GUIDE TO SOUND DATA AND ANALYTICAL PRACTICES FOR ON-FARM EXPERIMENTATIONS (OFE)

TABLE OF CONTENT

INTRODUCTION

ON-FARM RESEARCH DATA AND ANALYTICS

Kyveryga 2019 and Lacoste et al. (2021)framed the interest and recent development of on-farm research. How should we rethink the planning and analysis of on-farm experiments with farmers' interests in mind? How should we organize and curate data and metadata for this purpose? The following guidelines aim to sketch the good practices for that matter. They originate from the presentation of the Data and Analytics webinar on May 17, 2021, the third of a series of four in preparation of the #OFE2021 conference. These guidelines are planned to be evolutive. For suggestions and comments write to ofec@ispag.org.

WHY DATA?

- Understand, predict, control, and manage (Nix, 1984 [pp. 181-188])
- Actionable information for improved management that increases income, maximizes resources, and minimizes the impact on the environment
- Future farms, data-driven solutions (find data, interpret, aggregate, visualize, analyze)

DATA GENERATION OPTIONS ARE ON A CONTINUUM

- Continuum: Lab experiment ⇒ small plots ⇒ experimental farm ⇒ on-farm trial ⇒ farm field survey
- Away from the real world ⇒ close to the real farm managing world
- Low risk of confounding effects of factors on treatment effect ⇒ high risk of confusion on treatment effect
- High flexibility for designing the experiment ⇒ low flexibility
- Increasing farmer engagement along the continuum helps best practices adoption
- Data that can be recorded automatically during farming operations are more likely to be useful to farmers

DATA AND METADATA

MINIMUM SET

A minimum data and metadata set can be developed with a particular use in mind and is warranted for understanding of yield differences among years, fields, and farms. This minimum set should be easy to collect under field conditions and

should provide reasonable answers, possibly with different meanings depending on their use (models or scientists).

Level 1: Minimum input data that drive the crop production system

Environmental inputs

Weather

- Daily maximum temperature
- Daily minimum temperature
- Precipitation
- Solar radiation

Soil surface information

- Color
- Permeability
- Drainage

Soil profile information

- Water holding characteristics
- Nitrogen
- Organic matter
- Other nutrients

Management inputs

- Crop
- Name or description of cultivar or variety grown
- Planting date
- Row and plant spacing
- Irrigation (dates and amount of irrigation)
- Fertilizer (inorganic and organic; dates, amount, and type of fertilizer)
- Tillage operations (dates and type of implement, other applications)
- Harvest

Level 2: Observational data to understand difference among years, fields, or management practices (in addition to Level 1 data)

Yield components

- Yield (if yield is different, we might need additional information besides weather, soil, and crop management)
- Seed/grain number and seed/grain size, etc. at harvest

Final biomass

- Harvest index (ratio of yield over total biomass)

Phenology and development

- Flowering date, physiological maturity, harvest maturity, first seeds, etc.

LEVEL 3: ISSUE OR PROBLEM ORIENTED

Drought

- Soil moisture at different locations in the field and at different soil depths over time

Nutrients

- Soil measurements: Soil nutrients at different locations in the field and at different soil depths, potentially at different times during the growing season
- Plant nutrient measurements: Different plant components, especially leaves

DATA AND METADATA BASIC CONCEPTS

FAIR PRINCIPLES

- Findable
- Accessible
- Interoperable
- Reusable

References:

- Foundational paper
- Living document with useful tools
- Global Community Guidelines for Documenting, Sharing, and Reusing Quality Information of Individual Digital Datasets

ONTOLOGIES

Ontologies provide a common language for different kinds of data to be easily interpretable and interoperable allowing easier aggregation and analysis.

- Agronomy Ontology (AgrO)
- Crop Ontology (CO)
- Unit Ontology (UO)
- Phenotype and Trait Ontology (PATO)
- The Environment Ontology (ENVO)
- Chemical Entities of Biological Interest (CheBI)
- Socioeconomic Ontology (SEOnt)

International Consortium for Agricultural Systems Applications (ICASA)

- 30+ years of experience: balance necessary detail vs. approximations
- Completeness in specifying environment (e.g., initial soil conditions; daily weather)
- Completeness in specifying management (e.g., crops; rotations + intercrops; 600+ variables)
- Experiments (metadata, management, measurement)
- Weather data (individual weather station, daily values)
- Soil profile data (individual soil profiles, surface and profile data)

Generation of interoperable agronomic data

At collection

The Agronomy Field Management System (AgroFIMS) generates standardized field books to collect agronomic data that is born FAIR. It features

- Fieldbook design via ontology-based variables, terminology, and units in modules representing typical cycle of operations in agronomic trials
- Digital data collection with KDSmart, ODK or Field Book mobile app
- Data analysis via AgroFIMS statistical scripts (R-based) and reports
- Data archiving through easy upload to institutional repository

After collection

- Agricultural Data Research Network (ARDN) data sharing workflows
- VMapper allows mapping existing data to the ICASA standard

