	6154

Table 17: Confusion Matrix for NB Model (Test Set)

	Predicted Negative	Predicted Neutral	Predicted Positive
Actual Negative	973	77	600
Actual Neutral	92	492	1026
Actual Positive	175	85	2634

5.3.3.2 Logistic Regression (LR)

The LR model outperformed NB with 87.5% test accuracy (Table 18) versus NB's 66.6%. Cross-validation showed stable performance, with a mean accuracy of 87.4% (Table 19).

Table 18: Classification Metrics for LR model

Metric	Train Set	Test Set
Accuracy	0.987	0.875
Precision (Weighted)	0.987	0.877
Recall (Weighted)	0.987	0.875
F1 Score (Weighted)	0.987	0.875

Table 19: Cross-Validation Metrics for LR Model

Metric	Value
Score Metric	
CV Scores Mean	0.874 (+/- 0.011)
CV Scores Std	0.005
CV Folds Score	0.876, 0.881, 0.866, 0.871, 0.877

Table 20 shows higher precision, recall, and F1-scores for all sentiment classes. The confusion matrix in Table 21 highlights LR's accuracy in predicting negative, neutral, and positive sentiments.

Table 20: Classification Report for LR Model (Test Set)

Class/Metric	precision	recall	f1-score	support
Negative	0.872215	0.806667	0.838161	1650
Neutral	0.820571	0.891925	0.854762	1610
Positive	0.910354	0.905321	0.907831	2894
accuracy	0.875366	0.875366	0.875366	0.875366
macro avg	0.867714	0.867971	0.866918	6154
weighted avg	0.87664	0.875366	0.875267	6154

Table 21: Confusion Matrix for NB Model (Test Set)

	Predicted Negative	Predicted Neutral	Predicted Positive
Actual Negative	1331	164	155
Actual Neutral	71	1436	103
Actual Positive	124	150	2620

5.3.3.3 Model evaluation

Both models were fine-tuned using grid search with cross-validation:

- NB performed best with a smoothing parameter (α) of 0.01 and TfidfVectorizer with 2000 features.
- LR showed optimal performance with C = 10 and TfidfVectorizer set to 10,000 features.

Confusion matrices (Figure 29) show how models distinguished between Negative, Neutral, and Positive sentiments. LR outperformed NB, reducing misclassifications for neutral and negative sentiments. The comparison chart (Figure 30) highlights LR's superior metrics (88% for accuracy, precision, recall, and F1-score) versus NB's best precision of 70%.

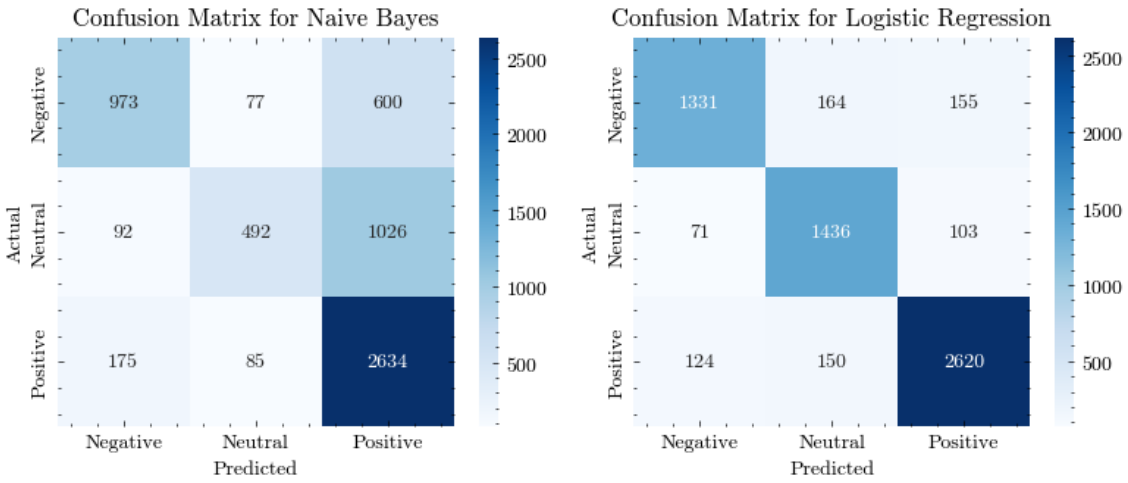


Figure 29: Confusion Matrices for Naive Bayes and Logistic Regression Models

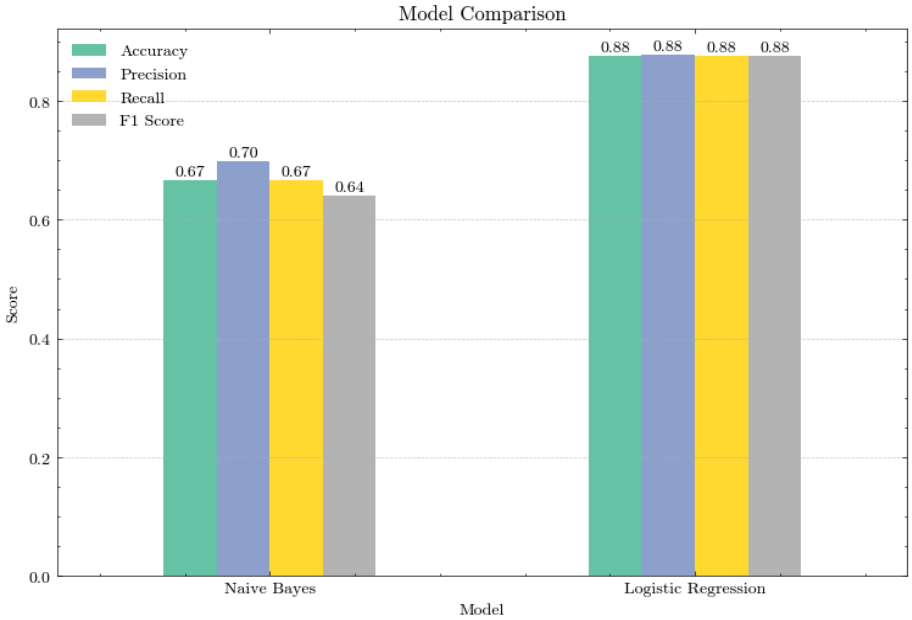


Figure 30: Model Metrics Comparison

Figure 31 and Figure 32 present the ROC and Precision-Recall curves used to evaluate models across decision thresholds. Logistic Regression (LR) outperforms Naive Bayes (NB), with ROC curves farther from the random-guess diagonal, indicating better class separation (Fawcett, 2006). Conversely, NB exhibits weaker discrimination. The Precision-Recall curves emphasize performance on imbalanced classes, demonstrating LR's superior handling, particularly for neutral sentiment, with higher precision-recall trade-offs (Saito and Rehmsmeier, 2015).

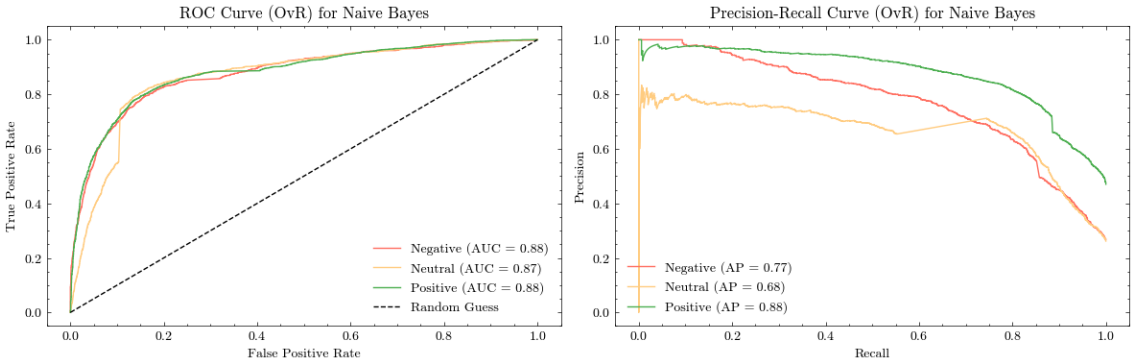


Figure 31: ROC and Precision-Recall Curves for Naive Bayes Model

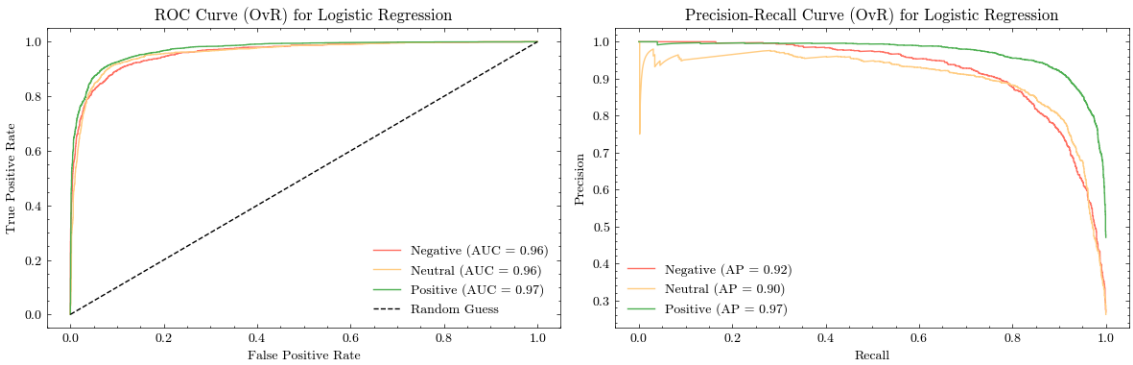


Figure 32: ROC and Precision-Recall Curves for Logistic Regression Model

5.3.4 Topic Modeling

Topic Modeling filtered posts and comments related to "organic," ensuring relevant content for analysis. Latent Dirichlet Allocation (LDA) was chosen for its ability to group terms by co-occurrence patterns. A Bag of Words approach created a document-term matrix, preferred over TF-IDF as LDA relies on raw term counts to capture co-occurrence, while TF-IDF may distort key relationships by downweighing frequent terms (Blei et al., 2003). Topic quality was assessed using coherence scores, a widely used metric for evaluating the interpretability of topic models (Röder, Both and Hinneburg, 2015). Models with 3, 4, and 5 topics were evaluated to determine the optimal number of topics. (Figure 33).

