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

EcoWheataly Team

April 4, 2025

1 Objective

This script is designed to extract, merge, and preprocess agricultural and economic data from the EcoWheataly JSON database for durum wheat farms in Italy. The goal is to generate a clean and structured panel dataset suitable for machine learning tasks, particularly clustering and predictive modeling.

2 Input

- `ecowheataly_database.json` – a nested JSON file containing farm-level data by year and crop.
- Analysis range: 2008 to 2022.

3 Structure of the Database

Each farm entry contains:

- General info (e.g., `farm_acreage`)
- Crop-specific data under species key (`durum_wheat`), such as:
  - Production, value, area, cost variables
  - Fertilizers (categorized by type)
  - Phytosanitary products (categorized by toxicity and type)

4 Extraction Workflow

4.1 Part 1: General Variables and Fertilizers

Extracts key indicators for each farm-year:

- Wheat production: `produced_quantity`, `PLV`, `crop_acreage`, etc.
- Economic data: `fert_costs`, `phyto_costs`, `human_costs`, etc.
- Fertilizer details: area treated, amount per hectare, unit cost, nitrogen/phosphorus/potassium content, divided by type.

4.2 Part 2: Fertilizers by Type

For each fertilizer type (`Mineral`, `OrganoMineral`, etc.), the following variables are extracted:

- `fert_area`, `whole_qt_ha`, `unit_cost`, `distribuited_value`
- Nutrients per hectare: `nitrogen_ha`, `phosphorus_ha`, `potassium_ha`

4.3 Part 3: Phytosanitary Products

For each phytosanitary type (`Herbicide`, `Insecticide`, `Fungicide`), the script:

- Extracts distribution quantity by toxicity class (0–4)
- Aggregates quantity per hectare for each class
- Stores data in a 3D array (`Phyto`) and flattens to a 2D DataFrame

5 Data Merging and Cleaning

- All extracted data frames (`df1`, `df2`, `df3`) are merged on `year` and `farm code`
- Missing phytosanitary data are filled with zeros (assumed not used)
- NaNs in other sections are handled and flagged for filtering
- Diagnostic statistics per year and farm are printed

6 Visualization

- Boxplots and trend plots of phytosanitary product usage by toxicity class
- Indicators such as zero-ratio for selected variables (e.g., `fert_costs`)

7 Output

- Cleaned and flattened DataFrame saved as `flat_df2.csv`
- Ready for clustering and machine learning pipelines

8 Preliminary Analysis of Results

After constructing the full panel dataset, an initial exploration was conducted to assess data completeness and potential signals for clustering.

8.1 Missing Data Overview

For each year, the script reports:

- Total number of farm-level observations
- Number of complete records with all key variables available
- Number of distinct farms producing `durum_wheat`

This provides insight into the evolution of data coverage and production activity over time.

8.2 Zero-Ratio Analysis

For selected economic and input-related variables such as:

- `PLV`, `fert_costs`, `phyto_costs`
- `Mineral_distribuited_value`, `Mineral_nitrogen_ha`

the share of zero values is computed. High zero-ratio values may indicate non-use, data quality issues, or real heterogeneity in input intensity.

8.3 Phytosanitary Product Trends

Toxicity-class-specific quantities are plotted over time for each phytosanitary category:

- Boxplots show distribution of `log(1 + quantity_ha)` per class and year.
- Line plots show mean quantities per class across years.

This visualization allows assessment of both intensity and spread of use by type and toxicity level.

8.4 Data Cleaning and Final Dataset

After filtering out incomplete observations, the cleaned dataset is stored in `flat_df2.csv`. This dataset contains:

- Time-consistent and farm-aligned information
- Explicit zeros for missing phytosanitary use (assumed non-use)
- Numerical consistency checks (e.g., removal of infinite values)

This forms the basis for downstream tasks such as clustering of farm strategies or predictive modeling.