Metadata related to positioning data

- ISO 19100 series of standards define structure and encode FAIR spatial data and metadata, including their position defined by coordinates. ISO 19115-1 is fundamental for specifying metadata related to spatial data
    - See: Table 12. End-user requirements support in current standards – Agriculture sector in "FAIRness of current standards for precise positioning data". Sections 2 and 3 discuss FAIRness of current standards specifically.
- The International Community Guidelines for Sharing and Reusing Quality Information for Earth Science Datasets aim to help stakeholders, such as (science) data centers, data repositories, data producers and publishers, data managers and stewards, etc: 1) to capture, describe, and represent quality information of their datasets in a way that is in line with the FAIR guiding principles; 2) to allow for the maximum discovery, trust, sharing, reuse and value of their datasets, and; 3) to enable global access to and integration of dataset quality information. Although developed for Earth Science datasets, principles and advice in the guidelines is applicable in

any application domain of these, including in agriculture and on-farm experimentation context.

DATA SHARING TOOLS

Data papers

- Paper attached to a public-access database
- Purpose is not to present new findings but to describe a dataset made fully available (origin, and reuse for all kinds of applications).
- Examples:
    - Su, Gabrielle and Makowski, A global dataset for crop production under conventional tillage and no tillage systems
    - Cernay, Pelzer and Makowski, A global experimental dataset for assessing grain legume production
    - Li, Ciais, Makowski and Peng, A global yield dataset for major lignocellulosic bioenergy crops based on field measurements

ISOFAST

- A good example of interactive summaries of on-farm strip trials. See Laurent et al. 2020

Gardian

- Global Agricultural Research Data Innovation Acceleration Network data discovery portal
    - Datasets
    - Publications
    - Data management toolkit
- Tools to collect standard compliant data
- Standard metadata and semantic

AgriMetrics

- Transactional data sharing, value generation and data governance

STATISTICAL ANALYSES OF OFE

ANALYSIS OF DATA FROM INDIVIDUAL FARMS

OFE DIFFERENCES WITH CONVENTIONAL

- Same process (hypothesis, design, data collection, analysis, review) as for conventional experimental farms but with different scales and objectives
- Farmer's focus is not so much the significance of effects (which treatment is best), but their quantitative importance (to help cost-benefit analyses).
- Overall or site-specific treatment effects can be studied, particularly with networks of OFE designs

DESIGNS

- Replications needed. For optimal number, see Alesso et al. (2019); Tanaka (2021)
- If just overall treatment effects are targeted, replicated strip trials are good enough. Experimental designs for assessing overall effects are considered by Marchant (2019), Alesso (2020), and Tanaka (2021). For designs for site-specific crop yield response where a stable spatial component is present, see Alesso et al. (2020)
- Randomization not determinant; a highly replicated systematic strip trial can be the best design in practice. If spatial trends within a field are consistent, randomization may not be needed. However, if we expect yields spatial trends to be unstable or skewed, the plots should be randomized. In practice, randomization can introduce difficulties with the quality of the implementation unless the farmers have the latest technology. Also, treatment design layout must match the data collection process (e.g., yield monitors). A lot of data is lost because of this mismatch
- Depending on the treatments to be tested and the design used, equipment and data limitations must be considered. An example is too short plot length for collecting yield data using a yield monitor

ANOVA IS NOT SUITED

- Assumption that observations are spatially independent does not apply to modern data collection. Also, the extremely high number of observations collected increases the likelihood of Type I error (false positive) and randomization/replications is no guarantee against that
- ANOVA shows large bias of estimates and narrow confidence intervals.
- ANOVA can give wrong estimates, more frequently and can also result in wrong proposed scenarios (Tanaka 2021)
- Simple linear models may be appropriate. If the focus of an experiment is on the causal relationships between an input and crop yield, a linear

regression model is a better choice than ANOVA. For example, a quadratic function can be used for modeling if a non-linear relationship is expected.

Spatial linear mixed models (LMM) instead of ANOVA

- A concern with spatial data is how the spatial correlation is removed from the trend. The estimates of the spatial correlation could double if the very subtle but consistent trends (the sorts that would normally be put on the account of natural variation) are not removed. It is important to model the spatial variation from areas within a field that is known to be reasonably uniform, and not try to simultaneously estimate treatments, design factors, and errors in a single fit
- Treatment effects (fixed) and simultaneous residual spatial effect (covariance parameter) constitute a less biased estimator
- Spatial LMM are closer to actual treatment effect and provide more stable estimate. Their confidence intervals are also more reliable
- However, general spatial LMM assume isotropic spatial variability. Prior knowledge of anisotropic elements (refer to previous observation, yield map or satellite imagery) should be either excluded from the experimental design or the analysis, or incorporated in spatial modeling