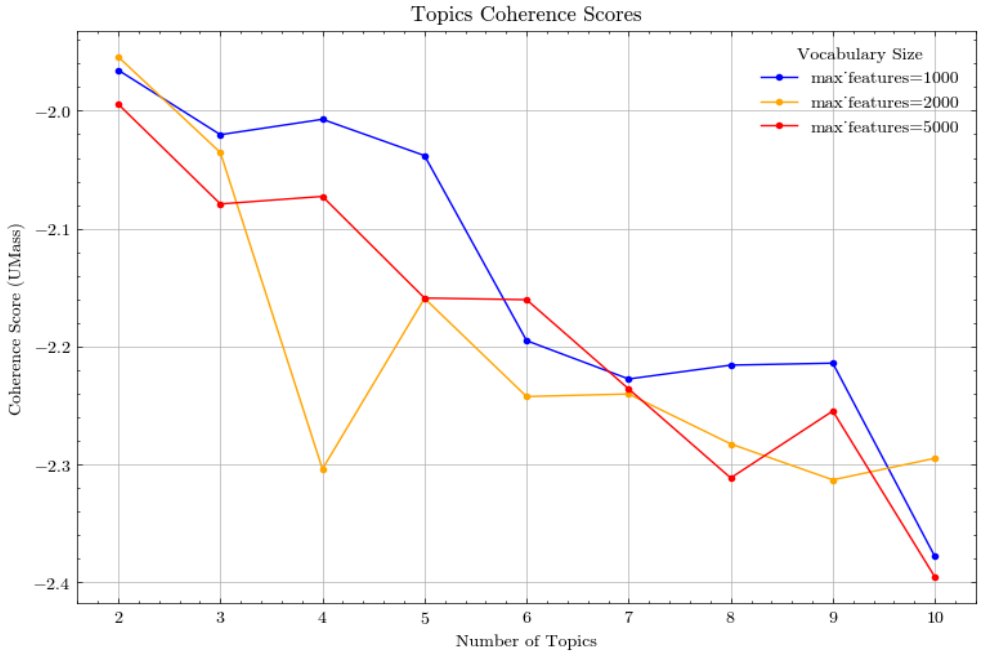


Figure 33: Coherence Scores for Different Numbers of Topics and Vocabulary Sizes

Table 22 shows the 5-topic model was chosen for effectively separating organic farming terms and minimizing unrelated content, assigning 3,382 posts and comments to the organic topic.

Table 22: Topic Coherence for Different Numbers of Topics

Number of Topics	Observations
3 topics	Significant dilution observed; organic-related terms mixed with unrelated content, such as politics.
4 topics	Improved topic isolation but retained some overlap among topics.
5 topics	Clearest separation, with terms like “organic,” “food,” and “meat” grouped into a distinct topic.

Figure 33 illustrates that models with 1000 terms achieve higher coherence scores; however, Table 23 demonstrates that the 5000-term model better captures the "organic farming" theme. The 1000-term model includes more generic terms such as "example" and "thank," which dilute topic specificity and relevance.

Table 23: Topic Words for Different Vocabulary Sizes

Vocabulary Size	Topic Words (Topic 5)
max_features=5000	organic, food, product, meat, buy, animal, think, good, eat, produce, use, price, farm, people, market, farmer, thing, plant, production, test
max_features=1000	organic, good, product, food, people, think, animal, buy, meat, mean, use, produce, thank, example, problem, high, thing, know, market, farm

The Word Cloud in Figure 34 highlights frequent terms in the "organic" topic, like "organic," "food," "product," and "meat," reflecting key discussion themes. Figure 35 shows positive sentiment dominates, consistent with the dataset's class balance.

Figure 34: Word Cloud for Organic Topic

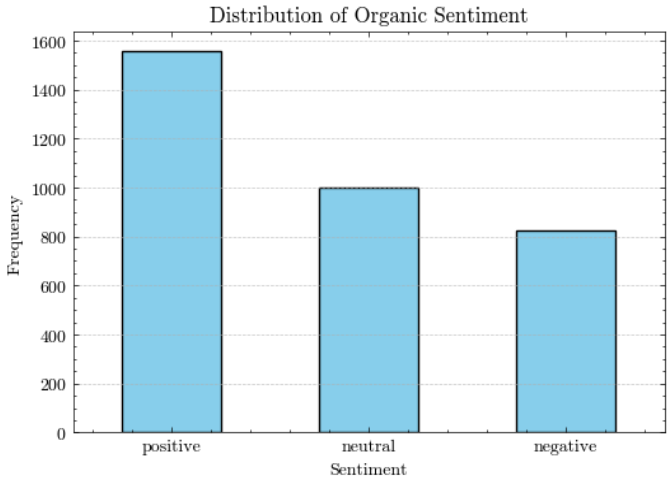


Figure 35: Distribution of Sentiment for Organic Topic

5.3.5 Findings

Positive sentiment has grown significantly since 2015, with sharp increases from 2019 to 2022, driven by policy initiatives and public awareness (Figure 36). Geographically, Austria and Denmark show strong positive sentiment, while smaller countries like Czechia and Estonia have minimal activity (Figure 37). Sentiment ratios in Figure 38 reveal Estonia leading in negative sentiment, followed by the Netherlands and Ireland.

