Big data, small explanatory power?

Experiences with cereal yield variability across the globe

J.V. Silva[1,2], J. van Heerwaarden[2], B. Assefa[1,2], K. Tesfaye[1], M.L. Velasco[3], A.G. Laborte[3], L. Spatjens[4], P. Reidsma[1], M.K. van Ittersum[1]

[1]CIMMYT, [2]Wageningen University, [3]IRRI, [4]Agrovision BV

December 9, 2020

Big data – the end of traditional agronomy?

- ❑ **Big data: “**high volume, velocity and variety of **information** to require specific analytical and technological **methods** for its transformation into value” (di Mauro et al., 2016).

Big data – the end of traditional agronomy?

❑ **Big data: “**high volume, velocity and variety of **information** to require specific analytical and technological **methods** for its transformation into value” (di Mauro et al., 2016).

❑ **Vision**: Farmer field data coupled with spatial explicit biophysical data, and advanced statistical and modelling tools, are useful to test M x E interactions in a cost-effective way.

Big data – the end of traditional agronomy?

- ❑ **Big data: “**high volume, velocity and variety of **information** to require specific analytical and technological **methods** for its transformation into value” (di Mauro et al., 2016).

- ❑ **Vision**: Farmer field data coupled with spatial explicit biophysical data, and advanced statistical and modelling tools, are useful to test M x E interactions in a cost-effective way.

- ❑ **Hope:** Big data can be the backbone of *ex-ante* policy assessments and can help prioritize management interventions to meet food availability and environmental targets in the future.

Big data – the end of traditional agronomy?

- ❑ **Big data: “**high volume, velocity and variety of **information** to require specific analytical and technological **methods** for its transformation into value” (di Mauro et al., 2016).

- ❑ **Vision**: Farmer field data coupled with spatial explicit biophysical data, and advanced statistical and modelling tools, are useful to test M x E interactions in a cost-effective way.

- ❑ **Hope:** Big data can be the backbone of *ex-ante* policy assessments and can help prioritize management interventions to meet food availability and environmental targets in the future.

- ❑ **Doubt:** To which extent can actual yields in farmers’ fields be predicted for different crops, farm types and farming systems across the world?

Objective

Assess the performance of statistical and machine learning techniques to predict actual yields in time and space, based on a wide range of biophysical and management factors.

