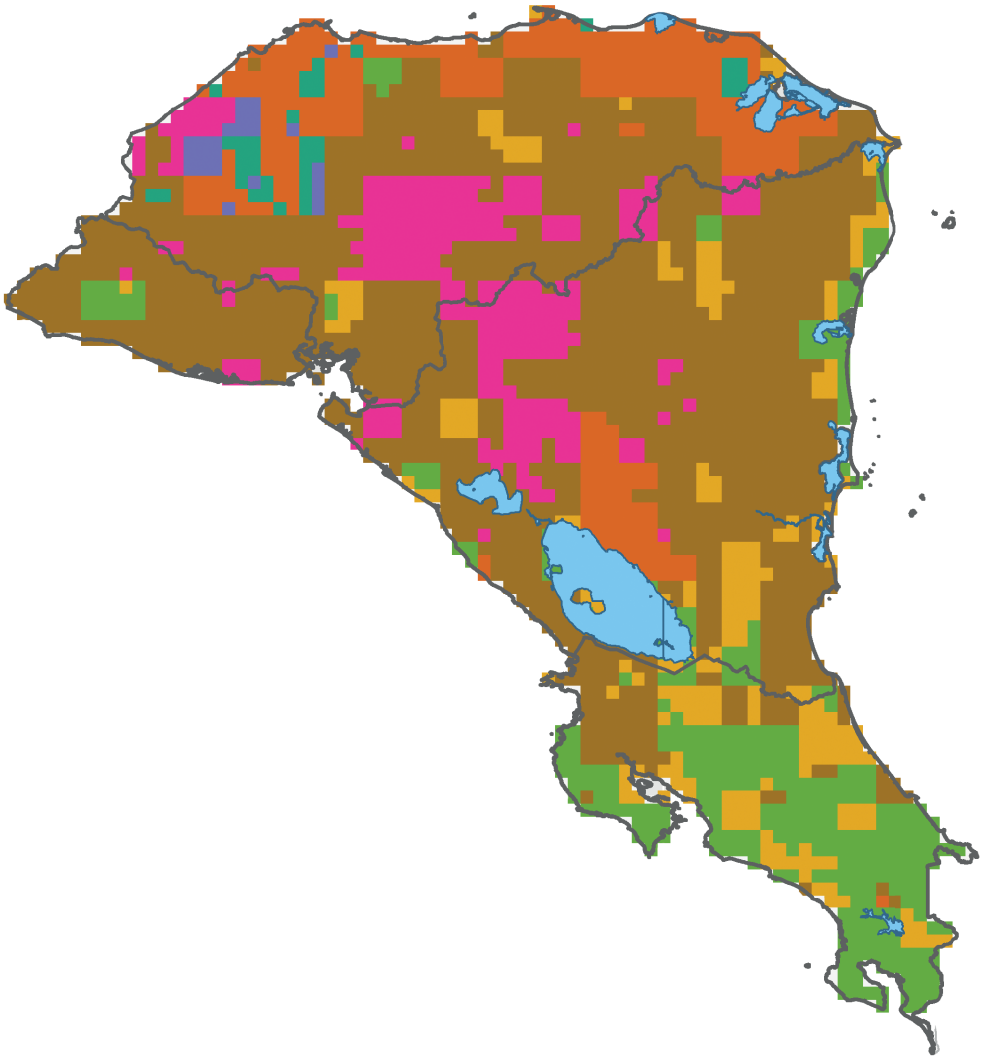
Data synthesis in crop variety evaluation: spatio-temporal characterization of environmental adaptation



David Brown Fuentes

Propositions

1. Environmental data enables spatio-temporal extrapolation of crop variety performance. (this thesis)

2. Model predictions of genotype performance are useless without uncertainty quantification. (this thesis)

3. Neglecting the role of nutrition in vitiligo pathogenesis hinders long-term results.

4. Scientific research in Latin America is hampered by cultural collectivism.

5. Major green-house gas emission reductions will be achieved by celebrating United Nations Climate Change Conferences outdoors.

6. Population-wide access to safe mobility and transportation is a fundamental pre-condition for social inclusion.

Propositions belonging to the thesis, entitled

Data synthesis in crop variety evaluation: spatio-temporal characterization of environmental adaptation

David Brown Fuentes

Wageningen, 20 December 2022

Data synthesis in crop variety evaluation: spatio-temporal characterization of environmental adaptation

David Brown Fuentes

Thesis committee

Promotor
Prof. Dr A.K. Bregt
Professor of Geo-information Science
Wageningen University & Research

Co-promotors
Dr S. de Bruin
Associate professor, Laboratory of Geo-information Science and Remote Sensing
Wageningen University & Research

Dr J. van Etten
Principal scientist and research director, Digital Inclusion
Alliance of Bioversity International and CIAT, Montpellier, France

Other members

Dr D.T.F. Endresen, University of Oslo Natural History Museum, Norway
Dr C.M. Gevaert, University of Twente, Enschede
Prof. Dr H.-P. Piepho, University of Hohenheim, Germany
Prof. Dr J.A. Ramirez-Villegas, Wageningen University & Research

This research was conducted under the auspices of the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC).

Data synthesis in crop variety evaluation: spatio-temporal characterization of environmental adaptation

David Brown Fuentes

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 20 December 2022
at 11 a.m. in the Omnia Auditorium.

David Brown Fuentes
Data synthesis in crop variety evaluation: spatio-temporal characterization of environmental adaptation,
171 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2022)
With references, with summary in English

ISBN: 978-94-6447-470-1
DOI: https://doi.org/10.18174/579737

Summary

Field crop production depends on environmental conditions, which vary across space and time. Climate change has increased the uncertainty about weather and environmental conditions. Therefore, farmers need information about which crop variety is adapted to local growing conditions. Also, farmers need to consider factors that go beyond productivity, such as market requirements and socioeconomic context. Such information about varieties is often lacking for farmers, especially in developing countries, and even for agricultural extension agents who assist farmers in decision-making. This information is also useful for crop improvement programs, which aim to develop new varieties with improved adaptation traits. The information relies on data from variety evaluation trials, which are resource-intensive exercises which usually generate a large amount of data. However, these data are often not fully taken into account to generate information that is useful for farmers, extension agents and breeders. Data synthesis could help to extract more value from existing data by aggregating trial datasets, enriching them with complementary data and generating actionable information. This thesis treats the development of new data synthesis methods that (1) account for environmental characteristics to provide better information supporting decision making in crop variety evaluation, and (2) make a more efficient use of crop variety evaluation data.

Chapter 2 is a literature review of the state of the art of data synthesis in crop variety evaluation, identifying the major constraints and knowledge gaps. Data from evaluations of crop varieties are produced every year around the world. The advent of digital technologies has increased the capacity of generating, processing, and storing data from agricultural experimentation. However, in some cases the data from agricultural experiments is incessantly piling up exceeding the analytical capacity. In other cases, the data is used once for a particular study and then remains dormant in storage devices. The literature review revealed that one of the main obstacles to integrate data from different variety evaluations is the heterogeneity among datasets, with different scales, data format and experimental designs. After reviewing the existing methods used in data synthesis, rank-based methods were identified as the most promising for the aggregation of heterogenous datasets and considering the environmental conditions in the analysis of crop variety performance. This is enabled by converting numerical measurements into rankings and applying statistical models for ranking data, which allows the use of environmental data as covariates.

Chapter 3 presents the application of the data synthesis approach to datasets produced with the triadic comparison of technologies (tricot) evaluating common bean (*Phaseolus vulgaris* L.) genotypes in four countries in Central America (Costa Rica, El Salvador, Honduras, and Nicaragua). In this case, the data synthesis has the aim to predict the best performing genotypes across the target region under different climatologies. The results show that it is possible to make such predictions by aggregating trial data from different organizations. Data were standardized to some extent because they were generated with the same approach

(tricot), but even then some data preparation is required, especially due to differences in genotypes names among trial datasets. A mapping approach was applied for the presentation of predictions, helping to visually interpret the results. A key contribution of this chapter is the quantification of uncertainty of model predictions within two complementary dimensions: (1) ranking probabilities and (2) environmental coverage of sampling data compared to the are covered by predictions.

Chapter 4 demonstrates how the data synthesis methodology can be applied to the analysis of data from different studies, which have different experimental designs, only partially overlap in evaluated varieties and have done measurements on different scales, using different measurement methods. This study focused on the reaction of *Musa* genotypes to black leaf streak disease (BLSD). This study converted several heterogeneous variables referring to a single phenomenon (BLSD incidence) to rankings. This is a strategy for data aggregation and analysis that has been earlier proposed and demonstrated by Simko and Linacre (2010); Simko and Pechenick (2010). The study presented in Chapter 4 further expanded this approach by analyzing the effect of climatic factors (precipitation, temperature, and relative humidity) on the reaction of *Musa* genotypes to BLSD. The results show that this kind of analysis allows to identify the differential response of the evaluated genotypes according to main climatic differences across regions. This study also demonstrates that making the trial data available through public repositories allows the integration and further (re)analysis, producing new insights supporting decision making.

Chapter 5 describes the R package *gosset*, developed to support data synthesis workflows in crop variety evaluations. The workflows implemented in the case studies of Chapters 3 and 4 were documented as R code and published in GitHub, accompanying the scientific papers of Chapters 3 and 4. Documenting the workflows and publishing the computer code allows its peer review, contributing to research reproducibility and scientific rigor, but also enables more efficient workflows allowing to reuse the computer code and follow a verified workflow. The documented workflows from the case studies were used to design a general workflow, identifying functionality that could be implemented as software tools. The main stages identified in the data synthesis workflow in crop variety evaluation are: (1) data management and preparation, (2) model validation, and (3) results presentation. The software tools were developed in the R environment for statistical computing and released as the R package *gosset* in the Comprehensive R Archive Network and its development version is available in GitHub, allowing code revision, reuse, and collaboration. The package *gosset* provides functionality that supports the three stages in the data synthesis workflow.

Chapter 6 describes the R package *ag5Tools*, which provides functionality to download and extract data from the Copernicus AgERA5 database, supporting stage 1 in the data synthesis workflow. Climatic data used as model covariates were fundamental applying the data synthesis approach in Chapters 3 and 4. The R package *ag5Tools* has been released in the Comprehensive R Archive Network and the code is available in GitHub.

Chapter 7 provides a general discussion of the main findings and the perspectives about the future work related to this thesis. This thesis contributes a new approach of data synthesis in the context crop variety evaluation, providing new methods that produce information to support decision making and at the same time an innovative way to optimize the use of data from agronomic and crop science experiments.

Table of Contents

Chapter 1 – General introduction

1.1 General introduction

Field crop production is intrinsically dependent on environmental conditions, which change across space and time. Climate change imposes challenging conditions to farmers, increasing uncertainty about weather, and the occurrence of extreme and devastating events (IPCC, 2022; Lesk et al., 2016). Decision making about which variety to plant is critical for farmers to successfully adapting crop production to environmental conditions (Bustos-Korts et al., 2019). Agricultural extension services aim to provide information and recommendation to farmers to improve their productivity (Anderson & Feder, 2007). A key element of those recommendations is which variety to plant, to achieve better results including adaptation to environmental conditions. Plant breeders develop new varieties aiming to deliver improved genetic material with better adaptation traits and higher productivity (Malosetti et al., 2013). Therefore, farmers, agricultural extension agents and plant breeders require information to support their decision making.

Crop varieties are evaluated mainly in field trials, often at several locations with different environmental conditions (Lecomte et al., 2010). Field evaluations of crop varieties are conducted around the world by both public and private organizations, producing substantial amounts of data (Yan, 2014b). The adoption of advanced digital technologies in agricultural experimentation has increased the capacity of generating field trial data (Janssen et al., 2017; Moore et al., 2021). For instance, by the time of publication of this thesis, the CGIAR platform GARDIAN accounts 26,474 datasets stored in 21 repositories. The increased capacity to generate, store and process crop variety trial data has enabled the application of analytical approaches based on ‘big data’, such as machine learning (Breiman, 2001), in contrast with the conventional statistical approaches which are mainly based on small-sample statistics. However, data from crop variety evaluations is often not fully utilized to generate information required by farmers, extension agents and plant breeders. Even when the trial data is used, it is often not re-used and then stored or even discarded without further utility. Trial data is costly to generate (Lecomte et al., 2010), requiring more efficient ways to extract more value per collected data unit. A suboptimal use of data creates a lack of trust at different levels in agricultural research. Participatory on-farm trials have the potential to provide more tailored solutions and recommendations to farmers. But if data collected in participatory research is not turned into useful information for farmers, they will not find a benefit in investing time and resources in more experiments. Furthermore, deficient information about variety performance may worsen the already slow variety adoption by farmers. Funding organizations will be averse to invest further resources generating and collecting data that is not converted into real solutions to farmers and the society in general.

Data synthesis can help to maximize the value from existing data integrating datasets from crop variety evaluations. In this thesis, data synthesis refers to the integration of data to generate new knowledge and insights (Carpenter et al., 2009; Pickett et al., 2007; Pillemer & Light, 1980), in contrast to other applications in which data synthesis refers to generating

synthetic data (Barth et al., 2018; Zhang, 2018). Systematic reviews and meta-analyses are the most common applications of data synthesis. Meta-analysis has been widely applied to human health, but also to other fields such as ecology, plant pathology and animal science (Philibert et al., 2012). In the case of crop variety evaluation, meta-analysis has focused mainly on variety performance in terms of yield and often without explicitly considering environmental factors. Decision making in crop variety evaluation involves several interconnected factors, such as climate, soil, pest and diseases, farmers preferences and needs, and market requirements. Data synthesis for crop variety evaluation should go beyond conventional meta-analysis, by integrating multiple types of data.

The integration of trial data from different sources is a complex task, because the data is often in different formats and measurement scales (Simko & Pechenick, 2010). Therefore, a data synthesis approach for crop variety evaluation requires a new methodological design. This new design should be accompanied by data availability, which is often constrained by deficient data sharing and management practices (Hyman et al., 2017; Moore et al., 2021). Data is often not made available by researchers or institutions because they do not find a direct benefit in it and it requires additional investments of time and resources (Diekmann, 2012; White & van Evert, 2008). Overcoming the resistance to data sharing requires tangible evidence of its benefits, instead of forcing its implementation. Demonstrations of feasibility of data synthesis, taking advantage of existing data and revealing new insights could encourage both scientists and institutions to actively engage in data sharing.

As a data-driven and computer-aided approach, data synthesis in crop variety evaluation requires well documented workflows, along with verifiable and reusable computer code. Documenting computer-based workflows enables reproducible and verifiable research results, but also can lead to the design of generalized workflows and required computer code. Generalized computer code can be implemented and released as software tools, supporting future applications of data synthesis in crop variety evaluation.

This thesis starts addressing the need of a new data synthesis approach by generating new insights in the current limitations and possibilities in the area of data synthesis applied to crop variety evaluations (Chapter 2), providing two case studies where the new data synthesis approach is applied (Chapters 3 and 4), and documenting the data synthesis workflow and developing software tools supporting future applications of the new data synthesis approach (Chapters 5 and 6).

1.2 Research objectives

The main objective of this thesis is to develop data synthesis methods producing location-specific information of crop variety performance, that helps decision makers involved in crop variety evaluation processes and make better use of collected data. To that aim, this thesis addresses the following specific research objectives:

1. To describe the state of the art of data synthesis methods used in crop variety evaluation identifying the major constraints and knowledge gaps.
2. To design and implement a data synthesis approach that characterizes environmental adaptation of crop varieties in time and space, and assess its strengths and limitations through two case studies.
3. To design and document reproducible synthesis workflows for crop variety trial data.

1.3 Thesis outline

This thesis is structured in 7 chapters. Chapter 1 presents the introduction, main research goal and research objectives. Figure 1.1 shows how Chapters 2 to 6 contribute to address the research objectives and how these objectives depend on each other.

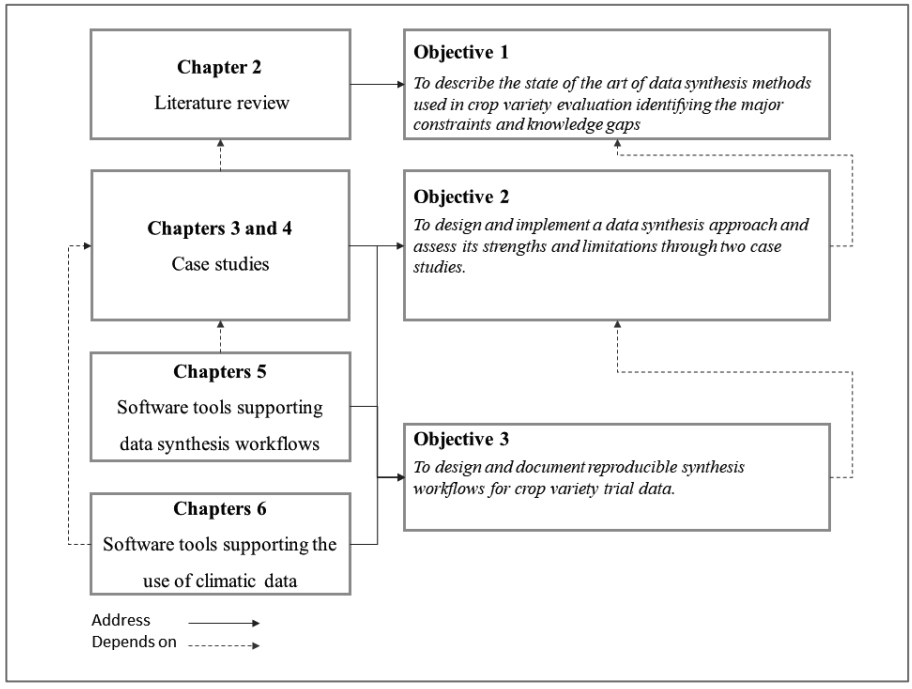


Figure 1.1 Relationship among Chapters 2 to 5 and the research objectives

Chapters 2 to 6 present how the research objectives were addressed. In Chapter 2 a literature review was conducted identifying the main elements required for data synthesis in crop variety evaluation, such as data types and models.

In Chapter 3, the data synthesis approach was applied to participatory on-farm trials established in four countries in Central America (Costa Rica, El Salvador, Honduras, and

Nicaragua), to predict the top-three performing genotypes across the study region. A spatial mapping approach was used to visualize the model predictions along with its estimated uncertainty.

In Chapter 4, the data synthesis approach was applied to trial data from multiple sources, with different scales, experimental designs and partial overlap among evaluated genotypes, to assess the effect of climatic factors on the reaction of *Musa* genotypes to black leaf streak disease.

Chapter 5 describes the R package gosset, developed to support the data synthesis workflow. The documented data synthesis workflows, implemented in Chapters 3 and 4, were used to design a general workflow which can be applied in future data synthesis. The core functionality identified in the generalized workflow was implemented as an open-source software and released for public access and free of costs.

Chapter 6 describes the R package *ag5Tools*, developed to support downloading and extracting climatic data from the AgERA5 dataset, which is fundamental for the use of climatic model covariates in data synthesis for crop variety evaluation approach.

Chapter 7 presents a general discussion summarizing the main findings and reflecting on the main weaknesses, finishing with a perspective of potential research lines for future work, based on the identified study weaknesses and remaining knowledge gaps.

Chapter 2 - Data synthesis for crop variety evaluation. A review

This chapter is based on:

Brown, D., Van den Bergh, I., de Bruin, S., Machida, L., & van Etten, J. (2020). Data synthesis for crop variety evaluation. A review. *Agronomy for Sustainable Development*, 40(4), 25. https://doi.org/10.1007/s13593-020-00630-7

Abstract

Crop varieties should fulfill multiple requirements, including agronomic performance and product quality. Variety evaluations depend on data generated from field trials and sensory analyses, performed with different levels of participation from farmers and consumers. Such multi-faceted variety evaluation is expensive and time-consuming; hence, any use of these data should be optimized. Data synthesis can help to take advantage of existing and new data, combining data from different sources and combining it with expert knowledge to produce new information and understanding that supports decision-making. Data synthesis for crop variety evaluation can partly build on extant experiences and methods, but it also requires methodological innovation. We review the elements required to achieve data synthesis for crop variety evaluation, including (1) data types required for crop variety evaluation, (2) main challenges in data management and integration, (3) main global initiatives aiming to solve those challenges, (4) current statistical approaches to combine data for crop variety evaluation and (5) existing data synthesis methods used in evaluation of varieties to combine different datasets from multiple data sources. We conclude that currently available methods have the potential to overcome existing barriers to data synthesis and could set in motion a virtuous cycle that will encourage researchers to share data and collaborate on data-driven research.

2.1 Introduction

Farmers, especially smallholders in developing countries, are facing ever more challenging production conditions and product requirements. Extreme weather events are on the rise as one of the effects of climate change (Coumou & Rahmstorf, 2012; Lesk et al., 2016). Emerging pests and diseases, as well as declining soil fertility, are also constraining farm productivity. Evolving crop production practices require the development of new genotypes that meet specific agronomic traits (Collard & Mackill, 2008). Markets are also evolving, and taste preferences need to be considered if new crop varieties are to easily find their way to the consumer (Dawson & Healy, 2018). Furthermore, there is growing knowledge of the different product needs and preferences relative to gender, which are influenced by their different roles in the value chain, differences in access to land and other inputs and differences in decision-making power (Christinck et al., 2017). There is also an increasing interest in more sustainable crop production systems, which would require a redesign of the whole food system and the role of players involved, including breeders (Lammerts van Bueren et al., 2018). Crop improvement aims to address the multiple challenges faced by farmers through delivering improved varieties (Malosetti et al., 2013). However, simply using the most recently released variety will not always lead to improvement, as breeding cannot address all requirements in all contexts. Decision-makers involved in crop improvement, including breeders, agronomists and farmers, evaluate multiple aspects and trade-offs relevant to the context in which they use the varieties. Crop variety evaluation is critical in decision-making in crop variety release, crop seed marketing or distribution and generating crop variety recommendations for farmers.

Crop variety evaluation is mainly conducted through field trials (Fig. 1), which are expensive and time-consuming (Kipp et al., 2014; Lecomte et al., 2010; Tenkouano et al., 2012). The limitations in resources, space, time and the required logistics in field trials also make it almost impossible to test all the varieties of interest in the same trial or in all the possible environments (Lecomte et al., 2010; Simko et al., 2012; Singh et al., 2014). Crop variety evaluation usually considers yield as the main trait while disease resistance and climate adaptation are secondary traits. Other characteristics of interest in crop variety evaluation, such as product quality and consumer preferences, are obtained through quality assessments and sensory evaluations, which are also expensive (Tomlins et al., 2004). An exception in terms of costs of data relevant to crop variety evaluation is climatic data, which acquiring costs have been decreasing due to advances in remote sensing and computational power.

Figure 2.1 NARITA hybrid field trials in Mbarara, Uganda, mulched with swamp grass to reduce weeds and soil moisture loss. Photo credit: Bioversity International/L. Machida

Crop variety evaluation has not always kept pace with the growing complexity of agricultural production and the growing availability of data. As a data-driven type of research, crop variety evaluation can benefit from multiple revolutions occurring in several fields such as genomics, phenomics, big data and machine learning(Bolger et al., 2019; Esposito et al., 2020; Tardieu et al., 2017; van Etten, Steinke, et al., 2017). These revolutions are driven by increased data storage and computing capacity, the availability of sensors, improved DNA sequencing technologies and new field data collection approaches, such as high-throughput and high-precision field phenotyping and crowdsourcing (Chawade et al., 2019; Esposito et al., 2020; Reynolds et al., 2020; Tardieu et al., 2017; van Etten, Beza, et al., 2019). This has caused not only a quantitative leap in data volumes but also a shift to 'big data' approaches that move beyond small-sample statistics to data analysis based on machine learning (Breiman, 2001; Ersoz et al., 2020; Thessen, 2016; van Etten, Steinke, et al., 2017). While there are multiple examples of useful applications of big data analysis in agriculture (Kamilaris et al., 2017; Liakos et al., 2018), such cases are still few compared to other industries (Kamilaris et al., 2017).

Specifically, crop variety evaluation has taken little advantage of the potential benefits of data synthesis. Data synthesis allows the combination of data from different sources, producing new information and knowledge to support decision-making (Carpenter et al., 2009; Pickett et al., 2007; Pillemer & Light, 1980; Wyborn et al., 2018).

Interest in the value from combining and (re)using datasets in agriculture has grown, supported by open data and data sharing initiatives (Leonelli et al., 2017). As new analytical technologies and methods become available, legacy data could be reanalyzed (Hampton et

al., 2013; White & van Evert, 2008). Data synthesis can improve the efficiency of data use in crop variety evaluation by combining and repurposing new and legacy data from field trials, environmental measurements, farmer requirements and consumer preferences (Figure 2.2). In the agricultural sciences, it has so far mainly taken the form of meta-analysis (Krupnik et al., 2019; Philibert et al., 2012).

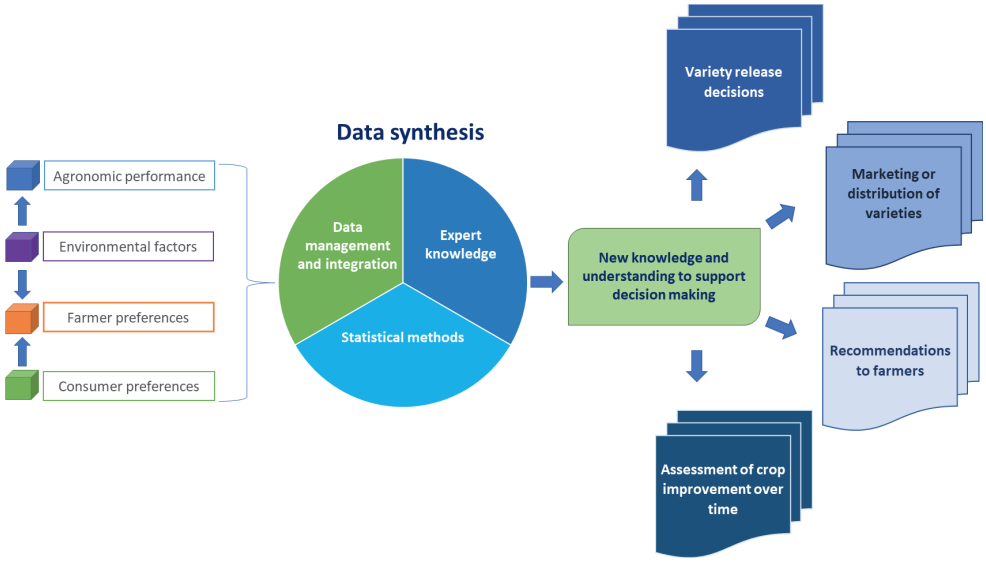


Figure 2.2 Different elements and processes involved in data synthesis for crop variety evaluation

Data synthesis can play a role in different functions of crop variety evaluation. The selection of genotypes to be released as cultivars can benefit from data synthesis to assess genetic gain (progress over time) (Streck et al., 2018), to benchmark against other breeding programs, to improve accuracy through multi-season assessments and to predict performance beyond the trial environments. The latter involves analyzing a combination of variety trial data and environmental data (Hyman et al., 2013). The analysis of trial data can be made more accurate when data from the last trial season is combined with historical variety performance data (Arief et al., 2015).

To release a new variety, breeders need to evaluate the proposed genotypes against existing varieties in a country or region. Data synthesis could facilitate enriching data from trials including the new varieties with data on the past performance of the older varieties to gain accuracy (Damesa et al., 2017). Seed companies need to assess variety performance to take seed production and marketing decisions. Service providers, such as agro-input suppliers, cooperatives, agricultural extension organizations and NGOs, need to make recommendations to farmers, considering the multiple dimensions of variety performance (and trade-offs between these dimensions) in different environments and under different types of crop management. Information from existing crop trials to formulate

recommendations is often used for this end. Such an analysis of existing data could benefit from data synthesis if the data that is available comes from different sources. Some service providers produce their own data about on-farm variety performance to generate recommendations or refine existing ones, which could also benefit from data synthesis to combine the new data with existing data. In many contexts, variety evaluation is done in a fragmented way (Rangarajan, 2002), which can preclude centralized coordination or standardization of data collection and weakens variety evaluation as each entity assesses genotype by environment interactions in a limited set of environments. Data synthesis could help to gain a better understanding of genotype by environment interactions across space and time. Flexible data synthesis could take advantage of heterogeneous data from different actors in the seed sector and provide value to the several functions that variety evaluation plays in each step of the crop improvement cycle.

A clearer perspective on data synthesis for crop variety evaluation is needed to achieve these potentials. Here, we review the literature relevant to data synthesis for context-specific decision-making in variety management. The objective of this article is to provide an overview of the required elements, current approaches and research gaps in data synthesis for crop variety evaluation, focusing on decision-making for variety pre-release and post-release. We limit ourselves to these later stages of the breeding process, and therefore, we do not cover the genomic and high-throughput phenotyping data. Even though these types of data are clearly part of the data revolution in crop improvement, they generally concern early and intermediate stages of the breeding process. We briefly refer to high-throughput field phenotyping data, as it has the potential to support later stages of the breeding process. In Section 2.2, we discuss the types and sources of data that are required. In Section 2.3, we discuss how data synthesis relies on proper data management, including sharing data across different trials and the compatibility of datasets. Data synthesis requires not only combining datasets to assess variety performance but also beyond assessing average performance, a careful analysis of how different genotypes respond to diverse environments and match the preferences of farmers, consumers and other stakeholders. Therefore, in Section 2.4, we review how data analysis is currently dealing with the end-users, their context and what is still lacking to evaluate crop varieties through a data synthesis approach. In Section 2.5, we review existing data synthesis approaches used in crop improvement and assess how they can be enhanced to include use context. In Section 2.6, we present our conclusions and recommendations.

2.2 Data required for crop variety evaluation

In this section, we describe the data types required by a data synthesis approach for crop variety evaluation. Field trial data are important to analyze the phenotypic response of a given genotype, to the environmental characteristics of the testing location and, in some cases, to management practices. Not only yield but also product quality is considered in variety evaluation. The evaluation of crop varieties also involves data about the preferences of

farmers obtained from participatory and on-farm trials, and consumers, obtained from sensory evaluations.

2.2.1 Agronomic performance data

Agronomic performance data are collected from field trials, which can be set up in several ways depending on the context and purpose. A rough classification of contexts includes (1) public international breeding programs (e.g., breeding programs within the CGIAR, (2) private breeding programs at commercial seed companies and (3) agricultural research at national or regional level, conducted by National Agricultural Research Systems, often in partnership with International Agricultural Research Organizations.

Field trials of breeding programs are usually known as performance trials or yield trials, given the importance of yield as the main trait (Acquaah, 2012). There are two main types of yield trials: (1) breeder trials and (2) official trials (Acquaah, 2012). Breeder trials aim to assess the performance of a set of genotypes to decide which ones should be released as cultivars (Priyadarshan, 2019). An official variety trial is part of the variety release and registration process, which varies among countries, but in most of the cases, it is conducted by an independent body, such as an official seed agency or under the jurisdiction of a variety release committee. Depending on the stage of the breeding process, the breeder's trials can be divided into preliminary yield trials (PYTs) and advanced yield trials (AYTs) (Priyadarshan, 2019). A PYT often concerns many genotypes (and few replications), whereas an AYT evaluates a small number of genotypes (selected from the PYT), with more replications over different environments, and during several years (Priyadarshan, 2019). In this review, we are focusing on data generated from AYT. Crop variety trials can also be established to test improved varieties to be recommended to farmers (Yan, 2014b). Both for breeding and to generate variety recommendations, field trials aim to evaluate varieties in different environments, where the environment is considered a combination of location and season (Acquaah, 2012). For this purpose, crop variety trials can be established mainly in three different levels of combination of location and season: (1) a single location in a single season, (2) multiple locations in a single season and (3) multiple locations in multiple seasons (Yan, 2014a). Crop variety trials can also be established in a centralized or decentralized system. The centralized approach involves on-station trials; it is the conventional approach for many crops and contexts. Decentralized methods include establishment of trials at farm locations, with different levels of participation from farmers. As not only biophysical factors determine the suitability of a variety, it has been suggested that the concept of environment should be extended to include the socioeconomic context of the target location (Desclaux et al., 2008). Participatory plant breeding and participatory varietal selection methods aim to better consider farmers' preferences and context in order to increase the adoption of improved varieties (Ceccarelli & Grando, 2007; Weltzien & Christinck, 2017). These approaches often include participatory on-farm trials, which may produce insights that are complementary to insights derived from conventional trials (Atlin et al., 2001; Coe, 2007; Coe, 2002). Although

farmers' participation is commonly associated with on-farm trials, farmers can also be involved in on-station trials. Participatory varietal selection (PVS) can be done through mother-baby trials, where the mother trial includes the full set of testing genotypes and the baby trial only includes a subset of test genotypes alongside the control genotype (Snapp, 2002; Virk et al., 2009). On-farm trials may produce unbalanced data, due to differences in the particular conditions of farmers' fields and the limited availability of seeds of the new varieties (Virk et al., 2009).

Data collected from field trial evaluations typically include trial design, the trial location, the date of establishment, trial management, evaluated genotypes and observations of the target traits (e.g., yield). Observations of the target trait can be either measured or estimated and should be ideally referenced to the observation date and to the phenological stage of the plant during observation (Billiau et al., 2012; Germeier & Unger, 2019; White et al., 2013). For instance, the second phase of the International Musa Testing Program (IMTP) used the following attributes (Orjeda, 2000): genotype, time from planting to shooting (days), time from shooting to harvest (days), time from planting to harvest (days), height of the pseudostem at shooting (cm), height of the following sucker at harvest (cm), bunch weight (cm), number of hands per bunch (hands), total number of fingers per bunch (fingers), average fruit weight (g) and leaf emission rate.

Technological innovations allowed the development of new methodologies for collecting data from field trials. These include high-throughput field phenotyping methods supported by satellite imagery or data from unmanned aerial vehicles (UAVs) and proximal phenotyping (Chawade et al., 2019).

Given the multiple context and evaluation objectives, each organization conducting crop variety trials may use its own experimental design and employ different methods and technologies for collecting, storing and publishing and/or sharing data. In Section 2.3, we review potential obstacles for data integration resulting from these differences. Furthermore, the diversity of goals and evaluation methodologies also produce different approaches to analyze collected data. We review these in Section 2.4.

2.2.2 Environmental data

To analyze the phenotypic response of genotypes to the testing environment, a fundamental step is to characterize the environment. The environment is the first source of yield variability in plant breeding trials (Chenu, 2015). Hence, environmental data are required to characterize the trial location and to understand its influence on the performance of tested genotypes in that particular location. It is known that the use of environmental data as model covariates analyzing multi-location trial data improves the degree of accuracy in the prediction of genotype performance (Piepho et al., 1998; Piepho, 2000). A recent study by van Etten, de Sousa, et al. (2019) demonstrated an improvement in variety recommendations using seasonal climate data as model covariates. Xu (2016) proposed to consider all environmental

factors affecting growth and production of plants through an approach called 'envirotyping'. Environmental data can be collected at trial sites directly. But even if environmental data were not collected during trials, geolocating trial sites allows enriching the dataset with existing environmental data. Adding environmental data to legacy trial data allows comparisons among trials conducted at different locations. Some types of environmental data are increasingly available through open and public repositories (Hyman et al., 2013). For instance, data on rainfall, temperature, elevation and soils are openly and freely available from open and public databases such as Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) (Funk et al., 2015), MODIS Land Surface Temperature (Wan et al., 2015), Hole-filled SRTM for the globe version 4 (Jarvis et al., 2008) and SoilGrids (Hengl et al., 2017). The European Centre for Medium-Range Weather Forecasts (ECMWF), through the Copernicus Climate Change Service (C3S), provides a comprehensive collection of climatic datasets, including the recently deployed 'Agrometeorological indicators from 1979 to 2018 derived from reanalysis', known as AgERA5. Available climatic data can be used to calculate climatic indices (Table 2.1), which were proven to be useful as model covariates in the analysis of crop variety trials (Kehel et al., 2016; van Etten, de Sousa, et al., 2019). Even though climatic data opens a wide range of possibilities for crop variety evaluation, the resolution of available data has to be carefully considered (Parkes et al., 2019).

Table 2.1 Temperature and precipitation indices commonly used as covariates in crop variety trial analyses. Adapted from van Etten, de Sousa, et al. (2019) and Kehel et al. (2016)

Environmental Index	Unit
Maximum daytime temperature	°C
Minimum daytime temperature	°C
Maximum nighttime temperature	°C
Minimum nighttime temperature	°C
Mean difference between daily maximum temperature and daily minimum temperature	°C
Number of days with maximum temperature > 30 °C	days
Number of nights with maximum temperature > 25 °C	days
Maximum length of consecutive days with precipitation < 1 mm	days
Maximum length of consecutive days with precipitation ≥ 1 mm	days
Number of days with precipitation > 5 mm	days
Number of days with precipitation > 10 mm	days
Maximum 1-day precipitation	mm
Maximum 5-day precipitation	mm

2.2.3 Food quality and consumer preference data

Sensory and nutritional evaluation has received more attention in recent decades, countering the narrow focus of crop improvement on yield, disease resistance and uniformity (Folta & Klee, 2016). At present, consumer markets are evolving, with consumers seeking additional product qualities such as nutritional and sensorial characteristics (Folta & Klee, 2016). Food quality involves both objective and subjective analyses, involving measurements of contents, texture as well as sensory analyses. Sensory evaluation is formally defined as 'a scientific discipline used to evoke, measure, analyze, and interpret reactions to those characteristics of foods and materials as they are perceived by the senses of sight, smell, taste, touch, and hearing' (Anonymous, 1975; Stone et al., 2012). Consumer preference data are obtained from sensory evaluations by panels of regular or specialized consumers, with different methods, such as descriptive analysis or rapid sensory evaluations (Dawson & Healy, 2018). Sensory and hedonic (i.e., related to pleasant or unpleasant) experiences cannot be measured directly and should be inferred from descriptive or numerical representations (hedonic scales) of subjects' responses (Lim, 2011). There are four main types of scales used in hedonic scaling, which are presented in Table 2.2.