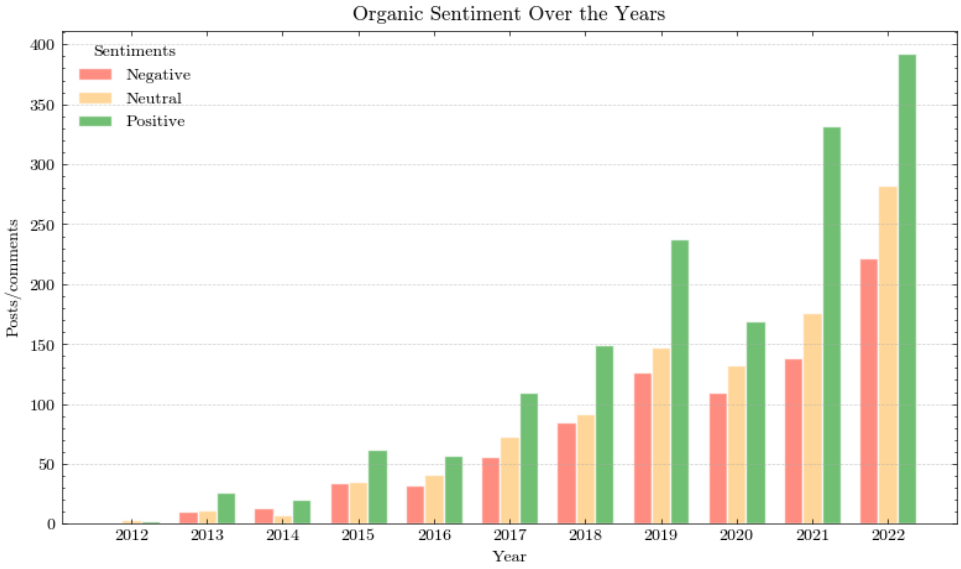


Figure 36: Organic Sentiment Trends Over the Years

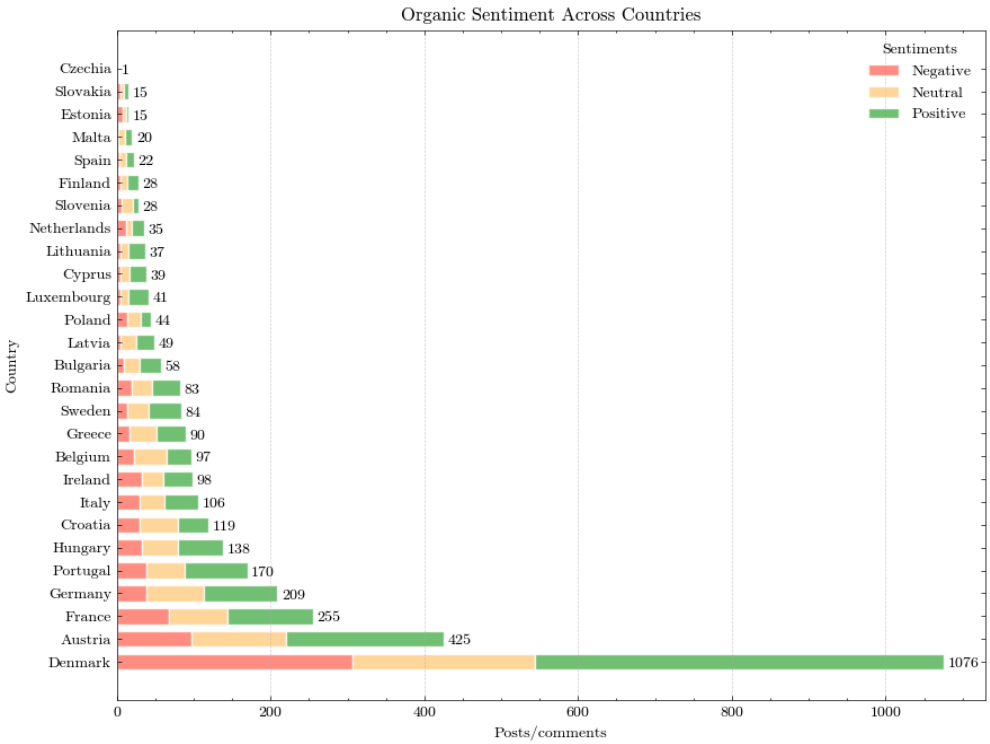


Figure 37: Organic Sentiment Across Countries

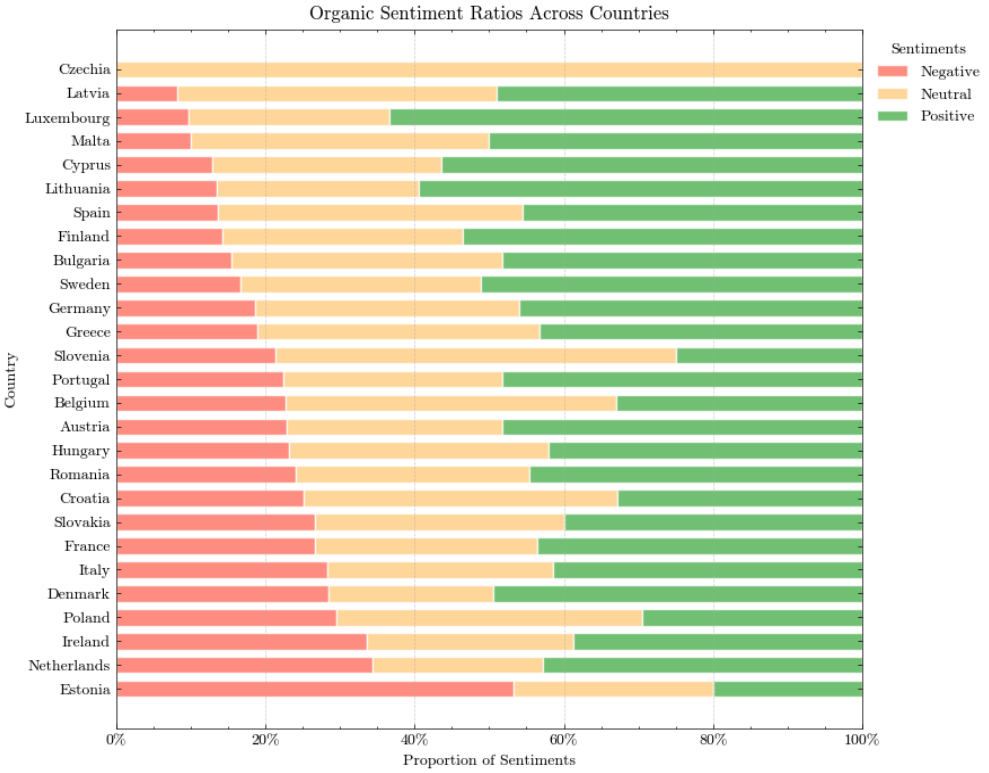


Figure 38: Organic Sentiment Ratio Across Countries

5.4 Clustering

The dataset, pre-processed with pivoted indicators, sentiment scores, and sentiment ratios, consists of 296 rows and nine columns for clustering analysis.

5.4.1 Imputation

Missing values disrupt clustering algorithms such as K-Means and Hierarchical Clustering, as these methods rely on complete data to compute distances and assign clusters (García-Laencina et al., 2010). To address this, all missing values (Figure 39) were resolved using appropriate imputation techniques based on the nature of the data and the extent of the gaps. For instance, linear interpolation was employed for time-series data with trends, ensuring continuity without introducing bias (Yang and Chiang, 2020). The Netherlands was excluded due to missing CAP funding data. Table 24 summarizes the strategies and rationale.

Table 24: Missing Data Strategies and Rationale

Column	Column Meaning	Missing Data Context	Strategy	Rationale for Strategy
tot_area	Total agricultural area under organic farming	Gaps in data collection over time	Linear interpolation within each country	Preserves temporal continuity by estimating missing values based on neighboring years
share_area	Share of agricultural area under organic farming	Gaps in data collection over time	Linear interpolation within each country	Maintains trends over time while estimating values accurately
tot_cap	Total public expenditure under CAP measures	No payments made in those years/regions	Set to 0 where missing	Reflects the absence of CAP payments for organic farming
eu_cap	EU-funded public expenditure only	No payments made in those years/regions	Set to 0 where missing	Ensures consistency, reflecting no EU funding was allocated.
area_cap	Physical area supported under CAP measures	No payments made in those years/regions	Set to 0 where missing	Matches the absence of total and EU CAP funding
tot_sentiment	Total sentiment score	No sentiment data collected	Set to 0 where missing	Signals that no sentiment data was gathered for those rows
positive_ratio	Positive sentiment ratio	No sentiment data collected; setting to 0 could imply negative sentiment	Replace missing/zero values with 0.5 (neutral)	Avoids bias by treating missing values as neutral rather than implying negativity

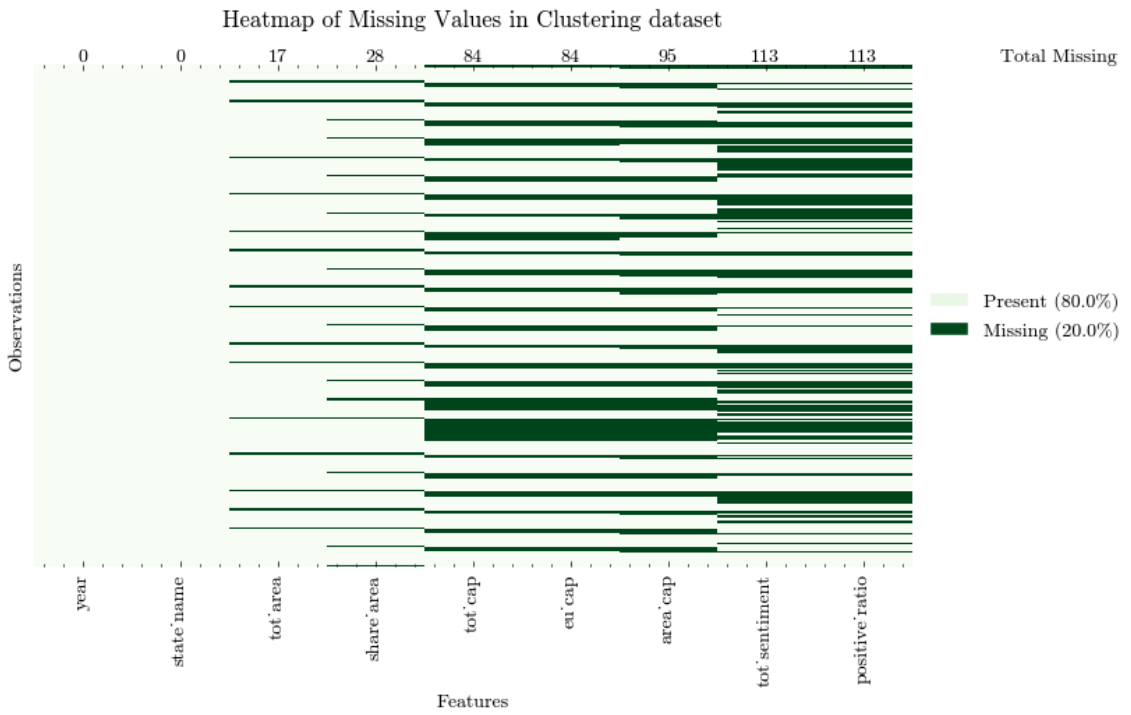


Figure 39: Heatmap of Missing Values in the Clustering Dataset

5.4.2 Correlation analysis

Spearman correlation is used because it measures monotonic relationships and is well-suited for data with skewed distributions (Mukaka, 2012), as visualized in the pair plot (Figure 40) and summarized in the heatmap (Figure 41). Correlation findings are detailed in Table 25.