> 10k field x year combinations

Maize and wheat in Ethiopia

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

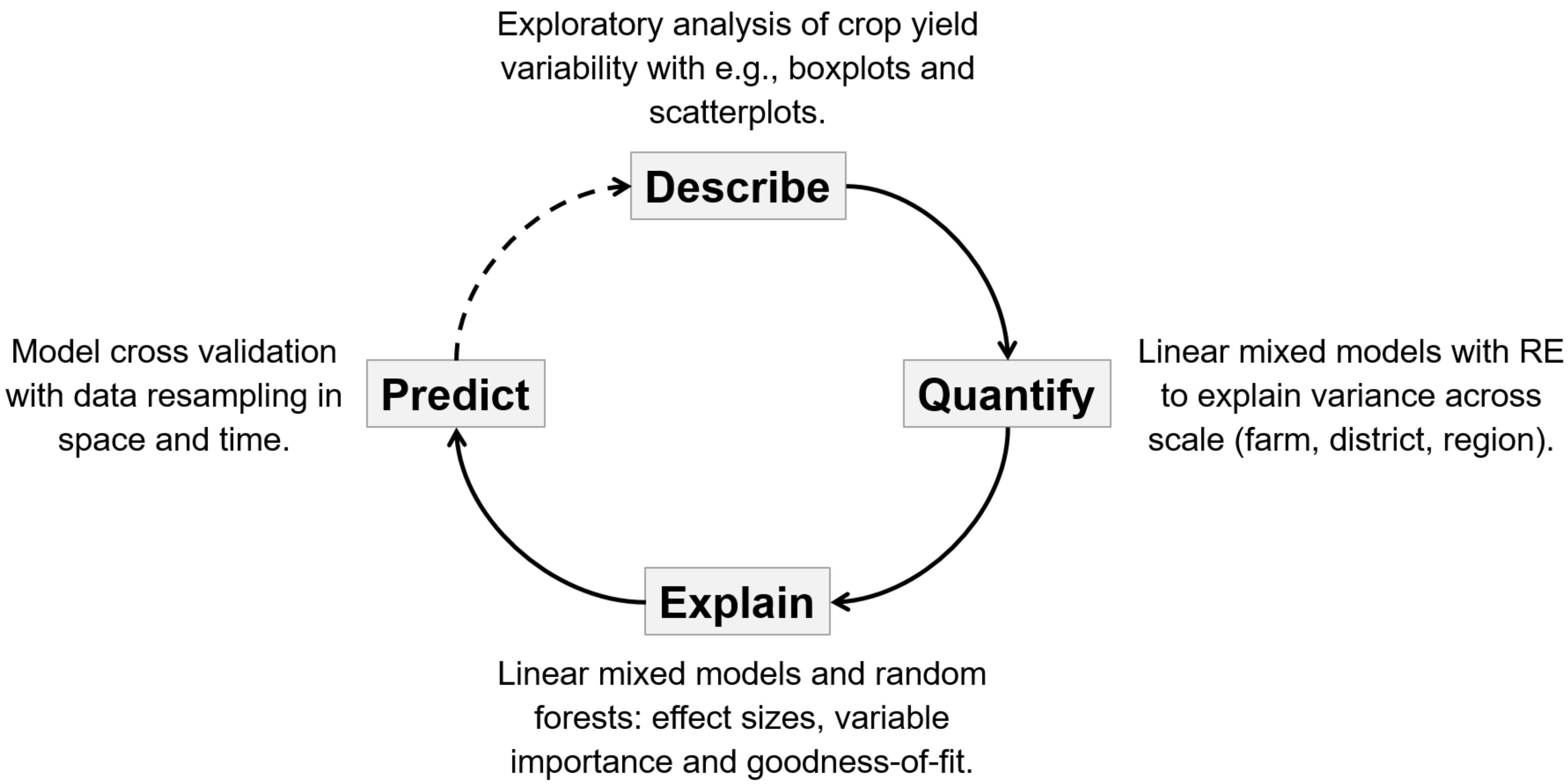
Rice in Central Luzon, Philippines

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

Methodological framework



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

Describe

Quantify

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

Explain

Linear mixed models and random forests: effect sizes, variable importance and goodness-of-fit.

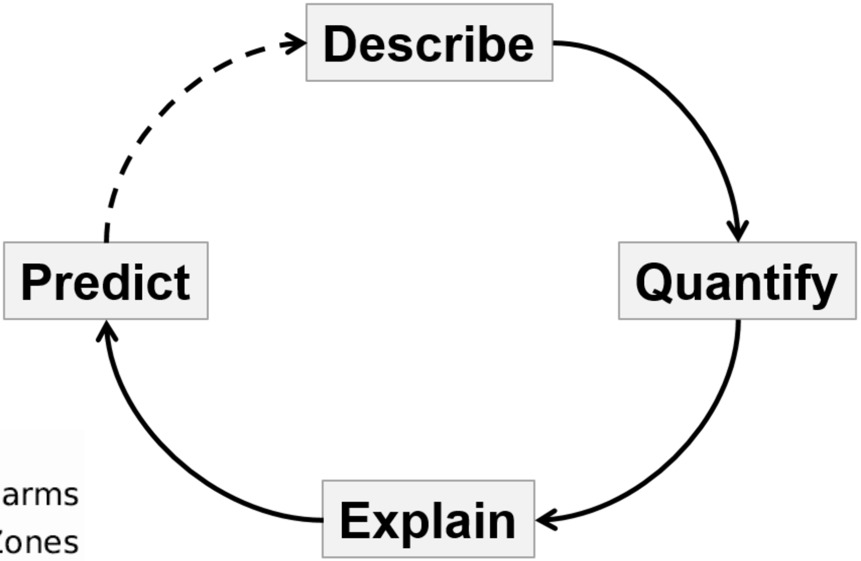
Predict

Model cross validation with data resampling in space and time.

Delaune et al.
Submitted (EJA)

Methodological framework

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



Describe

Quantify

Explain

Predict

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

Linear mixed models and random forests: effect sizes, variable importance and goodness-of-fit.

Model cross validation with data resampling in space and time.

