



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Università degli studi
"G. d'Annunzio"



Consiglio Nazionale
delle Ricerche



crea
Consiglio per la ricerca in agricoltura
e l'analisi dell'economia agraria

Missione 4 Istruzione e Ricerca

ECOWHEATALY

Evaluation of policies for enhancing
sustainable wheat production in Italy

PRIN MUR 2022

Clustering analysis for agent-based
simulation

Gianfranco Giulioni

Alessandro Ceccarelli



Introduzione: Il clustering ottimizza la segmentazione degli agenti, riducendo il rumore

- **Dati analizzati:**

- Il dataset **RICA** (Rete d'Informazione Contabile Agricola) include **8218** aziende agricole italiane e 15 anni di storico
- **L'analisi illustrativa** basata sul **clustering** include il solo anno **2016**, ovvero **1915** aziende che producono **grano duro**

- **Obiettivo strategico:**

- Definire N-cluster per **efficientare** la **modellazione ad agenti** e **gestire** il **rumore** nei **dati**

- **Architettura modulare:**

- **Framework flessibile** che combina **outlier detection** avanzati (Isolation Forest) e algoritmi di **clustering** (K-means, DBSCAN) per adattarsi a scenari dinamici
- Metriche in **input** per **clustering**, **rapportate** alla **resa** (quantità prodotta su superficie coltivata):
 1. Erbicidi (Herbicide_Qt Tox-0 - 4)
 2. Elementi totali ('nitrogen_ha', 'phosphorus_ha', 'potassium_ha')
 3. Efficienza operativa ('hours_of_machines_ha')



Una robusta data pipeline garantisce la rimozione degli outlier e la creazione di features rilevanti

- Il **filtering iniziale**, basato su tecniche di **anomaly detection**, ha:
 - Utilizzato **Isolation Forest** (ottimizzato per dati multivariati)
 - Eliminato gli **outlier**, preservando l'**integrità analitica** (da 1915 aziende a 1748; - 8.7%)
- La **feature engineering** si è **focalizzata** su:
 - Creazione di nuovi input (3)
 - Standardizzazione dell'*input space* (*StandardScaler*), per ottimizzare *distance-based metrics*

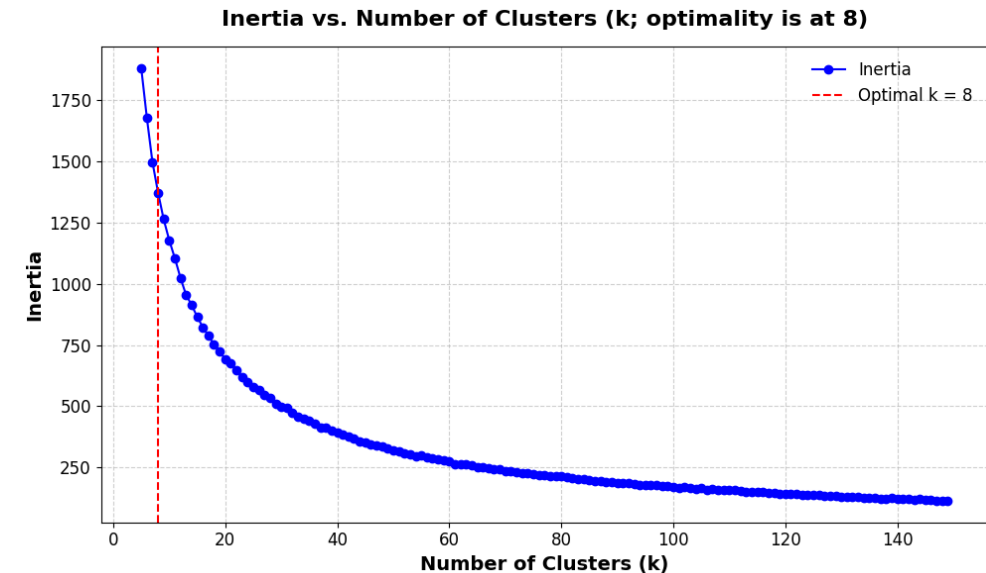
$$X'_i = \frac{X_i - \mu}{\sigma}$$

- **Robustezza operativa**:
 - Implementazione di **metodologie validate** (IsolationForest, GMM, DBSCAN e K-means) per mitigare rischi di **distorsione**