Table 25: Summary of Correlation Findings

Feature	Correlation Insights
tot_cap and eu_cap	Strongest correlation, confirming that EU contributions dominate CAP funding
tot_area	Moderately to strongly correlated with funding variables (tot_cap, eu_cap, area_cap), indicating the link between financial support and organic farming area
tot_sentiment	Moderately correlated with tot_cap (0.54) and tot_area (0.34), suggesting that discussion volume reflects organic farming activity and funding
positive_ratio	Weak correlation with other metrics, indicating that sentiment positivity has little impact on the analyzed metrics

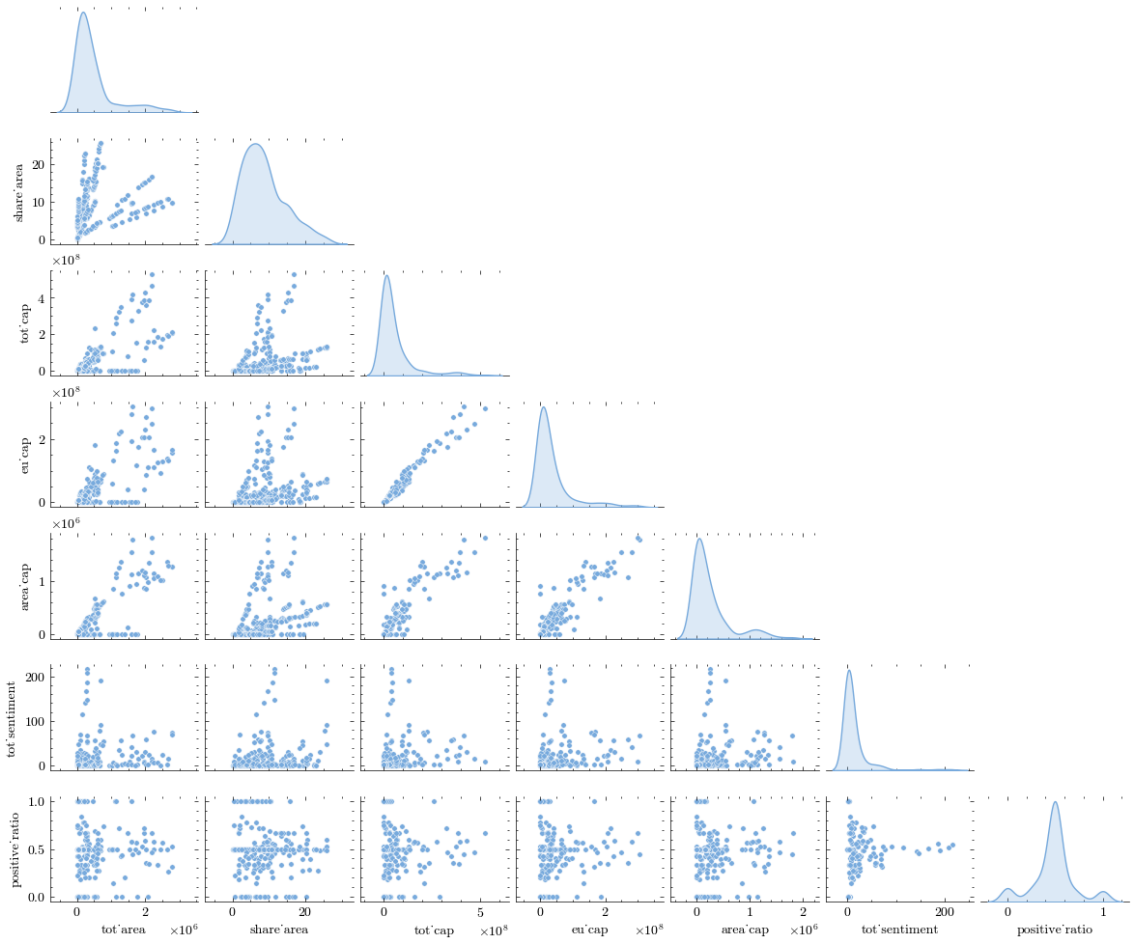


Figure 40: Pair Plot of Features

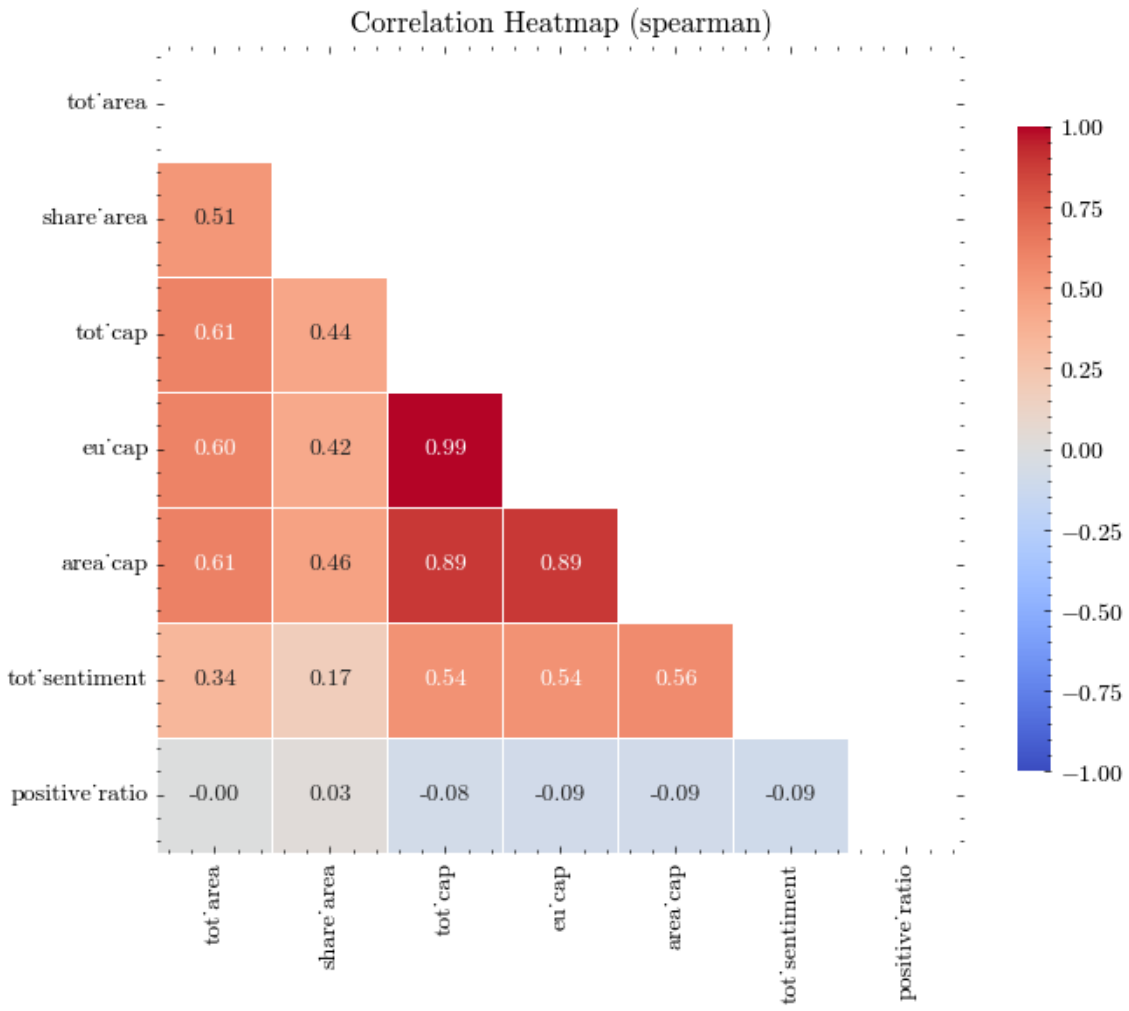


Figure 41: Spearman Correlation Heatmap

5.4.3 Feature selection

Feature selection, as detailed in Table 26, was performed to enhance the quality and interpretability of clustering by addressing redundancy and noise, following best practices in correlation analysis to remove uninformative or redundant features (Guyon and Elisseeff, 2003).

Table 26: Feature Selection Based on Correlation Analysis

Feature	Correlation/Reason	Action Taken	Rationale
tot_cap and eu_cap	Highly correlated (correlation = 0.99)	eu_cap removed	Avoids redundancy by keeping only one representative of strongly correlated funding metrics
area_cap	Strongly correlated with tot_cap and eu_cap	Removed	Prevents overemphasis on funding metrics in clustering
positive_ratio	Weakly correlated with all metrics	Removed	Contributes little to clustering and does not add meaningful differentiation

5.4.4 Aggregation

Data was aggregated by country using metrics like average organic farming area, CAP funding, sentiment scores, and percentage growth to capture trends while minimizing yearly fluctuations and over-representation (Hox et al., 2017). Boxplots (Figure 42, Figure 43, Figure 44) guided the strategy for handling skewed distributions and outliers, as detailed in Table 27.

Table 27: Aggregation Methods Summary

Feature	Reason for Aggregation Choice	Aggregation Method
Total Organic Area	Skewed data with outliers; median minimizes distortion.	Median
Share of Organic Farming	Percentages can be extreme; median ensures robustness.	Median
CAP Funding	High variability and outliers; median provides stability.	Median
Average Sentiment Score	Mean reflects the typical sentiment trend over time.	Mean
Annual Growth Percentage	Mean captures the average growth pattern effectively.	Mean

