# Big data, small explanatory power?

## Random forest modelling of cereal yield variability across contrasting farming systems

**João Vasco Silva***, Joost van Heerwaarden, Pytrik Reidsma,
Alice Laborte, Kindie Tesfaye, Martin van Ittersum

j.silva@cgiar.org
Agronomy-at-scale Data Scientist
Sustainable Agrifood Systems
CIMMYT-Zimbabwe

**FSD Symposium,
Marrakech, November 2022**

CIMMYT

# Background

- Big data as an **important asset** for agronomic research and decision, the end of traditional agronomy?

- Direct application to **explain and/or predict** crop yield variability in farmers’ fields across time and space – complex due to G x E x M

- Unclear **how useful** big data for farming systems in different stages of intensification. Yield variability, data quality?

- **Objective**: Assess the potential for on-farm production data to uncover systematic and predictable patterns in yield variation

CIMMYT

# > 10.000 farm-year combinations

**Maize and wheat in Ethiopia**

Sample: 6350 fields
Year: 2009/10 & 2013
Field size: < 1.5 ha
Source: CIMMYT Surveys

**Rice in Central Luzon, Philippines**

Sample: 2000 fields
Year: 2014 WS and DS
Field size: < 1.3 ha
Source: IRRI Surveys

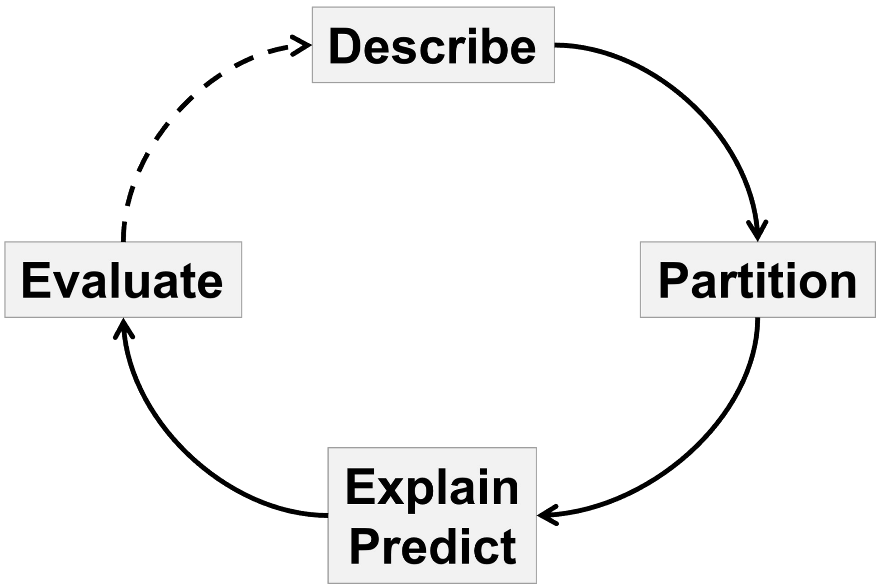
**Wheat and barley in the Netherlands**

Sample: 1770 fields
Year: 2015 – 2017
Field size: < 7.9 ha
Source: Agrovision Records

CIMMYT

# Methodological approach

Exploratory analysis of crop yield variability with e.g., boxplots and scatterplots.




Linear mixed models with RE to explain variance across scale (farm, district, region).

Random forest simulations: model performance and variable importance.

Model cross validation with data resampling in space and time.

CIMMYT

# Model formulation

**Predictive variables =** independent of growing season, time-invariant
WorldClim, GYGA climate zones, SoilGrids

**Explanatory variables =** growing-season specific, time-variant
farm survey variables, weather from AgERA5

| Model | Description | Explain | Predict | Variables (n) |
|---|---|---|---|---|
| M1*gps* | GPS coordinates only | X | X | **2** |
| M2*pc* | M1 + predictive climatic variables | | X | 2 + 22 = 24 |
| M3*pcs* | M2 + predictive soil variables | | X | 24 + 9 = 33 |
| M4*pcsf* | M3 + predictive survey variables | | X | 33 + 3 = 36 |
| M5*ec* | M1 + explanatory climatic variables | X | | 2 + 32 = 34 |
| M6*ecs* | M5 + explanatory soil variables | X | | 34 + 2 = 36 |
| M7*ecsf* | M6 + explanatory survey variables | X | | 36 + 17 = 53 |
| M8*pec* | M1 + pred. & expl. climatic variables | X | X | 2 + 54 = 56 |
| M9*pecs* | M8 + pred. & expl. soil variables | X | X | 56 + 11 = 67 |
| M10*pecsf* | M9 + pred. & expl. survey variables | X | X | 67 + 20 = **87** |