L'*elbow method* determina il numero ottimale di cluster analizzando l'inertia

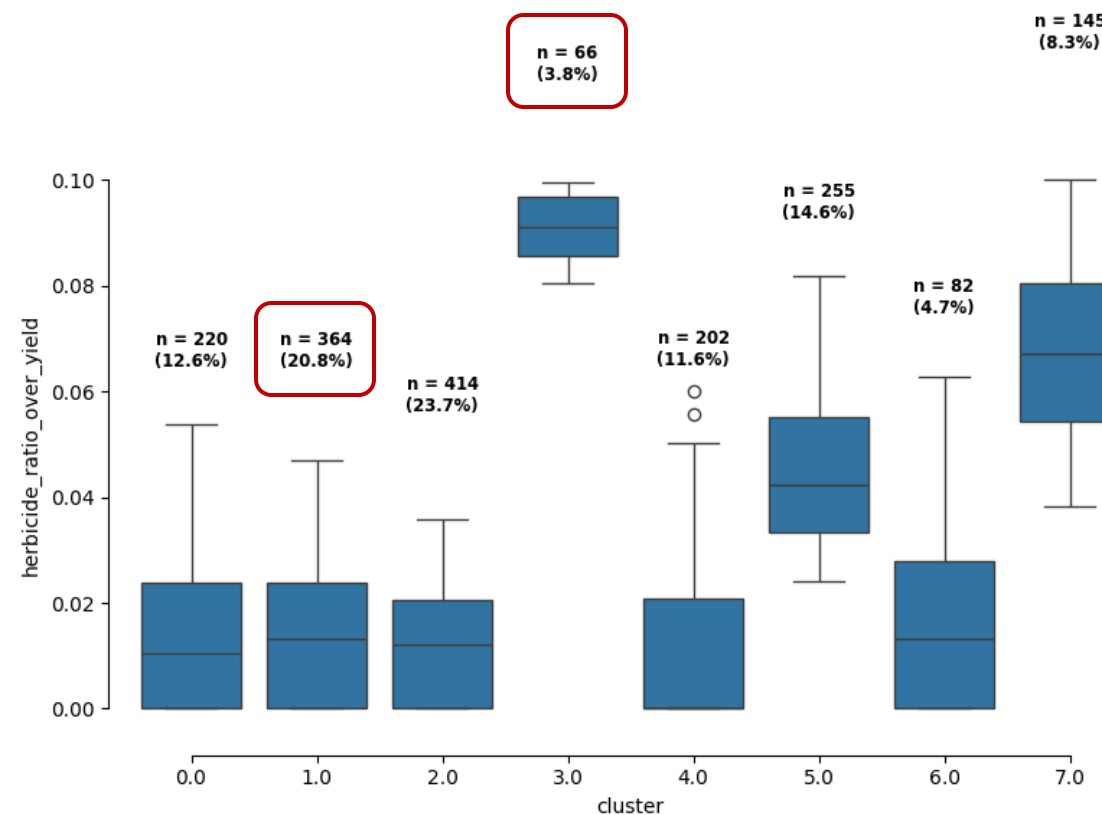
- L'algoritmo **K-means** è stato **ottimizzato**:
 - Utilizzando *l'elbow method* per bilanciare precisione e semplicità
 - Tecnicamente, si calcola la variazione dell'inerzia (il grado), cercando il valore di K per cui è massima
- Il **k ottimale (8)** è stato validato:
 - Plateau di inertia e coesione interna moderata
- **Decisioni data-driven**:
 - Metriche di performance (inertia, silhouette) integrate direttamente nel codice per validazione continua
- **Scalabilità**:
 - Framework pronto per l'integrazione con dati allargati



Rapporto erbicidi su resa (1/3)

- Cluster più **numerosi** (es. n=364, 20.8%) mostrano comportamenti **simili e moderati**
- Con l'**abbassarsi** della **numerosità** (es. n=66, 3.8%) emergono **squilibri** più **significativi**