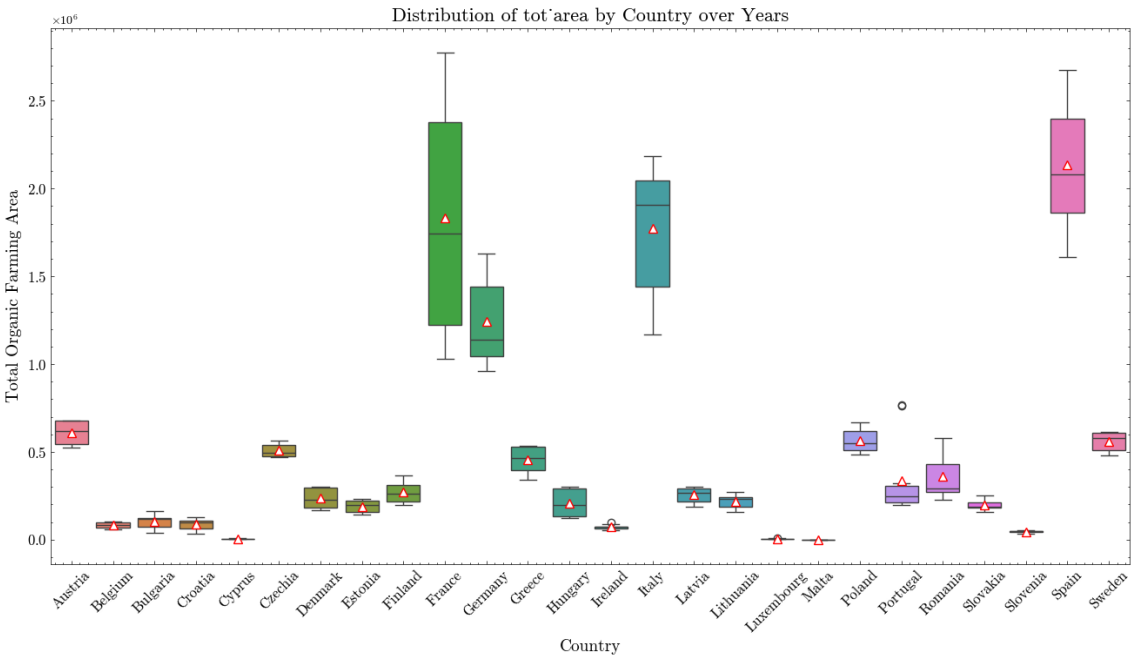


Figure 42: Distribution of Total Organic Farming Area by Country Over Years

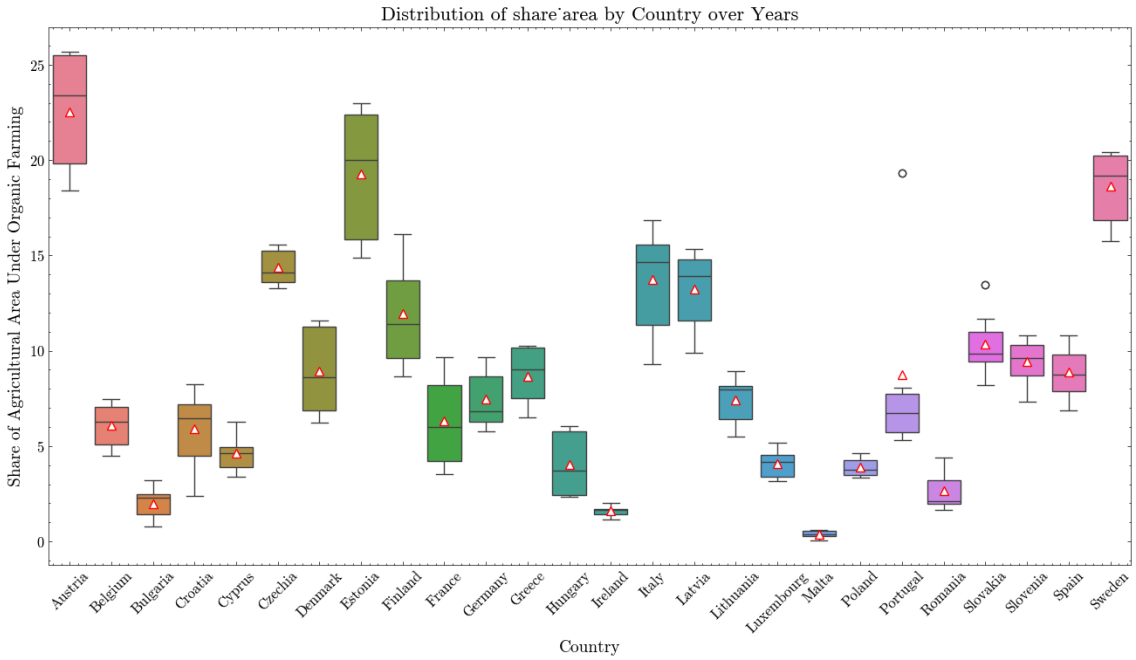


Figure 43: Distribution of Share of Agricultural Area Under Organic Farming by Country Over Years

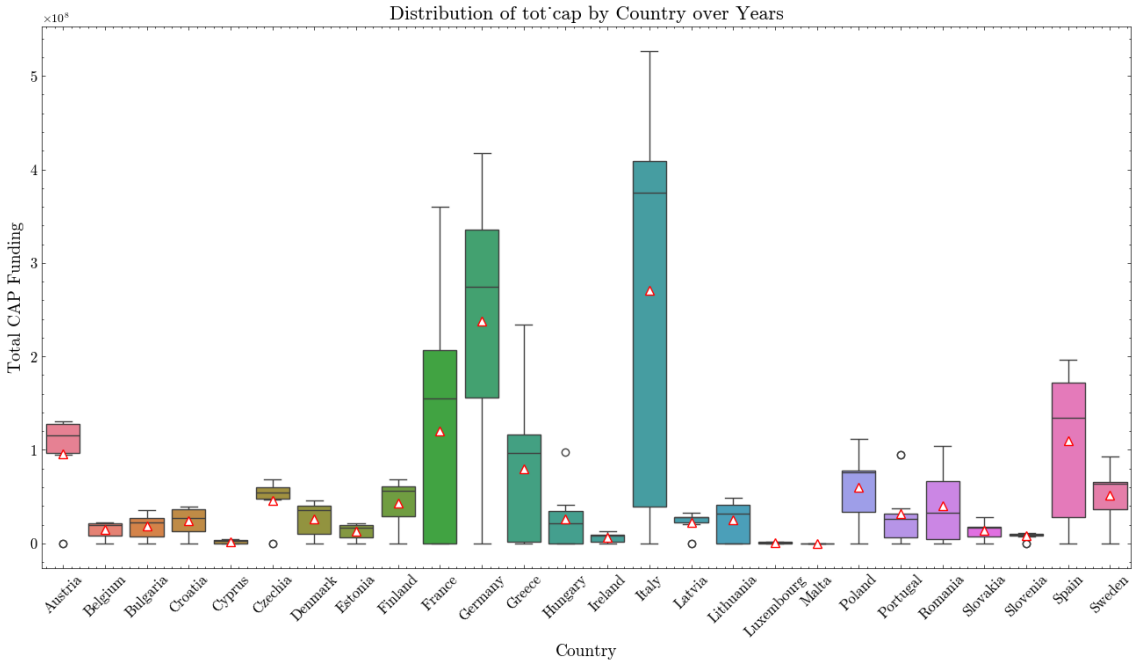


Figure 44: Distribution of Total CAP Funding by Country Over Years

Denmark's high total sentiment value, identified in EDA (5.3.5,) and Figure 43, was addressed with Winsorization, capping values at the 95th percentile (39.84). This method was chosen to preserve natural variability while reducing outlier influence.

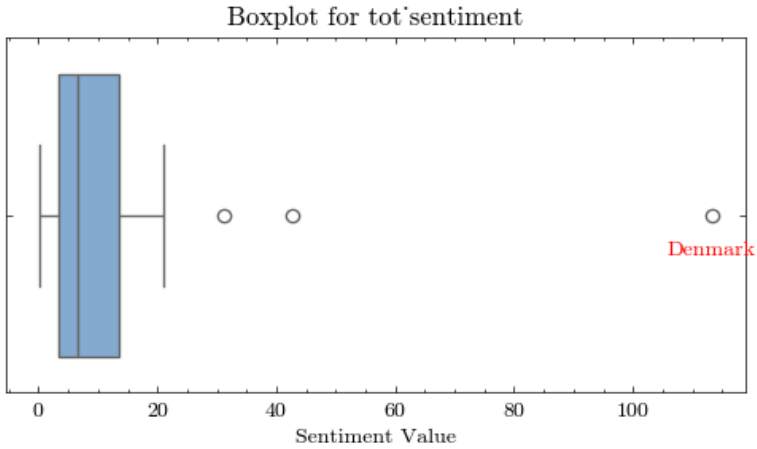


Figure 45: Boxplot with outliers in total sentiment

5.4.5 Feature scaling

Scaling ensures equal feature contribution to distance calculations in clustering by preventing domination from features with larger ranges (Han et al., 2011). Robust Scaling, using the median and IQR, was applied to handle skewed distributions and outliers (Kim et al., 2023).

5.4.6 Hierarchical clustering

Hierarchical clustering was chosen for its flexibility and visual interpretability, without needing a predefined cluster number (Murtagh and Contreras, 2012). Dendrograms (Figure 46) guided cluster selection. Ward linkage, tested alongside other methods (Table 28), was selected for minimizing intra-cluster variance and creating compact clusters.

Table 28: Description of Linkage Methods in Hierarchical Clustering

Linkage Method	**Description**
Single linkage	Minimizes the distance between the closest points in different clusters.
Complete linkage	Considers the furthest points between clusters.
Average linkage	Uses the average distance between all pairs of points across clusters.
Centroid linkage	Measures distances between cluster centroids.
Ward linkage	Minimizes the variance within clusters.