# Model evaluation

Cross-validation scheme with data resampling as follows:

1. **Traditional "out-of-bag"** (Breiman, 2001) for model fitted to pooled data

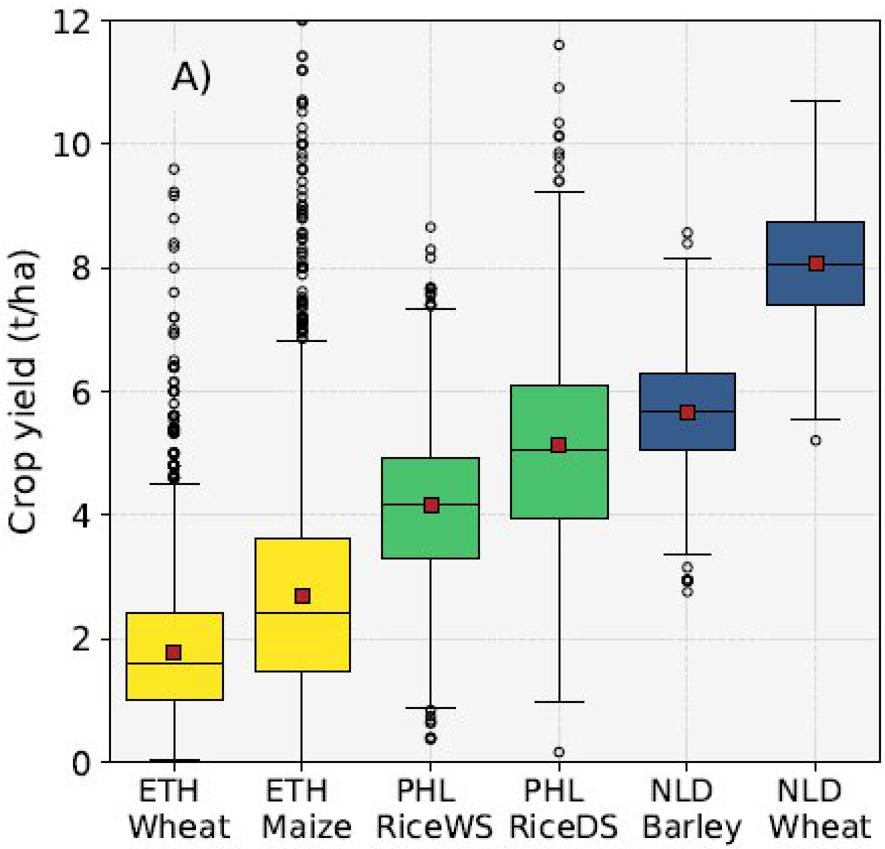
2. **Cross-validation over farms**:
   - 70% of farm-year combinations used for model training
   - remaining 30% for model evaluation ($R^2$ reported)

3. **Cross-validation over zones**:
   - 70% of admin provinces in the data used for model training
   - remaining 30% for model evaluation ($R^2$ reported)

4. **Cross-validation over years**:
   - 1 or 2 years (ETH and NLD) in the data used for model training
   - remaining year used for model evaluation ($R^2$ reported)

