Data Extraction, Preprocessing and Clustering Report: EcoWheataly – Durum Wheat Dataset

EcoWheataly Team

April 4, 2025

1 Objective

This report describes the construction of a clean, structured dataset from raw JSON data and the implementation of a clustering analysis pipeline on durum wheat farms in Italy. The final goal is to uncover meaningful farm-level typologies based on input use, costs, and productivity indicators, supporting simulation and policy evaluation in the EcoWheataly framework.

2 Database and Extraction Process

The input is a nested JSON file containing yearly farm-level information from 2008 to 2022. Each farm entry includes general data (like total farm acreage) and crop-specific details under the key `durum_wheat`, such as production quantities, crop areas, labor and machinery costs, as well as fertilization and phytosanitary product usage.

The extraction process involves several steps. First, general agricultural and economic variables are collected: produced quantity, PLV, crop acreage, hours of machinery use, and different cost components. Then, fertilizer-related information is extracted and structured both at the aggregate level and broken down by type (Mineral, OrganoMineral, Other, Micro-Mineral). Each type contains specific details such as distribution area, nutrient content, and costs.

Phytosanitary products are extracted by category (Herbicide, Insecticide, Fungicide) and classified by toxicity levels (0–4). These quantities are aggregated per hectare and stored in structured arrays, later converted into a unified 2D dataset.

3 Merging, Cleaning and Output Dataset

All collected data are merged by farm and year into a single DataFrame. Missing values in phytosanitary products are interpreted as non-use and replaced with zeros. Other NaNs are flagged and addressed through filtering. Outliers such as infinite values (e.g., from division by zero) are also removed to ensure numerical consistency. Basic statistics per year are printed to assess coverage and integrity.

The result is a flattened, cleaned panel dataset saved as `flat_df2.csv`, which serves as the input for machine learning and clustering tasks.

4 Preliminary Exploration

Before clustering, a preliminary exploration helps assess completeness and highlight potential dimensions of heterogeneity. For each year, we compute the number of available observations, the count of complete records, and the number of farms cultivating durum wheat. A zero-ratio analysis reveals the proportion of farms with zero values for key inputs like `fert_costs`, `phyto_costs`, or nutrient application, shedding light on possible differences in farming strategies or data sparsity.

Phytosanitary product use is further examined by plotting toxicity-class quantities over time. Both boxplots (showing the spread of log-transformed values) and line plots (showing yearly averages) help visualize changes in intensity and distribution across years and toxicity classes.

5 Clustering Analysis

Once the panel dataset is ready, we proceed with clustering to identify groups of farms that behave similarly. This unsupervised approach helps detect patterns in the way farmers use inputs and achieve yields, revealing potential typologies of input efficiency, intensification, or resource allocation.

To achieve this, three composite indicators were selected: the total phytosanitary product usage normalized by yield (`phyto_ratio_over_yield`), the combined nutrient use per yield (`ferti_ratio_over_yield`), and the hours of machinery use per hectare over yield (`hours_of_machines_ha_over_yield`). These indicators summarize multidimensional information and enable clustering in a compact feature space.

Outlier detection is performed using the Isolation Forest algorithm, which identifies and removes anomalous observations that could bias cluster centroids. The selected features are then standardized using the `StandardScaler` to ensure comparability in Euclidean space.

To find the optimal number of clusters, we use a combination of inertia (total within-cluster variance) and silhouette score (average cohesion and separation). By computing these metrics across a range of cluster numbers (from 5 to 150), we identify the elbow point and the silhouette peak. In our case, the analysis points to 5 as the optimal number of clusters.

After assigning each valid observation to a cluster using KMeans, descriptive statistics are computed for each group. These include the mean and standard deviation of each variable, as well as the size and proportion of each cluster.

6 Cluster Visualization

The following figures display the distribution of key variables across the resulting clusters, using boxplots to highlight intra-cluster variation and potential outliers. These charts provide insights into how different clusters compare in terms of input use and yield performance.