🟦 Pooled data

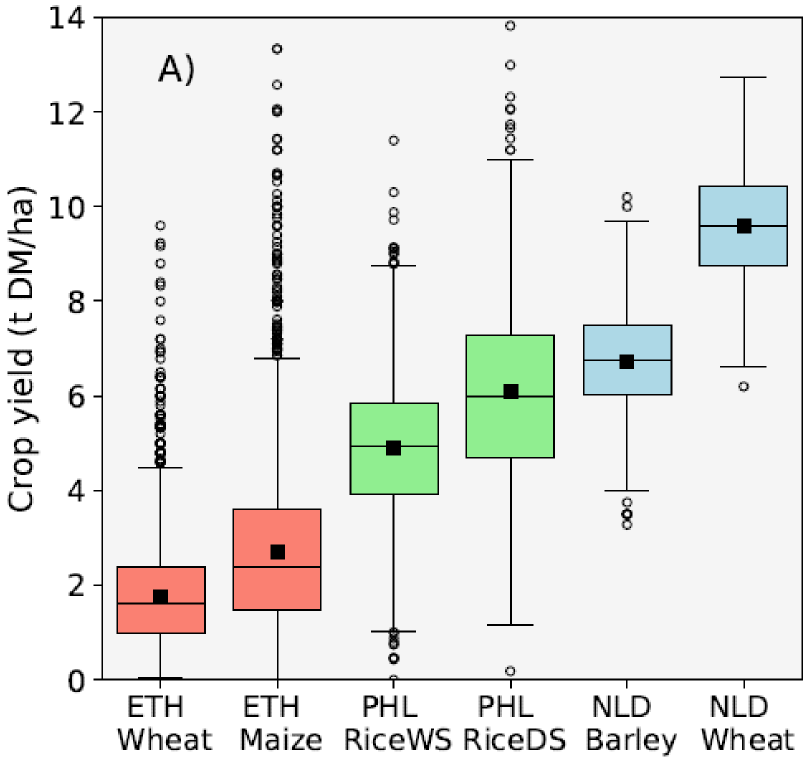
🟧 Cross-validation_Farms

🟩 Cross-validation_Zones

🟥 Cross-validation_Years

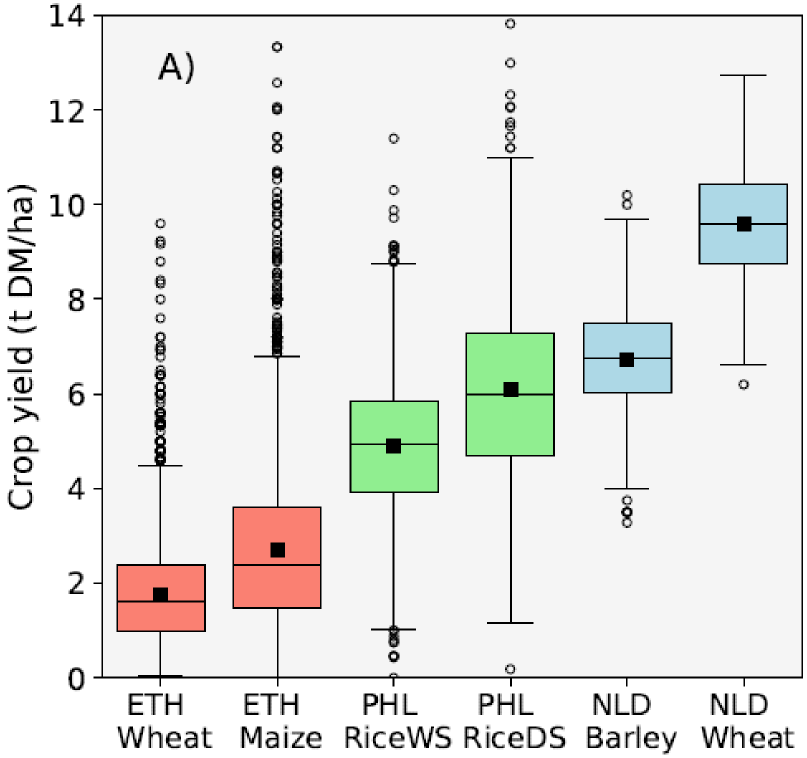
Delaune et al.
Submitted (EJA)

Yield variability and sources of variation



- Yield gaps are large in Ethiopia, intermediate in the Philippines and small in the Netherlands.