Table 2.2 Types of scales used in hedonic scaling. Adapted from Stevens (1946) and (Lim, 2011)

Scale	Basic empirical operation	Number usage	Permissible statistics	Example scale
Nominal	Determination of equality (categorization)	As labels	Non-parametric: number of cases; Mode	1: good, 2: bad
Ordinal	Determination of greater or less	To recognize the rank order	Non-parametric: median; percentiles	Rank rating
Interval	Determination of equality of intervals or differences	To represent degrees of differences	Parametric: mean; standard deviation	Category scale
Ratio	Determination of equality of ratios	To represent relative proportions	Parametric: log mean; standard deviation	Labelled affective magnitude scale (LAM); labelled hedonic scale (LHS)

The 9-point hedonic scale developed by Peryam and Girardot (1952) is the most widely used method for scaling consumer preference and food acceptability (Lim, 2011). It is composed of the following values and their correspondent description: 9, like extremely; 8, like very much; 7, like moderately; 6, like slightly; 5, neither like nor dislike; 4, dislike slightly; 3, dislike moderately; 2, dislike very much; and 1, dislike extremely (Peryam & Girardot, 1952). More recently, other scaling methods have been proposed, such as the labelled affective magnitude (LAM) scale (Schutz & Cardello, 2001) and labelled hedonic scale (LHS) (Lim et al., 2009). This diversity of measurement scales can pose a challenge to combine data from different sources, such as different laboratories testing food quality, or a sensory evaluation with farmers testing different varieties as part of a breeding process.

Agronomic performance, food quality and preference data are still expensive and complex to acquire and manage. In contrast, weather and soil data are increasingly available at significantly reduced costs. More effort is required to improve the efficiency in data use in the evaluation of crop varieties. The higher availability of weather and soil data can motivate and increase data reuse, repurposing legacy crop variety trial data by adding environmental data to extract new insights.

2.3 Data management challenges

As a data-driven research, data synthesis requires availability of the data to be reused. It also requires careful data management to integrate data of heterogeneous nature from different sources and formats, as described in Section 2.2. Here, we discuss the challenges in data

management that are relevant to data synthesis for crop variety evaluation, the main efforts to address these problems and further research needs.

2.3.1 Main barriers for data availability and integration

Data should be available to be integrated and then synthetized using formal statistical techniques. However, data are still rarely shared for reuse, especially in the case of raw data from variety trials. Diekmann (2012) and (Williams, 2012) found that researchers are often unwilling to share raw data out of concern about data being taken out of context, which could lead to incorrect results and misinterpretation. Authors may also oppose freely releasing data that cost them substantial work and resources, whereas they may be more willing to share data with colleagues Diekmann (2012).

In addition to cultural constraints, technical challenges arise for data sharing among both individuals and institutions. It has been a common practice for researchers and research centers to develop their own system for storing data, mainly because they do not trust global repositories with which they do not have a direct relationship and for which long-term support may not be guaranteed (Leonelli et al., 2017). This has resulted in a myriad of individual databases that are neither open nor compatible among research centers. This not only inhibits collaboration among scientists but also promotes duplication of efforts and increases costs, a luxury that the scientific community cannot afford in times of scarce economic resources for agricultural research.

In cases when data are available, data integration sometimes encounters problems due to lack of standardization in terms of syntax, semantics and structure. Crop variety trial datasets are often very heterogeneous in terms of quantity, quality, types and formats (Hyman et al., 2017; Leonelli et al., 2017). Individual trial designs and observational methods vary according to the specific purpose of trials (Section 2.2). This lack of standardization of crop variety trial data makes it difficult to compare results between trials and to reuse datasets with traditional tools (Leonelli et al., 2017; Rijgersberg & Top, 2000). Combining crop trial datasets often presents the following problems: (1) incomplete or inexistent overlap among evaluated accessions across trials, (2) measurements based on different rating scales and (3) the use of different methods for observing the same trait (Simko & Pechenick, 2010). Methods to solve the problem of measurement in different rating exist such as the threshold model (Hartung & Piepho, 2005), but it does not solve the problem of partial overlap among tested varieties in the different trials. In Section 2.5, we explore and evaluate how different data synthesis methods deal with this kind of problems.

The dearth of relevant data in the public domain limits the possibilities of data synthesis, as it provides a large initial cost of assembling, cleaning and reformatting data. For individual data synthesis efforts, this initial investment may be relatively very high, even though it could be worthwhile if data can be repurposed more than once. Furthermore, practices limiting data

reuse go against the aim of science of building universal knowledge, in which public funds play a fundamental role.

2.3.2 Efforts to overcome data management problems

Several international agricultural research organizations have made efforts to address the barriers of poor data availability and data format incompatibilities (Germeier & Unger, 2019; McLaren et al., 2005; Ritchie, 1995). Several global funders of agricultural research are increasingly seeking mechanisms to guarantee that research investments generate benefits for smallholder farmers in developing countries (Dalrymple, 2008). Global open data and sharing initiatives aim to facilitate data accessibility, allowing the generation of new knowledge (Wilkinson et al., 2016). With valuable contributions from diverse partner organizations, CGIAR centers have been developing information systems and platforms, aiming to integrate heterogenous data sources to support crop improvement research (Hyman et al., 2017; McLaren et al., 2005). Examples of this type of systems are the International Crop Information System (ICIS) and its derivatives, the International Rice Information System (IRIS) and the International Maize Information System (McLaren et al., 2005; Shrestha et al., 2010). The recently created CGIAR 'Platform for Big Data in Agriculture' aims to materialize the potential of big data–related methods and technologies to improve agricultural production. Outside the CGIAR system, agricultural researchers and organizations are also endeavoring to construct better ecosystems of data and methods. Table 2.3 contains a compilation of the main international initiatives on standardization and data sharing. Data standardization and sharing systems include online databases such as AgTrials, YamBase, CassavaBase and MusaBase, which all implement ontologies to standardize vocabularies and terminologies (see Table 2.3). Ontologies formally define the relationships among concepts within a given domain (Matteis et al., 2013). Similar approaches have also been proposed by other authors. For instance, Rijgersberg and Top (2000) proposed data model templates—a generalization of data models—to achieve a balance between standardization and flexibility. See Spyns et al. (2002) for an overview of specific differences between data models and ontologies. Germeier and Unger (2019) applied a modelling approach that goes further than data models, considering also statistical models in the implementation of a phenotyping information system. Efforts to standardize phenotyping data include the Minimal Information About Plant Phenotyping Experiment (MIAPPE) (Ćwiek-Kupczyńska et al., 2016; Krajewski et al., 2015; Papoutsoglou et al., 2020). Efforts to standardize data from field experiments include the International Consortium for Agricultural Systems Applications (ICASA) standard, initially developed by the International Benchmark Sites Network for Agrotechnology Transfer (IBSNAT) and updated by ICASA (White et al., 2013).

Table 2.3 Main international initiatives to increase data sharing and standardization in agriculture

Initiative	URL
AgTrials	http://agtrials.org
AgroPortal	http://agroportal.lirmm.fr
Breeding Application Programming Interface (BrAPI)	https://brapi.org
Breeding Management System	https://bmspro.io
CassavaBase	https://www.cassavabase.org
Crop Ontology	http://www.cropontology.org
GARDIAN	http://gardian.bigdata.cgiar.org
Global Open Data for Agriculture and Nutrition (GODAN)	https://www.godan.info
Global Trial Data Management System	https://research.cip.cgiar.org/gtdms
Integrated Breeding Platform	https://www.integratedbreeding.net
MIAPPE	https://www.miappe.org
MusaBase	https://musabase.org
Sol Genomics Network	https://solgenomics.net
YamBase	https://yambase.org

The FAIR (findability, accessibility, interoperability and reusability) guiding principles are intended for producing and publishing data, aiming to facilitate and enable data sharing and reuse (Wilkinson et al., 2016). To achieve these principles, a set of mechanisms is considered. These include unique and persistent identifiers, a standardized communications protocol and the use of domain-specific standards for both data and meta-data.

Despite these efforts, recent literature indicates that there are still serious challenges in implementing data standardization and sharing. About 85% of the more than 35,000 records in the AgTrials database contain only meta-data; hence, those interested in the underlying data should contact the original data provider (Hyman et al., 2017). Problems persist on both the supply and demand side. It has been found that researchers are often reluctant to use data produced by others because there is no guarantee about the quality (Diekmann, 2012). Data standards are now available (Table 2.3), but it is still difficult to persuade the agricultural research community to adopt the suggested standards. The lack of flexibility to adapt to scientific progress is indeed one of the arguments stated against standards (Germeier & Unger, 2019). For both new and legacy data, standardization requires considerable efforts, which will not immediately pay off, hindering its implementation. In addition, the efforts required for processing datasets with the accompanying meta-data to facilitate open access

and use by others are often not acknowledged (Diekmann, 2012; White & van Evert, 2008). From 112 surveyed users of the AgTrials platform, 34% considered that their data are incorrectly organized to be shared publicly (Hyman et al., 2017). In the cases where data are shared, it is mostly not through global repositories but rather as supplementary data on the journal website where associated articles were published (Williams, 2012). Sharing data as supplementary materials on a journal website would be adequate if journals followed commonly agreed guidelines, such as the FAIR principles. Although there has been an increase in data shared by researchers during the recent years, there is a still a lack of awareness and hence compliance with FAIR principles (Mark et al., 2018).

2.3.3 Further work to improve data availability and data integration

We identified barriers for data availability and data integration. The main efforts to stimulate data open access and repurposing have focused on compliance with standards and data sharing as a goal. Given the modest progress so far, we suggest that this focus be complemented with efforts to make data sharing more appealing, by the stimulation of data demand for data synthesis. This might set in motion a virtuous cycle of collaboration around data synthesis, providing clear incentives in the form of authorships on joint publications and citations to datasets (Figure 2.3).

Figure 2.3 A virtuous cycle set in motion by data synthesis

Therefore, we posit a need for simple methods that can deal with highly heterogeneous datasets to start to show the potential benefits of data synthesis (see Section 2.5). Ecology

research is a concrete example of how increasing the number of datasets, publications and collaboration among larger groups of scientists through meta-analysis can result in even larger collaborative initiatives which enhance the scope and potential impact of research (Cadotte et al., 2012). Data journals may help to boost this kind of data-centered collaboration (Candela et al., 2015). This could provide further motivation and feedback to continue the consensus building processes around data management.

2.4 Analysis of different types of data

Data synthesis requires the combination and analysis of different types of data as described in Section 2.2. Here, we review current approaches to combine those types of data and research gaps to achieve data synthesis for crop variety evaluation.

2.4.1 Multi-environment trial analysis

As we discussed in Section 2.2, the most basic evaluation of genotypes is the assessment of their agronomic performance, such as plant growth parameters, yield, plant reaction to diseases and tolerance to climate conditions (e.g., tolerance to drought, cold and flooding), among others. This type of evaluation is conducted through field trials, which can be established in different ways. Conventional trials are usually established in research stations under controlled conditions. Trials can also be established on farms and involve different levels of farmer participation, from limited participation as observers to full participation as citizen scientists (Ceccarelli, 2012; Ceccarelli et al., 2009; van Etten, Beza, et al., 2019). Combinations are also possible, such as on-station trials with some level of farmer participation. Regardless of the trial design, the idea is to establish trials at different locations for several growing seasons. The combination of location and time is known as the 'testing environment', and the trials are known as 'multi-environment trials' (METs). Multi-environment trials are conducted to evaluate the suitability of crop genotypes for different agroecological conditions (van Eeuwijk et al., 2005).

Genotype × environment (G × E) interaction is the relative difference in phenotypic response that a group of genotypes expresses depending on the environmental conditions (de Leon et al., 2016). Hence, G × E assessment requires the evaluation of a minimum of two different genotypes in at least two different environments (Kang, 1997). The phenotypic response of a genotype to the environment is described by a function known as the reaction norm (Bustos-Korts et al., 2019). When the reaction norm lines of evaluated genotypes in different environments are not parallel, there is presence of G × E (Bustos-Korts et al., 2019). Especially in conventional breeding, G × E is considered a challenge by breeders due to its implications for genotype selection (Kang & Gorman, 1989). Aiming for more specific adaptation, decentralized breeding programs take advantage of G × E instead of diminishing the effects (Ceccarelli, 1989).

There are different statistical models for G × E analysis. For an overview, we refer to recent reviews such as the work of Malosetti et al. (2013), van Eeuwijk et al. (2016) and Bustos-

Korts et al. (2019). Most of the statistical models for G × E analysis can be interpreted as (phenotypic) response functions for each genotype to environmental variables (van Eeuwijk et al., 2005). Therefore, G × E analysis represents a combination of two of the data types presented in Section 2.2: the agronomic performance data and the environmental data.

Data from multi-environment trials are usually summarized in two-way tables of means, with genotypes as rows, and environments as columns (Malosetti et al., 2013). Two major groups of statistical models for the analysis of G × E can be identified: (1) methods that only use the two-way table of means and environmental information are included only implicitly (usually as dummy variables) and (2) methods that use additional information, explicitly included as genotype and/or environment covariates (temperature, rainfall, etc.) (Malosetti et al., 2013). Examples of G × E models from the first group include additive models (ANOVA), regression on the mean (Finlay & Wilkinson, 1963; Yates & Cochran, 1938), additive main effects and multiplicative interaction (AMMI) models (Gauch, 1992) and the genotype + genotype × environment (GGE) model (Yan et al., 2000). Since they only require a two-way table of means as input, these models are considered to be good for descriptive and explorative purposes, but not for explaining G × E (Malosetti et al., 2013). The second group of models includes factorial regression, partial least squares regression, structural equation models and mixed effect models. Factorial regression allows the use of environmental or genotypic variables as covariates to explain G × E but has the limitation that only permits one dependent variable at a time (Vargas et al., 2007). Another limitation of factorial regression is its difficulty in dealing with multi-collinearity when several covariates are used (Vargas et al., 1999). For these cases, partial least squares regression is a more convenient approach, as it can easily handle multiple explanatory variables (Vargas et al., 1999). When the cause-effect analysis of G × E is aimed for, partial least squares regression becomes inadequate, and methods such as structural equation modelling are more suitable (Vargas et al., 2007).

Mixed-effect models are one of the most used approaches for analyzing G × E, and they are usually implemented using either single-stage or two-stage analysis (Möhring & Piepho, 2009). Single-stage models analyze data from individual plots, in which the residual effects and the G × E effects are estimated simultaneously (Smith et al., 2005). In contrast, two-stage models include a first stage in which design features and spatial variation are modelled using data from individual trials. Next, the second stage involves fitting an overall mixed model to the genotype by environment adjusted means obtained from stage 1 (Malosetti et al., 2013; Smith et al., 2005). The analysis can be extended to more than two stages, in which case the approach is more commonly known as stage-wise analysis (Damesa et al., 2017; Piepho, Möhring, et al., 2012). Although single-stage analysis is preferred from a theoretical point of view, two-stage analysis is less demanding in terms of computation requirements and provides similar results to single-stage when appropriate weights are selected (Malosetti et

al., 2013). Therefore, most of G × E models are implemented using a two-stage approach (Malosetti et al., 2013).

There are situations where non-parametric methods, such as rank-based methods, are more convenient (Elias et al., 2016). This type of non-parametric methods has been used mainly to rank genotypes at specific locations (Elias et al., 2016) and has a set of advantages that include no specific modelling assumption about the distribution of the effects and are easy to implement and interpret (Huehn, 1990). Non-parametric models are considered a useful option when the interest is focused on the ranking of genotypes rather than to evaluate the level of difference on performance between genotypes (Brancourt-Hulmel et al., 1997). In the context of selection in breeding and evaluation programs, Huehn (1990) considered the rank order of genotypes to be the most important information.

Other G × E models go beyond statistical analysis and integrate knowledge on crop physiology and expert assessments. For instance, Theobald et al. (2002) proposed the use of a Bayesian model to incorporate expert knowledge about the analyzed crop. In a bibliometric analysis, van Eeuwijk et al. (2016) identified an important growth in the application of both mixed models and crop growth models, especially after 2005. A crop growth model incorporates plant physiological aspects, along with the genotype and environment, in the analysis of interactions that produce a phenotype (van Eeuwijk et al., 2016). Furthermore, crop growth models also allow to consider the effect of cropping systems (intercropping, fertility management, etc.) on G × E (Jeuffroy et al., 2014).

One of the goals of crop growth models for variety evaluation in multi-environment trials is to improve the characterization of the environment (Jeuffroy et al., 2014). For example, Tesfaye et al. (2016) combined geospatial analysis with crop modelling (1) to characterize a maize growing environment in Southern Africa (Malawi, Mozambique, Zambia and Zimbabwe) and (2) to evaluate the variety performance of five new drought-tolerant varieties across the aforementioned region. The environmental characterization was conducted using a standardized precipitation index, and it focused on the frequency of drought occurrences rather than drought severity (Tesfaye et al., 2016). To evaluate the variety performance of new varieties, maize yields were simulated using the Crop Estimation through Resource and Environment Synthesis (CERES)-Maize model. Simulated relative yields of five drought-tolerant varieties outperformed the commercial check variety across many environments, but not in all (Tesfaye et al., 2016).

Jeuffroy et al. (2014) reviewed the use of crop growth models in variety performance prediction and concluded that although their use is increasing, they are still not mainstream for variety evaluation. For mechanistic models to have predictive power to distinguish between varieties, information is needed on the processes or the underlying genotypic factors that give rise to these differences. Some information can be derived from existing trial data through model fitting, but overfitting often occurs. Acquiring additional data to estimate crop model parameters directly is often costly or not possible retrospectively in a data synthesis

context. Jeuffroy et al. (2014) argue that cost-benefit considerations to assess the value of additional information are important. On the one hand, crop growth modelling generally focuses on a narrow set of variables (mostly yield). Yield is an important input into crop variety recommendations, but other aspects cannot be ignored, including the user perspective (see Section 2.4.2). Their relative complexity makes their application often difficult. One possibility is to use (at least initially) very simple crop models and build up their complexity gradually (Shorter et al., 1991). Another option is to generate intermediate variables that can be used in statistical models or machine learning approaches (see Feng et al. (2019)).

2.4.2 Evaluation in target environments and including user requirements

The current challenges in agricultural production are more likely to be addressed by locally adapted solutions that consider both environmental and socioeconomic information (van Etten, Beza, et al., 2019). The socioeconomic context of the target environment should be considered to match both farmer and consumer preferences. Socioeconomic data like human population, welfare and transportation infrastructure have been proposed for targeting genotypes to environments, but such recommendations mostly concerned logistics planning on germplasm deployment (Hyman et al., 2013). While this kind of socioeconomic data is indeed important, other types of data, such as consumer preferences, should also be considered. For example, Desclaux et al. (2008) proposed that, in addition to the usual biophysical and management factors, the environment should be a wider concept that also includes actors, markets, regulations and societal dynamics.

Participatory on-farm trials aim to take the variety trials closer to the target environments and user requirements (Ceccarelli & Grando, 2007). On-farm trials can provide much information, ranging from biophysical performance to economic assessment (Franzel & Coe, 2002). In this type of trials, the concept of environment in a G × E analysis is extended to include socioeconomic factors, besides the usual biophysical variables (Coe, 2002). Data collected from on-farm participatory trials is often in the form of ratings or rankings, requiring different statistical models to conduct G × E analysis (Coe, 2002). For example, Coe (2002) proposes to analyze ratings using proportional odds, and rankings with the Bradley-Terry model (Bradley & Terry, 1952). An extended Bradley-Terry model can incorporate covariates (Coe, 2002; Dittrich et al., 1998). As discussed before, the possibility of including covariates is especially relevant when location-specific information is to be extracted from the experimental data. It has been shown that environmental covariates can improve predictions of variety performance (Piepho et al., 1998; Piepho, 2000). For data synthesis, in cases when environmental data is not collected as part of the crop trial, environmental covariates can be linked to experimental data through geolocation, as shown by Lobell et al. (2011), van Etten, de Sousa, et al. (2019) and others.

An extended version of the Plackett-Luce model (Luce, 1959; Plackett, 1975) recently implemented by Turner et al. (2020) includes the use of model-based recursive partitioning

(Strobl et al., 2011), allowing the incorporation of covariates for predicting rank orders. Hence, subgroups of rankings with significantly different worth parameters are identified based on covariates (Turner et al., 2020). van Etten, de Sousa, et al. (2019) recently used this model to analyze data from on-farm participatory trials in three countries, successfully identifying environmental covariates that consistently influence variety performance across several seasons.

2.4.3 Multi-dimensional assessment for decision-making

An overall evaluation of varieties requires joint analysis of several traits, from both biophysical and socioeconomic perspectives. Different approaches have been proposed to handle multi-criteria prioritization on ranking varieties according to different traits of interest. Abeyasekera et al. (2002) developed a methodology to combine scores and rankings assigned by farmers evaluating bean (*Phaseolus vulgaris*) varieties, using a weighted index, which allows the farmer's preference to be captured, thus combining multiple criteria. The work of Abeyasekera et al. (2002) considered the farmers' preferences in terms of not only agronomic criteria (yield, pest resistance, etc.) but also non-agronomic criteria, such as taste, marketability and cooking time. Waldman et al. (2014) used choice experiment models to estimate farmers' preferences of perennial pigeon pea. Smith and Fennessy (2011) applied the PAPRIKA (Potentially All Pairwise RanKings of all possible Alternatives) method (Hansen & Ombler, 2008) to assess the relative importance of traits on the improvement of perennial pasture species. The PAPRIKA method asks participants to compare pairs of options (varieties) and select one. It assumes full transitivity to reduce the number of pairs compared. In other words, when $A > B$ and $B > C$, the model assumes $A > C$. An alternative method for priority setting is AgroDuos (Steinke & van Etten, 2017), which is similar to PAPRIKA but integrates the concept of gamification to increase participants' engagement (Deterding et al., 2011), while it does not require interactive updating of questions and can therefore be used without a digital device or Internet connection.

Farmers' comprehensive evaluations of the total value of a variety can also be derived from on-farm trials. For example, the 'tricot' (*triadic comparisons of technologies*) approach proposed by van Etten, Beza, et al. (2019) integrates farmers' feedback on variety evaluation as a ranking of varieties, based on their overall appreciation of the varieties. In the tricot approach, each farmer receives three packages of seeds, each with a different variety of the crop (van Etten, Beza, et al., 2019). Each farmer ranks the three varieties from best to worst, according to overall performance, considering traits such as pest resistance and yield (van Etten, Beza, et al., 2019). Rankings of varieties directly evaluated by farmers in on-farm trials are aggregated by rank aggregation models (see Section 2.5.1).

Recent work of van Etten, de Sousa, et al. (2019) is an example of how a rank-based model was applied to consider several criteria, such as disease resistance, yield and farmer preferences into a single judgement, in combination with local environmental conditions in the analysis crop variety trials. The work of van Etten, de Sousa, et al. (2019) includes three

independent studies in three countries: Ethiopia, India and Nicaragua. For brevity, we focus on the case of Nicaragua, where varieties of common bean were evaluated in 842 plots. An extended version of the Plackett-Luce model, implemented in the *PlackettLuce* package (Turner et al., 2020), was fitted for the trial data collected from on-farm trials, which were established following the tricot approach (van Etten, de Sousa, et al., 2019). The Plackett-Luce model estimates a worth parameter that represents the log probability of each evaluated element (a crop variety in this case) to be ranked first. Environmental conditions of the trial locations were included into the model using climatic indices (Table 2.1) as model covariates, through model-based recursive partitioning, implemented in the *PlackettLuce* package as Plackett-Luce trees. The use of climatic variables as model covariates led to the identification of environmental factors that influenced the probability of a variety performing better than the other varieties tested in the trials (see Figure 2.4 for an example).

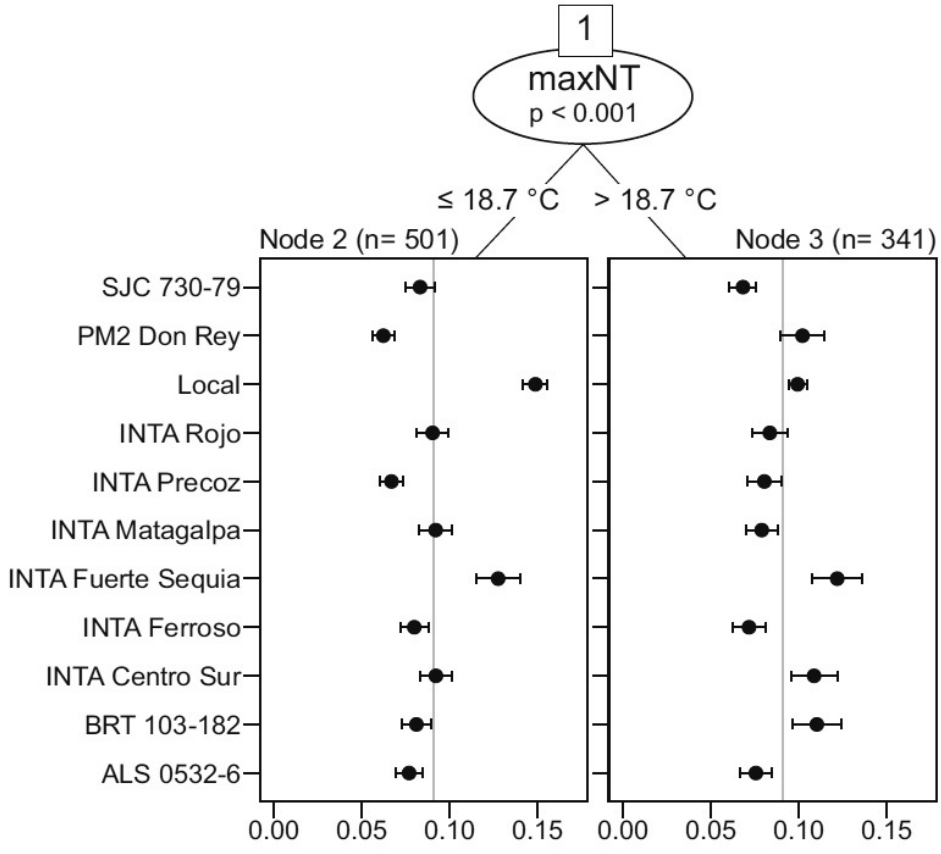


Figure 2.4 Plackett-Luce tree of farmer-participatory tricot trial data in Nicaragua. The probability of each variety to perform better than the others in the trial is presented on the horizontal axis. Grey vertical line represents the average probability of better performance (1/number of evaluated varieties). From the study of van Etten, de Sousa, et al. (2019, p. 4196, CC BY-NC-ND)

The evaluations mentioned above focus on average performance of varieties, but other approaches focus on the variation in performance across seasons to assess farmers' risks. It has been shown that multi-environmental trial data from several seasons can be used to propose variety portfolios to reduce risk and maximize farmers' profits (Nalley et al., 2009; Nalley & Barkley, 2010; Sukcharoen & Leatham, 2016). These studies all focus on yield as the main evaluation criterion. Fadda and van Etten (2019) proposed the adaptation of portfolio selection theory from financial asset management field, based on the portfolio management method developed by Dembo and King (1992). Instead of recommending a single variety, a portfolio of varieties is recommended based on calculations of the expected regret (Fadda & van Etten, 2019). This method does not require absolute (yield) data and can also be applied to ranking data. This is interesting for progress in data synthesis, as ranking methods can play a role in combining datasets from different sources (see Section 2.5.1).

2.5 Data synthesis approaches

In the previous section, we reviewed methods for the analysis of different data types used for crop variety evaluation. In addition to that, data synthesis involves integration of datasets from heterogenous sources. For instance, datasets come from several research programs, each one with diverse types of data formats, measurement units and experimental designs.

Data synthesis for crop variety evaluation has followed two main lines of research: rank aggregation and network meta-analysis. In the remaining part of this section, we review relevant examples from both rank aggregation and network meta-analysis, to finally weigh their advantages, disadvantages and existing gaps towards a data synthesis methodology for crop variety evaluation.

2.5.1 Rank aggregation methods

Rank aggregation methods are rank-based non-parametric statistical methods that allow for aggregation of results from individual studies to obtain one consensus ranking (Lin, 2010; Yu et al., 2019). They have been applied to several fields including advertisement research, psychology, Internet search engines and biological studies (Lin, 2010). Rank aggregation methods are suitable for high-level meta-analysis, where aggregation of different raw data is not feasible (Lin & Ding, 2009). They also provide more statistical power than individual analyses (Lin, 2010; Simko & Pechenick, 2010). This coincides with one of the widely argued characteristics of meta-analysis (Cohn & Becker, 2003).

Simko and Pechenick (2010) proposed to use rank aggregation methods to combine heterogenous data from independent plant breeding trials. Simko and Linacre (2010) demonstrated how the Rasch model (Rasch, 1960) can be used to combine heterogeneous data. The Rasch model is, in principle, very similar to the Luce model (Luce, 1959; Rasch, 1960) (Rasch 1960; Luce 1959).

Simko and Linacre (2010) presented four different real datasets as case examples; for brevity, we only focus on one of the datasets, containing data of potato chip quality evaluations. The analysis of this dataset implies two main constraints already mentioned in Section 2.3: (1) data measurements in different rating scales and (2) only partial overlap among tested varieties. Potato chip quality data were collected from online databases of 10 different laboratories. As we described in Section 2.2, assessment of food quality and consumer preferences can be done using several rating scales. In the case of quality assessments of potato chips, Simko and Linacre (2010) explained that it is a common practice that each laboratory uses a different rating scale such as one of the following: (1) a rating scale of 5, 9 or 10 categories (the number is subjectively selected by each laboratory; lower values indicate higher quality of potato chip); (2) a measurement of the potato chip color using specialized equipment with values ranging from 0 to 100 (higher readings indicate a lighter color of potato chips, which is a desired trait); and (3) a percentage of chips passing a given quality test defined by the laboratory. In this example, it was not specified which rating scale was used by each laboratory in each test, but indeed different ranges of values exist across the different tests.

The data of potato chip quality assessments collected from 10 different laboratories were aggregated into one dataset. The aggregated dataset contained 63 cultivars over 157 trials, with only partial overlap among evaluated cultivars. For instance, only one cultivar was evaluated in 154 trials, while only seven cultivars were evaluated in a single trial (Simko and Linacre 2010). The resulting matrix contains 994 data points, around 10% of the expected total data points (9891) that would have resulted if all the varieties had been evaluated in all the trials (Simko & Linacre, 2010). The original ordinal ratings are replaced for relative rankings (Simko & Linacre, 2010). The relative rankings were used to calculate an overall performance rating, by means of an extended version of the Rasch model (Linacre & Wilson, 1992; Simko et al., 2012). In this case, the extended version of the Rasch model allowed to compare 63 cultivars, even when not all were tested in the same trial.

Interestingly, rank aggregation has also found a direct application in variety trials, such as the work of van Etten, de Sousa, et al. (2019) presented in Section 2.4. The successful application of rank-based methods in both trial analysis and meta-analysis shows that this is an interesting way forward in data synthesis for variety evaluation.

2.5.2 Network meta-analysis

Commonly used meta-analysis methodologies, especially in the medical sciences, are often based on pairwise comparison of treatments, usually in the form of an intervention against a control or placebo (Lumley, 2002; Tonin et al., 2017). Network meta-analysis (Lumley, 2002) allows the comparison of multiple treatments, even when some of them have never been compared directly in trials (Tonin et al., 2017). Although network meta-analysis is commonly used in medical sciences, it has also been used recently in other fields such as

plant pathology (Madden et al., 2016). Network meta-analysis is also known by several other terms, such as 'multiple treatments meta-analysis' and 'mixed treatment comparison', which are often used interchangeably (Salanti, 2012).

As explained by Tonin et al. (2017), the approach evolved from the initial work of Bucher et al. (1997) on 'adjusted indirect treatment comparison', which was called 'network meta-analysis' after the improvements made by Lumley (2002), and later evolved to 'mixed treatment comparison' by Lu and Ades (2004). A distinctive characteristic of network meta-analysis is the case when both direct and indirect comparisons are available for a given pair of treatments. In this case, evidence from both direct and indirect comparisons is used to do a mixed treatment comparison (Figure 2.5), hence the alternative name (Dias & Caldwell, 2019). For more details about related terminology on mixed treatment comparisons, we refer to Salanti (2012) and Coleman et al. (2012).

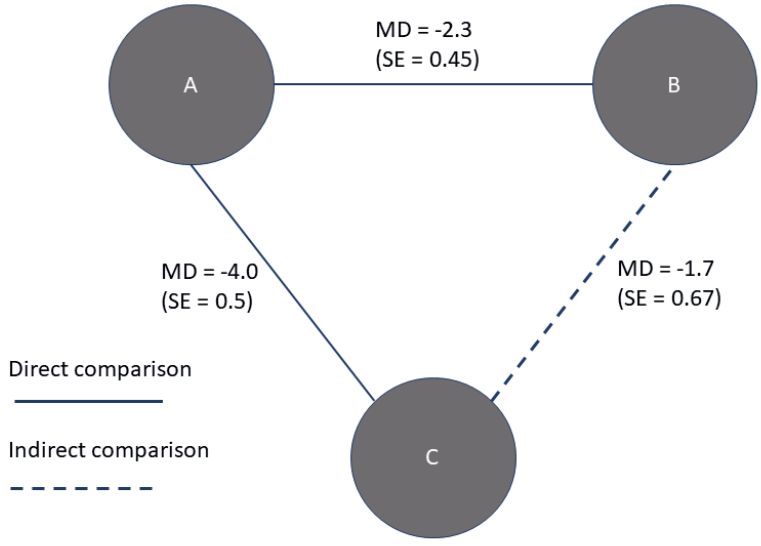


Figure 2.5 Example of a network of treatments (varieties) allowing direct and indirect comparisons. Adapted from Dias and Caldwell (2019). MD, mean difference

Network meta-analysis can be implemented with two different types of models: (1) contrast-based models, also known as conditional models, in which the treatment effects per trial are estimated as a contrast relative to a baseline treatment to subsequently analyze all the contrasts across studies, and (2) arm-based models, also known as unconditional models, in which the treatment summaries per trial are analyzed in a two-way linear mixed model

(Madden et al., 2016; Piepho, Williams, et al., 2012). Arm-based models are commonly applied for the analysis of multi-environment crop variety trials (Section 2.4) (Albert & Makowski, 2018; Damesa et al., 2017; Piepho, 1997). As explained in Section 2.4, it is possible to use single-stage or two-stage analysis in linear mixed models, although for network meta-analysis, a single-stage analysis might be constrained by the availability of data from individual primary studies, rather than usual summary results such as the estimated effect sizes (Madden et al., 2016).

Both frequentist and Bayesian approaches can be applied to network meta-analysis (Tonin et al., 2017), although the Bayesian approach seems to be more popular (Piepho, Williams, et al., 2012). Network meta-analysis usually includes the use of network diagrams, where the nodes represent the compared elements (e.g., treatments or varieties), and the lines (edges) connecting the nodes represent the direct comparison of elements, to evaluate network connectivity. This is relevant in network meta-analysis, especially because poorly connected networks might provide less reliable results compared to a strongly connected network (Tonin et al., 2017). It is also possible the computation of ranking probabilities for each treatment to be assigned a particular position in a ranking from best to worst (Tonin et al., 2017).

Based on yield data obtained from 28 published papers selected through a systematic literature review, Laurent et al. (2015) applied both direct and indirect comparisons in a meta-analysis for ranking crop species based on yield. Direct comparisons compare crops which were grown at the same site and in the same year, whereas indirect comparisons compare crops grown at different sites or in different years, using a third crop grown at all sites as a reference (Laurent et al., 2015). In this case, only results from experimental sites were considered (no farmers' fields), resulting in a database containing 856 records of yield for 36 crop species (Laurent et al., 2015). Mean yield was estimated using a linear mixed effect model, with a log transformation to normalize the yield data (Laurent et al., 2015). For the direct comparison, four crops species (*Miscanthus* × *giganteus*, *Panicum virgatum*, *Triticosecale*, *Salix*) were selected to be used as reference crops, as they were included in the higher number of comparisons with other crops for the same site-years (Laurent et al., 2015). A model was fitted for each reference crop using restricted maximum likelihood. Then, yield ratios of the mean yield of each evaluated crop (except reference crops) to the mean yield of a reference crop grown in the same site and year were calculated (Laurent et al., 2015).

Since direct comparison allows to compare only a limited number of species, indirect comparison was used to compare the yields of a crop of interest, *Miscanthus* × *giganteus*, to yields of crops that were not grown in the same site-years as the crop of interest. Three reference crops were selected for the indirect comparison: *Panicum virgatum*, *Triticosecale* and *Salix*. Therefore, *Miscanthus* × *giganteus* was compared to crops not grown in the same site-years, by indirect comparison using the reference crops, allowing to include more crop species than using direct comparison only.

Albert and Makowski (2018) recently published a paper describing the use of Bayesian mixed treatment comparison models for ranking crop species. According to Albert and Makowski (2018), the dataset used is the same as that analyzed in Laurent et al. (2015), although they also indicate that 639 yield observations were analyzed, which are less than the 856 yield data observations analyzed in Laurent et al. (2015). Mixed treatment comparison combines direct and indirect evidence (Dias & Caldwell, 2019). Five different models were fitted (Table 2.4), of which four were contrast-based models and one was an arm-based model. According to Albert and Makowski (2018), the Bayesian contrast-based models (1 to 4) are variants of the model presented by Dias et al. (2010), while the arm-based is a Bayesian two-way model. The model estimation was done using Markov chain Monte Carlo simulations, while model assessment was made using the deviance information criterion (DIC), in which the models with the lowest DIC are preferred (Albert & Makowski, 2018). Compared to the rankings obtained by Laurent et al. (2015) using direct and indirect comparison, the results are very similar for the two species with higher yields (*Pennisetum purpureum* and *Arundo donax*) when compared against *Miscanthus* × *giganteus*.

Table 2.4 Description of models used by Albert and Makowski (2018)

Model number	Model type	Effect	Variance	DIC
1	Contrast-based	Fixed	Common residual	912
2	Contrast-based	Random	Common residual	348
3	Contrast-based	Random	Species-specific residual	287
4	Contrast-based	Random	Study-specific residual	214
5	Arm-based	Two-way model		348

2.5.3 Assessment of available data synthesis methods

The methods reviewed above address the challenge of combining crop variety trial data from multiple and independent sources. In Section 2.3, we presented a set of challenges identified by Simko and Pechenick (2010) that arise when aiming to combine data from different trials. Here, we assess both rank aggregation and network meta-analysis as solutions to those problems. Additionally, we provide an overview of the relative strengths and weaknesses of data synthesis methods.