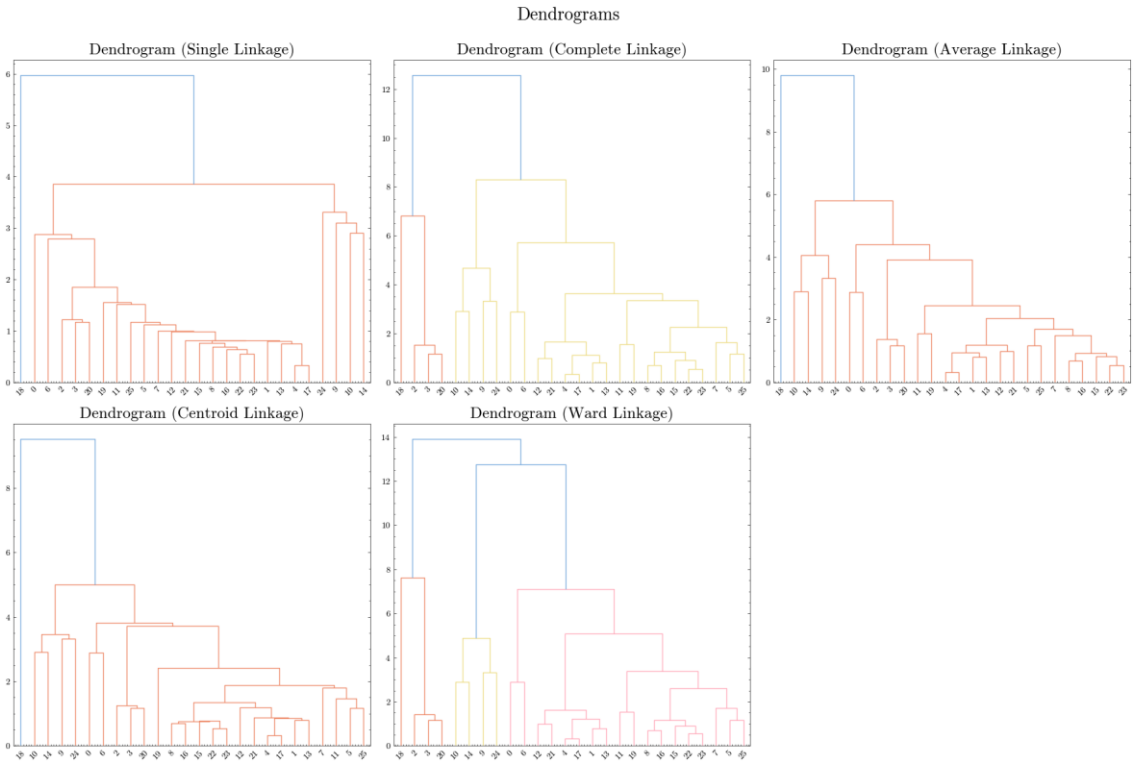


Figure 46: Dendrograms for Different Linkage Methods in Hierarchical Clustering

The Ward linkage dendrogram (Figure 47) set the cut-off distance at 70% of the maximum distance, based on a significant vertical gap. Three clusters were chosen, reflecting data patterns and outlined in Table 29.

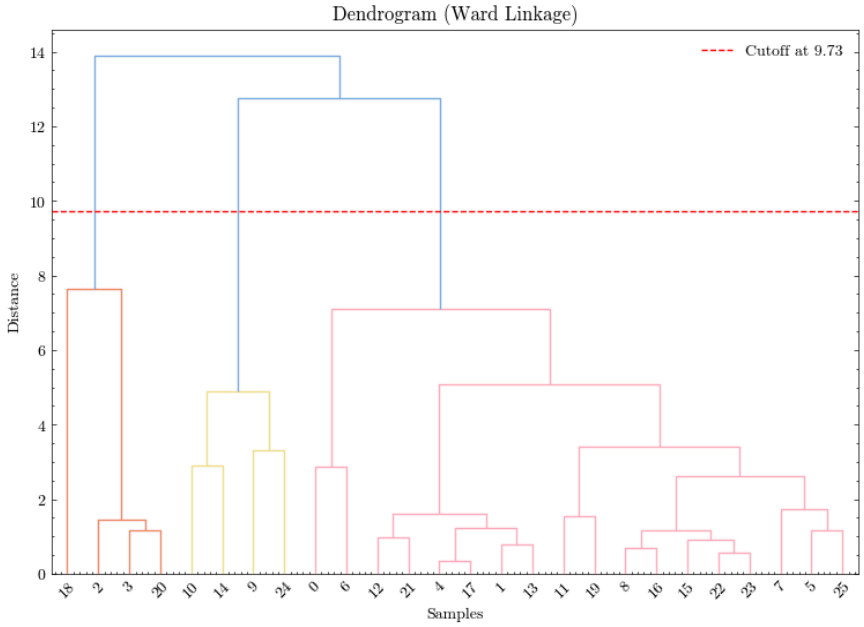


Figure 47: Dendrogram Using Ward Linkage

Table 29: Country Clusters Derived from Ward Linkage

Cluster	Country
Cluster 1	Bulgaria, Croatia, Malta, Portugal
Cluster 2	France, Germany, Italy, Spain
Cluster 3	Austria, Belgium, Cyprus, Czechia, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Latvia, Lithuania, Luxembourg, Poland, Romania, Slovakia, Slovenia, Sweden

5.4.7 Principal Component Analysis (PCA)

PCA was employed to identify the key features influencing clustering and to reduce dimensionality for enhanced interpretability (Jolliffe and Cadima, 2016). Figure 48 shows the first two components explain over 80% of the variance, increasing to 90% with the inclusion of a third component, thereby supporting the exclusion of the remaining components

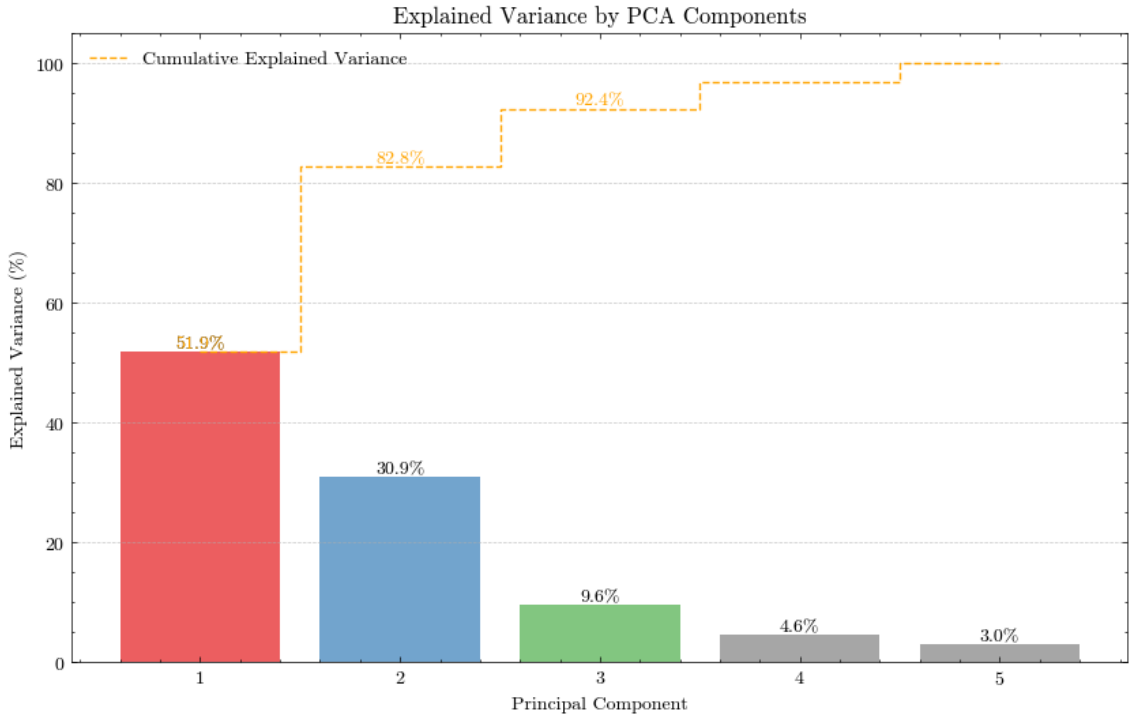


Figure 48: Explained Variance by PCA Components

PCA loadings reveal the contribution of each feature to the principal components. The heatmap in Figure 49 highlights annual growth and CAP funding as key drivers of regional distinctions, with public sentiment playing a smaller role.

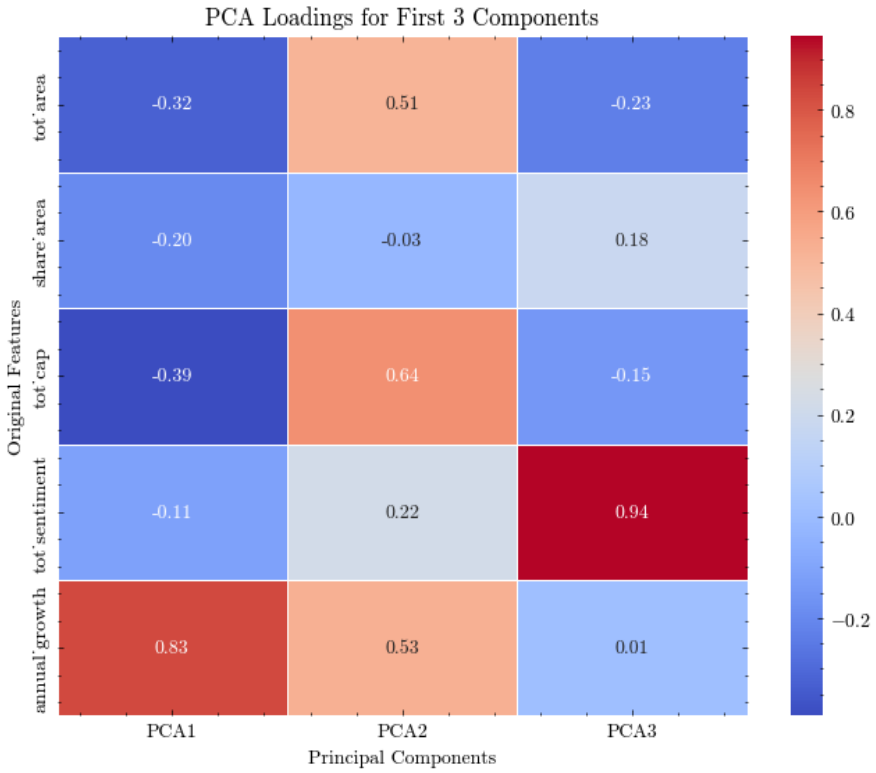


Figure 49: PCA Loadings Heatmap

Cluster profiling highlights the defining characteristics of each group, using key features like annual growth, CAP funding, agricultural area, and sentiment to uncover patterns in growth, resource allocation, and public perception. Table 30 summarizes the findings, while Figure 50 visualizes feature distributions across clusters.