ANALISI MONODIMENSIONALE DI CLUSTERING MULTIDIMENSIONALE

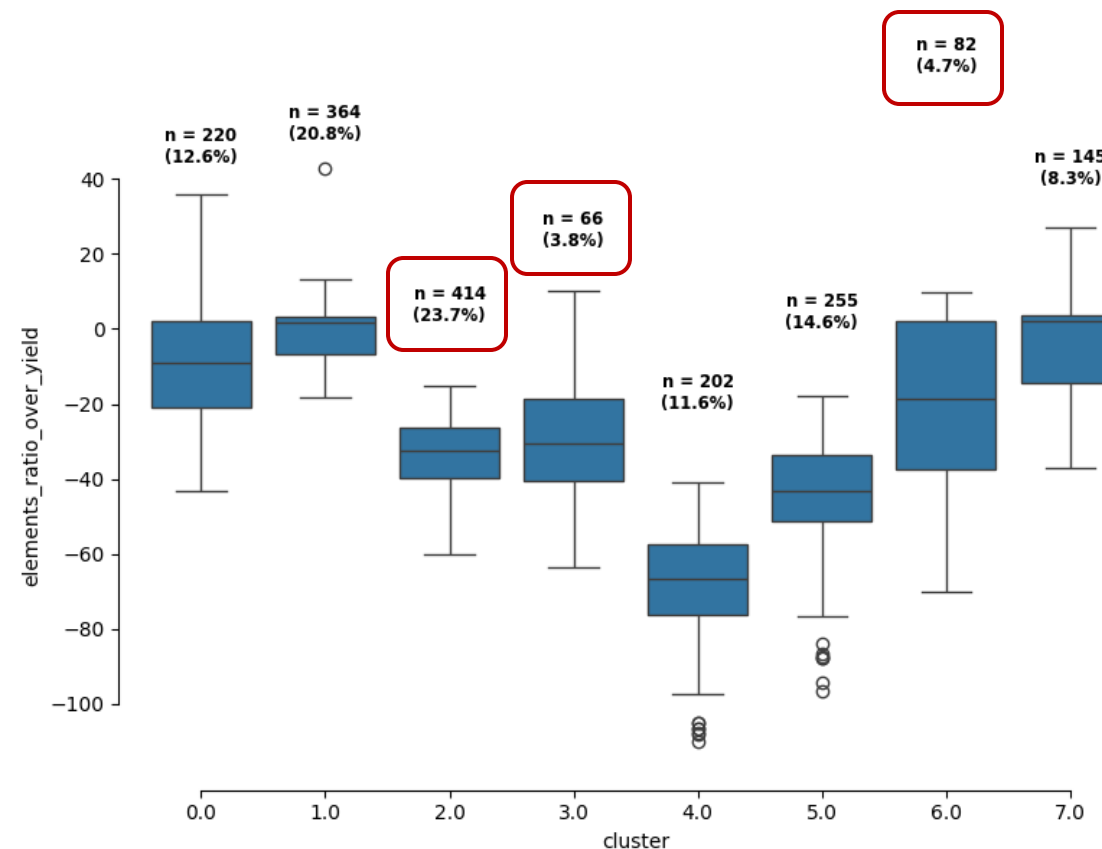




Rapporto elementi su resa (2/3)

- Gruppi più **estesi** (n=414, 23.7%) mostrano relazioni **comparabili e relativamente efficienti**
- Al diminuire del numero di aziende nel cluster, emergono gruppi con:
 - **Rapporti più rilevanti** (es. n=66, 3.8%)
 - **Variabilità maggiore** (es. n=82, 4.7%)