2.5.3.1 Partial overlap in evaluated accessions between trials

The problem of partial overlap in the varieties evaluated across trials can be solved by exploiting the capacity of rank aggregation methods to handle partially ranked lists, although the specific approach depends on the particular rank aggregation method. For example, some models are based on pairwise comparison such as the Bradley-Terry model, while others allow multiple comparisons, such as models based on the Plackett-Luce model. In the case of network meta-analysis, the problem of partial overlap is solved by indirect comparison.

For example, in Figure 5, items B and C are indirectly compared with A. Examples are Laurent et al. (2015) and Albert and Makowski (2018) who used reference crop species to allow the comparison of crop species not tested in the same trial.

2.5.3.2 Measurements based on different rating scales or different methods

Rank aggregation methods solve the problems of measurements in different rating scales or traits evaluated with different methodologies, replacing the original raw data from each trial by relative rankings (Simko et al., 2012; Simko & Piepho, 2011). In the work of Laurent et al. (2015) and Albert and Makowski (2018) this was no issue, however, because the data from the different yield studies were in the same units, tons of dry matter per ha per year (Laurent et al., 2015). Network meta-analysis and meta-analysis, in general, can deal with measurements in different units by estimating either the standardized mean difference or the response ratio (Borenstein et al., 2009; Makowski et al., 2019; Murad et al., 2019).

2.5.3.3 Relative strengths and weaknesses of data synthesis methods

There are a few studies applying either rank aggregation or network meta-analysis to crop variety evaluation. Future studies will need to consider the relative merits of each.

Network meta-analysis can provide absolute values (yield differences in tons per hectare), which is difficult to obtain with rank-based models. Even so, the item 'worth' estimated by the rank-based methods is linearly correlated with the underlying latent variable (for example, yield) (Coe, 2002; Fadda & van Etten, 2019). Also, in theory, it should be possible to combine ranking data and continuous variables in the same model (Böckenholt, 2004), but this is still challenging in practice, as such models have not been implemented in general use software.

A useful output that can be obtained from both rank aggregation and network meta-analysis is ranking probabilities, the probability of each variety to be ranked first. Ranking probabilities are related to the concept of reliability in plant breeding, the probability of outperforming a check variety (a reference; for example a previously released variety, commonly used variety or market leader). The concept of reliability was proposed by Eskridge (1990) in the context of crop improvement as a 'safety-first' approach, with subsequent applications by Eskridge and Mumm (1992) and Eskridge (1997).

Data synthesis approaches should consider the ease of use and interpretation by decision-makers in crop variety evaluation. In that sense, the complexity of network meta-analysis can lead to confusion on model implementation and interpretation (Madden et al., 2016). This complexity might be a barrier to its wider adoption as a tool for data synthesis in crop variety evaluation, just like the low level of expertise of users, is limiting the uptake of more sophisticated G × E analysis methods (Lecomte et al., 2010). Rank aggregation methods might be easier to implement but have implicit trade-offs such as information loss and less

power to detect existing differences if compared to parametric methods (Sabaghnia, 2016; Simko & Linacre, 2010; Whitley & Ball, 2002).

2.6 Conclusions and recommendations

We structure this section around three main statements based on our review, which derive conclusions from our main findings and translate these into recommendations.

2.6.1 Elements for a data synthesis approach are available and aligned around ranks and reliability

Based on our review, we assert that the main elements are available for data synthesis as an overarching approach that integrates different components, such as data, models and knowledge from experts (farmers and breeders), to efficiently extract useful information to support decision-making. We remarked in Section 2.5 data synthesis methods that have been tested, exist and can integrate well with existing trial analysis approaches. In particular, rank-based approaches fit within a conceptual framework to analyze variety superiority based on reliability (probability of outperforming a check). A rank-based framework would be able to make versatile use of data from different sources, without complex transformations or doubtful assumptions, and would facilitate the integration of objective measurements and preference data.

2.6.2 Data synthesis should progress from general to specific and from simple to complex

Explicit crop growth modelling has been proposed more than once as a way forward to integrate different types of data into a single conceptual framework for the evaluation of variety performance. However, model building starting from a detailed crop model is not parsimonious and does not build up complexity in a gradual way. For many crops, growth models are not available, hence requiring a large upfront investment in basic (eco)physiological research to enable model building. Also, as shown in Section 2.4, it seems that progress in this field is mainly theoretical, and that practical advances are limited. Even for the attempts that result in generalizable results, the focus is solely on yield and excludes user perspectives. For crop variety evaluation, it seems more logical to start with the ‘big picture’ and work down to the details based on better information indicating where the largest gain in accuracy can be obtained (Section 2.4). This may involve some type of explicit, physiological modelling, but perhaps of a limited number of aspects, not requiring a fully fledged crop growth simulation model. Therefore, we think that a further investment in simpler methods is warranted. This may be less stimulating from a basic research point of view but may give rise to new questions and priorities and give a better sense of the societal relevance and external validity of data synthesis efforts.

2.6.3 Use cases can spur further data sharing and model development

Our review shows that engaging the research community in data sharing is a major challenge (Section 2.3). Most efforts, however, have focused on the supply side: by encouraging/obligating researchers to share data and by providing the infrastructure to do so. While these efforts are certainly important, in crop science, few concomitant efforts have looked at the prospects of successful use of shared data that would drive citations of data papers, shape collaborations around data analysis, and increase researchers' motivation for further sharing. Success may at least partially be the result of a siphon effect: some early use cases can perhaps inspire other researchers to engage in sharing and start a virtuous cycle, as described in Section 2.3. Therefore, investment in a few use cases that use relatively simple methods to show the potential benefits of data synthesis for crop variety evaluation is needed. Our review has shown that those methods are available in principle (Section 2.5). Even so, they need a modest investment to be adapted and demonstrated for this field of application. Next steps would involve stepwise refinements to address components of variety performance that substantially improve the accuracy of predictions. Close collaboration with the decision-makers interested in such evaluations could also spur further interest in this area of research and demonstrate the relevance of further investment.

Acknowledgements

We thank Olga Spellman for the editorial support.

Funding

This research was undertaken as part of, and funded by, the CGIAR Research Program on Roots, Tubers and Bananas (RTB) and supported by the CGIAR Trust Fund contributors (https://www.cgiar.org/funders/). This research was also part of the PhD research project of David Brown, who received funding from the Bill & Melinda Gates Foundation through the Project 'Improvement of Banana for Smallholder Farmers in the Great Lakes Region of Africa'.

Chapter 3 - Rank-based data synthesis of common bean on-farm trials across four Central American countries

This chapter is based on:

Brown D, de Bruin S, de Sousa K, Aguilar A, Barrios M, Gómez M, Hernandez JC, Chaves N, Madriz B, Machida L, Mejía P, Mercado L, Pavón M, Rosas JC, Steinke J, Suchini JG, Zelaya V, van Etten J (2022) Rank-based data synthesis of common bean on-farm trials across four Central American countries. Crop Science. https://doi.org/10.1002/csc2.20817

Abstract

Location-specific information is required to support decision making in crop variety management, especially under increasingly challenging climate conditions. Data synthesis can aggregate data from individual trials to produce information that supports decision making in plant breeding programs, extension services, and of farmers. Data from on-farm trials using the novel approach of triadic comparison of technologies (tricot) are increasingly available, from which more insights could be gained using a data synthesis approach. The objective of our study was to present the applicability of a rank-based data synthesis approach to several datasets from tricot trials, to generate location-specific information supporting decision making in crop variety management. Our study focuses on tricot data from 14 trials of common bean (*Phaseolus vulgaris* L.) performed between 2015 and 2018 across four countries in Central America (Costa Rica, El Salvador, Honduras, and Nicaragua). The combined data of 17 common bean genotypes were rank-aggregated and analyzed with the Plackett-Luce model. Model-based recursive partitioning was used to assess the influence of spatially-explicit environmental covariates on the performance of common bean genotypes. Location-specific performance was predicted for the three main growing seasons in Central America. We demonstrate how the rank-based data synthesis methodology allows integrating tricot trial data from heterogenous sources to provide location-specific information to support decision making in crop variety management. Maps of genotype performance can support decision making in crop variety evaluation, such as variety recommendations to farmers and variety release processes.

3.1 Introduction

Reliable location-specific information supports better decision making in crop variety management, especially under increasing climate variability. On-farm trial data are expensive to obtain and limited by different factors, such as institutional reach into certain areas, availability of seeds, staff, and other resources. It is possible to aggregate data from individual trials to gain insights into variety performance at broader temporal and spatial scales. This can help to gain insights on using varieties across broader areas as climates shift, and to avoid simplistic assumptions about variety environmental adaptation based on rough adaptation zonation approaches. Data synthesis is needed for this and can lead to new insights into a genotype × environment interaction (G × E) under real farming conditions, as trial data can be combined with environmental data, which is increasingly available. Novel data synthesis approaches can extract information from crop variety evaluations to support critical decision making in crop variety management (Brown et al., 2020). In their review, Brown et al. (2020) have proposed rank-based methods as a way forward in data synthesis, because it allows for flexible aggregation of heterogenous data collected using different measurement scales. Rank-aggregation methods have been proposed for the meta-analysis of data from crop genetic resources evaluations by Simko and Pechenick (2010) and Simko and Linacre (2010), with further developments by Simko et al. (2012). This involves converting numerical data to relative ranks and applying a statistical model suitable to ranking data, such as the Plackett-Luce model (Luce, 1959; Plackett, 1975).

Rank-aggregation as a data synthesis method involves no data conversion if ranking data are analyzed, which are collected in on-farm trials, such as the triadic comparison of technology options (tricot) and comparable formats (Coe, 2002; van Etten, Beza, et al., 2019). The tricot methodology involves farmers participating as citizen scientists, evaluating sets of three genotypes in their own farms (van Etten, Beza, et al., 2019). Farmers growth the varieties in small trial plots and rank the varieties accordingly to different traits, such as yield, disease resistance, market value, and the overall performance of the genotypes (van Etten, Beza, et al., 2019). The use of rankings implies an inherent loss of information when compared to measurements in a continuous scale using specialized instruments (Coe, 2002). However, collecting data in ranking format allows the participation of a larger number of farmers and a reduction in the costs of the experiments compared with other participatory methods (Coe, 2002; van Etten et al., 2020). Since the data are in ranking format, it should be analyzed with an appropriate statistical model, such as the Plackett-Luce model (Luce, 1959; Plackett, 1975). The tricot methodology is increasingly used for different crops by several organizations in Africa and Latin America, producing considerable volumes of data. van Etten, de Sousa, et al. (2019) applied the Plackett-Luce model to analyze the variety performance of red common bean (*Phaseolus vulgaris* L.) varieties in Nicaragua. That analysis determined the influence of environmental factors on variety performance using model-based recursive partitioning in combination with the Plackett-Luce model (Turner et

al., 2020; Zeileis et al., 2008). The current study explores combining data from several trials executed by different organizations in multiple locations, following a data synthesis strategy as described in Brown et al. (2020). Applying a data synthesis approach to tricot data overcomes some of the limitations for data synthesis identified by Brown et al. (2020), such as incompatible data formats, scales, and experimental designs. Still, aggregating tricot data drawn from trials established by different organizations requires effort. Several elements can be adopted from the existing work on tricot, but an additional investment is required especially when characterizing uncertainty of model predictions, complementing the Plackett-Luce tree model. In this paper, we describe this new approach and apply it to a dataset on common bean from Central America.

Here, our main objective is to present the applicability of a rank-based data synthesis approach to several datasets from tricot trials, to generate location-specific information supporting decision making in crop variety management. The proposed approach is demonstrated with red common bean genotypes, which were evaluated by four teams of five different organizations in a series of tricot trials in four countries in Central America (Costa Rica, El Salvador, Honduras, and Nicaragua).

The specific objectives are to: (1) integrate data from tricot trials produced by different organizations at different locations and seasons; (2) identify the environmental factors affecting the performance of the evaluated genotypes; (3) predict the best performing genotypes for each main planting season in the study region; and (4) assess the uncertainty and applicability of model predictions.

3.2 Materials and Methods

3.2.1 Tricot Trial Data

We obtained data from 14 on-farm trials across Central America (Figure 3.1). Each trial is a set of incomplete blocks located on farms that test a set of genotypes in a single area and within the same agricultural season. The trials were executed by four teams of five different organizations working in the field across four Central American countries. Key characteristics of the datasets are provided in Table 3.1. All trials followed the tricot citizen science approach, described by van Etten, Beza, et al. (2019). This consists of an incomplete block design with blocks of size three, the use of ranking as a farmer-centric data collection approach, and the intensive use of digital tools to streamline the process. Each tricot plot is an incomplete block set of three red common bean genotypes, which were grown and evaluated by farmer citizen scientists. Genotypes include both released varieties and experimental lines. Packages of bean seeds were delivered to farmers without disclosing the names of the genotypes; bags with the genotypes were labelled as A, B and C. Each farmer assessed the three genotypes and provided feedback by ranking the genotypes. Farmers evaluated plant foliage, plant height, reaction to pests and diseases, drought tolerance, yield, market value, and taste, and also gave their overall judgment considering all the traits. In our work, we analyzed the ranking data generated from the overall performance of the genotypes.

The data collection card used is provided in the supplemental Figure S1. Each organization collected the data and then uploaded it to the ClimMob digital platform (van Etten, Quirós, et al., 2017).

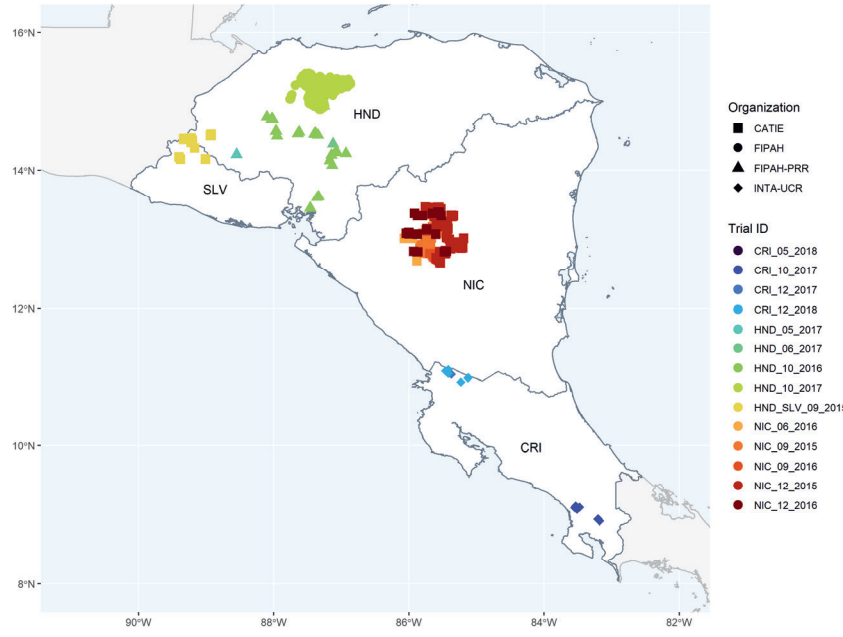


Figure 3.1 Location of tricot trials (identified by Trial ID) in Costa Rica (CRI), El Salvador (SLV), Honduras (HND), and Nicaragua (NIC). Symbol shapes indicate the organization that established and managed the trials

Table 3.1 Trial identifiers, managing organization, country, average planting date, and sample size of plots by trial

Trial ID	Organization [a]	Country	Planting date [b]	*n* [c]
CRI_05_2018	INTA – UCR	Costa Rica	2018-05-21	14
CRI_10_2017	INTA – UCR	Costa Rica	2017-10-19	46
CRI_12_2017	INTA – UCR	Costa Rica	2017-12-08	41
CRI_12_2018	INTA – UCR	Costa Rica	2018-12-15	23
HND_05_2017	FIPAH – PRR	Honduras	2017-05-22	87
HND_06_2017	FIPAH – PRR	Honduras	2017-06-05	17
HND_10_2016	FIPAH – PRR	Honduras	2016-10-16	37
HND_10_2017	FIPAH	Honduras	2017-09-20	714
HND_SLV_09_2015	CATIE	Honduras – El Salvador	2015-09-24	31
NIC_06_2016	CATIE	Nicaragua	2016-06-22	59
NIC_09_2015	CATIE	Nicaragua	2015-09-23	178
NIC_09_2016	CATIE	Nicaragua	2016-09-17	33
NIC_12_2015	CATIE	Nicaragua	2015-12-16	484
NIC_12_2016	CATIE	Nicaragua	2016-12-27	107

[a] INTA, Instituto Nacional de Innovación y Transferencia en Tecnología Agropecuaria, Costa Rica; UCR, Universidad de Costa Rica; CATIE, Centro Agronómico Tropical de Investigación y Enseñanza; FIPAH, Fundación para la Investigación Participativa con Agricultores de Honduras; PRR, Programa de Reconstrucción Rural – Honduras; [b] Median; [c] Number of plots in each trial.

The use of the ClimMob digital platform helped to standardize the data, making it compatible to be aggregated. However, the datasets still required some data curation before aggregating them to conduct the data synthesis. One important data preparation step involved checking variety names across datasets. In Central America, the same genotypes are generally released under distinct variety names in different countries (Rosas, Beaver, Beebe, et al., 2004). To allow data aggregation, the genotype names were translated into experimental line names (Table 3.2). The resulting dataset revealed partial overlap in the varieties tested across the different trials, but we removed genotypes that were tested only in one trial to avoid highly unbalanced comparisons across trials. This reduced the number of genotypes from 27 to 17. The data were aggregated using a rank-aggregation approach (Turner et al., 2020). Data were prepared with the R package *gosset* (de Sousa et al., 2022) for their use in the R package *PlackettLuce* (Turner et al., 2020).

In most of the cases, trial plots had geographic coordinates, as these were part of the tricot data collecting process. In cases when the geographic coordinates were not registered, we assigned the median value of geographic coordinates of remaining plots in the same community where the trial was conducted. Each trial plot data point should include the planting and harvest dates. To identify outliers of planting dates we set a threshold of 40 days above or below the median of each trial and replaced the outliers by the median planting date of the corresponding trial. We also identified outliers for the length of the growing period,

with 60 days as the lower limit and 120 as the upper limit and discarded any data point outside these limits.

Table 3.2 Experimental line names and variety names used in each country after variety release, along with references used to make the translation for red common bean genotypes evaluated in tricot trials in Central America

Experimental name	Variety name (Country*)	Reference
429 DFSZ 15094-39-4	INTA Ferroso (NIC)	Llano et al. (2013)
703-SM 15216-11-4-VB	Chepe (HND)	JC Rosas, personal communication, 11 May 2020
ALS 0532-6	Tolupan Rojo (HND)	Feed the Future Legume Innovation Lab and USDA (2018)
BCR 122-74	*Experimental line*	JC Rosas, personal communication, 11 May 2020
BFS 47	*Experimental line*	JC Rosas, personal communication, 11 May 2020
BRT 103-182	*Experimental line*	JC Rosas, personal communication, 11 May 2020
EAP 9508-93	Cedron (HND)	PRR-FIPAH (2019)
EAP 9510-77	Amadeus 77 (HND)	Rosas, Beaver, Beebe, et al. (2004)
	INTA Rojo (NIC)	
	Cabécar (CRI)	
	CENTA Sand Andrés (SLV)	
	IDIAP R3 (PAN)	
IBC 301-204	INTA Centro Sur (NIC)	Feed the Future Legume Innovation Lab and USDA (2018); Rosas and Escoto (2011)
	Paraisito Mejorado 1 (HND)	
IBC 302-29	Paraisito Mejorado 2 Don Rey (HND)	Feed the Future Legume Innovation Lab and USDA (2018)
IBC 308-24	Amilcar 58 (HND)	Feed the Future Legume Innovation Lab and USDA (2018)
MHC 2-13-49	*Experimental line*	JC Rosas, personal communication, 11 May 2020
MIB 397-72	Honduras Nutritivo (HND)	Rosas et al. (2016)
MPN 103-137	INTA Precoz (NIC)	JC Rosas, personal communication, 11 May 2020
SJC 730-79	Rojo Chorti (HND)	Feed the Future Legume Innovation Lab and USDA (2018); JC Rosas, personal communication, 11 May 2020
	CENTA EAC (SLV)	
SRC 2-18-1	DEORHO (HND)	Feed the Future Legume Innovation Lab and USDA (2018); JC Rosas, personal communication, 11 May 2020
	CENTA Nahuat (SLV)	
	INTA Matagalpa (NIC)	
SX 14825-7-1	INTA Fuerte Sequia (NIC)	Ferrufino (2014)
	Campechano JR (HND)	Feed the Future Legume Innovation Lab and USDA (2018)

*CRI, Costa Rica; HND, Honduras; SLV, El Salvador; NIC, Nicaragua; PAN, Panama.

3.2.2 Environmental data

The main abiotic limiting factors for common bean production are drought, heat stress, and low soil fertility (Beebe, 2012). In Central America, heat stress particularly limits production in the lowlands (Beebe, 2012; Beebe et al., 2011). We accessed publicly-available data repositories to obtain rainfall, temperature, and soil data. An initial set of climatic data were obtained from the "Agrometeorological indicators from 1979 to present derived from reanalysis" dataset, also known as AgERA5 (Hendrik Boogaard & Gerald van der Grijn, 2020; Copernicus Climate Change Service, 2020). Climatic variables and indices were computed following Kehel et al. (2016), Aguilar et al. (2005) and Challinor et al. (2016), using the R package *climatrends* (de Sousa, van Etten, et al., 2020). Table 3.3 describes variables and indices from Aguilar et al. (2005) and Kehel et al. (2016), which were used without major modifications. The climatic indices listed in Table 3.4 are based on Challinor et al. (2016), with thresholds adapted to common bean. Both temperature-based and rainfall-based climatic variables and indices were computed for the whole span of the growing season (i.e., from planting to harvest) of each trial plot. Additionally, temperature-based variables and indices were computed for the three phenological stages: vegetative, flowering, and grain development. Summarized variables and indices by trial location for the three phenological stages are provided in supplemental Table S2, to describe the climate variability among locations. The phenological stage definitions were according to de Medeiros et al. (2016) and Fernández de Córdova et al. (1986).

Table 3.3 Climatic indices that have a possible influence on bean variety performance. Adapted from Aguilar et al. (2005) and Kehel et al. (2016)

Covariate	Description	Unit
minDT	Minimum daytime temperature	°C
maxDT	Maximum daytime temperature	°C
minNT	Minimum nighttime temperature	°C
maxNT	Maximum nighttime temperature	°C
DTR	Diurnal temperature range: mean difference between daily maximum temperature and daily minimum temperature	°C
SU	Summer days: number of days with maximum temperature > 30 °C	°C
TR	Tropical nights: number of nights with maximum temperature > 25 °C	°C
WSDI	Maximum warm spell duration, consecutive days with temperature > 90th percentile	days
CSDI	Maximum cold spell duration, consecutive nights with temperature < 10th percentile	days
T10p	10th percentile of night temperature	°C
T90p	90th percentile of day temperature	°C
MLDS	Maximum length of consecutive days with precipitation < 1 mm	days
MLWS	Maximum length of consecutive days with precipitation ≥ 1 mm	days
R10mm	Number of heavy precipitation days 10 > = rain < 20 mm	days
R20mm	Number of very heavy precipitation days rain > = 20 mm	days
Rx1day	Maximum 1-day precipitation	mm
Rx5day	Maximum 5-day precipitation	mm
R95p	Total precipitation when rain > 95th percentile	mm
R99p	Total precipitation when rain > 99th percentile	mm
Rtotal	Total precipitation (mm) in wet days, rain > = 1	mm
SDII	Simple daily intensity index, total precipitation divided by the number of wet days	mm/days
SRF	Daily solar radiation flux	J

Table 3.4 Climatic sensitivity indices, thresholds, and references used to adjust the thresholds to common bean requirements. Adapted from Challinor et al. (2016)

Index	Description	Threshold (°C)	Reference [a]
5	High-temperature stress using daily mean temperature, expressed as the percentage number of days a certain threshold is exceeded.	Min = 19 Max = 25	Agtunong et al. (1992)
hts_max	High-temperature stress using daily maximum temperature.	Min = 26 Max = 32	Gross and Kigel (1994)
hse	Heat-stress event, expressed as the percentage of the number of days in which a certain threshold is exceeded for at least two consecutive days.	> 35	Gross and Kigel (1994)

[a] For threshold adjustment.

For soil variables (Table 3.5) we used data from SoilGrids250m version 2.0 from four depth layers, 0–5 cm, 5–15 cm, 15–30 cm, and 30–60 cm (Poggio et al., 2021). We selected the soil horizons following Ho et al. (2005). Soil water content was extracted from the Global High-Resolution Soil-Water Balance (Trabucco & Zomer, 2019), and averaged across the growing season of each trial plot.

Table 3.5 Description of soil variables retrieved from SoilGrids250m version 2.0 (Poggio et al., 2021)

Variable	Description	Units
cec	Cation Exchange Capacity of the soil	cmol(c)/kg
cfvo	Volumetric fraction of coarse fragments (> 2 mm)	cm3/100cm^3 (vol%)
clay	Proportion of clay particles (< 0.002 mm) in the fine earth fraction	g/100g (%)
nitrogen	Total nitrogen (N)	g/kg
phh2o	Soil pH in water	pH
sand	Proportion of sand particles (> 0.05 mm) in the fine earth fraction	g/100g (%)
silt	Proportion of silt particles (≥ 0.002 mm and ≤ 0.05 mm) in the fine earth fraction	g/100g (%)
soc	Soil organic carbon content in the fine earth fraction	g/kg
ocd	Organic carbon density	kg/m^3

3.2.3 Plackett-Luce Model

The statistical model applied to analyze the rankings of genotypes is an extension of the Plackett-Luce model (Luce, 1959; Plackett, 1975), implemented in the R package *PlackettLuce* (Turner et al., 2020). The Plackett-Luce model is a classic approach to analyze ranking data, based on Luce's axiom of choice (Luce, 1959; Turner et al., 2020). The version implemented in the *PlackettLuce* package is a generalization of the Plackett-Luce model, to allow handling of ties and partial rankings (Turner et al., 2020). For a given set *S* of *J* genotypes,

$$S = \{i_1, i_2, \ldots., i_j\}$$

The probability that an element i_j is selected from *S* is denoted by:

$$P(i_j|S) = \frac{\alpha_{i_j}}{\sum_{i \epsilon S} \alpha_i}$$

Where $\alpha_i \geq 0$ represents the worth of the genotype *i.* A genotype with higher worth value is more likely to be preferred over other items with lower worth. The worth parameter values are estimated by maximum likelihood (Turner et al., 2020). Considering genotypes, A, B, and C, A $\succ$ C $\succ$ B denotes that A is ranked higher than C, and C is ranked higher than B. To have finite maximum likelihood estimates, the network of wins and losses produced by the rankings needs to be strongly connected (Turner et al., 2020). A strongly connected network is when a path of wins and losses exists, either directly or indirectly, between every pair of items (Turner et al., 2020). Figure 3.2 shows the strongly connected network of genotypes evaluated in the tricot trials. The location of the nodes in Figure 3.2 is automatically determined by the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991). The Fruchterman-Reingold is a force-directed placement algorithm, which tries to optimize the location of the nodes for visualization purposes following two principles: (1) Connected nodes should be drawn next to each other, (2) Nodes should not be drawn too close to each other (Fruchterman & Reingold, 1991). Therefore, genotypes which were compared directly in the tricot trials will likely be nearer than those which were not. Figure 3.2 was made with the R package *GGally*, which uses the Fruchterman-Reingold algorithm implemented in the R package *sna* (Butts, 2020; Schloerke et al., 2021).

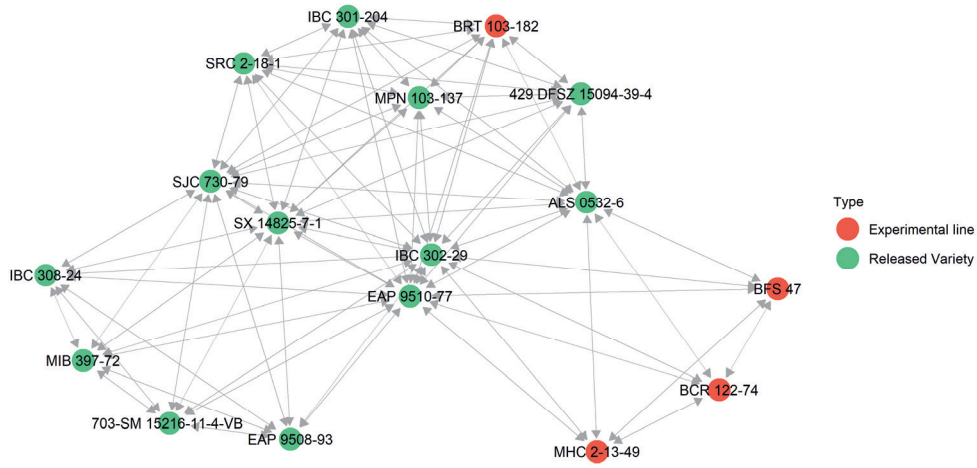


Figure 3.2 Connectivity network of genotypes evaluated in the tricot trials. The arrows indicate wins (outgoing) and losses (incoming) among genotype pairs. The graph drawing algorithm places directly compared genotypes close to each other (Fruchterman & Reingold, 1991).

3.2.4 Model-based Recursive Partitioning with Plackett-Luce trees

The original Placket-Luce model does not account for external factors that may influence the probability of an item to be preferred (Turner et al., 2020). To consider environmental factors in the model, we used an extension in which the Plackett-Luce model is combined with model-based recursive partitioning (Zeileis et al., 2008). It is implemented as Plackett-Luce trees in the R package *PlackettLuce* (Turner et al., 2020). The method involves the following four steps, from Turner et al. (2020) and Zeileis et al. (2008):

1) A Plackett-Luce model is fitted to the complete dataset.
2) The stability of worth parameter values, as influenced by the covariates, is assessed for each covariate.
3) If a significant instability is detected, the data is partitioned by the covariate with the strongest instability, based on a cut-point providing the highest improvement of the model fit.
4) Steps 1-3 are repeated for each branch of the tree until no more instabilities are detected, or if the resultant partitions are smaller than a given size threshold.

3.2.5 Model selection and validation

We first applied a forward variable selection with blocked cross-validation to select variables that are generalizable across the study region (Roberts et al., 2017). We used blocked cross-validation using trials as blocks, further referred to as Leave-One-Trial-Out cross-validation

(LOTO-CV). This partitioning structure aims to account for the geographical and temporal heterogeneity posed by the aggregation of several tricot trial datasets. We assume that each trial represents a particular combination of location and time, in some way equivalent to an environment. This is also motivated by the complex arrangement of planting seasons (Primera, Postrera, & Apante) across Central America (Table 3.6). Similar blocking strategies have been recently applied to validate predictive models of genotype performance (Neyhart et al., 2021).

Models implemented with the recursive partitioning framework can be tuned by adjusting the alpha hyperparameter that conditions the tree size (Hothorn et al., 2006). Low alpha values may result in low power for detecting dependencies between the covariates and the response variable (Hothorn et al., 2006). To overcome this, Hothorn et al. (2006) suggest setting a very large value for the alpha hyperparameter to assure that any dependence is detected. Therefore, we used a large alpha value ($\alpha = .9$) in the forward variable selection to ensure that most dependence is detected. Over-fitting is prevented by subsequent pruning the final tree, using the Akaike information criteria (AIC) (Akaike, 1974). We used the resulting Plackett-Luce tree to predict variety performance in the subsequent analysis steps (Section 3.2.7). A potential limitation of a single tree is its instability, which can be overcome using ensembles of trees (Strobl et al., 2009). On the other hand, a single tree is more interpretable than ensembles (Strobl et al., 2009). Here, we present our approach using a single tree to facilitate interpretation and conceptual clarity. In future work, potential instability can be addressed by using ensembles.

Since Plackett-Luce trees are fitted by maximum likelihood, we used the model deviance as the goodness of fit metric, computed on the hold-out data within the cross-validation procedure. To provide a more interpretable metric, we also computed McFadden's pseudo-R^2 (McFadden, 1973). As an accuracy metric, we calculated Kendall's W (Kendall & Smith, 1939) with the R package *DescTools* (Signorell et al., 2021). Kendall's W measures the concordance between the observed and the predicted rankings. As each trial has a different sample size, we calculated weighted averages of Kendall's W and McFadden's pseudo-R^2 using the test fold size as weights. To evaluate for remaining spatial structure not accounted for in the model, we subjected models with two other sets of covariates to cross-validation for comparison: (1) a model with selected covariates by the forward selection plus geolocation covariates, and (2) a model with only geolocation covariates. We followed van Etten, de Sousa, et al. (2019) for the selection of geolocation covariates: latitude, longitude, longitude + latitude and longitude – latitude (rotated axes).

Table 3.6 Planting seasons for common bean in Central America from different studies

Geographic area	Planting season name	Planting time frame	Study
Central America	Primera	April	(García-Solera & Ramírez, 2012)
	Postrera	August – September	
	Apante or Winter	December – March	
Nicaragua	Primera	May	(Gourdji et al., 2015)
	Postrera	September	
	Apante	November	
Honduras	Primavera or Primera	May 15 – June 20	(Escoto, 2013)
	Postrera or Segunda	End of August —October	
	Postrera tardía or Apante	November –January	
Costa Rica*	Huetar Norte	End of November to beginning of January	(Hernández Fonseca, 2009)
	Brunca 1	May	
	Brunca 2	End of September to end of October	
	Chorotega 1	November – December	
	Chorotega 2	September –October	
	Valle Central and Puriscal	September 15 – First week of October	
	Turrialba	December	
Costa Rica	Brunca 1	May	(Vargas et al., 2018)
	Brunca 2	October	
	Huetar Norte	November –December	

*In Costa Rica, seasons are usually named as Primera and Segunda, with changes in the time frame of Segunda depending on the region (JC Hernandez, personal communication, 4 May 2021).

3.2.6 Modeling and predicting planting dates

We extracted climatic indices to predict genotype performance for the growing period observed in each trial plot. To make predictions of variety performance for unobserved locations, the growing period for these locations needs to be predicted. Predefined planting calendars are often used in agricultural modeling, but this can lead to unrealistic results, as planting dates change across seasons. Farmers usually decide when to plant based on their experience and seasonal weather patterns. Here, we used survival analysis (Kleinbaum & Klein, 2012b) to estimate unobserved planting dates for each planting season in the study region. We fitted a Cox proportional hazard regression model with time-dependent covariates, an extension of the original Cox model (Cox, 1972; Therneau & Grambsch,

2000). Given a subject *i,* the hazard function assumed in the Cox proportional hazard regression model is:

$$\lambda_{i(t)} = \lambda_{0(t)} e^{\mathrm{X}_{i(t)}\beta}$$

Where λ_0 is the hazard baseline function, and $X_{i(t)}$ is the vector of time-varying covariates for subject *i* and β is the vector of coefficients (Therneau & Grambsch, 2000).

To fit the Cox regression model, we used the R package *survival* (Therneau, 2021). Survival analysis needs to define the observation period for modeling. For each season, we defined the start of the observation period as 1 April for the Primera season, 1 August for the Postrera season and 1 October for the Apante season. These dates correspond to roughly one month before each season is expected to begin. The end of each observation period is defined by the latest observed planting date of each season in the aggregated dataset. Previous studies have found that growing seasons follow rainfall patterns in Central America (Alfaro et al., 2018; García-Solera & Ramírez, 2012; Gourdji et al., 2015). In the case of the Primera season, farmers wait for the onset of the rainy season after the dry season. On the other hand, during the Segunda rainy season, farmers look for short dry periods that facilitate planting. Therefore, we selected the following variables that putatively influence bean planting dates: daily precipitation (DP), daily accumulated precipitation (DAP), and daily solar radiation flux (DSRF).

For each Cox regression model, one for each planting season, we first applied a stepwise model selection by AIC, using the function *step* available in base R (R Core Team, 2022a). To assess the model's goodness of fit we used the proportional hazard assumption test implemented in the function *cox.zph* of the *survival* package (Therneau, 2021), which is the approach proposed by Grambsch and Therneau (1994). The proportional hazard assumption test is passed when the *p* value of the chi-square statistic for each variable in the model is nonsignificant (Kleinbaum & Klein, 2012a). The prediction ability of the model is assessed by the *c* index, which is the probability of concordance of the observed survival against the predicted survival (Harrell, 2015; Harrell et al., 1982). We predicted survival curves with covariate data for the target locations, and for each of the past 20 years that were subsequently used in the prediction of genotype performance. From each predicted survival curve, we extract the number of days that intersect the survival curve at 0.25 survival probability, hence 0.75 probability to plant the day after that number of days. We choose this late cutoff to avoid undefined values in the predicted survival curve, which can be potentially caused by unfavorable weather conditions. This number of days is subsequently added to the start date of the corresponding observing period to obtain the planting date. The end date of the growing period is calculated by adding the average number of days of the growing season observed in the sampled trial data.

3.2.7 Predictions of Genotype Performance

Several studies have proposed using environmental covariates to account for G × E in the analysis of multi-environment trials (Piepho et al., 1998; van Eeuwijk et al., 1996). Predictions of genotype performance at new locations are less common, but recent studies have demonstrated it feasibility (Buntaran et al., 2021; van Etten, de Sousa, et al., 2019). To provide a visual representation of the spatio-temporal information generated by the model predictions, we applied a spatial mapping approach.

We defined the target region to predict the genotype performance as the whole area covered by Costa Rica, El Salvador, Honduras, and Nicaragua. We made a base raster layer as a template for the predictions, using the same spatial resolution of AgERA5, approximately 11 km × 11 km. Cell-wise predictions were produced for each planting season for the whole study region, covering the four countries included in the modeling stage. To obtain a temporal representation of current climatic conditions, we predicted genotype performance for each of the past 20 years (2000 – 2019 inclusive) and then averaged the predicted performance to have a representative prediction for each planting season. For each of the 20 years, we predicted the growing seasons using the Cox proportional hazard regression model described in Section 3.2.6. The climatic data were extracted for the periods corresponding to each of the predicted growing seasons (i.e., from planting to harvest), to compute the climatic variables and indices selected in the forward selection process, described in Section 3.2.5. We created a raster map with each cell containing the experimental line names of the top-three performing genotypes according to the averaged predictions.