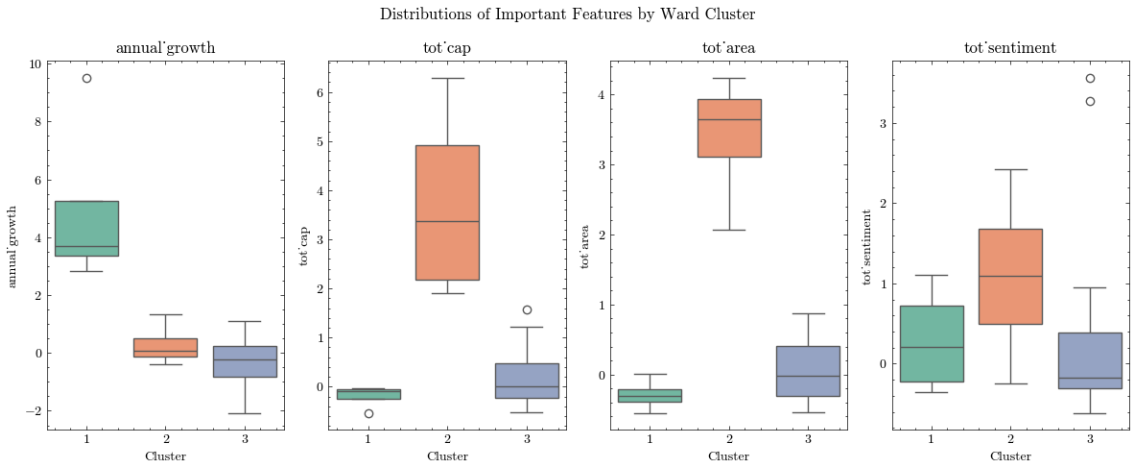


Figure 50: Clusters profiling

Table 30: Cluster Characteristics Summary

Cluster	Summary
Cluster 1	High growth rates, small organic areas, and low CAP funding. Represents emerging or fast-growing regions with minimal resources.
Cluster 2	Largest organic areas, highest CAP funding, and strong public sentiment. Likely established agricultural regions with strong financial and public support.
Cluster 3	Low growth rates, moderate organic areas, and mid-level CAP funding. Represents stable but slower-growing regions with limited public sentiment.

Clusters were visualized in 3D space using the first three PCA components, offering a clear representation of how regions are grouped based on growth, funding, and sentiment (Figure 49).

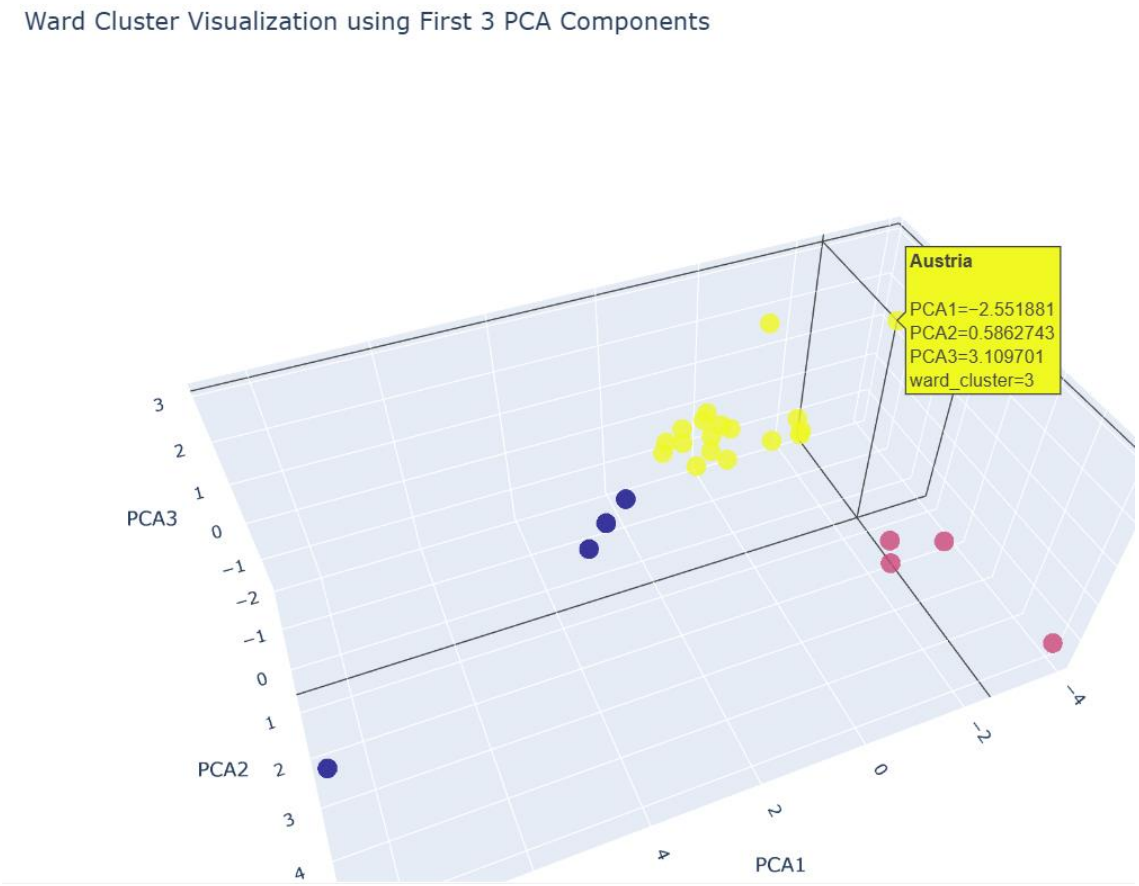


Figure 51: Interactive 3D visualization of Ward clustering using the first three PCA components

5.4.8 K-Means

K-Means clustering was applied to validate Ward's hierarchical clustering results. Both methods minimize intra-cluster variance, with K-Means using iterative centroid updates (Blomer et al., 2016). The optimal three clusters were identified using the Elbow Method and validated with Silhouette Scores, as illustrated in Figure 52 (Rousseeuw, 1987).

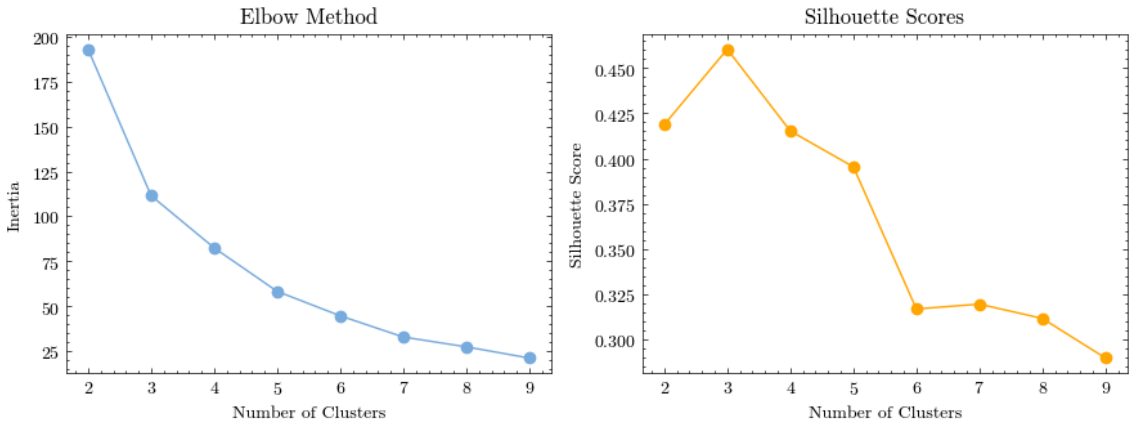


Figure 52: Optimal Number of Clusters Determined by the Elbow Method and Silhouette Scores

K-Means results closely matched Ward's method, except for Austria, which was reassigned from Cluster 3 to Cluster 1. The PCA projection (see figure) shows Austria near the cluster boundary, reflecting K-Means' sensitivity to centroids compared to Ward's variance-focused approach.