Frequentist vs. Bayesian approaches

- Frequentist
  - Pro
    - Easy to understand; moderate need for computational resources
    - Easy estimation of overall effects
  - Con
    - Difficult to estimate individual effects
    - Problem with local optima; difficult to know whether the assumption is wrong
    - Difficult to estimate the random effect
    - Between-trial variance in OFE networks sometimes poorly estimated
- Bayesian
  - Pro
    - Flexible modeling; incorporation of full uncertainties in all parameters
    - Prior information can be obtained from the literature or expert knowledge
  - Con
    - Difficult to automate evaluation

- May not be practical for farmer use due to the high need for computational resources
- Prior specification is not always straightforward

AVAILABLE SOFTWARE

Spatial LMM

- R: geoR (Ribeiro & Diggle 2007) No longer updated
- R: ASReml-R (www.vsni.co.uk/software/asreml/) Paid
- Note: Spatial LMM is very time consuming and R packages don't work well with large spatial datasets
- MATLAB: geostat (github.com/takashit754/geostat) Anisotropic model

Bayesian approach

- Stan (mc-stan.org) Cov function can be defined by the user
- R: INLA, spTimer, spBayes, mcmcGLMM, rjags…etc. Possible candidates. Cov function is predefined.

Site-specific crop response:

- R: spSVC, spgwr - Spatially varying coefficient models
- Python: Scikitlearn/Keras/Pytorch... Machine learning techniques

DEALING WITH TEMPORAL VARIATIONS

- Generally, each experiment is done for one year only
- As OFE can be assimilated as observational studies, it is difficult to acknowledge years' effects. One solution would be to use the LMM of coregionalization (Marchant and Lark, 2007) if the experiment can be repeated twice in the same field. Hence, the statistical difference among treatment effects between years could be assessed. If the experiments are conducted in different fields or multiple years (more than 3 years), there is no simple option. In such cases, meta-analyses or benchmarking methods would be warranted

ANALYSIS OF DATA FROM SEVERAL FARMS

Aggregation is useful to assess performance of different farming practices such as effects on yield, profitability, environment, multi-criteria

INDIVIDUAL AND MEAN EFFECT SIZE ESTIMATION (WITH CONFIDENCE OR PREDICTIVE INTERVALS)

Hossard et al. (2016); Laurent et al. (2020)

REGRESSION

Laurent et al. (2021)

MACHINE LEARNING METHODS

- ML normally strikes a good balance between bias and variance. However, this depends on the skill of the user. ML as often used, results in overfitting
- Training dataset ⇒ trained algorithm ⇒ test dataset ⇒ comparison of predicted vs. real values
- Options [see Chen et al. (2020)]
    - Regressions (standard, PLS, LASSO, Elastic net…)
    - SVM
    - Tree and random forest
    - Gradient boosting
    - Neural network
    - Deep neural network
    - Bayesian classification
- Specialized packages (R or Python)

TAKING DATA AVAILABLE FROM FARMERS AND PUBLIC SOURCES

- Integrating diverse data to create powerful predictive models
- FASF wHen2gO: farmer advisory for pesticides, combining data collection, field linked data, modeling and simple executable advice

ADDING VALUE WITH TRANSLATIONAL RESEARCH

- Model-based transfer learning, for example with earth observation data to train algorithm which can be tested on-farm

CROSS VALIDATION FOR MODEL SELECTION AND OPTIMIZATION

- The issue with spatial data is that they are correlated. If a random split is made to achieve two subsets, S1 and S2, it is likely that some of the data of S1 will be correlated with some data of S2. Thus, if the model is trained with S1 and then used to predict S2, the predicted data S2 is not independent from S1 and the validation may be biased (too optimistic).
- The best type of cross-validation depends on how the model is intended to be used. For example, if the intended use of the model is to predict new sites, it is advised to conduct a site-by-site cross-validation. If the intended use of the model is to predict new years on the same site, then a year-by-year cross-validation should be used. There are other types of cross-validation beyond site-by-site and year-by-year. The optimal choice always depends on the intended use of the model.

REFERENCES

Nix, H. 1984. Minimum data sets for agrotechnology transfer. In: Proceedings of the International Symposium on Minimum Data Sets for Agrotechnology Transfer, 21-26 March 1983, ICRISAT Center, India. Patancheru, A P. 502 324, India: ICRISAT. p 181-188.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

CONTRIBUTORS

COORDINATOR

Nicolas Tremblay, On-Farm Experimentation Community Co-lead, ISPA

OFE-C WEBINARS PRESENTERS

Medha Devare and Cheryl Porter: Enabling Digital Transformation of Agriculture by Overcoming the Agriculture Tower of Babel

Gerrit Hoogenboom: The Need and Resources for FAIR Data "Which Data to Collect and Why"

David Makowski: Data Agglomeration Among Farmers: Why and How?

Takashi Tanaka: Rethinking Experimental Designs and Data Analysis of On-Farm Experimentations

Richard Tiffin: Opportunities and Challenges for Translational Research from Big Data

OTHER CONTRIBUTORS

Norm Campbell

Ivana Ivanova

Anabelle Laurent

Alysa Gauci

Eiji Morimoto

Fernando Miguez

Corentin Leroux