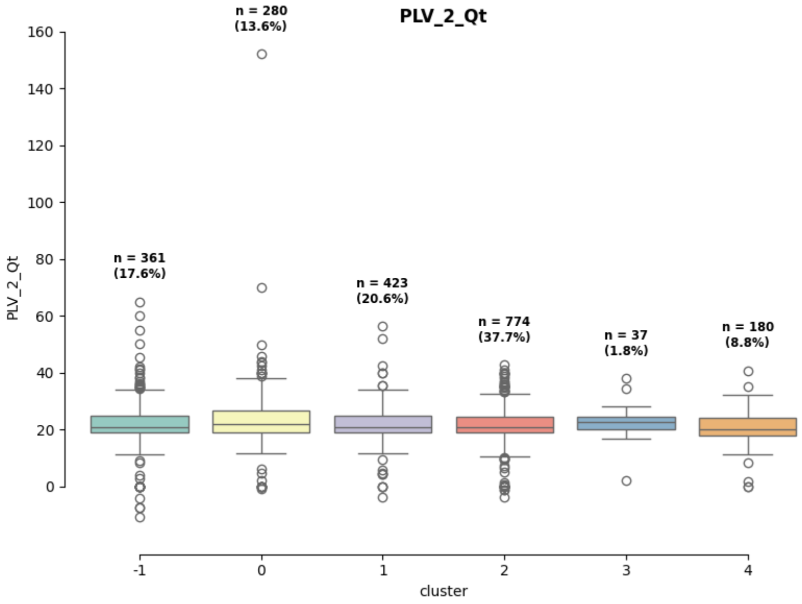

Figure 1: PLV per quintal across clusters

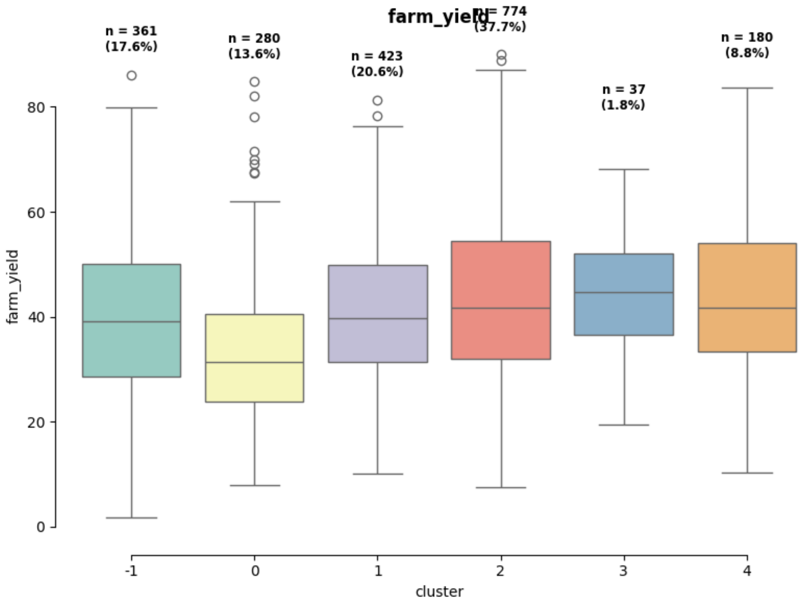

Figure 2: Farm yield per hectare across clusters

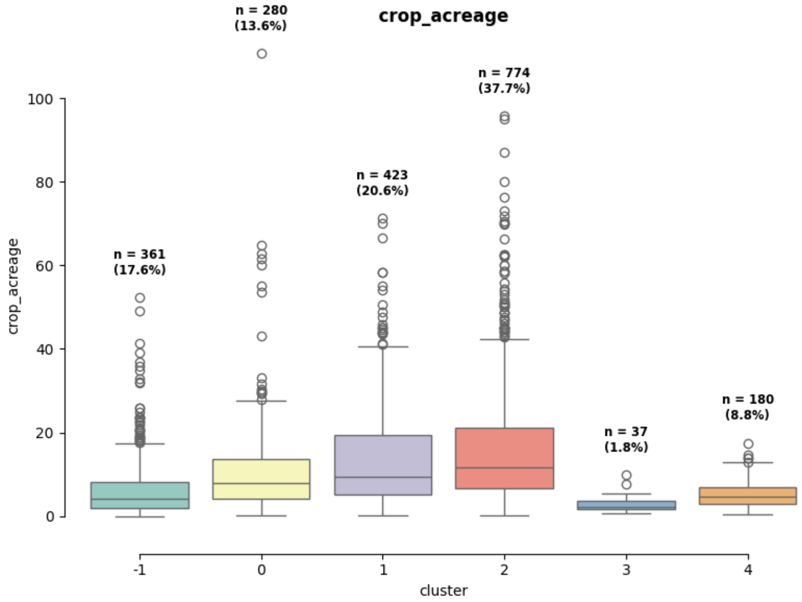

Figure 3: Crop acreage across clusters

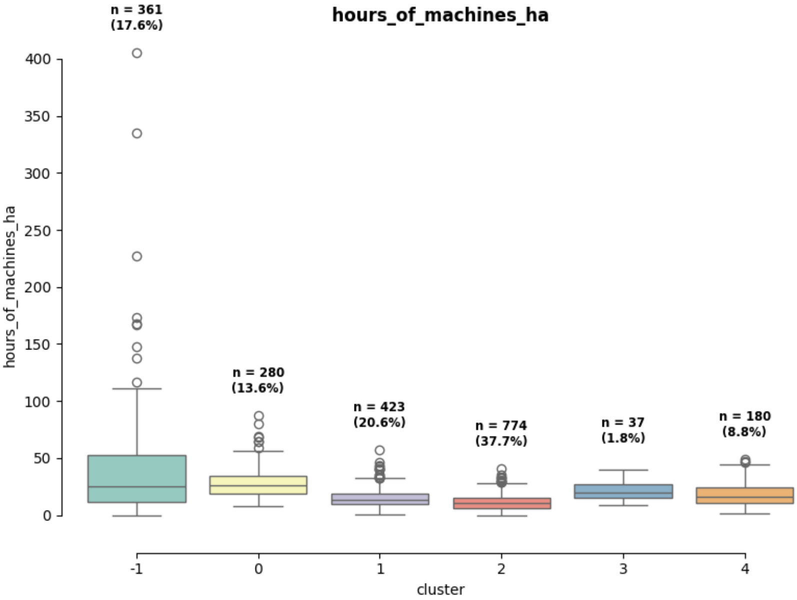

Figure 4: Machinery use per hectare across clusters

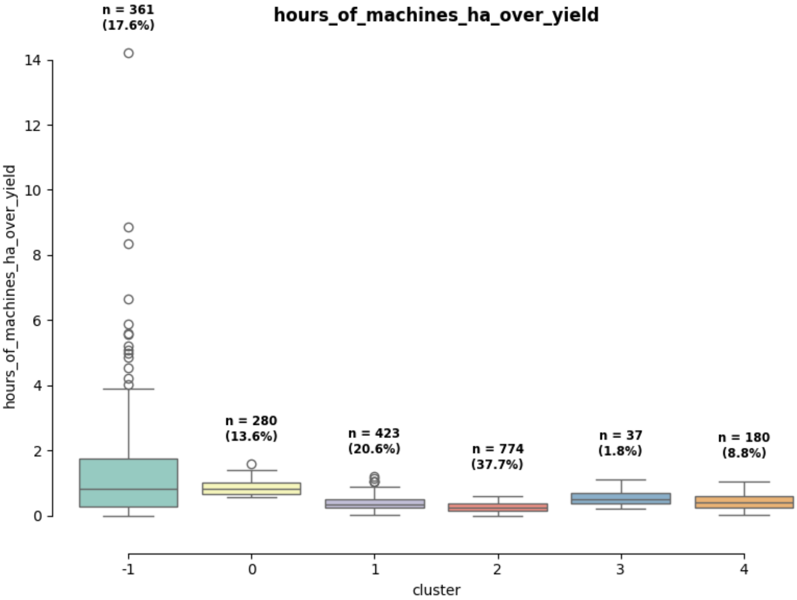

Figure 5: Machinery use over yield across clusters