CIMMYT MR

# Cereal yield variability

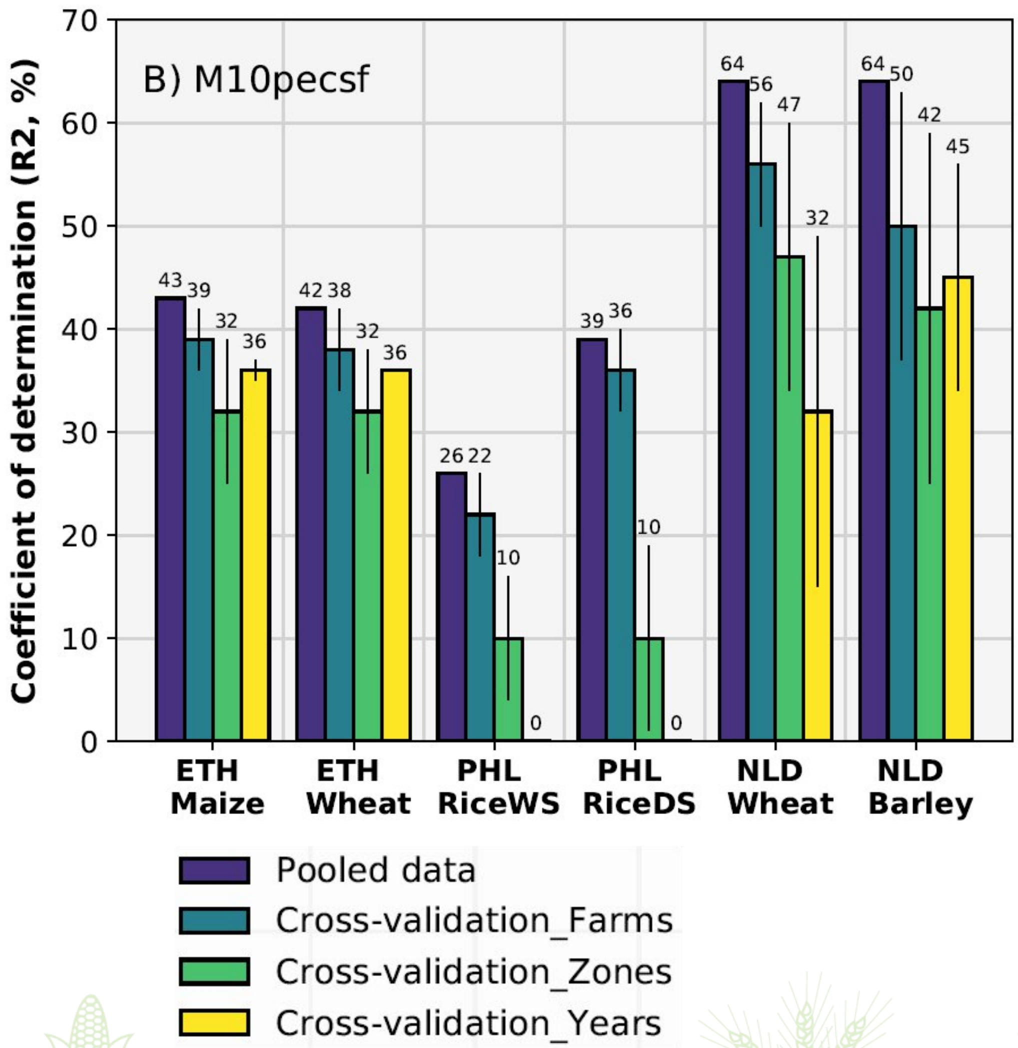



- Greater **yield variability** (standard deviation) for the lowest administrative unit in Ethiopia, followed by the Philippines, and the Netherlands
- **Random effects** accounted for 55% of residual variance in Ethiopia, 30% in the Philippines, and more than 70% in the Netherlands

# Explanatory power


- Farm survey variables improve explanatory power in **Ethiopia**
- Predictive climatic variables improve model performance in **the Philippines**
- Explanatory climatic variables improve model performance in **the Netherlands**

# Predictive power




- **Cross-validation across farms** reduces predictive power by 5-10% (except for barley) compared to the pooled data.
- **Cross-validation across space or time** reduces predictive power considerably compared to the pooled data, especially in the Philippines and in the Netherlands.

CIMMYT

# Take-home messages

1. 87 variables account for 65% of yield variability in the Netherlands and less than 45% in Ethiopia and in the Philippines

2. We need to understand better data quality, ‘missing predictors’, spatial and temporal extent of the data

3. Type of variables and cross-validation scheme have strong impact on model performance – system-specific or dataset-specific?

4. Big data from farmers’ fields may seem to explain yield variability, yet the same variables cannot be used to predict it – what value for big data then?

CIMMYT


# Thank you for your interest!

**João Vasco Silva, PhD**
j.silva@cgiar.org
Agronomy-at-scale Data Scientist
Sustainable Agrifood Systems
CIMMYT-Zimbabwe

# Variable importance

**Out of a total of 87 variables:**

- Nutrient management (explanatory survey variables) most important for cereal yield in **Ethiopia**
- Predictive climatic variables most important for rice yield in **the Philippines**
- Explanatory climatic variables most important for cereal yield in **the Netherlands**