Yield variability and sources of variation

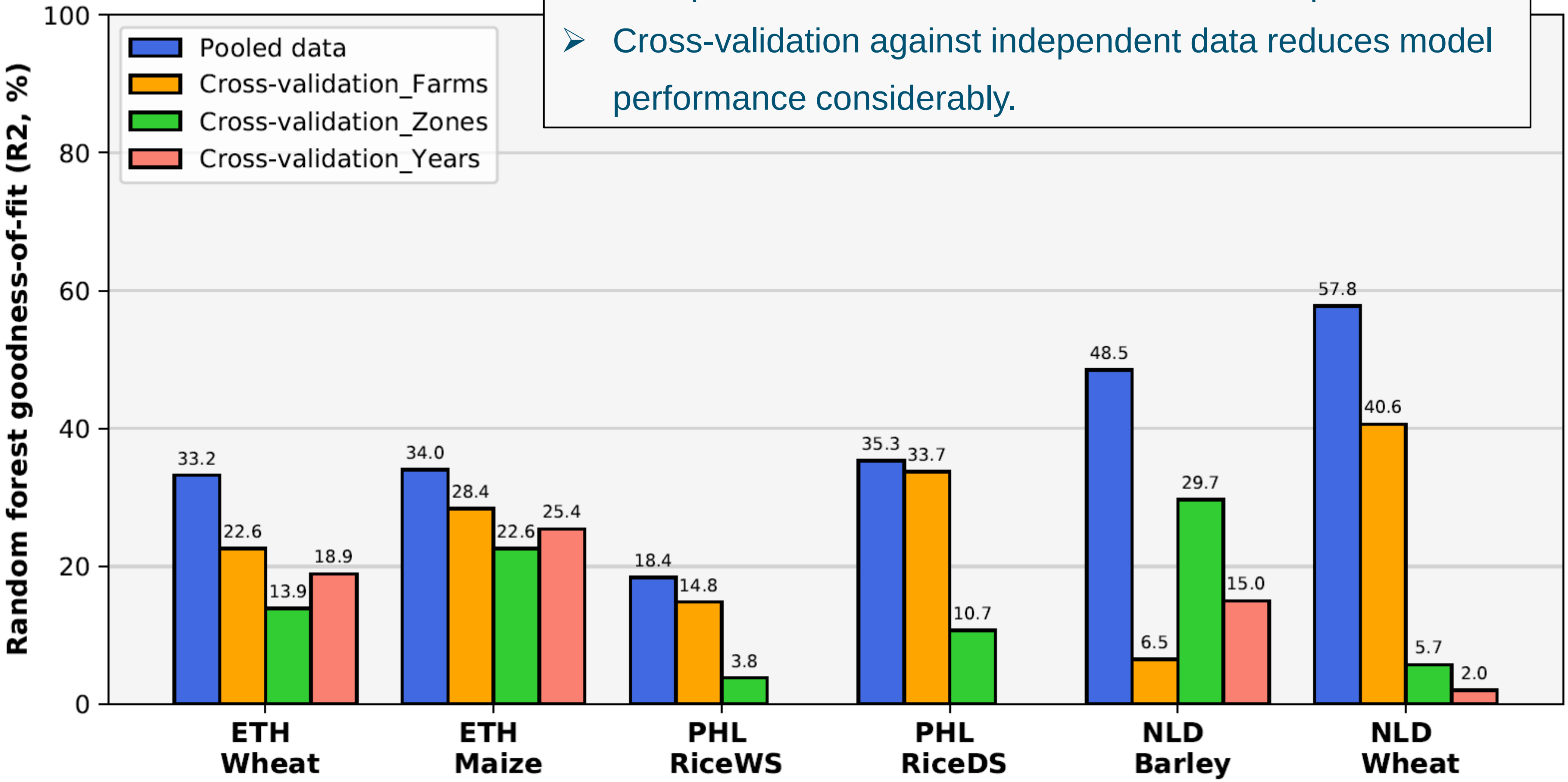


- Yield gaps are large in Ethiopia, intermediate in the Philippines and small in the Netherlands.

- Most yield variation is explained by **farm random effects** rather than by district or region random effects.

Model performance (R^2)

- ➢ Best performance achieved in the model with pooled data;
- ➢ Cross-validation against independent data reduces model performance considerably.



Conclusions

- Big data from farmers' fields are currently useful to describe cropping systems, but not to predict yield variability across time and space.

Conclusions

- Big data from farmers' fields are currently useful to describe cropping systems, but not to predict yield variability across time and space.

- Models with pooled data perform better for cereals in the Netherlands than for cereals in Ethiopia and the Philippines, where the scope and need for sustainable intensification is greatest.

Conclusions

- Big data from farmers' fields are currently useful to describe cropping systems, but not to predict yield variability across time and space.

- Models with pooled data perform better for cereals in the Netherlands than for cereals in Ethiopia and the Philippines, where the scope and need for sustainable intensification is greatest.

- Model performance declines when these are tested against independent data (beware of over-fitting!). Moreover, the performance of the cross-validated models was not consistent across crops or countries.

Conclusions

- Big data from farmers' fields are currently useful to describe cropping systems, but not to predict yield variability across time and space.

- Models with pooled data perform better for cereals in the Netherlands than for cereals in Ethiopia and the Philippines, where the scope and need for sustainable intensification is greatest.

- Model performance declines when these are tested against independent data (beware of over-fitting!). Moreover, the performance of the cross-validated models was not consistent across crops or countries.

- Improved data collection and robust agronomic frameworks are needed to materialize agronomy-at-scale approaches based on big data.

João Vasco Silva (PhD)

CIMMYT-Zimbabwe: j.silva@cgiar.org

Wageningen University: joao.silva@wur.nl

WAGENINGEN UR
For quality of life

IRRI International Rice Research Institute

CIMMYT
International Maize and Wheat Improvement Center