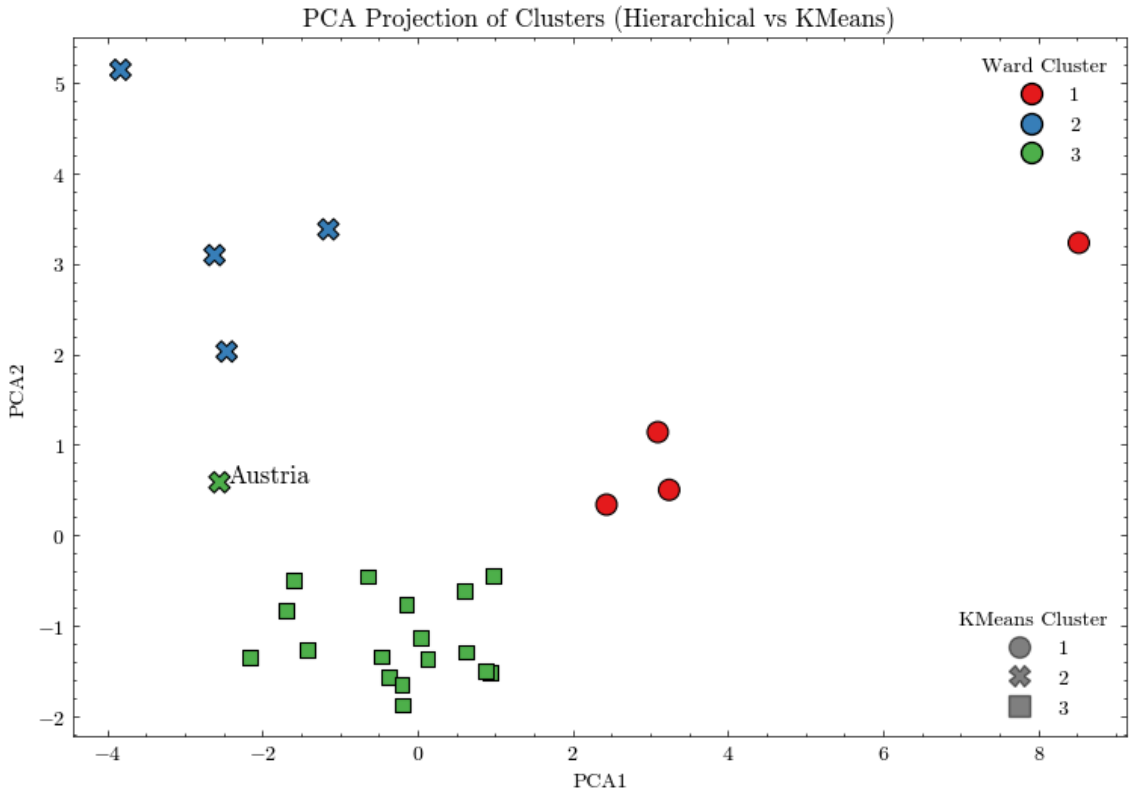


Figure 53: PCA Projection of Clusters using the first two PCA components (Hierarchical vs K-Means)

6. Conclusions

This report explored organic farming across Europe with Ireland as a baseline, examining the roles of CAP funding, public sentiment, and agricultural indicators in shaping adoption trends.

Ireland's organic farming sector shows steady but modest growth, with low CAP funding per hectare likely limiting progress compared to leaders like Austria and Sweden. Countries such as Austria and Portugal, with targeted CAP funding and robust policies, achieve higher adoption rates. Ireland's growth aligns more with stable regions showing moderate progress, highlighting the need for increased financial support and policy innovation. While public sentiment toward organic farming is growing, particularly in Ireland, its direct impact on adoption appears secondary to policy and financial incentives.

Future research should refine predictive models by incorporating factors like market demand and regional practices to improve forecasts of organic farming trends. Expanding sentiment analysis across platforms and timeframes can deepen insights into public opinion's role. Policy studies on successful countries like Austria and Bulgaria and longitudinal analysis of CAP reforms post-2022 can guide Ireland's strategies.

7. References

Bird, S., Klein, E., and Loper, E., 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3, 993–1022.

Blomer, J., Lammersen, C., Schmidt, M., and Sohler, C. (2016) *Theoretical analysis of the k-means algorithm—A survey*, Algorithm Engineering, 9220, pp. 81–116. doi: 10.1007/978-3-319-49487-6_3.

European Commission (n.d.a) *AgriData Portal: CAP Indicators*. Available at: https://agridata.ec.europa.eu/extensions/DataPortal/cap_indicators.html (Accessed: 23 November 2024).

European Commission (n.d.b) *Organic Production Sources.* Available at: https://agridata.ec.europa.eu/Qlik_Downloads/Organic-Production-sources.htm (Accessed: 23 November 2024).

Fawcett, T., 2006. *An Introduction to ROC Analysis*. Pattern Recognition Letters, 27(8), pp. 861–874.

García-Laencina, P.J., Sancho-Gómez, J.L. and Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: A review. Neural Computing and Applications, 19(2), pp. 263–282.

Guyon, I. and Elisseeff, A., 2003. *An introduction to variable and feature selection*. Journal of Machine Learning Research, 3, pp. 1157–1182.

Han, J., Kamber, M. and Pei, J., 2011. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann.

Hox, J.J., Moerbeek, M. and van de Schoot, R., 2017. *Multilevel Analysis: Techniques and Applications*. 3rd ed. Routledge.

Hutto, C.J. and Gilbert, E., 2014. VADER: *A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14).

Jolliffe, I.T. and Cadima, J., 2016. *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202.

Kim, Jae-Dong & Hwang, Ji-Hwan & Doh, Hyoung-Ho. (2023). *A Predictive Model with Data Scaling Methodologies for Forecasting Spare Parts Demand in Military Logistics*. Defence Science Journal. 73. 666-674. 10.14429/dsj.73.19129.

Medhat, W., Hassan, A., and Korashy, H., 2014. *Sentiment analysis algorithms and applications: A survey*. Ain Shams Engineering Journal, 5(4), pp. 1093–1113.

Mukaka, Mavuto. (2012). *Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research.* Malawi medical journal : the journal of Medical Association of Malawi. 24. 69-71.

Murphy, K.P., 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Murtagh, F. and Contreras, P., 2012. *Algorithms for hierarchical clustering: An overview.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp. 86–97.

Newbold, P., Carlson, W.L., and Thorne, B.M., 2019. *Statistics for Business and Economics*. 9th ed. Pearson.

Reddit Inc. (n.d.) *Developer terms*. Available at: https://redditinc.com/policies/developer-terms (Accessed: 23 December 2024).

Röder, M., Both, A. and Hinneburg, A., 2015. *Exploring the Space of Topic Coherence Measures*. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15), pp. 399–408.

Rousseeuw, P.J., 1987. Silhouettes: *A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, pp. 53–65.

Saito, T. and Rehmsmeier, M., 2015. *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. PLOS ONE, 10(3), p.e0118432.

Tufte, E.R. (2007). *The visual display of quantitative information.* 2nd ed., 5th printing. Cheshire, CT: Graphics Press.

Yang, L. and Chiang, J.A., 2020. *Use case and performance analysis for missing data imputation methods in big data analytics*. Proceedings of the 2020 International Conference on Computing and Data Engineering (ICCDE '20), pp.107-111. DOI: 10.1145/3379247.3379270.

Wooldridge, J.M., 2013. *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning.

Appendix A. CMEF Dataset

The European Commission details the role of organic farming in the Common Agricultural Policy (CAP) as follows (European Commission, n.d.b):

"The Common Agricultural Policy (CAP) recognises and supports the role of organic farming in responding to consumer demand for more environmentally friendly farming practices. Therefore, the Rural Development policy provides CAP specific support for farmers' conversion to organic production and/or the maintenance of farmers producing organically. Furthermore, in the first pillar, organic farms benefit from the green direct payment without the need to fulfil any further obligations because of their overall significant contribution to environmental objectives.

CMEF Indicators are part of the Common Monitoring and Evaluation Framework (CMEF), which is specifically designed to evaluate the CAP. CMEF indicators are a subset of CAP indicators focused on specific objectives like program performance, results, and impacts.

Dataset Info:

- CATS is the Clearance of Accounts Audit Trail System, i.e. the database used for audit, based on information received from Member States.
- The Declarations of expenditure for European agricultural fund (DOE) are quarterly notifications of Member States expenditure.
- The data series on organic farming start in 2012, when Eurostat questionnaire was changed to better fit the legal requirements imposed by the Regulation (EC) 889/2008.
- In the CAP programming period 2014 - 2020, there is an "Organic farming" measure eligible for rural development funding. Member States can support farmers' conversion to organic production and/or the maintenance of farmers producing organically. The level of aid is generally higher in the conversion phase.

Please note in particular that:

- France has experienced delays in making payments for the organic measure in 2015 and 2016.
- The Netherlands does not support organic farming through the EAFRD but relies on national funds.
- For Malta, the support target for the entire programming period 2014–2020 is low (28 hectares), and the measure has not yet been launched.
- In Italy, support for organic farming was low in 2015 because most of the Regional Development Programs were approved in the second half of 2015, leaving insufficient time to launch the organic farming measure.
- In Sweden, the share of organic area granted support is low because EU funding is provided only for conversion, not for the maintenance of organic farming.
- Similarly, in Romania, Bulgaria, and Spain, the share of organic area granted support is around 50%, as some regions only finance conversion."

The following indicators have been used in this analysis:

Type	Name	Parameter	Code	Unit	Source
Output Pillar II	OIR_06 Physical area supported (ha)	Physical area supported for M11	OIR_06_1.2	Supported hectare	CATS
Output Pillar II	OIR_01 Total public expenditure	Total: EU funds + national co-financing	OIR_01a_2.11	Euro	DOE
Output Pillar II	OIR_01 Total public expenditure	Total: EU funds	OIR_01b_2.11	Euro	DOE
Context	C.19 Agricultural area under organic farming	Total (certified plus in conversion)	CTX_SEC_19_1c	ha UAA	Eurostat: org_cropar
Context	C.19 Agricultural area under organic farming	Share (certified plus in conversion)	CTX_SEC_19_2	% total UAA	Eurostat: org_cropar and apro_cpsh1

Unit:

- ha UAA: Hectares of utilized agricultural area.
- % total UAA: Percentage of total agricultural land.
- Euro: Financial support measured in euros.
- Supported hectare: Land area in hectares supported under CAP Measure 11.