We provided the ranking probabilities of a genotype to be in its current position and not in any other position, in the event that the trial is repeated. To this end, we calculated the probability of each genotype to be in the top three or not. The procedure is summarized as follows:

1) We fitted a Plackett-Luce tree with the entire aggregated observed dataset.
2) For each of the resulting nodes in the tree obtained under 1, we extracted the worth estimates and the quasi-standard errors (Firth & De Menezes, 2004), using the *qvcalc* function, available from the *PlackettLuce* package.
3) For each node, we determined the probability for each rank position for each genotype. This can be done analytically; however, we used a Monte Carlo strategy for convenience. We drew one million samples for each genotype from a normal distribution centered on their worth, with the quasi-standard error as its standard deviation, using base R function *rnorm*. Then we converted the sampled worth values to ranks.
4) We calculated relative frequencies of rank from the simulated ranks.

5) For each of the cells in the target raster, we predicted in which node each cell falls for each season, using the model fitted in step 1.

6) Using the outputs from steps 4 and 5, we computed the probabilities for each genotype to be either in the top three or not.

Therefore, the top-three best-performing genotypes referred to previously are the three genotypes with highest ranking probabilities of being in the top three.

The R package *terra* was used for handling raster format data (Hijmans, 2021). The maps were plotted using the R packages *ggplot2* (Wickham, 2016) and *sf* (Pebesma, 2018). Data for mapping the administrative boundaries were obtained from the GADM database (Hijmans, 2010).

3.2.8 Uncertainty assessment of model predictions

To estimate the uncertainty of predicted rank probabilities, we calculated the normalized entropy of the rank probabilities for the genotypes with higher probability of being in the top three. We followed Wu et al. (2021), who characterize uncertainty in ranking probabilities using Shannon's entropy, defined as follows:

$$H(x) = -\sum_{i=1}^{n} P(x = i) \log_b P(x = i)$$

For the case of genotype rankings, P($x = i$) is the probability of genotype x to be ranked in position i in a given trial. In our case, we focused on the probability of genotypes being in the top-3, so we used the following equation.

$$H(x) = -(P(x \leq 3) \log_b P(x \leq 3) + P(x \geq 4) \log_b P(x \geq 4))$$

The unit of entropy depends on b, which is the base of the logarithm (Wu et al., 2021). We used the common base-2 logarithm, where the unit of entropy is the bit. The entropy is normalized to range between 0 and 1, by dividing the range of maximum and minimum entropy for the number of n elements in the ranking (Wu et al., 2021). An entropy value of 1 represents high uncertainty.

3.2.9 Applicability assessment of model predictions

To assess the applicability of model predictions from an environmental perspective, we calculated the area of applicability (AOA) as described by Meyer and Pebesma (2021). This provides geographic boundaries to separate areas where the relationship learned by the model with the training data can be extrapolated from those that cannot (Meyer & Pebesma, 2021). Furthermore, the AOA identifies the areas where the model performance estimated with

cross-validation applies (Meyer & Pebesma, 2021). The AOA is derived from the dissimilarity index (DI), which is the distance in the multidimensional space of the predictor variables between training data and new data from locations used for predictions (Meyer & Pebesma, 2021). The AOA is binary outcome calculated after applying a threshold on the DI. Cell points that surpass the threshold are labeled as 0, and otherwise as 1. The threshold on the DI is defined by the maximum DI of the training data in the cross-validation, considering only the data points that do not occur on the same fold. Hence the same blocking structure of the cross-validation folds is used in the calculation of the DI, and subsequently in the derived AOA. We calculated the AOA with the function *aoa* implemented in R package *CAST* (Meyer, 2018).

3.3 Results and Discussion

3.3.1 Model selection and validation

The forward variable selection procedure selected the following variables: WSDI, R20mm, T10p, hts_mean_19_flo. The model with environmental covariates produced better cross-validated values of deviance, pseudo-R^2 and Kendall's W values compared to the baseline model without covariates (Table 3.7). The models that include geolocation covariates produced a worse fit compared to the model with environmental covariates. Therefore, the model with environmental covariates effectively accounts for the spatial structure of the aggregated dataset.

Table 3.7 Deviance, pseudo-R^2 and Kendall's W comparing four models, no covariates, environmental covariates, covariates and geolocation, and geolocation only

Model	Deviance	Pseudo-R^2	Kendall's W
No covariates	6183	0.6679	0.5203
Environmental covariates	6033	0.6831	0.5215
Environmental covariates + geolocation	6393	0.6590	0.5146
Only geolocation	6350	0.6596	0.5122

The AIC-pruned tree using all the data (Figure 3.3) makes a split based on the warm spell duration index (WSDI), which is the number of days with temperature above the 90th percentile (de Sousa, van Etten, et al., 2020). Terminal nodes 2 and 3 in Figure 3.3 present the estimated worth values for the different resulting sub-sets of data after the split with variable WSDI. Node 3 presents the genotypic worth values for all plots with conditions of more than 4 days with temperature higher than the 90th percentile. In these warm conditions, the top three genotypes are: IBC 308-24, MHC 2-13-49 and SX 14825-7-1. The genotype IBC 308-24 was released in Honduras as Amilcar 58 (Feed the Future Legume Innovation Lab & USDA, 2018). In the case of genotype MHC 2-13-49, it is an experimental line developed mainly for resistance against web blight [*Thanatephorus cucumeris* (Frank) Donk] (Rosas, unpublished data, 2022). The genotype SX 14825-7-1 was released in Honduras in 2011 as "Campechano JR" and in Nicaragua as "INTA Fuerte Sequía". It was selected within

a community-based participatory plant breeding program in Honduras and it is well-adapted to high temperature conditions (Vargas et al., 2011).

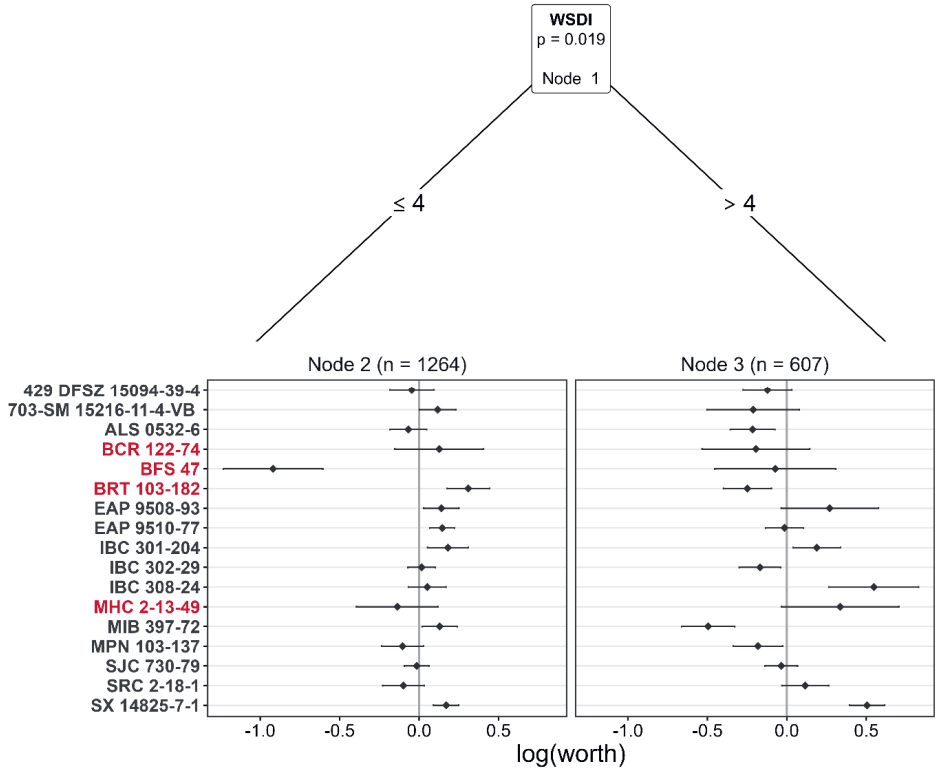


Figure 3.3 Plackett-Luce tree plot of the model fitted with aggregated data from all the trials. The x axis indicates the worth parameter estimates in logarithmic scale, which are the probabilities of each genotype to be ranked first. The bars represent quasi-standard error, and the central vertical gray bar is the zero intercept for each node. WSDI, warm spell duration index. Color codes for genotypes; red = experimental line, gray = released variety.

While some genotypes were not grown and evaluated together in the field, as shown in Figure 3.2, the Plackett-Luce tree model with environmental covariates allows to make those comparisons considering the different environmental conditions among trial locations.

Figure 3.3 shows the presence of G × E between the two sets of environments differentiated by the Plackett-Luce tree model. For example, the experimental line BRT 103-182 performed well in the less warm conditions (node 2) but poorly in the warmer conditions (node 3).

In general terms, the split made by WSDI is consistent with existing knowledge about the effect of high temperature on the performance of common bean genotypes (Beebe, 2012; Beebe et al., 2011). van Etten, de Sousa, et al. (2019) did a preliminary study of a subset of the data (the trials in Nicaragua only) and identified maximum night temperature (maxNT)

as the major factor influencing differences in common bean genotype performance. The interpretation of van Etten, de Sousa, et al. (2019) was that the main difference between varieties was their level of heat stress tolerance. We identified a similar effect of heat stress in our study. There is a weak, positive correlation between WSDI and maxNT ($r = 0.09$, $p = 1 \times 10^{-04}$). The current analysis has not identified additional environmental factors. The current dataset is larger than that of the Nicaragua study but also holds more varieties. Also, WSDI may be more generalizable across geographical space than maxNT or may capture other environmental influences beyond heat stress. In future applications, more covariates and interactions between them could be identified using machine learning methods (for example, using Plackett-Luce Forests, ensembles of trees).

3.3.2 Survival analysis to predict planting dates

The coefficients estimated by the Cox regression model for the three seasons are presented in Table 3.8. For instance, in the case of the Primera season model, a mm change in the daily accumulated precipitation is associated with around 0.5% increase in the probability of planting. The variable importance is described by the magnitude of the Z value, and the exponentiated coefficients provide the multiplicative effect of each covariate on the estimated risk (Therneau & Grambsch, 2000).

Table 3.8 Estimated Cox regression model coefficients for the three planting seasons

Season model	variable	coef	exp(coef)	se(coef)	Z	*p*
Primera	DAP	0.005	1.005	4.968×10^{-04}	10.83	2×10^{-16} ***
Postrera	DP	0.008	1.008	0.004	2.047	0.0407 *
	DSRF	8.930×10^{-08}	1.000	1.622×10^{-08}	5.505	3.69×10^{-08}***
Apante	DP	-0.051	0.95	0.013	-3.854	1.161×10^{-04}***

*Significant at the .05 probability level. **Significant at the .01 probability level. ***Significant at the .001 probability level.

For the Primera season, the stepwise model selection removed daily precipitation (DP), keeping only daily accumulated precipitation (DAP) as covariate in the model. In the case of the Postrera season, the stepwise model selection suggested all the three variables should be kept in the model. However, we detected a large violation of the proportional hazard assumption. We overcame this by removing DAP from the Postrera season model, keeping daily solar radiation flux (DSRF) and DP. This improved the model from the initial fit, with a modest improvement in the ability of prediction and only a moderate violation of the proportional hazard assumption (Table 3.9). The final model for the Apante season only has DP as covariate after applying the stepwise model selection.

Table 3.9 Results of the proportional hazard assumption tests for the three models (Primera, Postrera, Apante)

Season model	covariates	χ^2	df	*p*
Primera	DAP	1.76	1	0.18
	Global [a]	1.76	1	0.18
Postrera	DP	0.583	1	0.4450
	DSRF	10.223	1	0.0014
	Global [a]	12.901	2	0.0016
Apante	DP	6.12	1	0.013
	Global [a]	6.12	1	0.013

[a] Global is a global test with all variables.

The Cox regression model of the Primera season has a *c* index of 0.808 (SE = 0.03) — considered a good prediction ability — while the Postrera and Apante models have c index values of 0.563 (SE = 0.012) and 0.625 (SE = 0.013) respectively, which are relatively low, but still better than using a fixed planting date.

For the Primera season, it makes sense that accumulated precipitation influences farmers' decision to plant, because generally farmers wait until the onset of the rainy season. The Postrera season starts before the second peak of the bimodal rainfall distribution (Alfaro et al., 2018; García-Solera & Ramírez, 2012). As this occurs during the rainy season, farmers need to identify a time window of sunny days to plant. For the Apante season model, the influence of daily precipitation might be linked to the drier conditions of this season, hence farmers wait to notice some stability of precipitation to decide to plant.

We consider our results a good first approximation to model planting dates in function of observed data and climatic covariates. Previous studies have applied survival analysis to seasonal forecasting (Maia & Meinke, 2010). However, to the best of our knowledge, the approach presented here is the first application of a Cox model with time-dependent covariates to predict crop planting dates. Further refinement is required in this research field, such as adjustment in terms of optimal starting point of observation, required sample size, and selection of additional covariates. This was not the main objective of the present work, and the approach might be considered as an ad-hoc solution to the problem of obtaining planting dates to be used as inputs in the genotype performance model. We believe that our findings can support further exploration and development of the application of survival analysis to model planting dates, especially when the alternative is the use of unrealistic fixed planting dates.

3.3.3 Maps of Predicted Genotype Performance

The average estimated performance of common bean genotypes across the four countries in the study region for the Primera, Postrera, and Apante seasons is presented in Figures 3.4a, 3.5a, and 3.6a, respectively. Figures 3.4b, 3.5b and 3.6b present the normalized entropy for the ranking probabilities. Low entropy values represent areas where the predictions of ranking probability have low uncertainty. For instance, genotypes predicted to be in the top-three for a given area with low entropy have a high probability to be in the top three in a long series of repetitions of the same experiment. Figures 3.4c, 3.5c and 3.6c present the area of applicability (AOA), which differentiates areas where model predictions are supported by the sample data from those in which they are not. For the three seasons, a group of seven genotypes is in the top three (Table 3.10); within that group, ranking positions swap depending on the season and the location. A consistent pattern across the seasons is that areas in which the top three is either SRC 2-18-1 ≻ SX 14825-7-1 ≻ ALS 0532-6 or SRC 2-18-1 ≻ ALS 0532-6 ≻ SX 14825-7-1 have the lowest entropy and good AOA. Therefore, these two predicted rankings are the most reliable across seasons. The predictions are also consistent with the known traits of the genotypes. For instance, genotypes ALS 0532-6 and SX 14825-7-1 are tolerant to drought and heat, while genotype SRC 2-18-1 is tolerant to heat (Table 3.10). These three genotypes seem to perform well across the Central America Dry Corridor.

Differences in entropy values seem to be driven by the representativeness of genotypes in each of the trials. For instance, Costa Rica presented the highest levels of entropy compared with the rest of Central America. Genotypes SX 14825-7-1 and SRC 2-18-1 were not evaluated in Costa Rica. On the other hand, areas with relatively high entropy in El Salvador, Honduras, and Nicaragua are those in which BCR 122-74 is present in the top three, which was only evaluated in Costa Rica. ALS 0532-6 is in the top three in areas with lowest entropy, and was evaluated in 10 of the 14 trials (Figure 3.7).

The AOA values that indicate the areas in which the model predictions cannot be applied correspond to locations well known to be unsuitable for common bean production because of the unfavorable climatic conditions. Many of these locations are within conservation areas. For instance, Costa Rica's Talamanca Mountain range was identified as having no applicability of the models in the three seasons. In this case, the environmental constraints are the high altitude, cold temperatures and humid conditions (Oostra et al., 2008). Another example is the Indio Maíz Biological Reserve in Nicaragua. This area is very humid, with annual precipitation of more than 4000 mm, consisting of tropical forest and swampland (Jordan et al., 2019). We mapped all modeling results for demonstration reasons, also covering the area of no applicability, but in future applications these areas can be masked. Overall, the AOA maps show that the trials jointly cover most environmental conditions under which bean growing occurs in Central America. This is an indication of the potential of aggregating trial results across space and time to make predictions across the whole region.

In the case of the Primera season, in a large portion of the predicted area the top-three genotypes are SX 14825-7-1 ≻ SRC 2-18-1 ≻ BCR 122-74. The exception is Costa Rica, where genotype SRC 2-18-1 is not in the top three in a large part of the country. From an environmental perspective, the largest unrepresented area delimited by the AOA (Figure 3.4c) is Costa Rica's Talamanca Mountain range, described above.

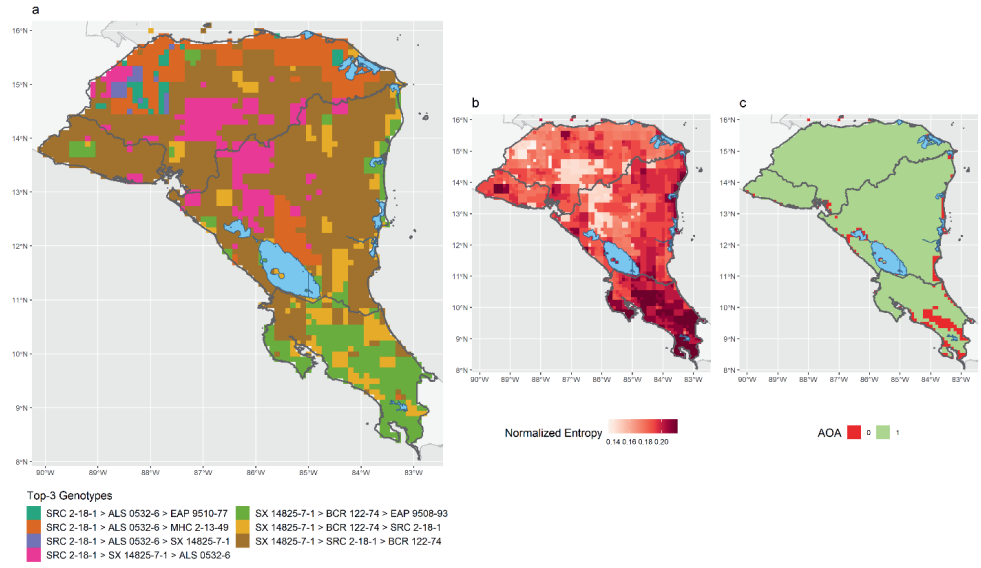


Figure 3.4a Map of genotypes with the higher probability of being in the top three across the study region for the Primera season. **Figure 3.4b** Normalized entropy (0-1) of the genotypes with higher probability of being in the top three; the legend scale is constrained to easily visualize the differences. **Figure 3.4c** Area of applicability (AOA) for the Primera season; areas in red denote no applicability of the model.

For the Postrera season, Figure 3.5a shows that areas in which the top three are SX 14825-7-1 ≻ SRC 2-18-1 ≻ BCR 122-74 are similarly large to those predicted for the Primera season, but the areas where the top three are SRC 2-18-1 ≻ SX 14825-7-1 ≻ ALS 0532-6 are larger in El Salvador, Honduras, and Nicaragua. This area has the lowest values of entropy, indicating that these genotypes have a high probability of not being outperformed by other genotypes (Figure 3.5b). The AOA for the Postrera season has a similar pattern compared to the Primera season, with an additional small area in Honduras with environmental conditions not covered by the trials (Figure 3.5c).

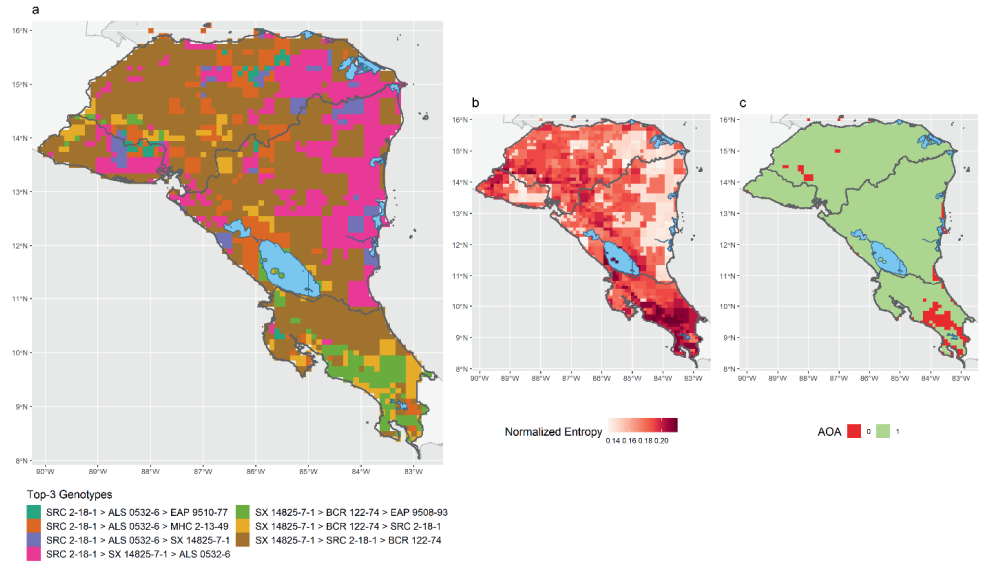


Figure 3.5a Map of genotypes with the higher probability of being in the top three across the study region for the Postrera season. **Figure 3.5b** Normalized entropy (0-1) of the genotypes with higher probability of being in the top three; the legend scale is constrained to easily visualize the differences. **Figure 3.5c** Area of applicability (AOA) for the Postrera season; areas in red denote no applicability of the model.

The map of predictions of the top-three genotypes for the Apante season shows a major difference with the Primera and Postrera seasons. In this case, in most of the area the top three are SRC 2-18-1 ≻ ALS 0532-6 ≻ MHC 2-13-49 (Figure 3.6a). Relative to Primera and Postrera, a much larger area has genotype MHC 2-13-49 ranking third in the top three. Still, the areas with the lowest entropy (Figure 3.6b) are those in which the top three are SRC 2-18-1 ≻ SX 14825-7-1 ≻ ALS 0532-6 (Figure 3.6a). Therefore, the main difference in terms of entropy for the Apante season seems to be presence of either MHC 2-13-49 or ALS 0532-6 in the top three. The genotype MHC 2-13-49 was only evaluated in Costa Rica (Figure 3.7). In areas where this variety is among the top three, entropy values are higher compared to areas in which ALS 0532-6 is among the top three. The latter variety has less uncertainty associated with it, as it was evaluated in 10 out of 14 trials across the four countries. The AOA for the Apante season is similar to the one for the Postrera season, with an increase in Honduras for areas in which the environmental conditions were not represented in the trials. These areas mainly correspond to the locations of the Pacayita Volcano Biological Reserve, the Opalaca Biological Reserve and the Cacique Lempira Señor de las Montañas Biosphere Reserve. Another important change is that most of the coastal areas delineated as with no applicability in the cases of Primera and Postrera, are classified as having good applicability in the Apante season. However, in the case of Apante, the coastal areas present high levels

of entropy. This higlights the importance of considering both uncertainty assesments — the entropy and the AOA — as they are complementary.

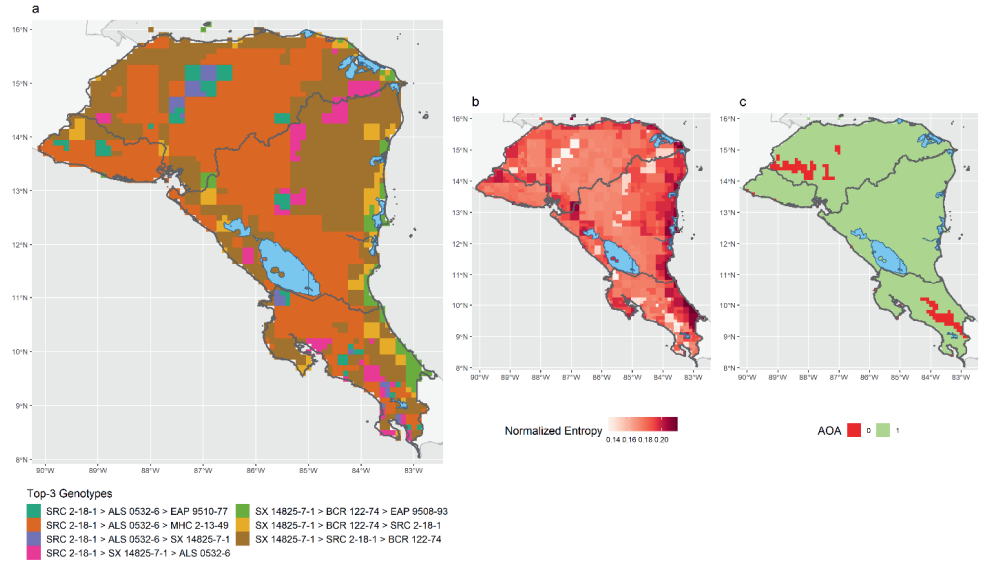


Figure 3.6a Map of genotypes with the higher probability of being in the top three across the study region for the Apante season. **Figure 3.6b** Normalized entropy (0-1) of the genotypes with higher probability of being in the top three; the legend scale is constrained to easily visualize the differences. **Figure 3.6c** Area of applicability (AOA) for the Apante season; areas in red denote no applicability of the model.

One use of the predictions from the data synthesis is to identify promising locations for new genotype evaluations. For instance, EAP 9508-93 has been released only in Honduras as "Cedrón" and the present study only included tricot trials in Honduras (Figure 3.7). Yet, EAP 9508-93 was found to belong to the top-three genotypes in most parts of Costa Rica (Figure 3.4a). Relatively high entropy in those areas indicates that a direct recommendation to introduce this genotype in Costa Rica is not warranted, but our findings suggest it is a relevant candidate for future evaluations, which are required to release this variety also in Costa Rica. Another example is ALS 0532-6, a relatively new genotype again released only in Honduras, but tested across the four countries. The low entropy in areas where this genotype belongs to the top three (in Costa Rica, El Salvador, Nicaragua) indicates relatively large certainty about its superior performance. Therefore, this genotype is considered a good candidate for introduction in these areas. The results of this study feed directly into decision-making on common bean breeding, which is well-coordinated in Central America (Reyes et al., 2016). Another enabling factor is that market preferences are relatively homogeneous in Central America, and food quality traits are considered by farmers in tricot trials (Supplemental Figure S1). Within countries, our findings can be used to recommend genotypes directly to farmers.

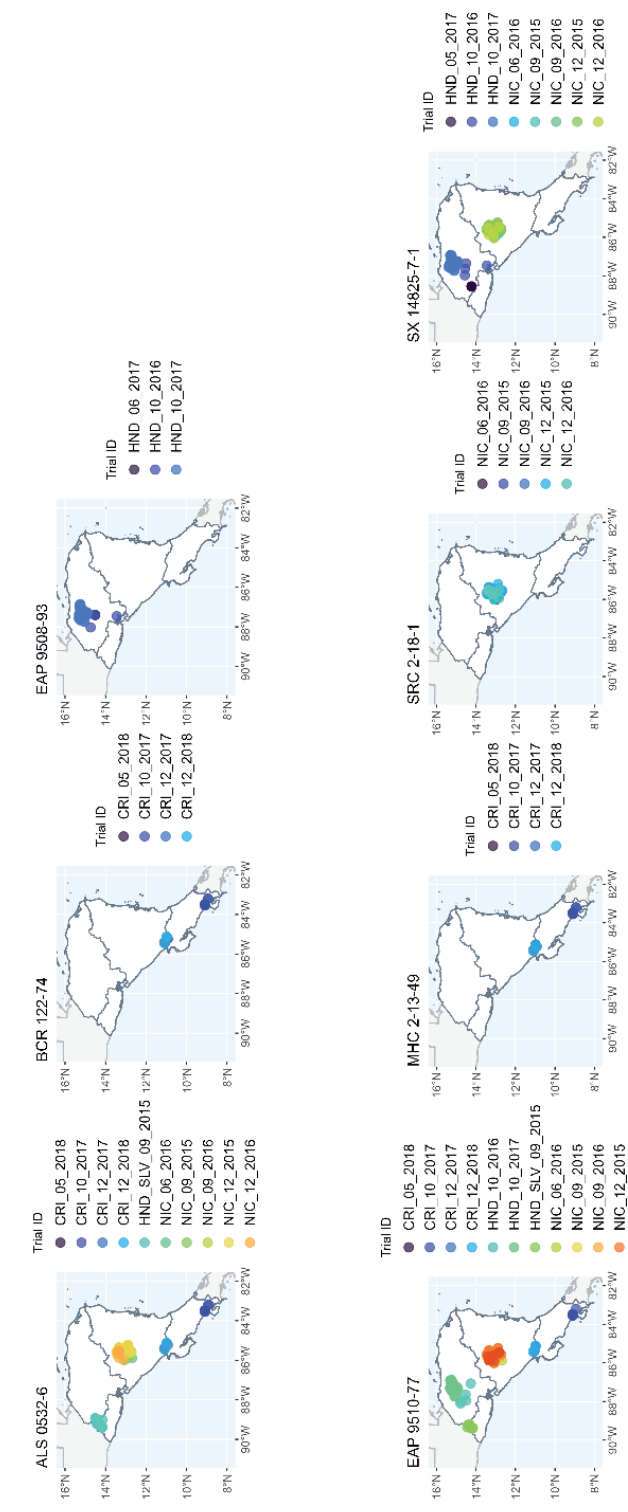


Figure 3.7 Presence of genotypes included in the top *three* in at least one of the trials.

Table 3.10 Experimental and variety names, year of release, and main traits of best-performing small red bean genotypes predicted by the model. Indices with near zero variability were removed.

Experimental name	Variety name (Country)	Released year	Traits	Reference
ALS 0532-6	Tolupan Rojo (HND)	2019	Resistant to Angular Leaf Spot [*Pseudocercospora griseola* (Sacc) Crous & Braun], *Bean Common Mosaic Virus* (BCMV), *Bean Golden Yellow Mosaic Virus* (BGYMV) and bean rust [*Uromyces appendiculatus* (Pers.:Pers.) Unger]. Tolerant to anthracnose [*Colletotrichum lindemuthianum* (Sacc. & Magnus) Lams.-Scrib.] and common bacteria blight [*Xanthomonas campestris* pv. *phaseoli* Smith (Dye)]. Good adaptability to drought, heat, and soils with low fertility.	Feed the Future Legume Innova Lab and USDA (2018); Rosas et al. (2019)
BCR 122-74	*Experimental line*		Resistant to *Bean Golden Yellow Mosaic Virus* (BGYMV) and *Bean Common Mosaic Virus* (BCMV).	Rosas, unpublished data, 2022
EAP 9508-93	Cedron (HND)	2007	Resistant to common bean rust [*Uromyces appendiculatus* (Pers.:Pers.) Unger] and anthracnose [*Colletotrichum lindemuthianum* (Sacc. & Magnus) Lams.-Scrib.], but susceptible to Angular leaf spot [*Pseudocercospora griseola* (Sacc.) Crous & Braun]. It is recommended for regions with altitude within 800–1200 masl, 800–1200 mm annual precipitation and 18–24 °C average temperature.	Rosas et al. (2009)
EAP 9510-77	Amadeus 77 (HND) INTA Rojo (NIC) Cabécar (CRI) CENTA San Andrés (SLV) IDIAP R3 (PAN)	2003	Resistant to *Bean Golden Mosaic Virus* (BGMV) and *Bean Common Mosaic Virus* (BCMV). Wide adaptation. It was developed as heat tolerant, with coastal regions of Central America as target locations.	Rosas, Beaver, Escoto, et al. (2(
MHC 2-13-49	*Experimental line*		Resistant to web blight [*Thanatephorus cucumeris* (Frank) Donk], *Bean Golden Yellow Mosaic Virus* (BGYMV) and *Bean Common Mosaic Virus* (BCMV).	Rosas, unpublished data, 2022
SRC 2-18-1	DEORHO (HND) INTA Matagalpa (NIC) CENTA Nahuat (SLV)	2007	Resistant to *Bean Common Mosaic Virus* (BCMV) and *Bean Golden Yellow Mosaic Virus* (BGYMV) Tolerant to high temperatures.	Rosas, unpublished data, 2022
SX 14825-7-1	INTA Fuerte Sequía (NIC) Campechano JR (HND)	2009 2011	Dark-red grain. Resistant to *Bean Golden Yellow Mosaic Virus* (BGMV) and *Bean Common Mosaic Virus* (BCMV). Good adaptation to drought, heat and low fertility soils.	Reyes et al. (2016) Vargas et al. (2011)

3.4 Conclusions

We demonstrated the use of a data synthesis approach to aggregate data from on-farm trials to produce location-specific information about common bean genotype performance across four countries in Central America.

The use of environmental covariates in the Plackett-Luce tree model allowed us to identify WSDI, a proxy for elevated heat, as the main abiotic factor influencing the genotype performance across the study region. The approach also allowed taking advantage of the different locations and growing seasons represented in the aggregated dataset, in contrast to what might be possible when single trials are analyzed in isolation.

The maps of predicted variety performance produced with our data synthesis approach can provide useful insights for (1) local plant breeding programs to target new locations for testing, and (2) extension agents to generate recommendations for farmers. The combination of entropy and the AOA allowed us to quantify the uncertainty of two different dimensions: the rankings and the environmental conditions. We expect our data synthesis methodology to be applicable to other crops and regions.

While the analysis of ranking data is less common in the evaluation of crop varieties, significative efforts have been made recently to facilitate both the collection and analysis of tricot trial data. For data collection, the digital platform ClimMob provides the required functionality to collect and store the data in a standardized format (van Etten, Quirós, et al., 2017). A challenge for the aggregation of trials from different countries is the different naming used for released varieties. For future cases, we concur with Rosas, Beaver, Beebe, et al. (2004) that coordinated regional releases, in contrast to individual releases, help to avoid this problem. For the large number of already released materials under different names, an open-access database with harmonized variety names might facilitate variety identification across countries. For instance, a solution has been proposed within the breeding management system Breedbase (Morales et al., 2022). Still, further developments are required to also provide a solution to a wider audience, including farmers. With regards to the modelling and analytical stage, most of the developments made within our work are being fed into the R package *gosset*, which complements the functionality of the package *PlackettLuce*. This will help researchers to apply the methods presented in this study to other datasets.

Our work also enables further synthesis of a larger dataset. As more tricot trial data becomes available, it could be aggregated and iteratively reanalyzed using the data synthesis approach. Two promising next steps are the use of ensembles of Plackett-Luce trees (Plackett-Luce Forests) and the integration of tricot and non-tricot on-farm trial data.

Acknowledgments

We acknowledge all the farmers who participated in the tricot trials and all the technical staff who helped in the establishment of trials and data collection. We acknowledge Omar Gallardo (FIPAH) for all his contribution with the establishment of the trials in Honduras.

We acknowledge Olga Spellman (Science Writing Service of the Alliance of Bioversity International and CIAT) for English editing of this manuscript. This research was supported by Cooperative Agreement AID-OAA-F-14-00035, which was made possible by the generous support of the American people through the US Agency for International Development. This work was implemented as part of the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), which is carried out with support from the CGIAR Trust Fund and through bilateral funding agreements. For details, please visit https://ccafs.cgiar.org/donors. The views expressed in this document cannot be taken to reflect the official opinions of these organizations.

Data availability

All the R code and the data required to replicate the results are available in GitHub: https://github.com/AgrDataSci/Data_synthesis_bean_tricot_CA

Software

We used R for all the statistical analysis (R Core Team, 2022a). In addition to R packages cited in the main text, we also used: *ag5Tools* (Brown & de Sousa, 2022), *caret* (Kuhn, 2022), *colorspace* (Zeileis et al., 2009), *dplyr* (Wickham, François, et al., 2022), *ggparty* (Borkovec & Madin, 2019), *janitor* (Firke, 2021), *network* (Butts, 2008), *RColorBrewer* (Neuwirth, 2022), *readr* (Wickham, Hester, et al., 2022), *stringr* (Wickham, 2019), *survminer* (Kassambara et al., 2021), *partykit* (Hothorn & Zeileis, 2015), *qvcalc* (Firth, 2020), *patchwork* (Pedersen, 2020), and *viridisLite* (Garnier et al., 2021).

Supplemental material

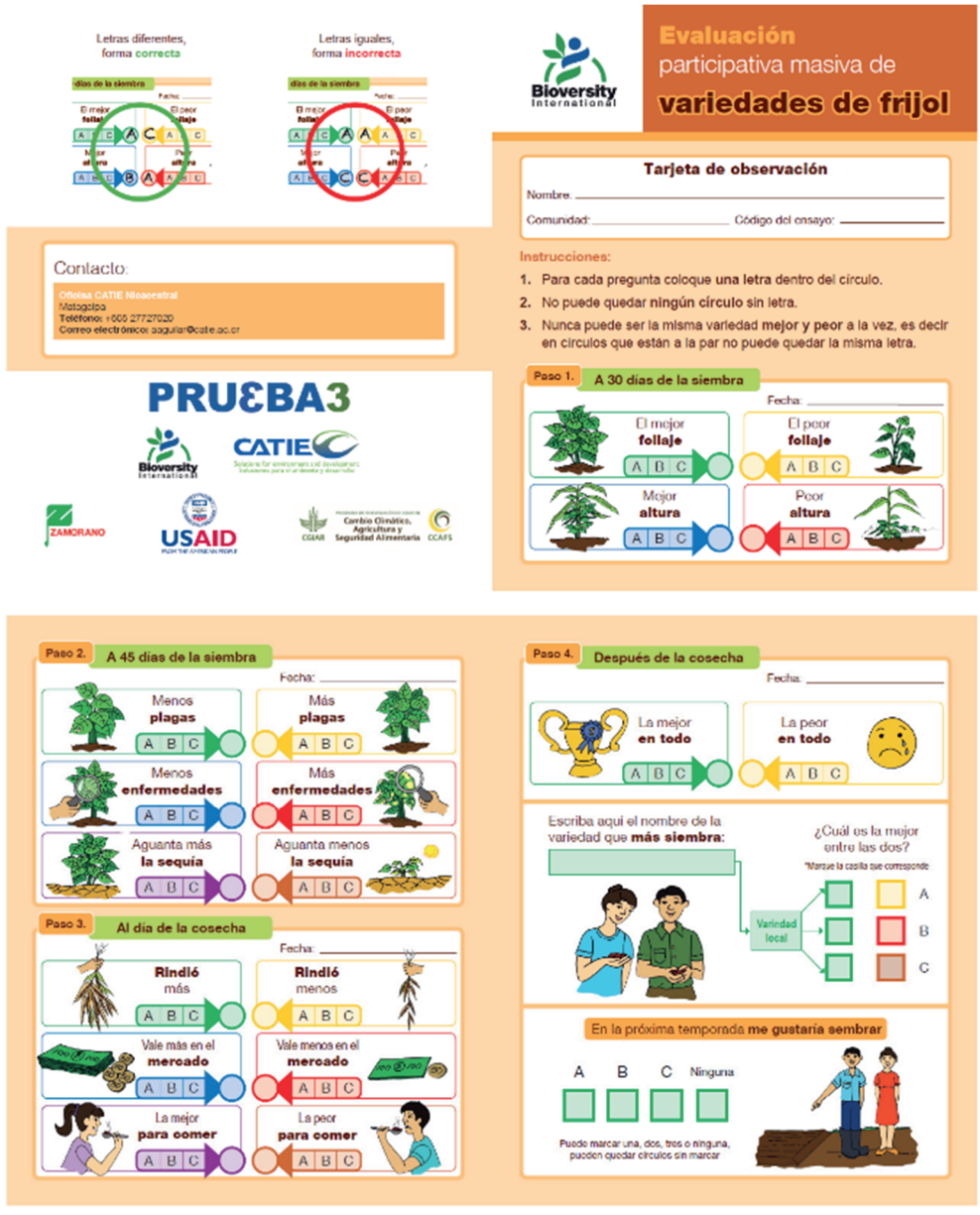


Figure S1 Data collection card used in tricot trials.