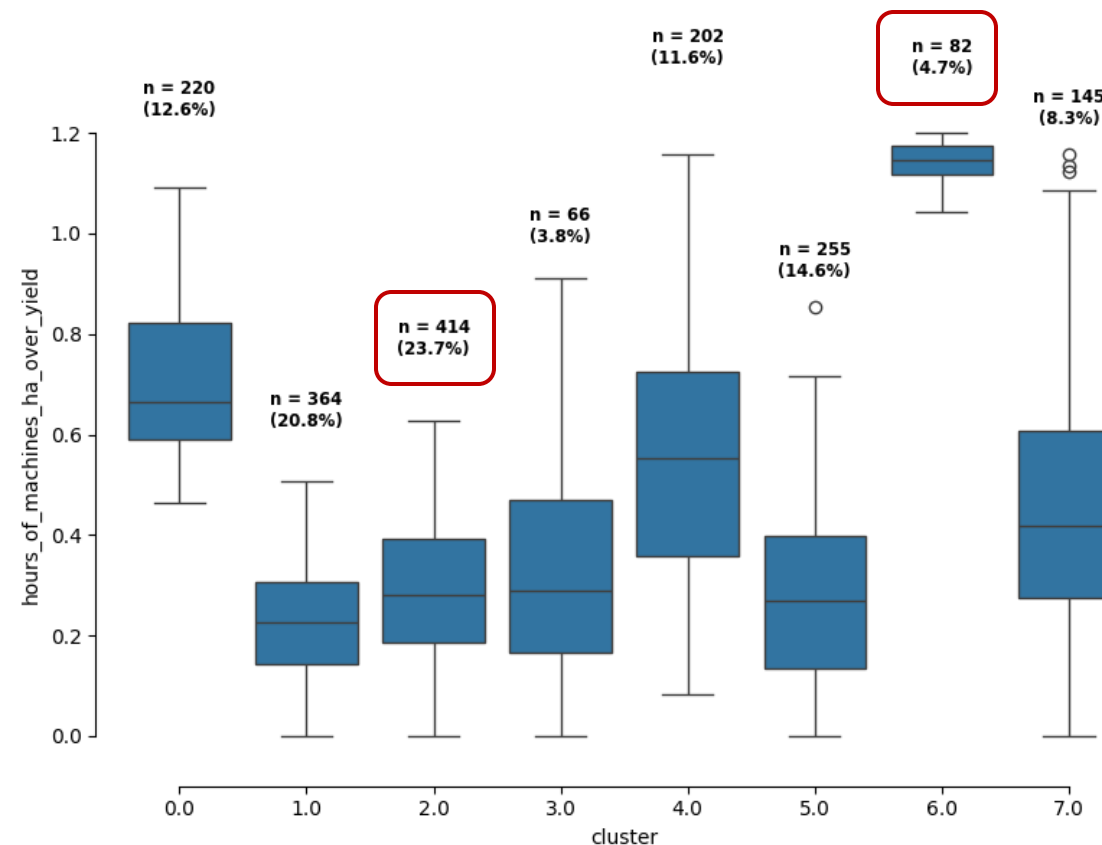
\ ANALISI MONODIMENSIONALE DI CLUSTERING MULTIDIMENSIONALE



Rapporto ore macchina su resa (3/3)

ANALISI MONODIMENSIONALE DI CLUSTERING MULTIDIMENSIONALE

- Il cluster principale (n=414, 23.7%) suggerisce un utilizzo **intensivo** ma **efficace** delle **macchine**, correlato a rese elevate
- Alcuni cluster (es. n=82, 4.7%) mostrano una **bassa** efficienza operativa, potenzialmente **associata a rese inferiori** o **metodi alternativi**





Implicazioni e raccomandazioni

- Le **simulazioni** possono essere **basate** sugli **N-agenti archetipici**, in ottica di:
 - **Riduzione del rumore**
 - **Ottimizzazione** computazionale
- **Integrazione futura:**
 - Estensione a **input allargati** (es. parametri territoriali o climatici) per aumentare la **robustezza** di anomaly detection e clustering)
- **Framework scalabile:**
 - Pipeline **pronta** per l'**adozione** in altri **progetti** agricoli con **minima** **riconfigurazione**



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Università degli studi
"G. d'Annunzio"


Consiglio Nazionale
delle Ricerche

 **crea**
Consiglio per la ricerca in agricoltura
e l'analisi dell'economia agraria

Clustering geo-visualization