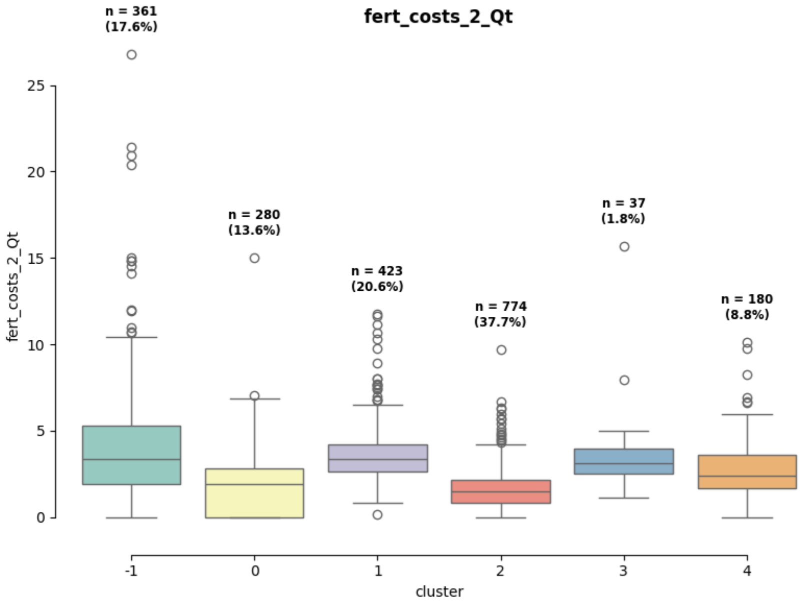

Figure 6: Fertilizer costs per quintal across clusters

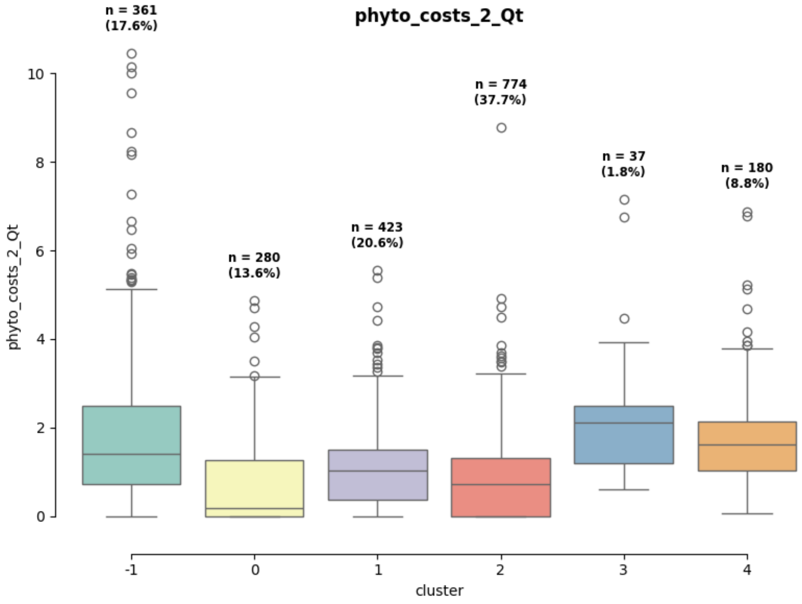

Figure 7: Phytosanitary costs per quintal across clusters

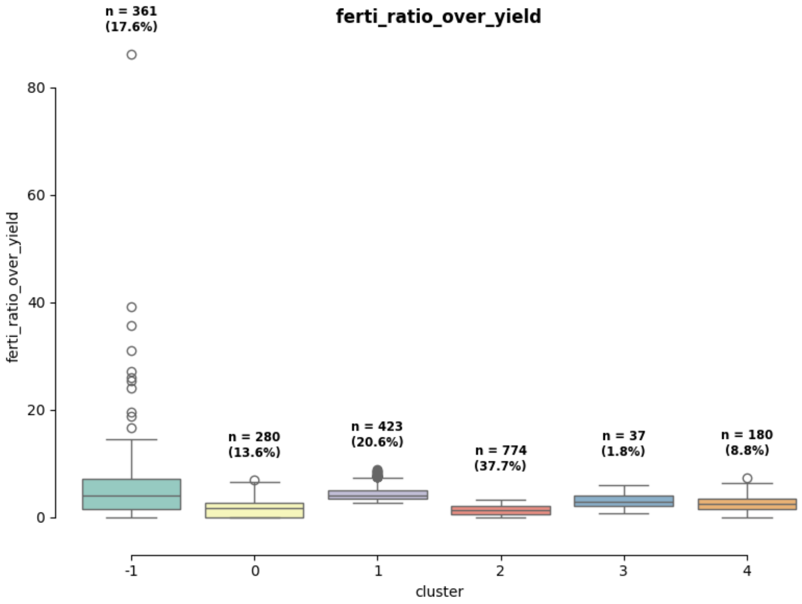

Figure 8: Fertilizer input ratio over yield across clusters

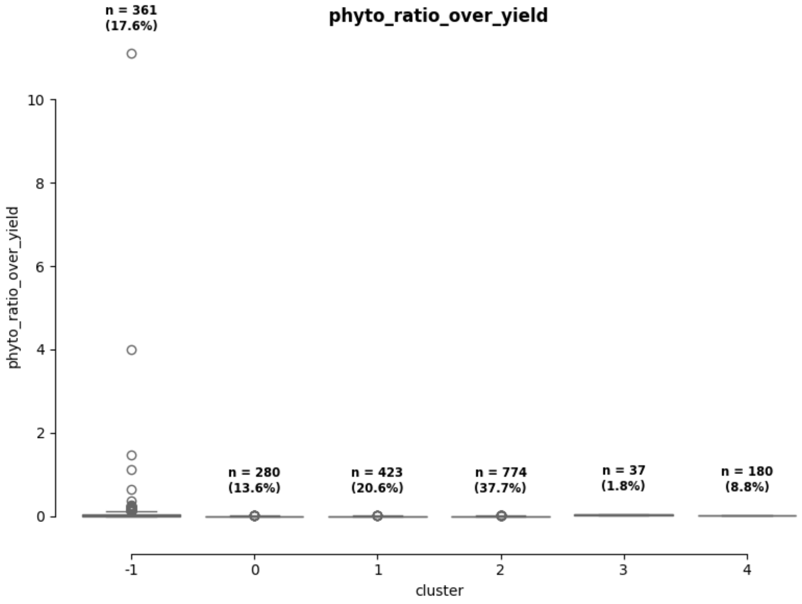

Figure 9: Phytosanitary input ratio over yield across clusters

7 Conclusion

The clustering procedure successfully segmented durum wheat farms into coherent groups based on resource use and productivity. The approach combines transparent preprocessing, robust outlier detection, and interpretable feature engineering. The resulting clusters offer a solid foundation for further simulation (e.g., agent-based models), targeted policy design, or exploratory analysis of input-output relationships in agriculture.