Table S2

Climatic indices for the three phenological phases, vegetative, flowering and grain filling, averaged by trial. Indices with near zero variability were removed.

Trial	Vegetative										Flowering											Grain filling								
	maxDT	minDT	maxNT	minNT	DTR	SU	WSDI	CSDI	T10p	T90p	maxDT	minDT	maxNT	minNT	DTR	WSDI	CSDI	T10p	T90p	hts_mean_19	hts_max_26	maxDT	minDT	maxNT	minNT	DTR	WSDI	CSDI	T10p	T90p
CRI_05_2018	28.1	24.9	21.2	20.6	5.0	0.0	1.0	2.0	20.8	27.6	27.8	25.3	21.0	20.6	5.0	1.0	1.3	20.6	27.1	1.0	0.9	27.9	25.1	21.3	20.6	5.0	2.0	2.0	20.6	27.4
CRI_10_2017	26.5	22.9	20.9	19.4	4.3	0.0	2.3	1.8	19.9	26.1	26.2	24.7	20.4	19.2	5.0	1.4	1.7	19.5	26.0	1.0	0.1	27.4	24.2	20.4	18.7	5.9	2.0	2.3	18.9	26.7
CRI_12_2017	27.7	24.4	22.5	20.1	4.1	0.0	2.0	2.3	20.6	27.0	27.4	24.5	21.9	20.6	4.3	1.6	1.7	20.8	26.9	1.0	0.5	28.0	24.2	22.0	20.3	4.8	1.8	2.6	20.8	27.7
CRI_12_2018	28.9	26.3	22.8	21.0	5.5	0.0	1.0	1.5	21.6	28.7	29.2	26.9	22.5	21.4	5.8	1.6	1.1	21.6	28.7	1.0	0.9	31.3	26.9	23.2	21.3	6.4	3.4	1.5	21.7	30.4
HND_05_2017	26.0	20.2	18.0	16.3	6.1	0.0	2.0	1.0	16.7	25.6	25.9	22.6	18.4	16.8	7.0	1.0	1.0	17.1	25.7	1.0	0.1	27.3	20.2	19.0	16.6	7.0	4.8	1.0	16.9	26.7
HND_06_2017	27.1	21.0	18.0	15.7	6.5	0.4	3.4	1.2	16.4	25.5	24.6	21.9	17.6	16.2	6.1	1.4	1.1	16.4	24.3	1.0	0.0	25.9	21.0	17.8	15.6	6.4	4.2	2.9	16.3	24.8
HND_10_2016	28.0	21.4	19.4	15.9	6.1	2.2	2.7	2.4	16.9	26.7	27.6	21.0	18.7	15.6	6.7	1.7	1.5	16.2	26.7	0.8	0.3	28.9	20.0	19.0	14.6	7.0	2.7	2.6	15.9	27.2
HND_10_2017	28.4	21.6	19.9	16.7	6.3	0.7	2.0	1.9	17.7	27.3	25.8	19.3	19.0	15.2	5.0	1.5	1.1	16.0	24.9	0.8	0.1	25.9	16.6	19.1	13.8	5.0	2.3	2.7	15.2	24.6
HND_SLV_09_2015	28.0	20.9	19.3	17.3	6.4	3.0	1.9	1.8	17.9	27.2	26.9	24.1	18.7	16.6	7.1	1.5	1.8	17.3	26.6	0.9	0.4	28.2	23.1	19.1	15.4	7.4	2.5	3.7	17.0	27.2
NIC_06_2016	28.7	22.9	21.2	19.5	6.2	0.3	1.0	1.9	19.9	28.4	29.1	25.5	20.9	19.6	7.1	1.0	1.5	19.9	28.6	1.0	0.9	29.4	24.7	21.4	19.5	7.0	2.9	1.9	19.8	29.0
NIC_09_2015	29.7	24.8	21.8	19.5	6.8	1.2	1.7	1.7	19.9	29.0	28.3	25.8	21.0	19.2	6.6	1.2	1.2	19.4	28.0	1.0	0.9	28.5	23.2	21.2	17.7	6.3	2.3	3.2	19.1	28.0
NIC_09_2016	30.4	26.4	22.1	20.0	6.5	1.8	1.8	2.0	20.7	29.6	28.8	24.8	21.5	19.4	5.9	1.0	1.7	20.0	28.4	1.0	0.8	29.0	24.9	21.7	18.4	6.1	3.3	1.5	19.0	28.4
NIC_12_2015	28.3	23.8	20.9	17.0	6.1	0.4	2.1	1.8	18.3	27.3	28.5	23.6	20.3	16.2	7.1	1.1	1.6	17.2	27.8	1.0	0.6	29.6	23.3	20.9	15.7	7.4	2.3	2.1	17.3	28.4
NIC_12_2016	28.3	21.3	19.7	15.0	6.0	0.4	3.4	1.9	16.8	26.2	28.3	24.5	19.5	15.7	7.7	1.8	1.2	16.9	27.6	1.0	0.5	31.0	23.5	19.9	14.9	7.8	4.3	2.1	16.6	28.4

Chapter 4 - Rank-based data synthesis of heterogeneous trials to identify the effects of climatic factors on the reaction of Musa genotypes to black leaf streak disease

This chapter is based on:

Brown, D., de Bruin, S., de Sousa, K., Abadie, C., Carpentier, S., Machida, L., & van Etten, J. (Under Review). Rank-based data synthesis of heterogeneous trials to identify the effects of climatic factors on the reaction of Musa genotypes to black leaf streak disease. *Agronomy Journal*.

Abstract

Synthesis of crop trial data can generate insights that are not available from the analysis of individual studies, but such synthesis is often constrained by the heterogeneity of data among studies. Rank-based data synthesis provides the flexibility to combine data of heterogeneous types and from different sources. We demonstrate the application of rank-based data synthesis of heterogeneous trial data to assess the effect of climatic factors on the reaction of several *Musa* genotypes to black leaf streak disease (BLSD). We aggregated data from the main public repositories of *Musa* trial data. We applied model-based recursive partitioning with the Plackett-Luce model, using climatic data as covariates. The model identified the maximum length of the dry spell as the main variable influencing differences in genotypic response to BLSD, dividing the aggregated trial dataset into humid and dry environments. We found differences in the reaction of genotypes to BLSD between these environments. In humid environments, NARITA 8 is the most resistant genotype, while in dry environments FHIA-01 is the best performing improved genotype. We also assessed reliability, which is the probability of outperforming the reference genotype (Calcutta 4). In humid environments NARITA 2, NARITA 8 and FHIA-01 had the highest reliability, while in dry environments only the landrace Saba surpassed 50% reliability.

4.1 Introduction

Crop variety trial data are essential for producing relevant information to support critical decision making in agronomy and crop science. These data are expensive to obtain, since field trials require a significant investment of time and resources. Research synthesis of heterogeneous studies can help to generalize conclusions and to better account for environmental variability, compared to the analysis of individual studies (Makowski, 2021). Furthermore, appropriate data synthesis methodologies can add value to existing crop trial data, by (1) integrating results from trials with different types of data and from different sources, (2) comparing a higher number of genotypes than would be feasible in individual field experiments, and (3) adding environmental data that were not available at the time at which each individual trial was conducted, providing new insights on the effect of environmental factors (genotype × environment interaction).

Current data synthesis methods often encounter obstacles related to poor data standardization (Eagle et al., 2017). Combining data from multiple trials with different experimental designs, measurement scales and data quality poses problems for data management and for subsequent statistical analysis (Simko & Pechenick, 2010). Simko and Pechenick (2010) proposed the use of rank-aggregation methods to allow for the combination of heterogeneous plant breeding data from different experiments. By converting diverse numerical measurements into rankings, trial data can be aggregated in one dataset and analyzed with appropriate statistical methods. Brown et al. (2020) suggested that the rank-aggregation approach could be extended to analyze heterogeneous data from crop trial evaluations under different environmental conditions by using climatic data to account for effects of abiotic stress on genotypic performance. van Etten, de Sousa, et al. (2019) analyzed farmer-participatory crop experiments in which field data were collected by farmers as rankings, following the 'tricot' approach (van Etten, Beza, et al., 2019), and then combined with environmental data. Brown et al. (submitted, 2022) demonstrate that data synthesis of tricot trial data of common bean (*Phaseolus vulgaris* L.) genotypes in Central America provides new insights to climate adaptation by predicting the performance of varieties beyond the locations in which they were tested.

Black leaf streak disease (BLSD), also known as black Sigatoka, is a fungal disease that affects banana leaves, causing necrotic lesions which gradually reduce the plant's photosynthetic capacity (Churchill, 2011). BLSD is caused by the fungus *Pseudocercospora fijiensis* (sexual morph: *Mycosphaerella fijiensis*) (Guzman et al., 2018). It is the most destructive leaf disease of bananas and is considered to be among the ten most destructive diseases in global agriculture (Pennisi, 2010). BLSD both reduces yields and fruit quality, and affects the wider environment because its control currently relies heavily on frequent applications of chemical fungicides (De Lapeyre de Bellaire et al., 2010). The cultivation of BLSD-resistant varieties is considered to be the most sustainable disease control method. Thus, the selection and release of such BLSD-resistant varieties have become crucial for

banana producers. Location-specific information on how different genotypes perform under different climatic conditions can support decision making that considers genotype by environment interactions (G × E) and is useful for *Musa* breeding programs.

Here, we apply the data synthesis approach (Brown et al., 2020) to a combination of data from various evaluations of *Musa* genotype reactions to BLSD and climatic data that are used as model covariates. To the best of our knowledge, this is the first application of such an approach to BLSD data originally collected in a non-ranking format. Multi-environment trials of *Musa* genotypes are complex to conduct because of their intricate requirements of space and time (Tenkouano et al., 2012). Our main objective is to demonstrate application of the data synthesis approach to unveil the effect of climatic factors on the reaction of several *Musa* genotypes to BLSD. Our interest is to explain rather than to predict a genotype's reaction to BLSD; the models are fitted and validated accordingly. The specific objectives are to (1) aggregate heterogeneous trial data previously deemed incompatible, (2) assess the effect of climatic variables and indices on the reaction of several genotypes to BLSD, and (3) identify the best performing genotypes across different locations.

4.2 Material and methods

4.2.1 *Musa* trial data

We retrieved data from the three main public data repositories currently storing field evaluation data on *Musa* genotypes: AgTrials, MusaBase and the Musa Germplasm Information System (MGIS) (BTI et al., 2018; Hyman et al., 2017; Ruas et al., 2017). From each repository, we selected data that fulfilled the following criteria:

1) Contains evaluations of *Musa* genotype reactions to BLSD
2) Contains geographic coordinates of trial locations
3) Contains the start and end dates of the evaluation period (i.e., planting date, and either shooting or harvesting date)
4) Genotypes evaluated at least partially overlap among trials

Table 4.1 presents the studies and projects in which the selected data were originally produced, the number of evaluated genotypes, number of locations, temporal extent and the data repository from which we retrieved the data.

Table 4.1 Description of the trials included in the data synthesis

Study/Project [1]	Data repository [2]	Number of locations [3]	Experimental design	Number of blocks	Number of cycles	Number of genotypes	Temporal extent
IMTP-1	AgTrials	6	Observation plots without replication		2	17	1990-1992
IMTP-2	AgTrials	8	RCBD	5	1	19	1995-1997
IMTP-3	AgTrials	7	RCBD	3-5 [4]		40	1999-2003
Orlando Narváez (2004)	AgTrials	3	CRD		1	13	2003-2004
Irish et al. (2013)	MGIS	1	RCBD	4	2	19	2008-2009
Irish et al. (2019)	MGIS	1	RCBD	4	2	15	2014-2016
WP4-BBB-Project	MusaBase	5	RCBD	4	3	32	2016-2019

[1] IMTP, International Musa Testing Program; WP4-BBB, Working Package 4 Breeding Better Bananas Project. [2] MGIS, Musa Germplasm Information System, currently does not store trial data, but provides a curated selection of literature on phenotypic evaluations of *Musa* genotypes. [3] Some locations are repeated across studies; unique locations are presented in Figure 1. [4] Varies among trials within the same study; information corresponds to the original design but not all data were available for data synthesis (e.g., one cycle missing).

Figure 4.1 shows the 22 unique locations of the trials. To provide a general description of the climatic conditions at the trial locations, Table 4.2 presents the monthly mean temperature, precipitation and relative humidity, averaged from monthly climatologies for the reference period 1990 to 2020, which roughly covers the time span of the aggregated dataset (Copernicus Climate Change Service, 2018).

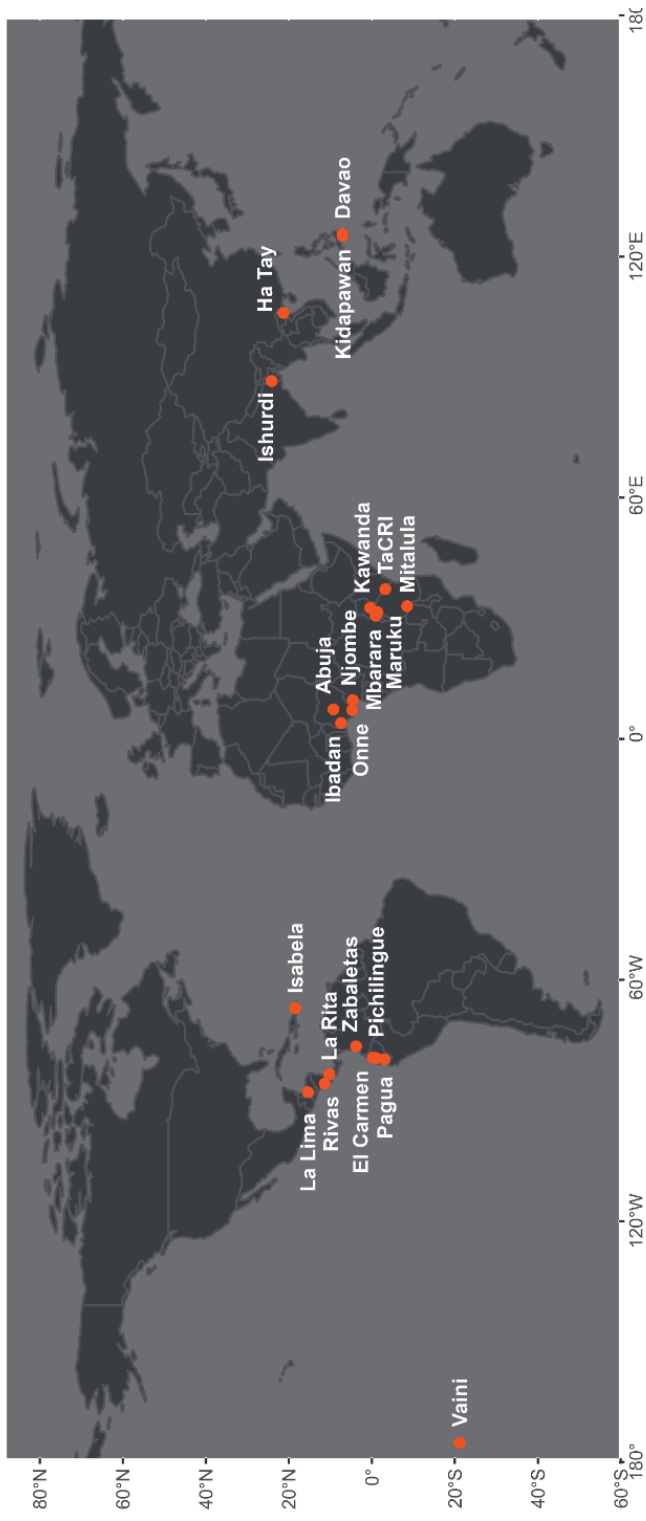


Figure 4.1 Unique locations of trials included in the aggregated data set

Table 4.2 Location name, geographic coordinates, and monthly mean climatic data (temperature, precipitation and relative humidity) for the period 1990–2020 at unique trial locations. Climatic data from Copernicus Climate Change Service (2018)

Country	Location	Longitude (°)	Latitude (°)	Temperature (°C)	Precipitation (mm)	Relative humidity (%)
Bangladesh	Ishurdi	89.033	24.133	25.57	146.38	77.04
Cameroon	Njombe	9.650	4.583	25.83	217.32	87.46
Colombia	Zabaletas	-76.517	3.817	19.40	407.45	88.99
Costa Rica	La Rita	-83.450	10.267	25.01	275.99	85.91
Ecuador	Pichilingue	-79.483	-1.100	23.84	508.66	90.27
Ecuador	El Carmen	-79.317	-0.233	23.32	232.44	87.61
Ecuador	Pagua	-79.769	-3.074	22.04	548.46	89.27
Honduras	La Lima	-87.933	15.417	24.42	166.41	86.89
Nicaragua	Rivas	-85.799	11.356	26.35	122.65	80.97
Nigeria	Abuja	7.333	9.267	26.45	134.08	64.04
Nigeria	Onne	7.167	4.767	26.20	251.58	87.43
Nigeria	Ibadan	3.900	7.433	26.09	119.72	80.79
Philippines	Davao	125.600	7.083	26.65	147.48	80.27
Philippines	Kidapawan	125.154	7.014	25.27	239.50	83.60
Puerto Rico	Isabela	-67.051	18.472	25.75	93.83	80.92
Tanzania	Mitalula	33.000	-8.500	24.57	51.18	55.59
Tanzania	Maruku	31.500	-1.333	20.93	91.26	82.07
Tanzania	TaCRI	37.246	-3.244	17.41	200.53	79.42
Tonga	Vaini	-175.167	-21.167	24.14	142.26	76.80
Uganda	Kawanda	32.600	0.417	21.08	150.45	86.99
Uganda	Mbarara	30.617	-0.933	20.91	70.96	72.74
Vietnam	Ha Tay	105.983	21.300	24.20	157.40	78.95

4.2.2 Converting data to rankings

To analyze the data using the Plackett-Luce model (Luce, 1959; Plackett, 1975), the data were transformed from numerical BLSD measurements to a ranking format. Four different metrics for BLSD evaluation were used in the original trial data (Table 4.3). The scores for these metrics were used to rank the genotypes within each trial, where the order of the ranking depended on the metric used in a trial. For instance, in evaluations where the youngest leaf spotted (YLS) was reported, genotypes with the highest values were ranked first. Conversely, for rating scales such as disease infection index (DII), representing the diseased surface per plant, genotypes with the lowest values were ranked first. If multiple measurements were

made in the same evaluation, we selected the metric for which the most data were available and which resulted in a more homogeneous aggregated dataset. In the final aggregated dataset, 33 of the rankings were derived from the underlying variable YLS, 3 from disease development time (DDT), 41 from DII, 24 from the number of standing leaves (NSL), and 9 from the number of functional leaves (NFL).

Table 4.3 Description of metrics commonly used for BLSD evaluations, adapted from BTI et al. (2018), Jones and Tézenas du Montcel (1994) and Orjeda (1998)

Metric	Description
Disease development time (DDT)	Number of days from infection (occurring at the appearance of the unfolded leaf) to the appearance of 10 or more necrotic mature lesions.
Disease infection index (DII)	$DII = \frac{\sum nb}{(N-1)T} \times 100$ N = number of grades in the scale n = number of leaves in each grade b = grade T = total number of leaves scored
Number of functional leaves (NFL)	Total number of leaves per plant with more than 50% green area.
Number of standing leaves (NSL)	Total number of standing leaves per plant, starting from the highest unfolded leaf, regardless of infection status.
Youngest leaf spotted (YLS)	The first (from top to bottom) fully unfolded leaf with 10 or more necrotic lesions.

If the trial data were reported at plant or block level, we treated each block as an individual experiment and the summarized results at block level generated a ranking. On the contrary, where trial data were available only as summarized results, the ranking was constructed using the averaged values. Each evaluation cycle generated a separate ranking. For example, trials evaluating over two cycles (i.e., mother and first ratoon) produced two separate rankings. The evaluation also depends on decisions made when the trial was designed. For instance, YLS is usually registered at shooting but in some cases, it may have been registered only at harvest or at both. These differences were considered in order to define the period for which climatic data were obtained for modelling (see Section 4.2.3). In the aggregated dataset, the evaluation period for all YLS, DDT, and NFL and for one DII was from planting to shooting. For the rest of the data, where the BLSD metric was DII, DSI and NSL, the evaluation period was from planting to harvest.

To convert the data from numerical measurements to ranks, we used the function *rank_numeric()* from the R package *gosset* (de Sousa et al., 2022). The rankings were constructed separately for each evaluation and then aggregated into a sparse matrix, where

the columns are the evaluated genotypes, and the rows correspond to the evaluations. After aggregating all the evaluations, the number of rankings was 110, with 62 genotypes evaluated. Information about the evaluated genotypes is presented in Table 4.4. There are three types of genotype: (1) a crop wild relative, (2) landraces, and (3) improved genotypes either through breeding (hybrids) or selection of somaclonal variants.

Guzman et al. (2018) described two types of reactions and three types of interactions of *Musa* genotypes to and with BLSD:

1) Incompatible interaction characterized by high resistance (HR) or hypersensitivity (phenotype 1). This expression of high resistance is characterized by the blockage of disease development at an early stage and YLS is not observed.
2) Compatible interaction with two types of reaction:
    2.1 Partial resistance (phenotype 2) expressed by slow disease evolution (from first streak symptoms to spots) and a reduction in pathogen reproduction. The YLS is high. There is a large progression of response with phenotype 2 from resistance to almost complete susceptibility.
    2.2 Susceptibility (phenotype 3) expressed by rapid disease evolution; YLS is low.

The BLSD reactions indicated in Table 4 follow the types of reaction described above. The classification of any genotype into one of the three phenotypes results from evaluating the host reaction under field conditions in comparison with references of known resistance phenotypes. We used Calcutta 4 as the reference in our analysis; it is frequently used as a highly resistant reference in BLSD evaluations.

Table 4.4 Main characteristics of *Musa* genotypes included in the data synthesis

Name	Status	BLSD reaction [1]	Genome	Subgroup/Type	Breeding program	Information source
Agbagba	Landrace	S	AAB	Plantain	*NA*	(Ferris et al., 1999)
BITA-2	Hybrid	PR	ABBB	Plantain	IITA	(Ning et al., 2007); (Noupadja & Tomekpé, 2001)
BITA-3	Hybrid	PR	AAAB	Plantain	IITA	(Bakry et al., 2009); (Ortiz & Vuylsteke, 1998)
Burro Cemsa	Landrace	PR	ABB	Bluggoe	*NA*	(Alvarez, 1997)
Cachaco	Landrace	PR	ABB	Bluggoe	*NA*	(Guzman et al., 2018)
Calcutta 4	Wild	HR	AAw	*M. acuminata* subsp. *burmannicoides*	*NA*	(Guzman et al., 2018)
CRBP-39	Hybrid	PR	AAAB	Plantain	CARBAP	(Cohan et al., 2003)
EMB 403	Hybrid	PR	AAAB		EMBRAPA	(Hernández Núñez, 1995)
FHIA-01	Hybrid	PR	AAAB	Pome	FHIA	(Irish et al., 2013)
FHIA-02	Hybrid	PR	AAAA		FHIA	(Karamura et al., 2012)
FHIA-03	Hybrid	PR	AABB		FHIA	(Karamura et al., 2012)
FHIA-17	Hybrid	PR	AAAA		FHIA	(Irish et al., 2013)
FHIA-18	Hybrid	PR	AAAB	Pome	FHIA	(Irish et al., 2013)
FHIA-20	Hybrid	PR	AAAB	Plantain	FHIA	(Sakyi-Dawson et al., 2008)
FHIA-21	Hybrid	PR	AAAB	Plantain	FHIA	(Irish et al., 2013)
FHIA-23	Hybrid	PR	AAAA		FHIA	(Orjeda, 2000)
FHIA-25	Hybrid	PR	AAB		FHIA	(Njukwe et al., 2010)
French Sombre	Landrace	S	AAB	Plantain	*NA*	(Guzman et al., 2018)
Grande Naine	Landrace	S	AAA	Cavendish	*NA*	(Guzman et al., 2018)
Gros Michel	Landrace	S	AAA	Gros Michel	*NA*	(Guzman et al., 2018)

Kisansa	Landrace	S	AAA	EAHB	*NA*	(Kimunye, Were, et al., 2021); (Nyombi et al., 2009)
Lakatan	Landrace	U	AA		*NA*	IMTP-2 data
Mbwazirume	Landrace	S	AAA	EAHB	*NA*	(Kimunye, Were, et al., 2021)
Nakitembe	Landrace	S	AAA	EAHB	*NA*	(Karamura & Karamura, 1994); (Kimunye, Were, et al., 2021)
NARITA 2	Hybrid	PR	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021)
NARITA 4	Hybrid	PR	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021)
NARITA 6	Hybrid	PR	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021)
NARITA 7	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 8	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 9	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 10	Hybrid	S	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 11	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 12	Hybrid	S	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 13	Hybrid	S	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 14	Hybrid	PR	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 15	Hybrid	S	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 16	Hybrid	PR	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 18	Hybrid	S	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 19	Hybrid	S	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 20	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 21	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 22	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)

NARITA 23	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 24	Hybrid	S	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021; Tushemereirwe et al., 2015)
NARITA 25	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 26	Hybrid	PR	AAA	EAHB	NARO-IITA	(Tushemereirwe et al., 2015)
NARITA 27	Hybrid	PR	AAA	EAHB	NARO-IITA	(Kimunye, Jomanga, et al., 2021)
Ndizi Uganda	Landrace	S	AAA	EAHB	*NA*	(Karamura et al., 2012); (Kimunye, Were, et al., 2021)
Niyarma Yik	Landrace	S	AA		*NA*	(Guzman et al., 2018)
PA 03-22	Hybrid	S	AAAB	Pome	EMBRAPA	(Irish et al., 2013)
Pisang Berlin	Landrace	S	AA	Inarnibal	*NA*	(Guzman et al., 2018)
Pisang Ceylan	Landrace	PR	AAB	Mysore	*NA*	(Guzman et al., 2018)
Pisang Jari Buaya	Landrace	S	AA	Pisang Jari Buaya	*NA*	(Kimunye, Were, et al., 2021)
PITA-16	Hybrid	S	AAB	Plantain	IITA	(Irish et al., 2013)
PV 03-44	Hybrid	PR	AAAB	Pome	EMBRAPA	(Irish et al., 2013)
Rose	Landrace	PR	AA	*M. acuminata* subsp. *malaccensis*	*NA*	(Kimunye, Were, et al., 2021)
Saba	Landrace	PR	ABB	Monthan	*NA*	(Guzman et al., 2018)
SH-3436-9	Somaclonal variant	PR	AAAA	Highgate	INIVIT	(Orjeda, 2000)
SH-3640	Hybrid	S	AAAB		FHIA	(Irish et al., 2013)
Valery	Landrace	S	AAA	Cavendish	*NA*	(Guzman et al., 2018)
Williams	Landrace	S	AAA	Cavendish	*NA*	(Guzman et al., 2018)
Yangambi Km5	Landrace	HR	AAA	Ibota Bota	*NA*	(Guzman et al., 2018)

[1] HR, highly resistant; PR, partially resistant; S, susceptible; U, unknown; *NA*, not applicable. EAHB landraces were assumed to be susceptible. NARITAs were assumed to be partially resistant following Tushemereirwe et al. (2015), except those indicated as susceptible by Kimunye, Jomanga, et al. (2021).

4.2.3 Climatic data

To account for climatic effects on the plant host reaction to BLSD, we used climatic data retrieved from the AgERA5 database (Hendrik Boogaard & Gerald van der Grijn, 2020; Copernicus Climate Change Service, 2020). We downloaded the following variables: daytime maximum temperature, nighttime minimum temperature, daily precipitation flux, and relative humidity at 06:00, 09:00, 12:00, 15:00 and 18:00. The initial set of climatic covariates (temperature, precipitation and relative humidity) were chosen based on guidelines for *Musa* disease evaluation (Orjeda, 1998) and previous studies relating BLSD to climatic factors (Churchill, 2011; Jacome & Schuh, 1992; Yonow et al., 2019). The climatic data were extracted for the period of each evaluation that generated a ranking. This corresponds to either the time from planting to shooting or from planting to harvest, depending on what was reported in each evaluation. If more than one cycle was reported, the start time for the ratoon cycle was computed as the end time of the previous crop cycle. For example, for a second cycle in which YLS at shooting is reported, the evaluation period of the mother is from planting to shooting, whereas for the first ratoon cycle the evaluation time is from the shooting of the mother to the shooting of the first ratoon cycle. All relative humidity variables were averaged over the evaluation period of each trial (i.e., either planting to shooting or planting to harvest). Precipitation and temperature variables were used as inputs with R package *climatrends* (de Sousa, van Etten, et al., 2020) to compute climatic indices (Table 4.5).

Table 4.5 Climatic variables and indices used as model covariates. Indices were calculated for the evaluation period of each trial (e.g., planting to shooting)

Variable	Description	Unit
MLDS	Maximum length of dry spell (consecutive days with precipitation < 1 mm)	day
MLWS	Maximum length of wet spell (consecutive days with precipitation ≥ 1 mm)	day
r10mm	Number of heavy precipitation days (10 ≤ rain < 20 mm)	day
r20mm	Number of very heavy precipitation days (rain ≥ 20 mm)	day
R95p	Total precipitation when rain > 95th percentile	mm
R99p	Total precipitation when rain > 99th percentile	mm
Rtotal	Total precipitation (mm) on wet days (rain ≥ 1 mm)	mm
Rx1day	Maximum 1-day precipitation	mm
Rx5day	Maximum 5-day precipitation	mm
SDII	Simple daily intensity index (total precipitation divided by the number of wet days)	mm/day
rhum_06h	Daily relative humidity at 06:00, averaged over the evaluation period	%
rhum_09h	Daily relative humidity at 09:00, averaged over the evaluation period	%
rhum_12h	Daily relative humidity at 12:00, averaged over the evaluation period	%
rhum_15h	Daily relative humidity at 15:00, averaged over the evaluation period	%
rhum_18h	Daily relative humidity at 18:00, averaged over the evaluation period	%
CSDI	Cold spell duration index (maximum consecutive nights with temperature <10th percentile)	day
DTR	Diurnal temperature range (mean difference between daily maximum temperature and daily minimum temperature)	°C
maxDT	Maximum daytime temperature	°C
maxNT	Maximum nighttime temperature	°C
minDT	Minimum daytime temperature	°C
minNT	Minimum nighttime temperature	°C
SU	Summer days (number of days with maximum temperature > 30 °C)	°C
T10p	10th percentile of night temperature	°C
T90p	90th percentile of day temperature	°C
TR[1]	Tropical nights (number of nights with maximum temperature > 25 °C)	°C
WSDI	Warm spell duration index (maximum consecutive days with temperature >90th percentile)	day

[1]Removed from the model because it has near zero variability.

4.2.4 Plackett-Luce trees

The Plackett-Luce model (Luce, 1959; Plackett, 1975) is a statistical model for ranking data. Its implementation in the R package *PlackettLuce* (Turner et al., 2020) is an extension of the original model that allows ties and partial rankings, although ties of order four or higher are

difficult to analyze. We briefly describe the Plackett-Luce model and recursive partitioning here. For a more detailed description of its implementation in R, refer to Turner et al. (2020). For a given set of items, the model estimates the worth of each item, which represents the probability that each item is selected over the rest in the set. To account for context-specific differences among rankings, the *PlackettLuce* package is coupled with the model-based recursive partitioning framework of the R package *partykit* (Hothorn & Zeileis, 2015; Zeileis et al., 2008). The combination of the recursive partitioning algorithm (Zeileis et al., 2008) and the Plackett-Luce model (Turner et al., 2020) proceeds as follows:

1) The model is fitted to the entire dataset.
2) The influence of each covariate on the stability of worth parameters is assessed.
3) If there is significant instability, the dataset is split by the covariate with the strongest instability, according to a threshold that provides the greatest improvement in model fit.
4) The algorithm stops either when no more instabilities are detected or when the resulting partitions are smaller than a pre-specified minimum threshold. Otherwise, steps 1-3 are repeated.

The threshold in step 3 is specified by the α parameter in the Plackett-Luce tree model function, which defines the threshold significance level at which to admit a split (Zeileis et al., 2008). We used a Bonferroni-corrected α value of 0.05. In step 4, the threshold refers to the minimum number of observations required in a node. In our case, we set the minimum node size as 35% of the aggregated dataset. In predictive mode, this threshold might be considered as too conservative. However, we are modelling in explanatory mode and very small nodes might be uninformative. Furthermore, the model requires some level of connectivity among compared items, which might not be feasible with very few observations per node. Further details about connectivity are provided in Section 4.2.5.

4.2.5 Network connectivity

To compute finite maximum likelihood estimates (MLE) and standard errors with *PlackettLuce*, the network of items should be strongly connected, which means that a win-lose relationship should exist between each pair of items included in the rankings (Hunter, 2004; Turner et al., 2020). From the initial aggregated dataset, we removed genotypes that were weakly connected to others. To guarantee the representativeness of genotypes among trials and avoid extremely biased comparisons, we only kept genotypes that were present in at least 5% of the trials. This threshold was defined through an iterative search of the minimum amount of data that guarantees model convergence and minimizes data loss. It resulted in 37% of the genotypes being discarded. When a Plackett-Luce tree is fitted, a network that is initially strongly connected might become weakly connected if the dataset is divided by the recursive partitioning algorithm. The *PlackettLuce* package provides a solution for networks that are not strongly connected, through the inclusion of pseudo-rankings (Turner et al., 2020). These are symmetric wins and losses between an artificial item

and each of the real items, which make the network strongly connected. Even though our initial network was strongly connected (Figure 4.2) after removing the weakly connected genotypes, we used the pseudo-rankings mechanism to guarantee the connectivity of the network after partitioning splits. The use of pseudo-rankings also reduces the variance and bias of the worth parameter estimators (Turner et al., 2020). The number of pseudo-rankings to the artificial item is set by the *npseudo* parameter in the Plackett-Luce model; we set *npseudo* = 2. Network connectivity is also important to guarantee tree stability, as detailed in Section 2.7. Cultivars Williams and Gros Michel are commonly used as susceptible references in BLSD evaluations. Pisang Ceylan is often used as a partially resistant reference. Figure 4.2 shows how these reference genotypes enabled the comparison of genotypes which were not compared to each other in the same trial. For instance, the FHIAs and NARITAs were not compared directly in any of the aggregated trials.

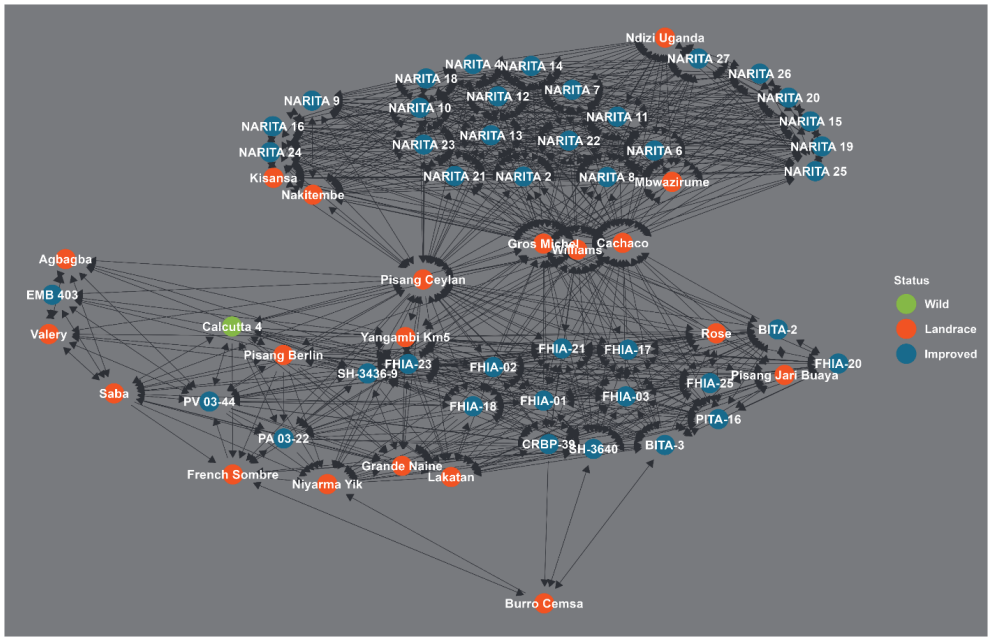


Figure 4.2 Network of win (outgoing arrows) and lose (incoming arrows) relationships between each pair of evaluated *Musa* genotypes. Incoming and outgoing arrows between the same pair indicate different outcomes at independent evaluations.

4.2.6 Handling tied ranks

In the present analysis, we found many cases in which the evaluated genotypes were tied, especially in lower rank positions. The Plackett-Luce model implemented in the R package *PlackettLuce* can handle ties up to an order of four (Turner et al., 2020). Unfortunately, the large number of ties in our data prevented model convergence. Therefore, from the subset of tied genotypes in a single rank, we removed all except one. To avoid bias, the genotype to be

retained was randomly selected. We found that changing which genotype was retained had a negligible influence on the overall results.

4.2.7 Stability assessment of the Plackett-Luce tree

A well-known limitation of recursive partitioning as used in Plackett-Luce trees is model instability, which means that a small change in the input data can cause large changes in the output (Breiman, 1996; Philipp et al., 2018; Strobl et al., 2009). Stability is fundamental in the explanatory models to guarantee that similar model results (i.e., semantic similarity of the learned predictor-response relationship) are obtained using the same algorithm on a different random sample from the same data generation process (Philipp et al., 2018). We assessed stability through data sampling using the framework proposed by Philipp et al. (2018) and implemented in the R package *stablelearner* (Philipp et al., 2016). The method consists of fitting an ensemble of trees, each with a resampled instance of the original dataset, and counting the number of times each variable generates a split (Philipp et al., 2016). If the variables selected for splitting in the original tree are consistently selected across the ensemble of trees, then the original tree is considered stable (Philipp et al., 2016). We created 1000 subsamples, each using 80% of the original data (sampling without replacement). We selected this sampling method to generate sufficiently large learning samples and to ensure strong network connectivity (see Section 4.2.5). We fitted 1000 Plackett-Luce trees and reported the relative frequencies of each variable being selected as the best splitting variable by the recursive partitioning algorithm. The relative frequencies of variable selection for splitting are reported using a histogram to facilitate interpretation.

4.3 Results

4.3.1 Plackett-Luce tree

The model selected the variable MLDS (maximum length of dry spell, consecutive days with precipitation < 1 mm) as the best splitting variable, partitioning the dataset in two. While MLDS is a rainfall-derived index, it is only very weakly correlated with rainfall indices such as Rtotal (r = -0.06) and r10mm (r = -0.03). However, MLDS has a strong negative correlation with rhum_09 (r = -0.75). Figure 4.3 shows how the environments of the two nodes selected by the model differ in terms of length of dry spell and relative humidity, expressed by the variables MLDS and rhum_09. There are no major differences between the environments in terms of rainfall variables such as r10mm and Rtotal. Therefore, the Plackett-Luce model discriminated two main environmental conditions, humid environments (node 2) and dry environments (node 3). The Plackett-Luce tree model that resulted from fitting the aggregated data is presented in Figure 4.4.

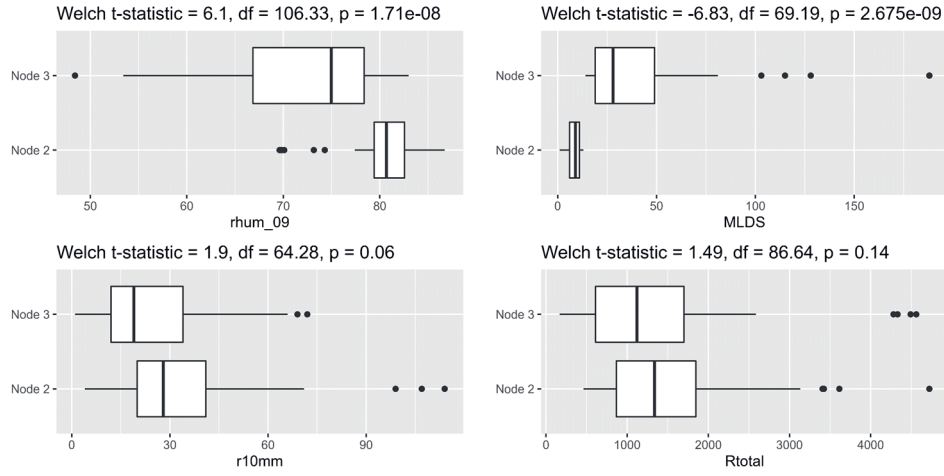


Figure 4.3 Differences between the two nodes resulting from the Plackett-Luce tree model, in terms of relative daily humidity at 09:00 (rhum_09), maximum length of dry spell (MLDS), number of days with heavy precipitation (r10mm) and total precipitation on wet days (Rtotal).

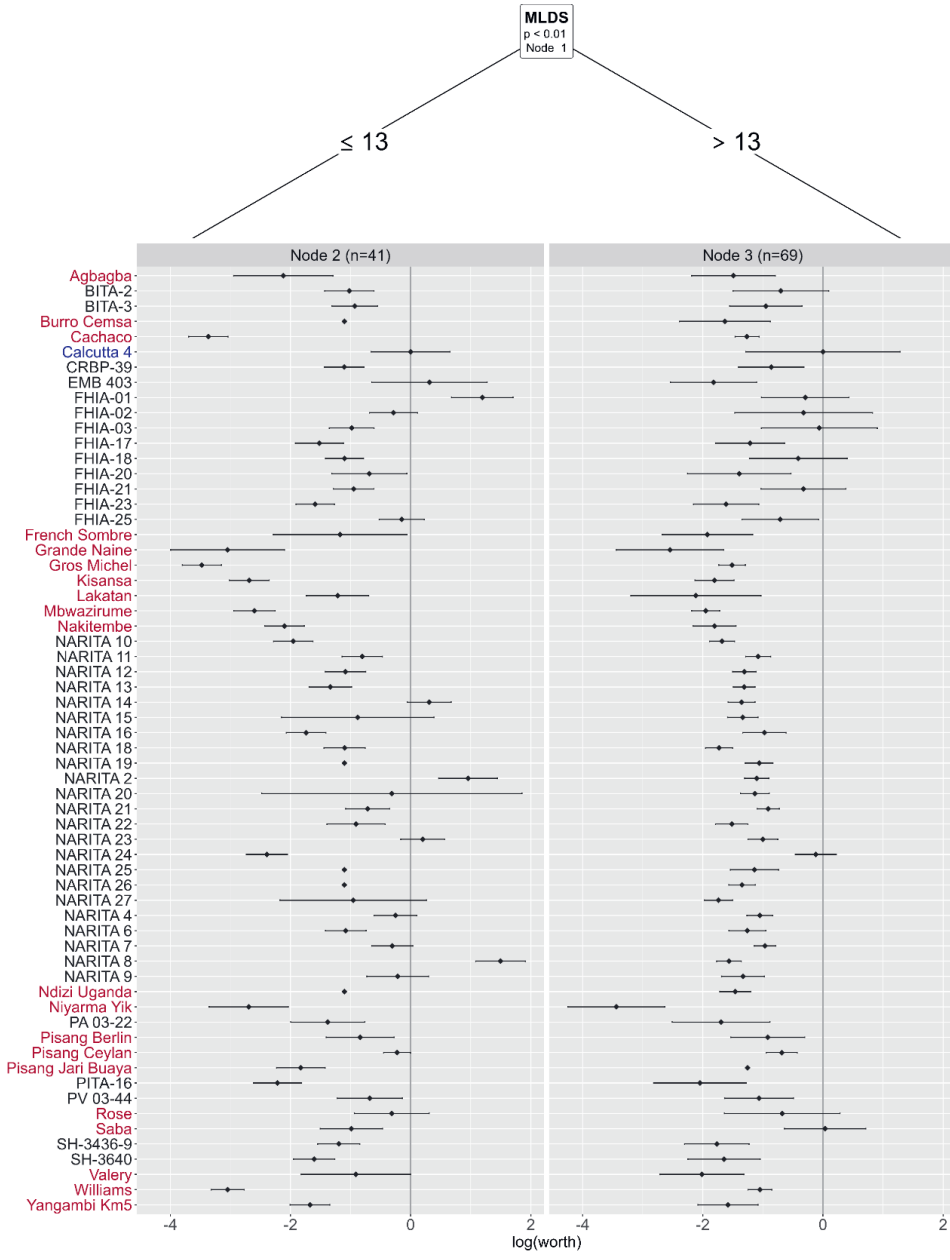


Figure 4.4 Plackett-Luce tree of *Musa* genotypes split into two nodes by variable MLDS (maximum length of dry spell) during the BLSD evaluation period. The probability of each genotype being ranked first is presented on the *x*-axis on a logarithmic scale. The probability for Calcutta 4 is zero as it served as the BLSD-resistant reference. The vertical gray lines show the zero intercept. Horizontal black bars represent quasi-standard errors of each estimated probability. The *y*-axis shows genotype names, with color indicating genotype status (blue, wild; red, landrace; gray, improved genotype).

Figure 4.5 presents a visualization of the Plackett-Luce tree model that is complementary to Figure 4.4. While some genotypes seem to outperform the reference (Calcutta 4), only NARITA 8 does so with statistical significance (Tables 4.6 and 4.7). However, there are statistically significant differences among genotypes.

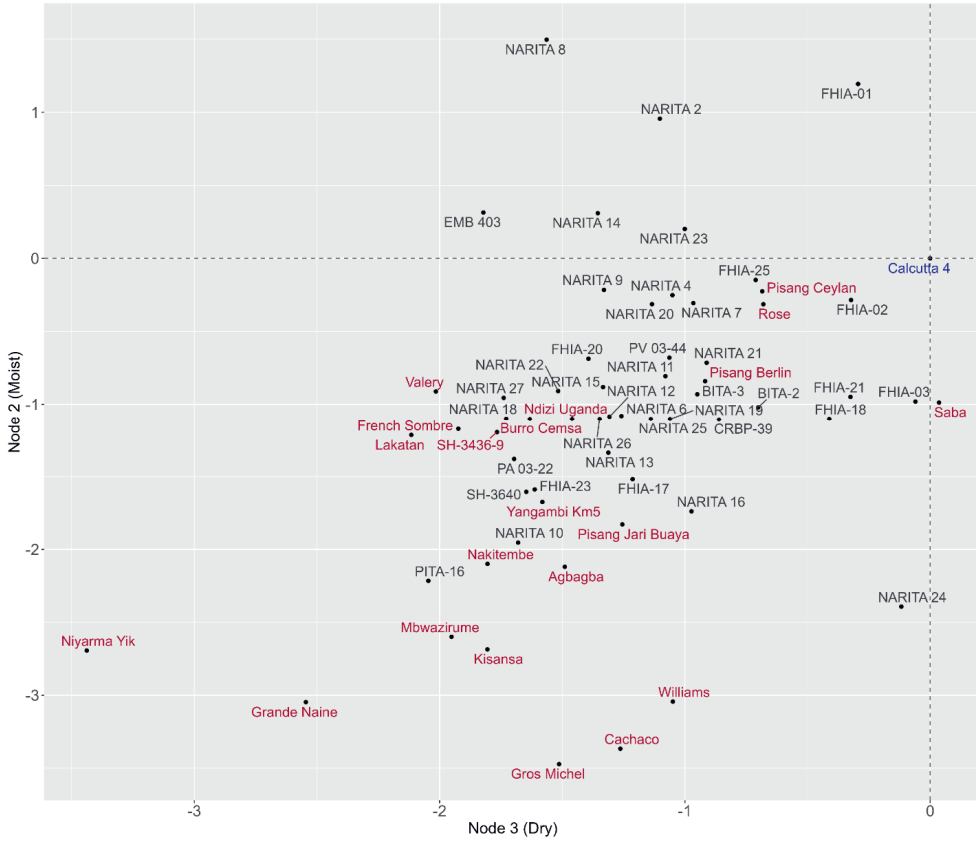


Figure 4.5 Worth estimates (on logarithmic scale) of each genotype in each of the nodes from the Plackett-Luce tree model. Color indicates genotype status (blue, wild; red, landrace; gray, improved genotype).

The horizontal black bars in Figure 4.4 represent quasi-standard errors (Firth & De Menezes, 2004) and reflect the unbalanced distribution of genotype evaluations across the environments identified by the Plackett-Luce tree model (humid and dry) (Figure 4.6).

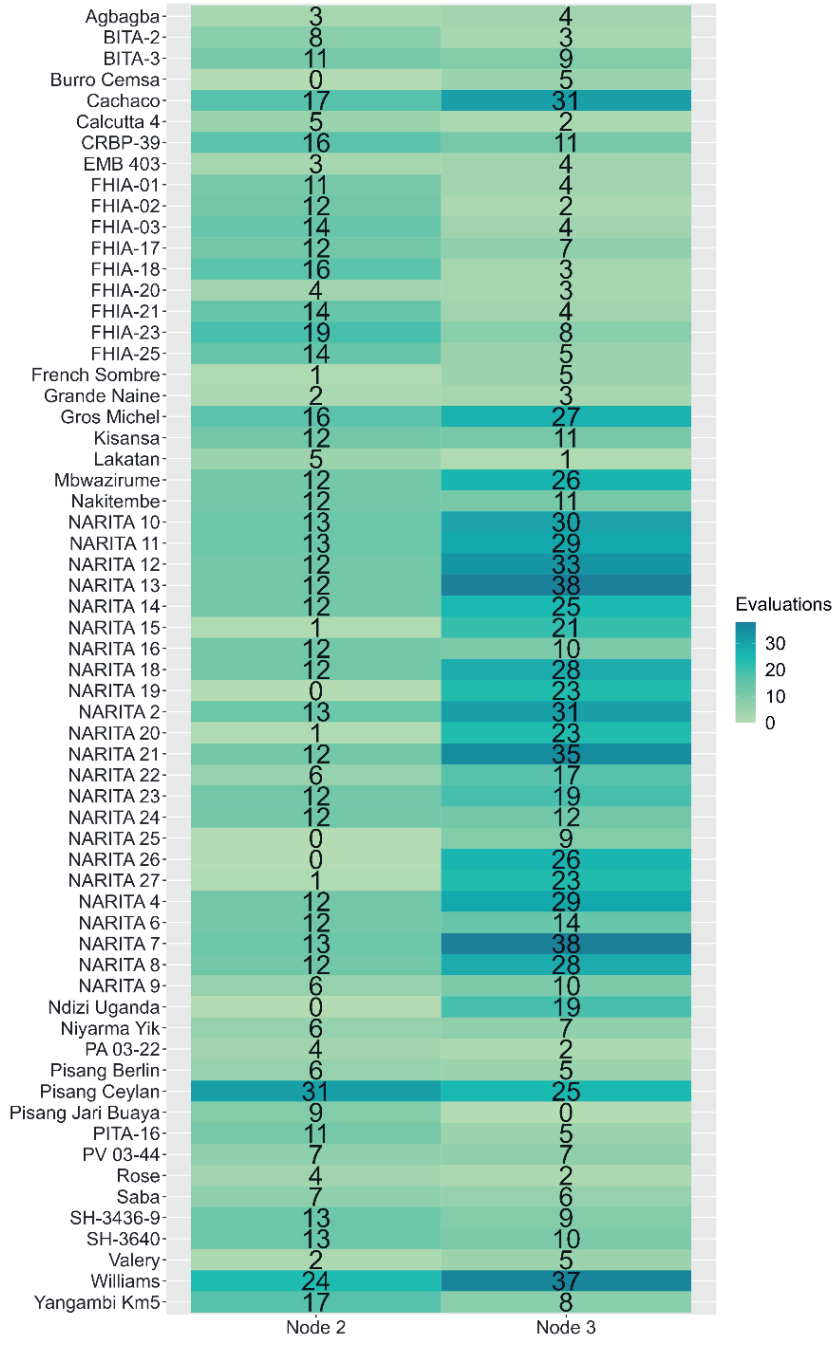


Figure 4.6 Number of times that each genotype has been evaluated in each environment (node 2 = humid; node 3 = dry).

The (log) worth estimates for the Plackett-Luce tree model are presented in Table 4.6 for node 2 (humid environments) and in Table 4.7 for node 3 (dry environments). In addition to the difference in estimated worth values, we calculated reliability, which is the probability of each genotype outperforming the check or reference genotype (Eskridge & Mumm, 1992). The reliability estimates are conservative because of the shrinkage effect of using pseudo-rankings (Section 4.2.5).

Table 4.6 Worth estimates and reliability for the top ten ranked genotypes in humid environments (node 2). The estimate denotes the log-worth probability of a genotype being ranked first. The worth of genotype Calcutta 4 is zero as it serves as the reference

Genotype	Estimate	Std. error	z value	Pr(>|z|)	quasiSE	quasiVar	Reliability	relSE
NARITA 8	1.496	0.810	1.848	0.065	0.415	0.172	0.817	0.070
FHIA-01	1.193	0.800	1.492	0.136	0.511	0.262	0.767	0.103
NARITA 2	0.956	0.851	1.123	0.261	0.489	0.240	0.722	0.108
EMB 403	0.314	1.140	0.275	0.783	0.959	0.920	0.578	0.234
NARITA 14	0.309	0.785	0.394	0.693	0.365	0.133	0.577	0.091
NARITA 23	0.201	0.786	0.256	0.798	0.364	0.133	0.550	0.091
Calcutta 4	0.000	0.000	NA	NA	0.657	0.432	NA	0.159
FHIA-25	-0.148	0.738	-0.200	0.841	0.377	0.142	0.463	0.091
NARITA 9	-0.216	0.865	-0.249	0.803	0.514	0.265	0.446	0.121
Pisang Ceylan	-0.225	0.679	-0.332	0.740	0.222	0.049	0.444	0.054

Significance levels ***, 0.001; **, 0.01; *, 0.05; relSE, standard error of reliability.

Table 4.7 Worth estimates and reliability for the top ten ranked genotypes in dry environments (node 3). The estimate denotes the log-worth probability of a genotype being ranked first. The worth of genotype Calcutta 4 is zero as it serves as the reference

Genotype	Estimate	Std. error	z value	Pr(>|z|)	quasiSE	quasiVar	Reliability	relSE
Saba	0.036	1.394	0.026	0.979	0.677	0.458	0.509	0.164
Calcutta 4	0.000	0.000	NA	NA	1.285	1.651	NA	0.283
FHIA-03	-0.060	1.586	-0.038	0.970	0.964	0.929	0.485	0.221
NARITA 24	-0.117	1.350	-0.087	0.931	0.341	0.116	0.471	0.083
FHIA-01	-0.294	1.455	-0.202	0.840	0.729	0.531	0.427	0.163
FHIA-02	-0.322	1.705	-0.189	0.850	1.146	1.313	0.420	0.233
FHIA-21	-0.325	1.413	-0.230	0.818	0.705	0.497	0.419	0.156
FHIA-18	-0.411	1.496	-0.275	0.783	0.813	0.661	0.399	0.171
Rose	-0.680	1.589	-0.428	0.669	0.962	0.925	0.336	0.174
Pisang Ceylan	-0.685	1.299	-0.527	0.598	0.258	0.066	0.335	0.055

Significance levels ***, 0.001; **, 0.01; *, 0.05; relSE, standard error of reliability.

4.3.2 Stability assessment

The results of fitting 1000 Plackett-Luce trees by subsampling 80% of the data are presented in Figure 4.7. In 855 of the 1000 Plackett-Luce trees, the splitting variable was MLDS, as shown in the case of the single tree presented in Figure 4.4. Therefore, the original tree can be considered stable.

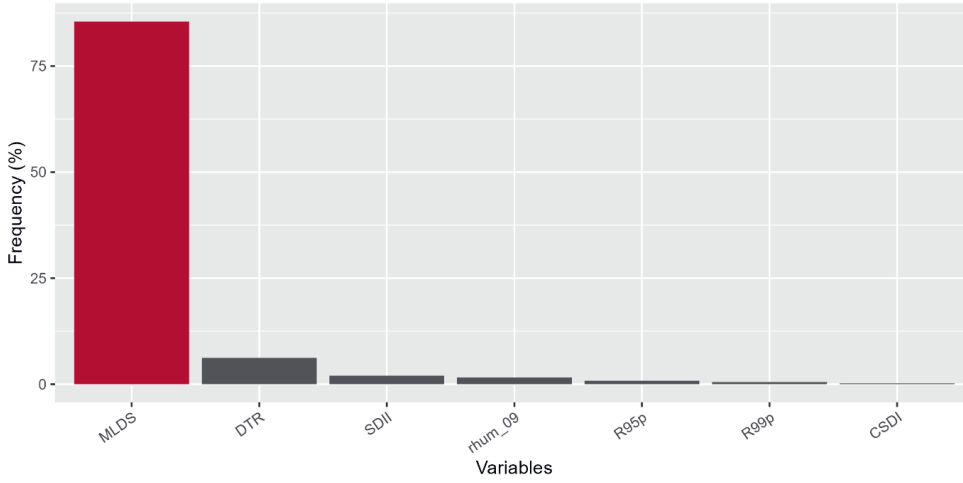


Figure 4.7 Relative frequencies with which a variable is selected for a split in each of the 1000 runs of the ensemble of trees. The red bar indicates the variable selected in the original tree. Only variables that were selected at least once are shown.

4.4 Discussion

The Plackett-Luce tree model (Figure 4.4) identified the maximum length of dry spell (MLDS) as the most important climatic variable determining *Musa* genotypes' reactions to BLSD. MLDS is a rainfall-derived climatic index, expressed as the number of days with precipitation less than 1 mm. While it is derived from rainfall, it is strongly negatively correlated with relative humidity (r = -0.75). The Plackett-Luce model partitioned the aggregated dataset into two contrasting environments: humid environments (node 2) in which the dry spell is less than or equal to 13 days, and dry environments (node 3) with a dry spell of more than 13 days. Humid environments (node 2) are assumed to be more favorable than dry environments (node 3) to BLSD development (Churchill, 2011; Guzman et al., 2018). Humidity is indeed required during various steps in the BLSD infection cycle, such as for infection efficacy (spore penetration in stomates), lesion growth on leaves and fungus sporulation (Guzman et al., 2018). Therefore, humid environments (node 2) are considered appropriate climatic conditions in which to evaluate differences in BLSD (Perez-Vicente et al., 2021). However, dry environments cannot be considered to be disease-free, as differences were found among genotypes. Environmental conditions act on both the host (banana plant) and the pathogen (*P. fijiensis*). From our results, it is not possible to discriminate between

the two but only to formulate plausible hypotheses to be further investigated or considered in future research.

From the Plackett-Luce tree, we identified a contrasting reaction of genotypes across environments. For example, NARITAs 8 and 24 have relatively extreme and opposite responses to BLSD in humid (node 2) and dry (node 3) environments. In these cases, it might be that the genetic component of resistance in these genotypes has a different expression in different environments (Craenen & Ortiz, 1997) or at a particular stage (e.g., sporulation) of the disease cycle (Abadie et al., 2003). In the case of NARITA 8, its seems that its resistance component is expressed in humid (node 2) but not in dry (node 3) environments. On the other hand, NARITA 24 appeared susceptible in humid environments (node 2), while in dry environments (node 3) its performance did not differ from that of the reference, Calcutta 4. Our results might help to explain the contradictory results among previous studies in which NARITA 24 was evaluated. For instance, in evaluations of NARITAs by (Tushemereirwe et al., 2015), NARITA 24 was among the best performing genotypes with respect to BLSD reaction. In contrast, it performed poorly in two locations in Uganda in evaluations conducted by (Kimunye, Jomanga, et al., 2021). Another contrasting result between environments is FHIA-03, which performed better in dry environments than in humid environments. We concur with Kimunye, Were, et al. (2021) that the effect of environmental factors on the reaction of genotypes to BLSD should be further investigated at different locations.

While Calcutta 4 was used as the highly resistant reference, failure of a genotype to outperform Calcutta 4 cannot be considered poor performance. Calcutta 4 has qualitative resistance, which blocks disease development at an early stage (Guzman et al., 2018). In contrast, quantitative resistance allows disease development but seems to be more durable than qualitative resistance (Guzman et al., 2018). Therefore, we advise against interpreting our results through dichotomization of whether or not a genotype outperforms the resistant reference. In dry environments (node 3), differences between the reference, Calcutta 4, and each of the genotypes included in the top ten are not statistically significant (Table 4.7). Hence, we cannot say that the top ten genotypes performed differently from the reference, but they did perform better than the rest of the genotypes.

Water is the most limiting abiotic factor for banana growth (Turner, 1995). We hypothesize that plant growth could be negatively affected in dry conditions (node 3), as banana plants react very early to water deficit (Eyland et al., 2022) and require constant rainfall for normal development (Turner et al., 2007). The genotype FHIA-01 performed well in both humid and dry environments, which is in agreement with both its partial resistance to BLSD and its tolerance of extended periods of deficient rainfall (Rowe & Rosales, 1993). Furthermore, of the top ten best performing genotypes in dry environments, six contain the *Musa balbisiana* (B) genome, which has been identified to contribute to drought tolerance (Thomas et al., 1998; van Wesemael et al., 2019; Vanhove et al., 2012). Abiotic stresses can affect plant reaction to diseases (Bostock et al., 2014). In the case of *Musa*, there is evidence that

nutritional deficiencies and poor soil conditions predispose BLSD infection (Guzman et al., 2018). Our results suggest that predisposition to BLSD caused by drought should be further explored.

Evaluation metrics based on leaf number (e.g., YLS) strongly depend on plant growth (Guzman et al., 2018). Slower plant growth in dry environments could induce some bias in rankings in which the disease variable is linked to leaf number, such as YLS, NFL and NSL. In spite of this potential bias, the Plackett-Luce tree model seems to adequately detect the different reactions of genotypes with partial resistance among environments (G × E), while it is also consistent with the expected performance of susceptible genotypes, such as Grand Naine, Gros Michel and Niyarma Yik.

Our results could help to select genotypes of interest to breeders for further testing. For example, NARITA 8, FHIA-01 and NARITA 2 were not only the best performing genotypes in humid environments, but also the most reliable in outperforming the reference Calcutta 4 (Table 6). This can provide impetus for breeding programs to include certain genotypes in new evaluations, based on their overall reliability in addition to their worth (ranking probability). In humid environments, only genotypes EMB 403, FHIA-01, and NARITAs 2, 8, 14, 20 and 23 have more than 50% reliability, while in dry environments, only the landrace Saba marginally surpassed 50% reliability. Reliability is especially relevant given the lengthy process of developing and releasing a *Musa* genotype. For instance, FHIA-21 took 30 years from crossing to release (Tenkouano et al., 2019). Our results support the use of both ranking-probabilities and reliability as criteria for selecting genotypes for further testing in new locations, minimizing the risk of investment.

4.5 Conclusions

Our work presents the first application of the data synthesis approach in combining heterogeneous trial data and using climatic data as model covariates to discover *Musa* genotypes' reactions to BLSD. We aggregated data from 110 evaluations in 31 heterogeneous trials evaluating *Musa* genotypes for reaction to BLSD. The rank-based data synthesis methodology enabled the comparison of 62 genotypes, aggregated from trials established at 22 unique locations, with a temporal range from 1990 to 2019. The large number of banana genotypes analyzed in our study would typically be challenging to compare in a single advanced multi-location testing trial. We have demonstrated how field trial environmental conditions can be reconstructed using publicly available climate datasets even where locally-sensed weather data are lacking. In our work, the large temporal range for which the AgERA5 data is available (1979 to present) allowed the addition of climatic data as model covariates for all the data selected for the study. The MLDS, a precipitation-derived climatic index, was found to be the best splitting variable in the Plackett-Luce tree model. We found that humidity is the main climatic factor driving differential reactions of genotypes to BLSD. Our results support previous evidence that genetic components of resistance to BLSD are triggered under

different environmental conditions, leading to different genotypic response patterns. We have provided insights to support the use of reliability in selecting genotypes for further evaluation.

One limitation of our study is that despite the aggregation of several trials, the sample size of the aggregated dataset is still relatively small, due to the relative scarcity of *Musa* evaluation data in public repositories. Increased availability of trial data in public repositories is therefore required for further applications of the data synthesis approach to efficiently (re)use crop trial data. We expect our findings to motivate plant breeding programs to share their data in public repositories, to enable future reanalysis with extended versions of the dataset used in our study.

Acknowledgments

We thank Harri Washington, consultant to Alliance of Bioversity International and CIAT Science Writing Service, for language editing of this paper.

Funding sources

This research was undertaken as part of, and funded by, the CGIAR Research Program on Roots, Tubers and Bananas (RTB) and supported by CGIAR Trust Fund contributors https://www.cgiar.org/funders/.

Data availability

All the R code and data required to reproduce the analysis are publicly available in the public repository https://github.com/AgrDataSci/Data_Synthesis_Musa_Trials_BLSD.

Supporting software

We used the following R (R Core Team, 2022a) packages for data management and preparation: *readr*, *readxl, janitor*, *dplyr, gosset* and *caret* (de Sousa et al., 2022; Kuhn, 2022; Wickham, François, et al., 2022). Climatic variables were obtained using the packages *ag5Tools* and *climatrends* (Brown & de Sousa, 2022; de Sousa, van Etten, et al., 2020). The statistical modelling was performed with the packages *PlackettLuce*, *stablelearner*, *gosset* and *qvcalc* (de Sousa et al., 2022; Firth, 2020; Philipp et al., 2016; Turner et al., 2020). The figures were made with packages *ggplot2*, *ggparty, patchwork* and *GGally* (Borkovec & Madin, 2019; Pedersen, 2020; Schloerke et al., 2021; Wickham, 2016). The geospatial data were handled with packages *sf* and *terra* (Hijmans, 2021; Pebesma, 2018).

Chapter 5 - *gosset*: An R package for analysis and synthesis of ranking data in agricultural experimentation

This chapter is based on:

de Sousa, K., Brown, D., Steinke, J., & van Etten, J. (Under Review). gosset: An R package for analysis and synthesis of ranking data in agricultural experimentation. *SoftwareX*.

Abstract

To derive insights from data, researchers working on agricultural experiments need appropriate data management and analysis tools. To ensure that workflows are reproducible and can be applied on a routine basis, programmatic tools are needed. Such tools are increasingly necessary for rank-based data, a type of data that is generated in on-farm experimentation and data synthesis exercises, among others. To address this need, we developed the R package *gosset*, which provides functionality for rank-based data and models. The gosset package facilitates data preparation, modelling and results presentation stages. It introduces novel functions not available in existing R packages for analyzing ranking data. This paper demonstrates the package functionality using the case study of a decentralized on-farm trial of common bean (*Phaseolus vulgaris* L.) varieties in Nicaragua.

5.1 Motivation and significance

Participatory on-farm experimentation approaches are reaching scale in agricultural research (De Roo et al., 2017). Participatory experiments often collect data as rankings, a format that is less common in other agricultural research settings (Coe, 2002). A recently developed approach for on-farm experimentation, triadic comparison of technologies (tricot), makes intensive use of data in ranking format (van Etten, Beza, et al., 2019). Also, a newly proposed approach for synthesizing crop variety evaluation data largely depends on the analysis of ranking data (Brown et al., 2020).

The analysis of ranking data requires the use of appropriate statistical models such as the Plackett-Luce model (Luce, 1959; Plackett, 1975) or the Bradley-Terry model (Bradley & Terry, 1952). Functionality for fitting these models is available in R with the packages *BradleyTerry2* (Turner & Firth, 2012) and *PlackettLuce* (Turner et al., 2020) respectively. However, extended functionality was required for the entire data science workflow, which usually includes: (1) Data preparation and cleaning, (2) modelling and validation, and (3) results presentation. For (1) gosset provides functions for converting and preparing data into a ranking or pairwise format required by the packages *PlackettLuce* and *BradleyTerry2* respectively. For (2), gosset provides functions for model selection and validation using cross-validation. In the case of (3), enhanced functionality for plotting model results is provided by the gosset package.

5.2 Software description

The R package *gosset* provides functionality supporting the analysis workflows in agricultural experimentation, especially for rank-based approaches. The package is available in the Comprehensive R Archive Network (CRAN) and can be installed by executing `install.packages("gosset")`. The package is named in honor of William Sealy Gosset, known by the pen name ‘Student’. Gosset was a pioneer of modern statistics in small sample experimental design and analysis. As a beer brewer at Guinness, he developed practical approaches to experimentation to compare barley varieties and beer brewing practices (Ziliak, 2019).

5.2.1 Software Architecture

The R package gosset is structured following the guidelines described in the manual for creating R add-on packages (R Core Team, 2022b). This structure consists of files DESCRIPTION, LICENSE, NAMESPACE and NEWS, and directories data, dev, docs, inst, man, R, and vignettes. The package functions were developed following the S3 methods style (R Core Team, 2022b) and are contained in the R sub-directory.

5.2.2 Software Functionalities

Data management and preparation

- `rank_binomial` transforms a ranking object into a binary comparison, as required by package *BradleyTerry2* (Turner & Firth, 2012).

- `rank_numeric` converts numeric values into rankings. The parameter ascending = indicates if the rankings should be made considering the numeric values in ascending order. The default is `asceding = FALSE`. This function is useful when the data have been collected as numerical observations, for instance, in an experiment measuring crop yield.

- `rank_tricot` transforms data in tricot format into *PlackettLuce* rankings.

Modelling

- `AIC` computes the Akaike Information Criterion (Akaike, 1974) for a Bradley-Terry model (Turner & Firth, 2012) or a Plackett-Luce model (Turner et al., 2020).

- `btpermute` deviance-based forward variable selection (Lysen, 2009) procedure for Bradley-Terry models.
- `crossvalidation` performs k-fold cross-validation, where k could be specified by the user. The default is 10-fold. Folds can be provided as a vector for a custom cross-validation, such as blocked cross-validation.

- `forward` executes forward variable selection with cross-validation.

- `kendallTau` computes the Kendall-tau rank correlation coefficient between two rankings (Kendall, 1938).

- `kendallW` computes Kendall's W (coefficient of concordance) among observed rankings and those predicted by the Plackett-Luce model (Kendall & Smith, 1939).

- `pseudoR2` computes goodness-of-fit metrics, such as McFadden's pseudo-R2 (McFadden, 1973).

Visualization and results presentation

- `compare` is a visualization approach to compare measures from two different methods (Bland & Altman, 1986).
- `plot` provides an alternative S3 method to `plot.pltree()` method implemented by the *PlackettLuce* package.

- `regret` computes the regret coefficients, the loss under the worst possible outcome; a common heuristic in risk assessment strategy (Loomes & Sugden, 1982).

- `reliability` computes the probability of a set of items outperforming a reference item; a common heuristic in plant breeding (Eskridge & Mumm, 1992).

- `worth_bar` creates a bar plot of the estimated worth for each evaluated item.

- `worth_map` creates a heatmap plot of the estimated log-worth for all items considering each of the evaluated traits.

5.3 Illustrative example

To demonstrate the functionality of the *gosset* package, we use the nicabean dataset, which was generated with decentralized on-farm trials of common bean (*Phaseolus vulgaris* L.) varieties in Nicaragua over five seasons (between 2015 and 2016). Following the tricot approach (van Etten, Beza, et al., 2019), farmers were asked to test in their farms three varieties of common bean. The varieties were randomly assigned as incomplete blocks, each representing 3 varieties out of a total set of 10 varieties. Each farmer assessed which of the three varieties in one incomplete block had the best and worst performance in eight traits (vigor, architecture, resistance to pests, resistance to diseases, tolerance to drought, yield, marketability, and taste). The farmers also provided their overall appreciation of the varieties, by indicating which variety had the best and the worst performance based on the overall performance considering all the traits. To analyze the data, we use the Plackett-Luce model implemented in the R package *PlackettLuce* (Turner et al., 2020).

The nicabean dataset is a list with two data frames. The first, trial, contains the trial data with farmers' evaluations, ranked from 1 to 3, with 1 being the higher ranked variety and 3 the lowest ranked variety for the given trait and incomplete block. The rankings in this dataset were previously transformed from tricot rankings (where participants indicate best and worst) to ordinal rankings using the function `rank_tricot()`. The second data frame, covar, contains the covariates associated to the on-farm trial plots and farmers. This example will require the packages *PlackettLuce* (Turner et al., 2020), *climatrends* (de Sousa, van Etten, et al., 2020), *chirps* (de Sousa, Sparks, et al., 2020) and *ggplot2* (Wickham, 2016).

```
library("gosset")
library("PlackettLuce")
library("climatrends")
library("chirps")
library("ggplot2")

data("nicabean", package = "gosset")
dat <- nicabean$trial
covar <- nicabean$covar
traits <- unique(dat$trait)
```

To start the data analysis, we transform the ordinal rankings into the Plackett-Luce rankings format (a sparse matrix) using the function `rank_numeric()`. We run iteratively over the traits adding the rankings to a list called R. Since the varieties are ranked in an ascending order, with 1 being the higher ranked and 3 the lower ranked, we use the argument `ascending = TRUE` to indicate which order should be used.

```
R <- vector(mode = "list", length = length(traits))

for (i in seq_along(traits)) {
     dat_i <- subset(dat, dat$trait == traits[i])
     R[[i]] <- rank_numeric(data = dat_i,
                            items = "item",
                            input = "rank",
                            id = "id",
                            ascending = TRUE)
}
```

Then, using the function `kendallTau()` we assess the Kendall tau (τ) coefficient (Kendall, 1938). This approach can be used, for example, to assess what traits influence farmers' choices or to prioritize traits to be tested in a next stage of tricot trials (e.g., a lighter version of tricot with no more than 4 traits to assess). We use the overall appreciation as the reference trait and compare the Kendall tau with the other 8 traits.

```
baseline <- which(grepl("OverallAppreciation", traits))

kendall <- lapply(R[-baseline], function(X){
  kendallTau(x = X, y = R[[baseline]])
})

kendall <- do.call("rbind", kendall)

kendall$trait <- traits[-baseline]
```

The Kendall correlation (Table 5.1) shows that farmers prioritized the traits yield ($\tau = 0.749$), taste ($\tau = 0.653$) and marketability ($\tau = 0.639$) when assessing overall appreciation.

Table 5.1 Kendall tau correlation between 'overall performance' and the other traits assessed in the Nicaragua bean on-farm trials

trait	kendallTau
Vigor	0.439
Architecture	0.393
ResistanceToPests	0.463
ResistanceToDiseases	0.449
ToleranceToDrought	0.411
Yield	0.749
Marketability	0.639
Taste	0.653

Then, for each trait, we fit a Plackett-Luce model using the function `PlackettLuce()` from the package of the same name. This will allow us to continue the trial data analysis using the other functions in the package *gosset*.

```
mod <- lapply(R, PlackettLuce)
```

The `worth_map()` function can be used to visually assess and compare item performance based on different characteristics. The values represented in a `worth_map` (Figure 5.1) are log-worth estimates. From the breeder or product developer perspective the function `worth_map()` offers a visualization tool to help in identifying item performance based on different characteristics and select crossing materials.

```
worth_map(mod[-baseline],
          labels = traits[-baseline],
          ref = "Amadeus 77") +
          labs(x = "Variety",
               y = "Trait")
```

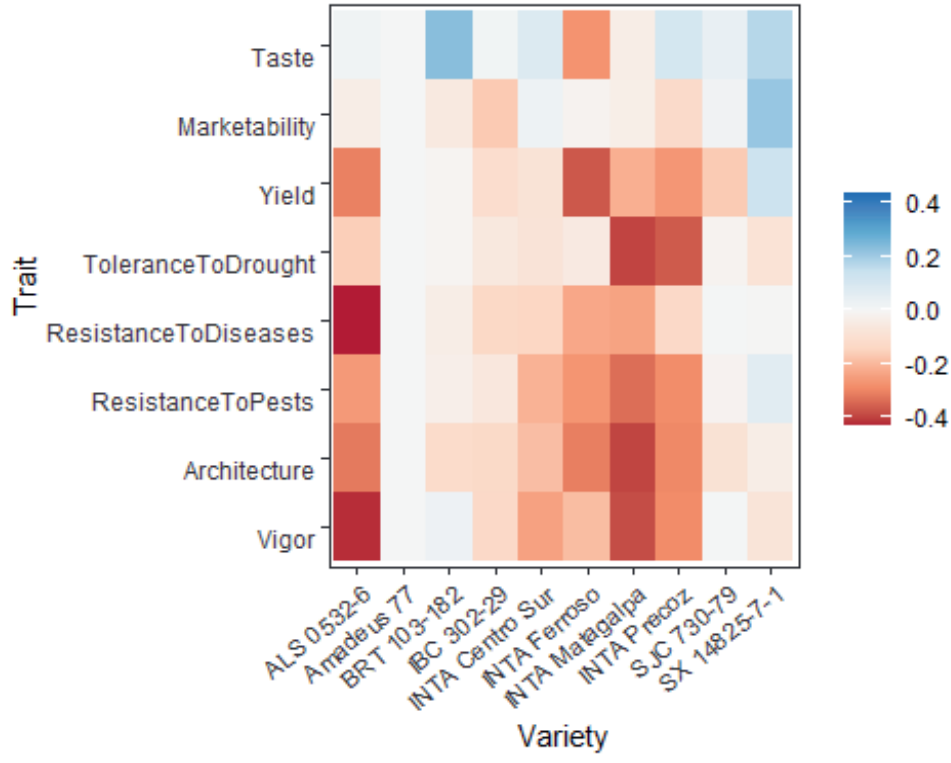


Figure 5.1 Trait performance (log-worth) of bean varieties in Nicaragua. Variety 'Amadeus' is set as reference (log-worth = 0). Blue values indicate a superior performance of varieties for a given trait, compared to the reference. Red values indicate a variety with weak performance for the given trait, compared to the reference.

To consider the effect of climate factors on yield, we use agro-climatic covariates to fit a Plackett-Luce tree. For simplicity, we use the total rainfall (Rtotal) derived from CHIRPS data (Funk et al., 2015), obtained in using the R package *chirps* (de Sousa, Sparks, et al., 2020). Additional covariates can be used in a Plackett-Luce tree, for example using temperature data from R packages *ag5Tools* (Brown & de Sousa, 2022) or *nasapower* (Sparks, 2018).

We request the CHIRPS data using the package *chirps*. Data should be returned as a matrix. This process can take some minutes to be implemented.

```
dates <- c(min(covar[, "planting_date"]),
           max(covar[, "planting_date"]) + 70)

chirps <- get_chirps(covar[, c("longitude","latitude")],
                     dates = as.character(dates),
                     as.matrix = TRUE,
                     server = "ClimateSERV")
```

We compute the rainfall indices from planting date to the first 45 days of plant growth using the function `rainfall()` from package *climatrends* (de Sousa, van Etten, et al., 2020).

```
newnames <- dimnames(chirps)[[2]]
newnames <- gsub("chirps-v2.0.", "", newnames)
newnames <- gsub("[.]", "-", newnames)

dimnames(chirps)[[2]] <- newnames

rain <- rainfall(chirps, day.one = covar$planting_date, span = 45)
```

To be linked to covariates, the rankings should be coerced to a 'grouped_rankings' object. For this we use the function `group()` from *PlackettLuce*. We retain the ranking corresponding to yield.

```
yield <- which(grepl("Yield", traits))

G <- group(R[[yield]], index = 1:length(R[[yield]]))
```

Now we can fit the Plackett-Luce tree with climate covariates.

```
pldG <- cbind(G, rain)

tree <- pltree(G ~ Rtotal, data = pldG, alpha = 0.1)
```

The following is an example of the plot (Figure 5.2) made with the function `plot()` in the *gosset* package. The functions `node_labels()`, `node_rules()` and `top_items()` can be used to identify the splitting variables in the tree, the rules used to split the tree and the best items in each node, respectively.

```
plot(tree, ref = "Amadeus 77", ci.level = 0.9)
node_labels(tree)
node_rules(tree)
top_items(tree, top = 3)
```

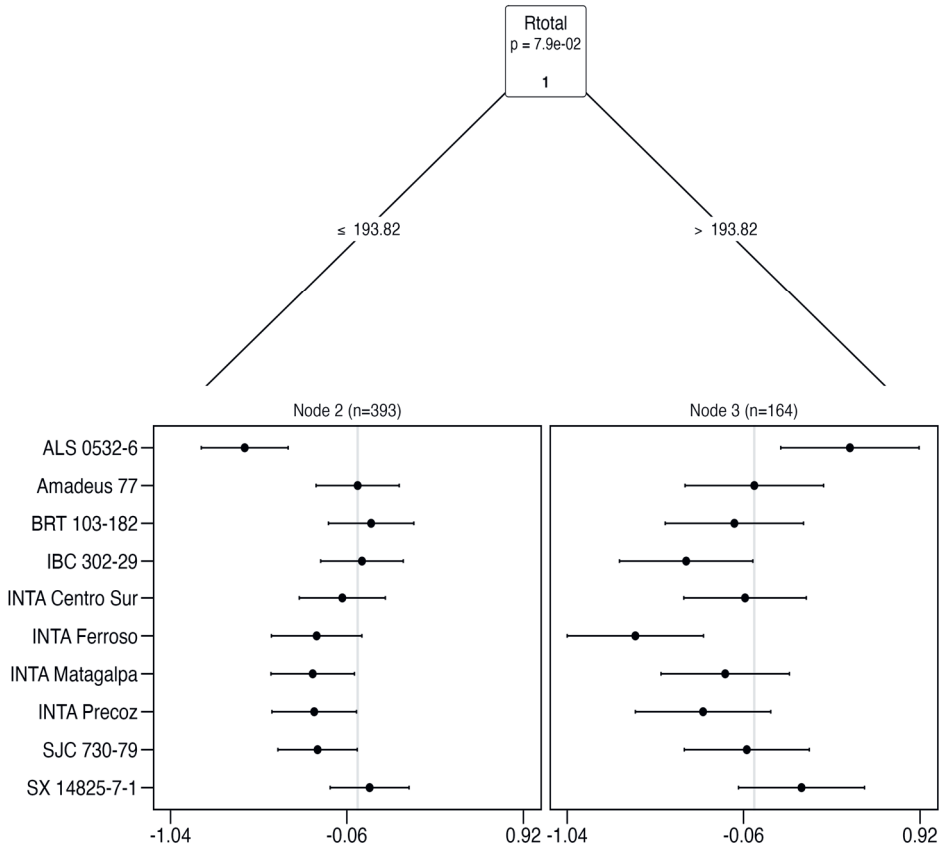


Figure 5.2 Effect of total rainfall (Rtotal) on yield of common beans in on-farm trials. Agroclimatic variables are obtained from planting date over the first 45 days of plant growth. The x axis presents log-worth, the log-probability of outperforming the other varieties in the set.

We can use the function `reliability()` to compute the reliability of the evaluated common bean varieties in each of the resulting nodes of the Plackett-Luce tree (Table 5.2). This helps in identifying the varieties with higher probability of outperforming a check variety (Amadeus 77). For the sake of simplicity, we present only the varieties with reliability ≥ 0.5.

```
reliability(tree, ref = "Amadeus 77")
```

Table 5.2 Reliability of common bean varieties based on yield performance under different rainfall conditions from planting date to the first 45 days of plant growth. Variety Amadeus 77 is set as reference

node	item	reliability	reliabilitySE	worth
2	Amadeus 77	0.500	0.035	0.114
2	BRT 103-182	0.519	0.036	0.123
2	IBC 302-29	0.506	0.035	0.117
2	SX 14825-7-1	0.517	0.033	0.122
3	ALS 0532-6	0.630	0.056	0.177
3	Amadeus 77	0.500	0.058	0.104
3	SX 14825-7-1	0.565	0.053	0.135

The results show that three varieties can marginally outperform Amadeus 77 under drier growing conditions (Rtotal $\leq$ 193.82 mm) whereas two varieties have a superior yield performance when under higher rainfall conditions (Rtotal $>$ 193.82 mm) compared to the reference. This approach helps in identifying superior varieties for different target population environments. For example, the variety ALS 0532-6 shows weak performance in the whole yield ranking, however for the sub-group of higher rainfall, the variety outperforms all the others. Combining rankings with socio-economic covariates could also support the identification of superior materials for different market segments.

A better approach for assessing the performance of varieties can be using the "Overall Appreciation", since we expect this trait to capture the performance of the variety not only for yield, but for all the other traits prioritized by farmers (Table 5.1). To assess this, we use the function `compare()` which applies the approach proposed by Bland and Altman (1986) to assess the agreement between two different measures. We compare overall vs yield. If both measures completely agree, all the varieties should be centered to 0 in the axis Y.

```
Overall <- PlackettLuce(R[[baseline]])
Yield <- PlackettLuce(R[[yield]])
compare(Overall, Yield) +
  labs(x = "Average log(worth)",
 y = "Difference (Overall Appreciation - Yield)")
```

The chart (Figure 5.3) shows no complete agreement between overall appreciation and yield. For example, variety SX 14825-7-1 shows superior performance for overall appreciation when compared with yield. Looking at the log-worth in the heat map of Figure 1, we can argue that the superior performance of the given variety is also related to taste, marketability and diseases resistance. This performance, however, was not captured when assessing only yield.

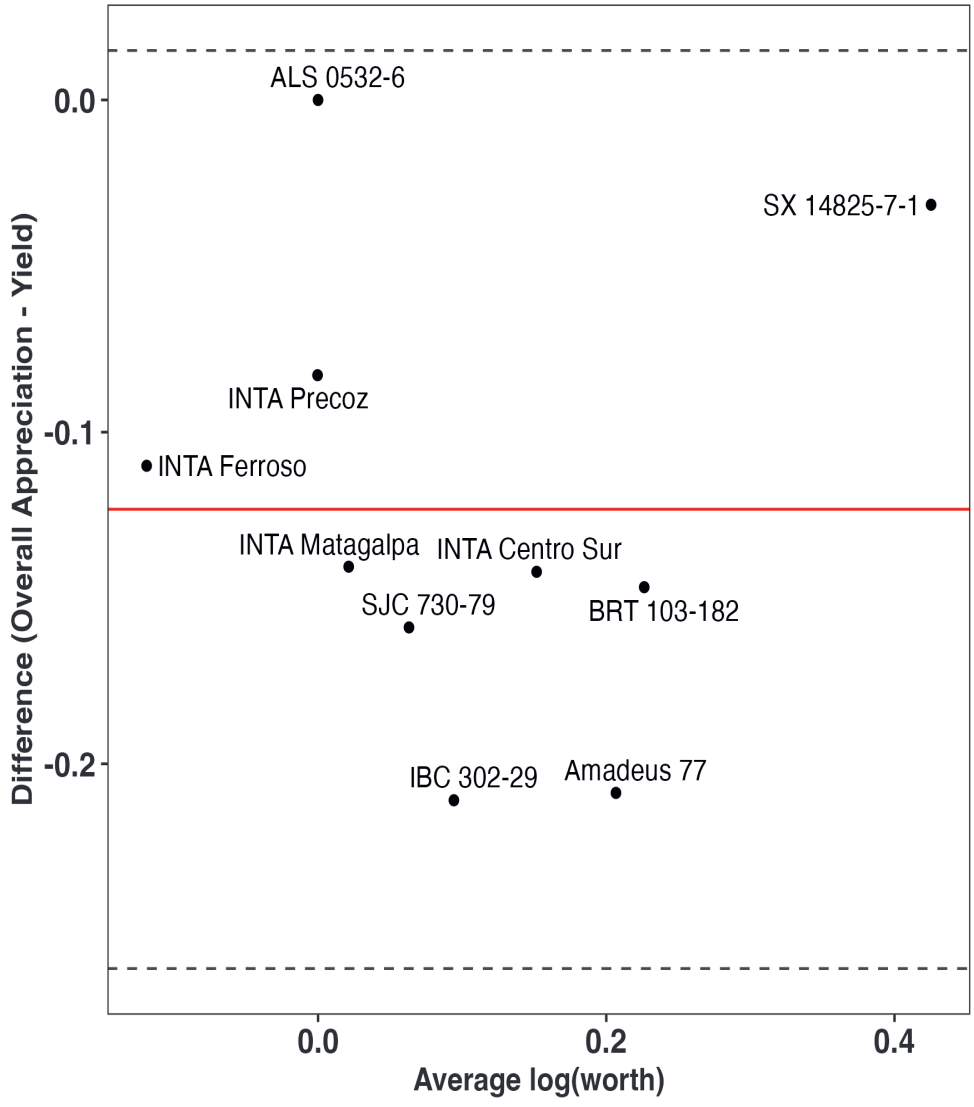


Figure 5.3 Agreement between overall appreciation and yield for crop variety performance in on-farm trials.

Here we present a simple workflow to assess crop variety performance and trait prioritization in decentralized on-farm trials with the tricot approach. A more complex workflow would also utilize other functions available in gosset, Examples include: (1) a forward selection combined with `crossvalidation()` to improve model robustness, or (2) model selection with `btpermute()` to consider all possible permutations in Bradley-Terry models, or (3) a risk analysis using regret() to support the selection of varieties, or also (4) using `rank_numeric()` to combine legacy data and deal with heterogeneous data from different

trials. All of these were previously implemented and validated elsewhere (Brown et al., 2021; de Sousa et al., 2021; Moyo et al., 2021; Steinke et al., 2019; van Etten, de Sousa, et al., 2019).

5.4 Impact

Reproducible and efficient workflows are fundamental in scientific research (Lowndes et al., 2017). The gosset package provides functionality that was not previously available from other R packages, and which enabled scientific studies based on the analysis of ranking data. This functionality allows reproducibility and greater efficiency of the entire workflow. The utility of the gosset package has been demonstrated by enabling studies based on the analysis of decentralized on-farm trial data and/or heterogeneous data from different sources. For instance, van Etten, de Sousa, et al. (2019), Moyo et al. (2021) and de Sousa et al. (2021) applied the Plackett-Luce model in combination with recursive partitioning (Turner et al., 2020; Zeileis et al., 2008). In these studies, the gosset package supported data preparation, model validation and results presentation tasks.

5.5 Conclusions

We described the functionality of the R package gosset to support the synthesis and analysis of ranking data. The package provides functions that are not available in existing R packages for analyzing ranking data. We provided an illustrative example covering the main functionality across the stages involved in the analysis workflow.

Acknowledgements

We acknowledge Vincent Johnson (Science Writing Service of the Alliance of Bioversity International and CIAT) for English editing of this manuscript. The nicabean data was generated through the Cooperative Agreement AID-OAA-F-14-00035, which was made possible by the generous support of the American people through the US Agency for International Development (USAID). The package was developed as part of the CGIAR Research Program (CRP) on Climate Change, Agriculture and Food Security (CCAFS) and the (CRP) on Roots, Tubers and Bananas (RTB), which were carried out with support from the CGIAR Trust Fund and through bilateral funding agreements (details are at https://www.cgiar.org/funders). New analytical approaches were developed during the projects Accelerated Varietal Improvement and Seed Systems in Africa (AVISA) and 10000FARMS supported by the Bill and Melinda Gates Foundation. The views expressed in this document cannot be taken to reflect the official opinions of these organizations.

Chapter 6 – *ag5Tools:* An R package for downloading and extracting agrometeorological data from the AgERA5 database

This chapter section is based on:

Brown, D., de Sousa, K., & van Etten, J. (Accepted for publication). ag5Tools: An R package for downloading and extracting agrometeorological data from the AgERA5 database. *SoftwareX*.

Abstract

Agrometeorological data is important in agricultural research, especially for investigating genotype by environment interactions. The AgERA5 dataset from the Copernicus Climate Data Store provides free and public access to daily agrometeorological data, from 1979 to present, with ready to use variables tailored for agricultural and agro-ecological studies. We developed the R package *ag5Tools*, which provides a simplified interface for downloading and extracting AgERA5 data. The package facilitates extracting time-series data for sets of geographic points in a format that can be conveniently used in statistical models applied in agricultural research. The use of the package is demonstrated with a synthetic dataset of multi-location trials in Arusha, Tanzania.

6.1 Motivation and significance

The use of climatic data as model covariates in the analysis of multilocation trials enables extracting location-specific insights, such as targeted recommendations of crop varieties (Buntaran et al., 2021; van Etten, de Sousa, et al., 2019). Several statistical and machine learning models allow incorporating climatic data as model covariates. The lack of accessibility to climatic data from local weather stations at the required temporal and spatial resolution has been an obstacle for its application (Ramirez-Villegas & Challinor, 2012). Recently, several climatic datasets with global coverage have been made freely available to the public, enabling agricultural researchers to incorporate this kind of data in their analysis.

The AgERA5 dataset provides ready-to-use agrometeorological indicators from 1979 to present for agricultural and agro-ecological research studies (Hendrik Boogaard & Gerald van der Grijn, 2020). It is derived from on the European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric re-analyses of the global climate (ERA5) data (Hendrik Boogaard & Gerald van der Grijn, 2020; Hersbach et al., 2020). It has a global coverage, with a temporal coverage from 1979 to present at a daily temporal resolution, and a spatial resolution of 0.1° × 0.1° (approximately 11 km × 11 km at the equator).

The AgERA5 dataset provides 22 variables (Table 6.1) tailored for agronomic research (Hendrik Boogaard & Gerald van der Grijn, 2020). It allows the users to get all the required variables from a single climate dataset with a homogeneous spatial resolution.

Table 6.1 Variables and statistics available for download from the AgERA5 dataset

Variable	Statistic	Time	Unit
10m wind speed	24 hours mean		$m\ s^{-1}$
2m dewpoint temperature	24 hours mean		K
2m relative humidity		06:00	%
		09:00	
		12:00	
		15:00	
		18:00	
2m temperature	24 hours maximum		K
	24 hours mean		
	24 hours minimum		
	Day time maximum		
	Day time mean		
	Nighttime mean		
	Nighttime minimum		
Cloud cover	24 hours mean		
Liquid precipitation duration fraction			
Precipitation flux			$mm\ day^{-1}$
Snow thickness	24 hours mean		cm
Snow thickness LWE	24 hours mean		cm
Solar radiation flux			$J\ m^{-2}\ day^{-1}$
Solid precipitation duration fraction			
Vapor pressure	24 hours mean		hPa

Information retrieved from: https://doi.org/10.24381/cds.6c68c9bb

The AgERA5 dataset is freely available online for downloading from the Copernicus Climate Data Store (CDS). The data can be downloaded using the CDS web interface but depending on the amount of data required, this interface might become unpractical. For example, there is a limit of 100 items, which means that only around 3 months of daily data can be downloaded in each request. Functionality for programmed downloading data from the CDS is provided by the CDS Application Programming Interface (API) (https://cds.climate.copernicus.eu/api-how-to). The CDS API is developed and supported by the ECMWF. Currently, the officially-supported API client is available only as a Python library (https://pypi.org/project/cdsapi/). The CDS API can also be used with the online CDS Toolbox. However, even when the official Python CDS API is used, the restrictions for downloading the previously-mentioned data still apply. R (R Core Team, 2022a) users can

access a wide range of ECMWF datasets, including AgERA5, through the package *ecmwfr* (Hufkens et al., 2019). The wide range of accessibility to ECMWF products provided by the package *ecmwfr* is indeed convenient for users that require several datasets in their modelling workflows. To provide this cross-dataset compatibility, several parameters are available in the package *ecmwfr*. However, for users whose main interest lies in only one climatic dataset, this large number of parameters available in the package *ecmwfr* may be confusing. For instance, new users might feel overwhelmed by just finding if a parameter is indeed required for a download request of the AgERA5 dataset or not. Therefore, when the modelling workflow relies on mainly one climatic data product, such as the AgERA5 dataset, a product-specific tool might be more convenient. Furthermore, the data limit of 100 items also applies to download requests using the *ecmwfr* package.

The AgERA5 data is provided by the CDS as Network Common Data Form (NetCDF-4) files. This type of file can be easily read and handled in R by packages like terra (Hijmans, 2021), especially if the data will be used in raster format, either as a single or multilayer SpatRaster object. However, when data is required as a point-based time series for the locations of interest, the corresponding files should be searched by date and climatic variable, which can be a tedious task, especially when the required workflow includes several time-series of different meteorological variables and statistics. When different climatic products from different sources are used, it is often the case that they are in different spatial resolutions and coordinate reference systems. Since the AgERA5 dataset provides a large number of variables tailored for agricultural research, the need for mixing datasets from different sources, and hence potential disagreement among them, is largely reduced.

6.2 Software description

We developed the R package *ag5Tools* to facilitate agricultural researchers to download and extract AgERA5 data. The package is aimed at supporting data analytics and synthesis workflows, such as the analysis and modelling of on-farm crop trials data, to assess the effect of climatic factors on a trait of interest (e.g., yield or disease resistance). In many of these workflows, the data is often required in a point-based format, such as R numeric vectors or data.frame objects.

6.2.1 Software architecture

The *ag5tools* package was developed following the R add-on packages guidelines and applying the S3 methods style (R Core Team, 2022b). The package architecture consists of seven main sub-directories: data, dev, docs, inst, man, R, and vignettes. The root directory contains the files DESCRIPTION, LICENSE, NAMESPACE and NEWS.

For the development of the package *ag5Tools*, we have used several open and free software, such as R and packages devtools, fs, terra, reticulate, and sf (Hester et al., 2021; Hijmans, 2021; Pebesma, 2018; R Core Team, 2022a; Ushey et al., 2022; Wickham et al., 2021). The

downloading functionality of package *ag5Tools* uses the Python library *cdsapi* (ECMWF, 2019).

Since the package is published in *The Comprehensive R Archive Network* (CRAN), installing it can be made from R by executing install.packages("ag5Tools"). The source code is available in the GitHub repository https://github.com/AgrDataSci/ag5Tools and the development version can be installed from there by executing devtools::install_github("agrdatasci/ag5Tools", build_vignettes = TRUE). Once installed, it can be loaded into a typical R session by library(ag5Tools). The package automatically checks and configures the local environment to fulfill the requirements for downloading data from the CDS services. The only pre-requisite is that the user should be registered with the CDS and has retrieved his or her user key. After that, the user should store the key in a file in a local hard drive, which will be retrieved automatically by the *ag5Tools* package.

6.2.2 Software functionalities

6.2.2.1 Download data

The package *ag5Tools* provides functionality for downloading the full set of variables and statistics available from the AgERA5 dataset (Table 6.1). The users can make a download request through the function *ag5_download*, which is internally parsed by the R package reticulate to the Python library cdsapi. Those internal dependencies are internally managed by the *ag5Tools* package and do not require the intervention of the user. One advantage of the *ag5Tools* package is that it also sidesteps the current limitation of download request of the CDS platform, which does not allow requesting more than 100 elements. Therefore, *ag5Tools* users can request one or more years of data, without worrying about this limitation. Since *ag5_download* is specific for the AgERA5 dataset, it requires less input parameters from the user compared to other tools. For instance, parameters such as dataset name, dataset type and file format are handled internally by the function ag5_download, providing a simplified programming interface.

6.2.2.2 Extract data

Each NetCDF files (file extension .nc) downloaded from the Copernicus Climate Change Service contains AgERA5 data for a specific day. These files can be easily read with the R package terra (Hijmans, 2021). When the data is required as numeric vector or as a data.frame, for multiple point locations and different time frames, extracting the data could be a complex task for non-expert users. The *ag5_extract* function provides a simple interface that facilitates the extraction of AgERA5 data, automatically searching each of the required files in the local hard drive. The *ag5_extract* is a generic function, which encapsulates different methods depending on the input parameters and the corresponding output (Table 6.2).

Table 6.2 Description of the methods, input parameters and outputs for the function *ag5_extract*

Method	Parameters	Output
***ag5_extract*.numeric**	***coords*:** numeric vector of length = 2 of the form (lon, lat), or a data.frame with required columns ***path*:** character indicating the path for the folder containing the AgERA5 files ***dates*:** character the dates for extracting the specified variable, a vector of length 1 extracts a single date, while a vector of length 2 indicates the start and end dates. ***variable*:** character indicating the AgERA5 variable to extract, see details for available options ***statistic*:** character, required only for some variables. ***time*:** only for variable Relative-Humidity-2m ***celsius*:** logical, if TRUE converts the temperature values from the degrees Kelvin to degrees Celsius. Only for variables "Temperature-Air-2m" and "2m_dewpoint_temperature".	A numeric vector with length equal to the number of dates between the first and second date of the input parameters *dates*. The vector names correspond to the requested dates. If only one date is provided the function returns a numeric vector of length = 1.
ag5_extract.data.frame [1]	***coords***: a *data.frame* with required columns ***start_date*:** character indicating the column name for the start of period of time to be extracted. ***end_date*:** character indicating the column name for the end of period of time to be extracted. ***lon:*** character indicating the name of the column containing the longitude values in the input *data.frame* ***lat:*** character indicating the name of the column containing the latitude values in the input *data.frame*	A list of named numeric vectors, each one corresponding to the rows in the input *data.frame.*

[1] Parameters path, variable, statistic, time and celsius are also required but omitted for brevity

6.3 Illustrative examples

6.3.1 Download data

"Example code 1" below shows the code required for downloading the maximum daytime temperature data for years 2000 to 2005 using the function *ag5_download*. The request required six parameters, whereas twelve would be required using the ecmwfr package.

```
#Example code 1 ----------------------------
library(ag5Tools)
ag5_download(variable = "2m_temperature",
             statistic = "day_time_maximum",
             day = "all",
             month = "all",
             year = 2000:2005,
             path = "C:/custom_target_folder/")
#-----------------------------------------
```

The data are downloaded to the location indicated by the path argument in the function call. Within this path, a subfolder is created for each year contained in the download request. The data is downloaded as temporary zip file named agera5_download.zip which is automatically uncompressed and deleted by the *ag5Tools* package after copying the files to the corresponding folder. The downloaded and extracted files after uncompressing the zip file are already named by the CDS using their nomenclature system. In the example above, we explicitly indicated that we wanted to download all days and months for each of the selected years, but specific days or months can also be requested. Depending on the variable, some arguments need to be specified while others do not. In the previous example, the variable 2m_temperature needs specification of the statistic day_time_maximum. In the case of relative humidity (2m_relative_humidity), a statistic should not be indicated, but indicating the time is mandatory. Example code 2 shows how to download relative humidity for times 6:00 and 18:00 for the same years as in the previous example.

```
#Example code 2 -----------------------------
ag5_download(variable = "2m_relative_humidity",
             time = c("06_00", "18_00"),
             day = "all",
             month = "all",
             year = 2000:2005,
             path = "C:/custom_target_folder/")
#-----------------------------------------
```

6.3.2 Extracting data

To demonstrate the functionality of *ag5_extract*, we use a synthetic dataset of 100 locations randomly generated across Arusha, Tanzania (Figure 6.1).

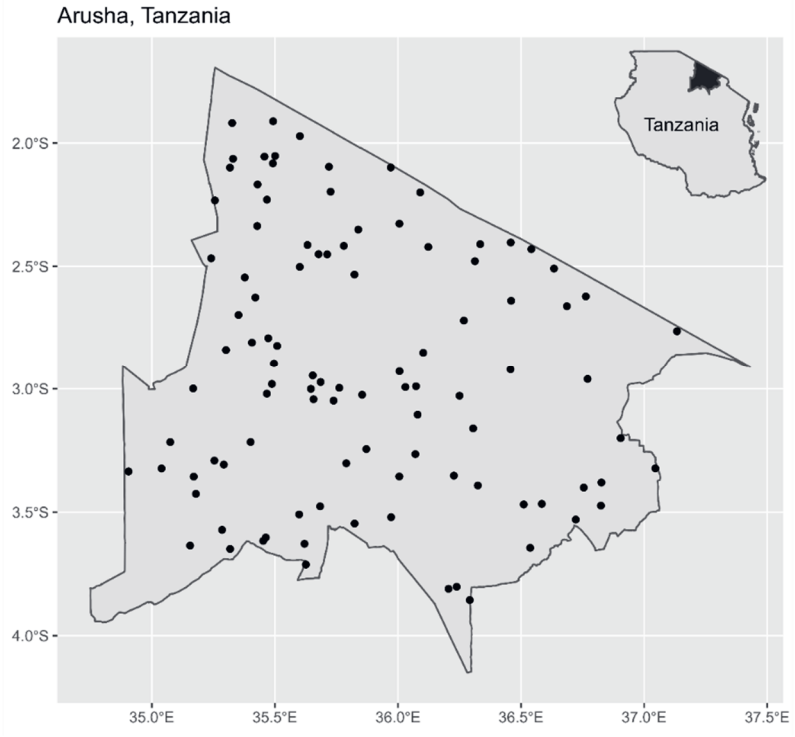


Figure 6.1 Location of the data points randomly generated in Arusha, Tanzania.

For this example, we will focus only on variable maximum daytime temperature. Table 6.3 presents the first 10 data points of the example dataset.

Table 6.3 First 10 data points of the synthetic data example, presenting geographic coordinates, along with planting and harvest dates

Longitude	Latitude	Planting date	Harvest date
35.726	-2.197	4/22/1991	8/20/1991
36.102	-2.851	1/24/1990	5/24/1990
35.463	-3.603	3/6/1991	7/4/1991
36.292	-3.856	10/10/1990	2/7/1991
35.453	-3.616	1/22/1990	5/22/1990
35.401	-3.216	10/19/1990	2/16/1991
35.170	-3.356	3/22/1990	7/20/1990
35.601	-2.502	10/14/1990	2/11/1991
36.537	-3.645	3/6/1990	7/4/1990
35.488	-2.981	4/27/1991	8/25/1991

If the data presented in Table 6.3 is stored in an *R data.frame* object, the following code can be used to extract the maximum daytime temperature data with the *ag5_extract* function.

```
#Example code 3 ----------------------------------------
arusha_maxDT <- ag5_extract(coords = arusha_data,
                            path = "D:/agera5_data/",
                            variable = "Temperature-Air-2m",
                            statistic = "Max-Day-Time",
                            start_date = "planting_date",
                            end_date = "harvest_date",
                            celsius = TRUE)
#-------------------------------------------------------------
```

AgERA5 temperature data is provided in degrees Kelvin. In our example, we set the parameter *celsius = TRUE*, to extract the data in degrees Celsius. We set the parameters *start_date* and *end_date*, as the column names in the *data.frame* have different column names, *planting_date* and *harvest_date* respectively. However, if the column names of the *data.frame* corresponding to dates are named as *start_date* and *end_date*, those parameters could be omitted. In Example code 3, the parameters *lon* and *lat* were omitted from the function call, because the column names match the default function parameters. If the column names in the input *data.frame* do not match the parameters, the column names corresponding to *lon* and *lat* should be provided as parameters in the function call. When the *coords* parameter is provided as a *data.frame*, the function returns a list of *data.frames*, each one containing a time series for each of the data points (the rows in the original *data.frame*), where column names are each of the dates from start_date to end_date. If the data is aimed at being used directly as model covariates, we need to compute the required aggregate metric (e.g., mean) for each time series. Following the same example trial dataset, Table 6.4 shows the data extracted for variables maximum daytime temperature (maxDT), minimum night temperature (minNT), precipitation (prec), solar radiation (srf), and relative humidity at time

09:00 a.m. (rhum_09). Since the variables are downloaded as daily observations, we computed the mean corresponding for time from planting to harvest of each trial data point of the synthetic dataset.

Table 6.4 Climatic variables daily maximum daytime temperature (maxDT), minimum night temperature (minNT), precipitation (prec), solar radiation flux (srf), and relative humidity at time 09:00 a.m. (rhum_09), extracted for the trial data points of the example dataset and averaged for the corresponding planting to harvest period

maxDT (°C)	minNT (°C)	prec (mm)	srf (J)	rhum_09 (%)
24.04	14.59	1.36	18778210.69	74.62
27.44	19.34	3.74	19964845.93	71.71
25.77	16.43	1.71	21830741.07	65.16
27.47	16.82	2.61	21344995.40	60.05
25.32	17.03	3.74	21727487.53	69.16
20.60	11.02	11.48	21106727.64	70.30
25.73	18.80	1.58	22175333.94	68.14
28.97	17.85	0.73	23368231.39	60.19
24.81	16.47	2.50	17724179.31	77.39
24.89	14.72	1.00	20505541.25	56.97

The data shown in Table 6.4 is ready to be used as model covariates in a statistical model. Also, the data extracted using ag5_extract can be used to calculate additional climatic variables or indices not directly available from Ag5ERA. The package *climatrends* (de Sousa, van Etten, et al., 2020) provides functionality for computing a range of climatic indices.

The *ag5_extract* function can also be used to extract data for one point location and one date or a time series for one point location. In the case of one point location, the argument *coords* should be provided as a vector of length = 2, in the form c(lon, lat). For example, using the coordinates of the first row in Table 6.3, the argument *coords* would be c(35.726, -2.197). This functionality might be useful in the case where climatic characterization of a single site is required. For instance, the example code 4 shows an example of daily precipitation data extracted for the first location of the synthetic dataset. If the data is extracted for only one date, the argument dates should be a vector of length = 1, and either a character or *Date* object. On the other hand, if a time series is required for just one location, the argument dates should be a vector of length = 2, where the first value indicates the start date and the second the end date of the series.

```
#Example code 4 ----------------------------------------
arusha_prec_01 <- ag5_extract(coords = c(35.726, -2.197162),
                              dates = c("1991-04-22",
                                        "1991-08-20"),
                              variable = "Precipitation-Flux",
                              path = "D:/agera5_data/")
#-----------------------------------------------------------
```

6.4 Impact

The study of the effects of environmental factors on any genotype's performance is important in agronomy and crop science research. For instance, in breeding trials, the environment represents the main source of yield variability (Chenu, 2015). The use of climatic data as model covariates can support the generation of location-specific insights in crop variety evaluations (Buntaran et al., 2021; de Sousa et al., 2021; van Etten, de Sousa, et al., 2019). The AgERA5 dataset provides an alternative data source when climatic data have not been collected in the field trials or when it is not available from local weather stations. This is even more relevant when a study involves several locations at regional or global scale, with disparities in terms of climate data availability. Given its wide time span (1979 to present) the AgERA5 provides an important source of climatic information for modelling purposes. Repurposing and reanalyzing legacy crop variety evaluation data, as described by Brown et al. (2020), is an example in which this wide time span is useful. For instance, the *ag5Tools* package was used by Brown et al. (2022) for downloading and extracting climatic data, which were used as covariates for modelling and predicting genotype performance. The *ag5Tools* package has been released in the Comprehensive R Archive Network (CRAN) and currently has more than 1600 downloads.

6.5 Conclusions

In this software paper, we have described the functionalities of the R package *ag5Tools*, for downloading and extracting AgERA5 data. As far as we know, this is the only R package with tools for downloading and extracting data exclusively designed for the AgERA5 dataset. The package is available freely for downloading at CRAN: https://cran.r-project.org/package=ag5Tools. We provided examples on how to download and extract AgERA5 data. Additional examples and information can be found on the package website https://agrdatasci.github.io/ag5Tools/. Since the development version is hosted in GitHub, current functionality problems or new feature requests can be managed by opening an issue in the GitHub repository.

Acknowledgements

We acknowledge the constructive criticism and useful suggestions made by Dr. Sytze de Bruin. We also acknowledge Vincent Johnson (Science Writing Service of the Alliance of Bioversity International and CIAT) for English editing of this manuscript.

Chapter 7 - General discussion

7.1 Main findings

This section summarizes the main findings of Chapters 2 to 6, which address the research objectives presented in Chapter 1.

7.1.1 Research objective 1: To describe the state of the art of data synthesis methods used in crop variety evaluation identifying the major constraints and knowledge gaps.

This objective was addressed in Chapter 2 through a literature review, exploring the main elements required for data synthesis in crop variety evaluation, and assessing the available data synthesis methods applied to crop variety evaluation. The literature review revealed that a new data synthesis approach for crop variety evaluation must integrate data, models and expert knowledge from farmers and breeders. To provide location-specific and contextualized information, the data synthesis requires data of agronomic performance, environmental factors, and food quality and consumer preference data. In the case of data, the major challenges are related to availability and integration of data. It was found that data availability is constrained by cultural and technical factors. An example of a cultural factor constraining data availability, is researchers lacking interest to share their data (Diekmann, 2012). Technical constraints include individual storage solutions developed by different research institutions, resulting in closed databases which are incompatible among research centers (Leonelli et al., 2017).

The first conclusion from the literature review was that rank-based methods have the required flexibility to aggregate data from different sources but especially in different formats and scales. It was also found that previous experiences support the rank-based methods as the way to go to combine farmers preferences with environmental data as model covariates. For instance, the work of van Etten, de Sousa, et al. (2019) demonstrated that environmental model covariates can be added to the statistical analysis of ranking data. Another point in favor of rank-based methods is that they seamlessly fit within the conceptual framework of reliability assessment, which is the probability of outperforming a check variety (Eskridge & Mumm, 1992).

The second conclusion from the literature review was that data synthesis should progress from general to specific and from simple to complex, as an alternative to fully detailed crop growth models, which has been proposed to advance the integration of different types of data into a single crop variety evaluation framework. In that sense, the literature review suggested the use of simpler methods involving statistical analysis or relatively limited crop-physiological modelling.

The third conclusion of the literature review was that real applications of the data synthesis approach in crop variety evaluation could motivate further data sharing and model

development. Chapters 3 and 4 demonstrate that the data synthesis approach can provide the versatility to be applied to standardized and non-standardized data, and for different purposes within the crop variety evaluation framework.

7.1.2 Research objective 2: To design and implement a data synthesis approach that characterizes environmental adaptation of crop varieties in time and space, and assess its strengths and limitations through two case studies.

Chapter 2 provided a general framework of a data synthesis approach in crop variety evaluation. Chapters 3 and 4 apply this general framework to design and implement a data synthesis approach to two case studies, each one with distinct contexts and goals (Table 7.1).

Table 7.1 Main differences between case studies applying the data synthesis approach in Chapters 3 and 4

Chapter	Crop	Modelling mode	Geographical extent	Original data type	Evaluations	Type of trials
3	Common bean	Predictive	Regional	Tricot (rankings)	Overall appreciation	Participatory on-farm trials
4	Bananas and Plantains	Explanatory	Global	Numerical measurements	Reaction to black leaf streak disease	Conventional on-station trials

In Chapter 3, the data synthesis approach was applied to data from tricot trials established in four countries in Central America. In this study, environmental (soils and climate) data were used as covariates in a Plackett-Luce tree model. The Plackett-Luce tree model with environmental covariates were used previously by van Tilborg (2018) and van Etten, de Sousa, et al. (2019). Chapter 3 extended those previous works in different aspects, such as model validation, spatial mapping and, quantifying the uncertainty of model predictions. The new data synthesis approach takes advantage of the different locations and times of each experiment, by adding environmental covariates which were extracted for the specific locations and time extent of each experimental plot. Spatial mapping is often used for communicating crop suitability or the outcome of species distribution models, but it is rarely used for representing predictions of the best performing genotypes across a region. The genotypes with the highest probability to be in the top-three ranking, as predicted by a Plackett-Luce tree model with environmental covariates, were mapped at cell level for the four countries considered in the study: Costa Rica, El Salvador, Honduras, and Nicaragua. Spatial maps are helpful to present location-specific information about crop variety performance, which can be used by regional or national researchers to either plan new evaluations or recommend the introduction of a variety in a country. For instance, experts in Honduras working in participatory improvement of common bean already expressed their

intention to use the information provided by the data synthesis for selecting the set of genotypes to be tested in new experiments (Marvin Gomez, personal communication, February 15, 2022). Providing more than one top performing genotype may facilitate decision makers to have a narrowed set of good options but still have some margin to adapt their choice to the local context. For instance, if the decision making is for variety recommendations, and one of the top-three is not available the other two can be recommended to farmers. On the other hand, having more than one option may also contribute implementing variety diversification and portfolio management approaches for minimizing risks associated weather variability, pest and diseases, and market fluctuations (Lin, 2011; Nalley et al., 2009; Sukcharoen & Leatham, 2016; van Etten et al., 2020). The choice of using three and not any other number is partly arbitrary for demonstrative purposes, and other arrangements could be made in future applications, such as top-five for example.

Chapter 4 demonstrates how data from different trials, established at different locations, time and testing genotypes that only overlap partially can be aggregated and analyzed to produce new insights. The data integrated in Chapter 4 were not originally rankings but numerical measurements requiring conversion to rankings to allow aggregation and analysis. This contrasts with the data used in Chapter 3, which were already collected as rankings in the tricot trials. In Chapter 1 it was stated that data synthesis in crop variety evaluation should go beyond the usual meta-analysis. In Chapter 2, the findings define data synthesis as an overarching approach that integrates data, models, and expert knowledge. This was demonstrated in Chapter 4, integrating data retrieved from public data repositories and scientific papers, deviating from traditional meta-analytic methods, which consist of a systemic literature review and a subsequent statistical analysis (Makowski et al., 2019). The data synthesis conducted in Chapter 4 integrated data from trials established at 22 different locations around the world, allowing to compare 62 *Musa* genotypes. As in the case of Chapter 3, the Chapter 4 also faced the challenge of inconsistencies in the genotype names, but in the case of Chapter 4 these were mainly caused by misspelling or mistyping variety names or by using the breeding name or code in one trial dataset and the name of release in other. The study case developed in Chapter 4 demonstrates that existing trial data can be repurposed and reanalyzed adding environmental covariates, which in many cases were not considered in the original analysis of each study. For instance, the aggregated dataset has a temporal range from 1990 to 2019, combining so-called legacy data with data from very recent experiments. This was possible by using location and time specific climatic data from the AgERA5 dataset. For instance, the climatic data was extracted for each location of the trials, and for each of the locations the time frame corresponds to the evaluation time (e.g., planting to shooting). This allows to characterize the effect of climatic factors on the reaction of the genotypes to BLSD in space and time. The results shown that the maximum length of the dry spell (MLDS) is a differential factor for the reaction of *Musa* genotypes to black leaf streak disease (BLSD). Although the length of the dry spell is a rainfall derived climatic index, it is strongly negatively correlated with relative humidity. The results of Chapter 4

expand the knowledge about how the environment affects the reaction of genotypes to BLSD, by (1) identifying MLDS as the main influencing factor in the reaction of 62 *Musa* genotypes to BLSD in the aggregated trials, and (2) characterizing the response of each of the 62 genotypes to BLSD under two contrasting environments as defined by MLDS. The results are also consistent with earlier findings, as relative humidity is important for the development of BLSD. The research reported in Chapter 4 further included a reliability assessment, which assessed the probability of a variety to outperform the check variety (Eskridge & Mumm, 1992). While the ranking probabilities estimated by the Plackett-Luce model represent the probability of each variety to be ranked first in particular environment, reliability represent the probability that a testing variety outperforms the reference or check variety across the environments (Eskridge & Mumm, 1992; Turner et al., 2020). Eskridge and Mumm (1992) proposed the use of reliability for the context of a breeding program, especially when a breeder aims to apply the safety first behavior, in which the main interest is in achieving an acceptable performance for each of the evaluated varieties (Eskridge, 1990; Eskridge & Mumm, 1992). In the study case of Chapter 4, if a breeding program wants to select a genotype for further testing, both ranking probabilities and reliability could be used as complementary information. In that case, it would not be necessary for a variety to be ranked first to advance to the next evaluation stage, but only to surpass a threshold of reliability agreed by the breeding program (e.g., 70%). This is also important if additional information about traits other than the main evaluated trait is available, such as quality traits (e.g., taste).

7.1.3 Research objective 3: To design and document reproducible data synthesis workflows for crop variety trial data.

This research objective is addressed by Chapters 3, 4, 5 and 6. The data synthesis workflows applied in Chapters 3 and 4 were documented as R code and made available in public repositories (i.e., GitHub). These study-specific workflows were used to design a general workflow of the data synthesis in crop variety evaluation. The generalized functionality was implemented as software tools. The design of a general workflow was an iterative process involving three main steps: (1) For each case studies, document the entire workflow as R code and made it available in public repositories, (2) identify functionality that could be generalized and implemented as R packages and releasing them for public access and use. The general workflow consists of the following stages:

a) Data management and preparation.
b) Modelling and validation.
c) Results presentation.

Currently, in the R statistical environment, the most suitable alternative for statistical modelling of ranking data in combination with environmental covariates is the R package *PlackettLuce* (Turner et al., 2020). In this model implementation, covariates are allowed by

the recursive partitioning framework, implemented in the R package *partykit* (Hothorn & Zeileis, 2015; Zeileis et al., 2008, 2010). The combined functionality of those two packages partially covers the modelling stage but lacks functionality required for model assessment. Furthermore, functionality required for data management and preparation and for results presentation were not available or difficult to customize. To address this functionality gap, the gosset package were developed. The initial development of the *gosset* package dates to times before the start of the development of this thesis, mainly to support the study made by van Etten, de Sousa, et al. (2019). As the research activities of this thesis progressed, new requirements were detected and subsequently implemented in the *gosset* package.

The case studies required climatic data with the higher possible spatial resolution and daily temporal resolution. However, a key criterion for selecting the climatic data source was the temporal coverage. Only the AgERA5 (Hendrik Boogaard & Gerald van der Grijn, 2020; Copernicus Climate Change Service, 2020) dataset fulfilled all the requirements but accessing it from the R environment was not straightforward. Accessing and retrieving the climatic data from the R environment is convenient, but also makes reproducible that part of the workflow. Therefore, the *ag5Tools* package was developed to provide a user-friendly interface for downloading and extracting climatic data from the AgERA5 dataset, supporting research activities that are not covered by *PlackettLuce* and *gosset*.

The packages *gosset* and *ag5Tools* have been released in the Comprehensive R Archive Network (CRAN), facilitating its installation but also guaranteeing that they were developed following the official guidelines for developing R packages (R Core Team, 2022b).The development of the packages *gosset* and *ag5Tools* is framed into a public collaborative effort, where scientists and digital developers from different organizations participate developing digital tools supporting agricultural research. This digital collaborative ecosystem is currently hosted in GitHub (https://github.com/AgrDataSci) and has several packages, tools, and code available for free and public access. Similar collaborative approaches have been applied to improve efficiency, transparency, collaboration, and reproducibility in other scientific research areas such as ecology (Lowndes et al., 2017).

7.2 Research limitations and remaining gaps

This section presents the research limitations and how they affect the main findings, along with the remaining research gaps.

In Chapter 2, four potential areas in crop variety evaluation were delineated in which the data synthesis approach can provide knowledge and insights supporting decision making: (1) variety release decisions, (2) marketing or distribution of varieties, (3) recommendations to farmers, and (4) assessment of crop improvement over time. Chapters 3 and 4 demonstrated how the data synthesis approach can indeed contribute useful information to areas 1 to 3. However, (4) is not explicitly demonstrated in the cases developed in this thesis. Such progress over time is usually measured as genetic gain and is valuable information for the assessment of breeding programs.

In Chapters 1 and 2, it was stated that choosing the right variety goes beyond productivity traits and socioeconomic factors should be considered as well. The tricot trials analyzed in Chapter 3 incorporate the preferences of farmers with regards of market value and taste of common bean genotypes as part of their overall evaluation of the genotypes, which was the focus of this study. However, farmers' preferences represent current conditions, which has limitations. For example, this does not consider consumer preferences in the future, or in markets that are not currently targeted by farmers. Data to address this gap is often not available and adding market insights to the data synthesis approach presented here involves methodological challenges. It will require that information about varieties would be linked to target market locations, which may be remote from production locations when crops are exported. The inclusion of this type of information may also require more complex models, especially to handle prioritization of traits and multiple criteria from both farmers and consumers, which may be in conflict with each other.

In Chapter 3, a Plackett-Luce tree model with environmental covariates was used to predict the top-three best performing genotypes across the study region. A forward variable selection using blocked cross validation selected four variables which were reduced to one in the final AIC-pruned tree. This restrictive modelling approach was used to achieve a simplified model. This ensured that the model was interpretable but also restricted it in taking advantage of all the information present in the set of covariates. A less restrictive modelling approach would be a machine learning approach, such as ensembles of Plackett-Luce trees, allowing to fully explore the information contained in the environmental covariates. Furthermore, a well-known weakness of tree-based models is their instability (Breiman, 1996; Philipp et al., 2018; Strobl et al., 2009) and ensembles of trees have been proposed to overcome the instability problem (Strobl et al., 2009). Currently the main constraint for ensembles of Plackett-Luce trees is the uncertainty quantification, implemented for the single tree setting but not yet for ensembles.

In Chapter 4, two different environments were identified from the aggregated dataset of trials evaluating the reaction of *Musa* genotypes to black leaf streak disease. While there is a differential response of genotypes to disease on the different environments, there is no information available about differences of disease intensity among environments. The model results allow to elaborate hypotheses for further investigation but does not allow to discriminate how the environmental conditions act on the host and the pathogen. Another limitation of the study presented in Chapter 4 is the limited sample size of the aggregated dataset, which is caused mainly by the relatively low availability of data of *Musa* genotypes evaluations in public repositories.

In Chapter 3 uncertainty quantification using entropy was a key element for transforming model predictions into useful information. While most of functionality required in the data synthesis workflow is implemented in the R package *gosset*, uncertainty quantification is currently not yet available.

7.3 Reflection

This section reflects on how this thesis contributed to expand the body of knowledge of data synthesis in crop variety evaluation, considering the different elements involved in this research approach: data, models, and expert knowledge.

7.3.1 Conceptualizing data synthesis in crop variety evaluation

This thesis contributed a new conceptualization of data synthesis in crop variety evaluation, placing the characterization of variety environmental adaptation as an integral part of the methodology. Focusing on spatio-temporal characterization, the information provided by the new data synthesis approach is closer to the conditions of the farming systems for which the decision making should be made. This has implications in the requirements of data and the statistical models that can be used for data synthesis. To characterize the environmental adaptation in time and space, the environmental data is extracted for the location of each trial plot and for the specific temporal extent of the experiments. Therefore, the trial data must contain the temporal extent (e.g., planting and harvest dates) and geographic coordinates of each trial plot. On the one hand requiring having location and time extent could be a constraint to use trial datasets lacking this information. On the other hand, the capability of the data synthesis approach to characterize the environmental adaptation of genotypes in time and space may serve as motivation to researchers and organizations to value the importance of good data management practices, including recording basic information of the trials such as the location and the temporal extent.

7.3.2 Trial data management and sharing

This thesis demonstrated that good data practices, including consistent standardization of field trial data collection, supported by digital technologies pay off in the long term to produce useful insights. However, not all future field variety evaluation will adhere to a single data collection standard or experimental design. Furthermore, there is vast amount of variety evaluation data already collected. For those cases, this thesis also provided a feasible alternative to reuse the data to extract useful information. For both contexts, a critical factor is the availability of data either on public repositories or shared directly among scientists. The ideal case would be that datasets are published in public repositories and with a unique and persistent identifier to make them citable. Knowing that their dataset will be properly acknowledged with the corresponding citation, would be a good motivation for scientists or institutions to share their data and invest time documenting it. But it should be recognized that there are or will be cases in which the data is shared directly among scientists. For these cases, co-authorship on the data synthesis work reusing the data seems a feasible mechanism to motivate and reward data sharing. However, this should be done based on collaboration and not only in terms of just handling over the data. The party that is sharing the data will likely know specific details about the experimental design, the variables measured, local names of genotypes, which might be unclear or unknown to the party conducting the data synthesis. This is an example on how data sharing and integration of expert knowledge

complement and motivate each other as fundamental parts of collaborative research in data synthesis.

7.3.3 Modelling environmental adaptation of genotypes in time and space

The modelling approach used in the data synthesis methodology is based on the aggregation of data, not for the simple sake of density of data points, but especially to take advantage of the diversity of locations, time frames and genotypes of each trial. Those differences are often aimed when multi-location or multi-environment trials are designed and established. Field trials are expensive and time consuming and will be difficult for an organization alone to reach the same scale allowed by the data synthesis. The use of environmental covariates for the specific time and location of each experiment is the basis for the assessment of environmental adaptation of varieties. In that sense, further refinement could be made in terms of pre-selection of model covariates, including approaches such as feature engineering (Kuhn & Johnson, 2013). Expert knowledge from breeders and crop-modelers could be further considered at the modelling stage of the data synthesis, to pre-select meaningful covariates from the biological point of view. An important step made in the development of the data synthesis methods, especially when used for predictive modelling, is the integration of uncertainty quantification of model predictions in the data synthesis workflow. In that way, decision makers can assess the reliability of the information provided.

7.3.4 Expert knowledge and collaborative research

This thesis contributed to demonstrate that collaborative research with intense participation of farmers provides an excellent context to (re)valorize data collected from field trials. For farmers, engaging in research activities and then get practical information from it may provide a stronger motivation not only to keep participating in future experimentation activities, but also to reflect on the importance of the data they helped to produce and collect, not only for themselves but for other farmers and for the society. This could also help to raise awareness about the importance of not only providing information per se but also information in a format tailored to its intended audience. However, collaborative research is not exclusive for experimentation but can also occur at posterior stages, by sharing and synthesizing data which were produced independently. As mentioned in the previous sections, scientists sharing the data have knowledge about the context of experiments that generated the data. Farmers, breeders, and crop modelers will know from their experience about abiotic factors affecting crop variety performance, which could be considered to pre-select meaningful covariates in the modelling stage. Furthermore, experts from other disciplines such as product designers could contribute helping to design information products to tailored for the intended audience. Therefore, collaborative research is a pivotal element of data synthesis, especially on the basis of expert knowledge integration, but also as the enabling factor motivating data sharing.

Further efforts are still required to streamline the data synthesis approach in crop variety evaluation, but the first steps taken by this thesis provide a motivating environment to move forward.

7.4 Future research

This section provides a perspective on potential research lines that could be developed to fill the remaining knowledge gaps.

7.4.1 Assess genetic gain made by breeding programs

Genetic gain is frequently used to assess the success of crop improvement programs, especially to prioritize investments (Cobb et al., 2019; Kholova et al., 2021). Genetic gain is a measure of the efficiency of a breeding program, expressing the genetic response to selection as change over time. This can be measured by doing a regression with crop trial data, taking as response variable yield or another target variable, and as explanatory variable the year of release of the different varieties. The use of rank-based methods provides an interesting option to assess genetic gain in breeding programs. Regression on rankings data can be done with the Plackett-Luce regression model (Yıldız et al., 2020), which was recently implemented in the R package *PlackettLuce* (Turner et al., 2020). The current implementation of the Plackett-Luce regression model with item covariates is experimental but can already be used to create a proof of concept. A difficulty would be that genetic gain would be expressed in a log-probability of exceeding the performance of a certain check variety (the concept of reliability) or the entire set of varieties, rather than the usual variable, yield gain. This already indicates that this potential research line does not only involve fitting a model but also sharing and assessing the results with breeders, iteratively exploring how to interpret and frame the results, to adjust and improve statistical models, and designing appropriate data visualizations. Hence, it requires interdisciplinary work among statisticians, data analysts and plant breeders.

7.4.2 Improve predictions of genotype performance using ensembles of Plackett-Luce trees

As mentioned in section 7.2, ensembles of Plackett-Luce trees could better exploit all information contained in the environmental covariates, by allowing a larger number of covariates being selected by the model. Implementing ensembles of Plackett-Luce trees will take the data synthesis approach even closer to machine learning, which will be more relevant as more trial data become available, enabling the integration of larger amounts of data. While machine learning models are not restricted to large sample datasets, they can handle larger datasets compared to conventional statistical data models (Breiman, 2001). Furthermore, for more complex applications using more covariates such as market and socioeconomic data, machine learning approaches may be more appropriate compared to more restrictive settings such as conventional statistical models, which have problems dealing high dimensional data (Breiman, 2001). Therefore, implementing ensembles of Plackett-Luce trees is suggested as

a way forward to improve spatially-explicit predictions of genotype performance considering environmental factors.

7.4.3 Multi-objective optimization

In Chapter 4, the analysis considered only one factor, reaction to BLSD, to elucidate the best performing genotypes. In Chapter 3, farmers considered several traits to provide their overall assessment of common bean genotypes. However, choosing the right variety involves several factors to be considered jointly. Furthermore, a more complex situation occurs when those factors are in conflict among them. Multi-objective optimization methods (Chiandussi et al., 2012) could help to account for multiple and conflicting traits in crop variety evaluation. Further steps should be made to integrate multi-criteria decision making within the data synthesis approach. Currently, data synthesis for crop variety evaluation relies on the Plackett-Luce model for ranking data. Some multi-objective optimization approaches are also based on implementations of the Plackett-Luce model (Lan et al., 2014). Therefore, it will be worth exploring the integration of multi-objective optimization with the data synthesis approach using the Plackett-Luce model.

7.4.4 Tailored information products

The disconnection between scientists and farmers may be an obstacle for building trust of farmers in research findings (Moore et al., 2021). Participatory research methods, such as those based on citizen science, have the potential to create trust among scientists, extension services and farmers (Moore et al., 2021). The data synthesis presented in Chapter 3 is based on the aggregation of tricot trials, which are participatory on-farm trials based on citizen science. The information generated by the case study presented in Chapter 3 can provide useful information for breeding programs or extension services but is still far from being readily presented to farmers. Additional efforts are required to design user-friendly information products. Currently undergoing initiatives can contribute to this aim, applying participatory and co-creation approaches, such as user-centered design (Ortiz-Crespo et al., 2020). Those new information products can help to build trust in the data synthesis approach, providing the farmers with actionable information they helped to create. In that sense, implementing a user-centered design for designing tailored information products based on the outputs of the data synthesis approach, is a potential research line for closing the gap of information products supporting farmers' decision making.

References

Abadie, C., Elhadrami, A., Fouré, E., & Carlier, J. (2003). *Efficiency and durability of partial resistance against black leaf streak disease* Mycosphaerella leaf spot diseases of bananas: present status and outlook, San Jose, Costa Rica.

Abeyasekera, S., Ritchie, J. M., & Lawson-McDowall, J. (2002). Combining ranks and scores to determine farmers' preferences for bean varieties in southern Malawi. *Experimental Agriculture*, *38*(1), 97-109. https://doi.org/10.1017/S0014479702000182

Acquaah, G. (2012). Performance Evaluation for Crop Cultivar Release. In G. Acquaah (Ed.), *Principles of Plant Genetics and Breeding* (2nd ed., pp. 491-510). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118313718.ch26

Agtunong, T., Redden, R., Mangge-Nang, M., Searle, C., & Fukai, S. (1992). Genotypic variation in response to high temperature at flowering in common bean (*Phaseolus vulgaris* L.). *Australian Journal of Experimental Agriculture*, *32*(8), 1135-1140. https://doi.org/10.1071/EA9921135

Aguilar, E., Peterson, T. C., Obando, P. R., Frutos, R., Retana, J. A., Solera, M., . . . Mayorga, R. (2005). Changes in precipitation and temperature extremes in Central America and northern South America, 1961–2003. *Journal of Geophysical Research: Atmospheres*, *110*(D23). https://doi.org/10.1029/2005jd006119

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Albert, I., & Makowski, D. (2018). Ranking crop species using mixed treatment comparisons. *Research Synthesis Methods*, *0*(0). https://doi.org/doi:10.1002/jrsm.1328

Alfaro, E. J., Chourio, X., Muñoz, Á. G., & Mason, S. J. (2018). Improved seasonal prediction skill of rainfall for the Primera season in Central America. *International Journal of Climatology*, *38*(S1), e255-e268. https://doi.org/10.1002/joc.5366

Alvarez, J. (1997). *Introduction, evaluation, multiplication and dissemination of FHIA hybrids in Cuba. Infomusa*.

Anderson, J. R., & Feder, G. (2007). Chapter 44 Agricultural Extension. In R. Evenson & P. Pingali (Eds.), *Handbook of Agricultural Economics* (Vol. 3, pp. 2343-2378). Elsevier. https://doi.org/10.1016/S1574-0072(06)03044-1

Anonymous. (1975). *Minutes of Division Business Meeting*. Chicago: IFT

Arief, V. N., DeLacy, I. H., Crossa, J., Payne, T., Singh, R., Braun, H.-J., . . . Dieters, M. J. (2015). Evaluating Testing Strategies for Plant Breeding Field Trials: Redesigning a CIMMYT International Wheat Nursery. *Crop Science*, *55*(1), 164-177. https://doi.org/10.2135/cropsci2014.06.0415

Atlin, G. N., Cooper, M., & Bjørnstad, Å. (2001). A comparison of formal and participatory breeding approaches using selection theory [journal article]. *Euphytica*, *122*(3), 463-475. https://doi.org/10.1023/a:1017557307800

Bakry, F., Carreel, F., Jenny, C., & Horry, J.-P. (2009). Genetic Improvement of Banana. In S. M. Jain & P. M. Priyadarshan (Eds.), *Breeding Plantation Tree Crops: Tropical Species* (pp. 3-50). Springer New York. https://doi.org/10.1007/978-0-387-71201-7_1

Barth, R., Ijsselmuiden, J., Hemming, J., & Henten, E. J. V. (2018). Data synthesis methods for semantic segmentation in agriculture: A Capsicum annuum dataset. *Computers and Electronics in Agriculture*, *144*, 284-296. https://doi.org/10.1016/j.compag.2017.12.001

Beebe, S. (2012). Common Bean Breeding in the Tropics. In J. Janick (Ed.), *Plant Breeding Reviews* (pp. 357-426). https://doi.org/10.1002/9781118358566.ch5

Beebe, S., Ramirez, J., Jarvis, A., Rao, I. M., Mosquera, G., Bueno, J. M., & Blair, M. W. (2011). Genetic Improvement of Common Beans and the Challenges of Climate Change. In R. J. R. S.S. Yadav, J.L. Hatfield, H. Lotze-Campen and A.E. Hall (Ed.), *Crop Adaptation to Climate Change* (pp. 356-369). https://doi.org/10.1002/9780470960929.ch25

Billiau, K., Sprenger, H., Schudoma, C., Walther, D., & Köhl, K. I. (2012). Data management pipeline for plant phenotyping in a multisite project. *Functional Plant Biology*, *39*(11), 948-957. https://doi.org/https://doi.org/10.1071/FP12009

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Böckenholt, U. (2004). Comparative Judgments as an Alternative to Ratings: Identifying the Scale Origin. *Psychological methods*, *9*(4), 453-465. https://doi.org/10.1037/1082-989X.9.4.453

Bolger, A. M., Poorter, H., Dumschott, K., Bolger, M. E., Arend, D., Osorio, S., . . . Usadel, B. (2019). Computational aspects underlying genome to phenome analysis in plants. *The Plant Journal*, *97*(1), 182-198. https://doi.org/10.1111/tpj.14179

Boogaard, H., & van der Grijn, G. (2020). *Agrometeorological indicators from 1979 to present derived from reanalysis* Copernicus Climate Change Service. https://doi.org/10.24381/cds.6c68c9bb

Boogaard, H., & van der Grijn, G. (2020). *Data Stream 2: AgERA5 historic and near real time forcing data* (Product User Guide and Specification, Issue. http://datastore.copernicus-climate.eu/documents/sis-global-agriculture/C3S422Lot1.WEnR.DS2_ProductUserGuideSpecification_v2.2.pdf

Borenstein, M., L. V. Hedges, Higgins, J. P. T., & Rothstein, H. R. (2009). When Does it Make Sense to Perform a Meta-Analysis? In *Introduction to meta-analysis* (pp. 357-364). John Wiley & Sons.

Borkovec, M., & Madin, N. (2019). *ggparty:"ggplot" Visualizations for the "partykit" Package*. In (Version 1.0.0) https://CRAN.R-project.org/package=ggparty

Bostock, R. M., Pye, M. F., & Roubtsova, T. V. (2014). Predisposition in Plant Disease: Exploiting the Nexus in Abiotic and Biotic Stress Perception and Response. *Annual Review of Phytopathology*, *52*(1), 517-549. https://doi.org/10.1146/annurev-phyto-081211-172902

Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, *39*(3/4), 324.

Brancourt-Hulmel, M., Biarnès-Dumoulin, V., & Denis, J. B. (1997). Points de repère dans l'analyse de la stabilité et de l'interaction génotype-milieu en amélioration des plantes. *Agronomie*, *17*(4), 219-246. https://doi.org/10.1051/agro:19970403

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*(6), 2350-2383, 2334. https://doi.org/10.1214/aos/1032181158

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199-215. www.jstor.org/stable/2676681

Brown, D., Carpentier, S., de Bruin, S., de Sousa, K., Machida, L., & van Etten, J. (2021). *Combining Legacy Data from Heterogeneous Crop Trials to Identify Genotype By Environment Interactions Using Model-Based Recursive Partitioning* [Poster].

ASA, CSSA, SSSA International Annual Meeting, Salt Lake City, UT. https://scisoc.confex.com/scisoc/2021am/meetingapp.cgi/Paper/137905

Brown, D., de Bruin, S., de Sousa, K., Aguilar, A., Barrios, M., Gómez, M., . . . van Etten, J. (2022). Rank-based data synthesis of common bean on-farm trials across four Central American countries. *Crop Science*. https://doi.org/10.1002/csc2.20817

Brown, D., & de Sousa, K. (2022). *ag5Tools: Toolbox for Downloading and Extracting Copernicus AgERA5 Data*. In (Version 0.0.1) CRAN. https://cran.r-project.org/package=ag5Tools

Brown, D., Van den Bergh, I., de Bruin, S., Machida, L., & van Etten, J. (2020). Data synthesis for crop variety evaluation. A review. *Agronomy for Sustainable Development*, *40*(4), 25. https://doi.org/10.1007/s13593-020-00630-7

BTI, IITA, NARO, & CGIAR-RTB. (2018). *MusaBase* https://musabase.org/

Bucher, H. C., Guyatt, G. H., Griffith, L. E., & Walter, S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology*, *50*(6), 683-691. https://doi.org/https://doi.org/10.1016/S0895-4356(97)00049-8

Buntaran, H., Forkman, J., & Piepho, H.-P. (2021). Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: accuracy and precision. *Theoretical and Applied Genetics*, *134*(5), 1513-1530. https://doi.org/10.1007/s00122-021-03786-2

Bustos-Korts, D., Romagosa, I., Borràs-Gelonch, G., Casas, A. M., Slafer, G. A., & van Eeuwijk, F. (2019). Genotype by Environment Interaction and Adaptation. In R. Savin & G. A. Slafer (Eds.), *Crop Science* (pp. 29-71). Springer New York. https://doi.org/10.1007/978-1-4939-8621-7_199

Butts, C. T. (2008). network: a Package for Managing Relational Data in R. *Journal of Statistical Software*, *24*(2). https://www.jstatsoft.org/article/download/v024i02/217

Butts, C. T. (2020). *sna: Tools for Social Network Analysis*. In (Version 2.6) CRAN. https://CRAN.R-project.org/package=sna

Cadotte, M. W., Mehrkens, L. R., & Menge, D. N. L. (2012). Gauging the impact of meta-analysis on ecology [journal article]. *Evolutionary Ecology*, *26*(5), 1153-1167. https://doi.org/10.1007/s10682-012-9585-z

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, *66*(9), 1747-1762. https://doi.org/10.1002/asi.23358

Carpenter, S. R., Armbrust, E. V., Arzberger, P. W., Chapin, F. S., III, Elser, J. J., Hackett, E. J., . . . Zimmerman, A. S. (2009). Accelerate Synthesis in Ecology and Environmental Sciences. *BioScience*, *59*(8), 699-701. https://doi.org/10.1525/bio.2009.59.8.11

Ceccarelli, S. (1989). Wide adaptation: How wide? [journal article]. *Euphytica*, *40*(3), 197-205. https://doi.org/10.1007/bf00024512

Ceccarelli, S. (2012). Plant breeding with farmers – a technical manual. In (pp. xi + 126). Aleppo, Syria: ICARDA.

Ceccarelli, S., & Grando, S. (2007). Decentralized-participatory plant breeding: an example of demand driven research [journal article]. *Euphytica*, *155*(3), 349-360. https://doi.org/10.1007/s10681-006-9336-8

Ceccarelli, S., Guimarães, E. P., & Weltzien, E. (2009). *Plant breeding and farmer participation*. Food and Agriculture Organization of the United Nations.

Challinor, A. J., Koehler, A. K., Ramirez-Villegas, J., Whitfield, S., & Das, B. (2016). Current warming will reduce yields unless maize breeding and seed systems adapt

immediately. *Nature Climate Change*, *6*(10), 954-958. https://doi.org/10.1038/nclimate3061

Chawade, A., van Ham, J., Blomquist, H., Bagge, O., Alexandersson, E., & Ortiz, R. (2019). High-Throughput Field-Phenotyping Tools for Plant Breeding and Precision Agriculture. *Agronomy*, *9*(5), 258. https://www.mdpi.com/2073-4395/9/5/258

Chenu, K. (2015). Chapter 13 - Characterizing the crop environment – nature, significance and applications. In V. O. Sadras & D. F. Calderini (Eds.), *Crop Physiology (Second Edition)* (pp. 321-348). Academic Press. https://doi.org/10.1016/B978-0-12-417104-6.00013-3

Chiandussi, G., Codegone, M., Ferrero, S., & Varesio, F. E. (2012). Comparison of multi-objective optimization methodologies for engineering applications. *Computers & Mathematics with Applications*, *63*(5), 912-942. https://doi.org/10.1016/j.camwa.2011.11.057

Christinck, A., Weltzien, E., Rattunde, F., & Ashby, J. (2017). *Gender Differentiation of Farmer Preferences for Varietal Traits in Crop Improvement: Evidence and Issues* [Working paper]. C. G. a. A. R. Network.

Churchill, A. C. L. (2011). Mycosphaerella fijiensis, the black leaf streak pathogen of banana: progress towards understanding pathogen biology and detection, disease development, and the challenges of control. *Molecular Plant Pathology*, *12*(4), 307-328. https://doi.org/10.1111/j.1364-3703.2010.00672.x

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., . . . Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theoretical and Applied Genetics*, *132*(3), 627-645. https://doi.org/10.1007/s00122-019-03317-0

Coe, R. (2007). Analysing Data from Participatory On-farm Trials. *The African Statistical Journal*, *4*.

Coe, R. D. (2002). Analyzing ranking and rating data from participatory on-farm trials. In M. R. Bellon & J. Reeves (Eds.), *Quantitative analysis of Data from participatory methods in plant breeding* (pp. 44-65). CIMMYT. https://hdl.handle.net/10568/76948

Cohan, J.-P., Abadie, C., Tomekpé, K., & Tchango Tchango, J. (2003). Agronomic performance and resistance to black leaf streak of the hybrid 'CRBP-39'. *InfoMusa*, *12*(1), 29-32.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological methods*, *8*(3), 243-253. https://doi.org/10.1037/1082-989X.8.3.243

Coleman, C. I., Phung, O. J., Cappelleri, J. C., Baker, W. L., Kluger, J., White, C. M., & Sobieraj, D. M. (2012). *Use of mixed treatment comparisons in systematic reviews* [Methods Research Report](12-EHC119-EF). www.effectivehealthcare.ahrq.gov/reports/final.cfm

Collard, B. C. Y., & Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1491), 557-572. https://doi.org/doi:10.1098/rstb.2007.2170

Copernicus Climate Change Service. (2018). *Essential climate variables for assessment of climate variability from 1979 to present*. https://cds.climate.copernicus.eu/cdsapp#!/dataset/ecv-for-climate-change

Copernicus Climate Change Service. (2020). *Agrometeorological indicators from 1979 to present derived from reanalysis* AgERA5. https://doi.org/10.24381/cds.6c68c9bb

Coumou, D., & Rahmstorf, S. (2012). A decade of weather extremes [Perspective]. *Nature Climate Change*, *2*, 491. https://doi.org/10.1038/nclimate1452

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187-220. http://www.jstor.org.ezproxy.library.wur.nl/stable/2985181

Craenen, K., & Ortiz, R. (1997). Effect of the bs1 gene in plantain-banana hybrids on response to black sigatoka. *Theoretical and Applied Genetics*, *95*(3), 497-505. https://doi.org/10.1007/s001220050589

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., . . . Krajewski, P. (2016). Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods*, *12*(1), 44. https://doi.org/10.1186/s13007-016-0144-4

Dalrymple, D. G. (2008). International agricultural research as a global public good: concepts, the CGIAR experience and policy issues. *Journal of International Development*, *20*(3), 347-379. https://doi.org/10.1002/jid.1420

Damesa, T. M., Möhring, J., Worku, M., & Piepho, H.-P. (2017). One Step at a Time: Stage-Wise Analysis of a Series of Experiments. *Agronomy journal*, *109*(3), 845-857. https://doi.org/10.2134/agronj2016.07.0395

Dawson, J. C., & Healy, G. (2018). Flavour Evaluation for Plant Breeders. In I. Goldman (Ed.), *Plant Breeding Reviews*. https://doi.org/10.1002/9781119414735.ch5

De Lapeyre de Bellaire, L., Fouré, E., Abadie, C., & Carlier, J. (2010). Black Leaf Streak Disease is challenging the banana industry [Article]. *Fruits*, *65*(6), 327-342. https://doi.org/10.1051/fruits/2010034

de Leon, N., Jannink, J.-L., Edwards, J. W., & Kaeppler, S. M. (2016). Introduction to a Special Issue on Genotype by Environment Interaction. *Crop Science*, *56*(5), 2081-2089. https://doi.org/10.2135/cropsci2016.07.0002in

de Medeiros, G. A., Daniel, L. A., & Fengler, F. H. (2016). Growth, Development, and Water Consumption of Irrigated Bean Crop Related to Growing Degree-Days on Different Soil Tillage Systems in Southeast Brazil. *International Journal of Agronomy*, *2016*, 8065985. https://doi.org/10.1155/2016/8065985

De Roo, N., Andersson, J. A., & Krupnik, T. J. (2017). On-farm trials for development impact? The organisation of research and the scaling of agricultural technologies. *Experimental Agriculture*, *55*(2), 163-184. https://doi.org/10.1017/S0014479717000382

de Sousa, K., Sparks, A. H., Ashmall, W., van Etten, J., & Solberg, S. Ø. (2020). chirps: API Client for the CHIRPS Precipitation Data in R. *Journal of Open Source Software*, *5*(51). https://doi.org/10.21105/joss.02419

de Sousa, K., van Etten, J., Brown, D., & Steinke, J. (2022). *gosset: Tools for Data Analysis in Experimental Agriculture*. In (Version 0.3) https://cran.r-project.org/package=gosset

de Sousa, K., van Etten, J., Poland, J., Fadda, C., Jannink, J.-L., Kidane, Y. G., . . . Dell'Acqua, M. (2021). Data-driven decentralized breeding increases prediction accuracy in a challenging crop production environment. *Communications Biology*, *4*(1), 944. https://doi.org/10.1038/s42003-021-02463-w

de Sousa, K., van Etten, J., & Solberg, S. Ø. (2020). *climatrends: Climate Variability Indices for Ecological Modelling*. In (Version 0.1.6) https://CRAN.R-project.org/package=climatrends

Dembo, R. S., & King, A. J. (1992). Tracking models and the optimal regret distribution in asset allocation. *Applied Stochastic Models and Data Analysis*, *8*(3), 151-157. https://doi.org/10.1002/asm.3150080305

Desclaux, D., Nolot, J. M., Chiffoleau, Y., Gozé, E., & Leclerc, C. (2008). Changes in the concept of genotype × environment interactions to fit agriculture diversification and decentralized participatory plant breeding: pluridisciplinary point of view [journal article]. *Euphytica*, *163*(3), 533. https://doi.org/10.1007/s10681-008-9717-2

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, Tampere, Finland.

Dias, S., & Caldwell, D. M. (2019). Network meta-analysis explained. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, *104*(1), F8-F12. https://doi.org/10.1136/archdischild-2018-315224

Diekmann, F. (2012). Data Practices of Agricultural Scientists: Results from an Exploratory Study. *Journal of Agricultural & Food Information*, *13*(1), 14-34. https://doi.org/10.1080/10496505.2012.636005

Dittrich, R., Hatzinger, R., & Katzenbeisser, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(4), 511-525. https://doi.org/doi:10.1111/1467-9876.00125

Eagle, A. J., Christianson, L. E., Cook, R. L., Harmel, R. D., Miguez, F. E., Qian, S. S., & Ruiz Diaz, D. A. (2017). Meta-Analysis Constrained by Data: Recommendations to Improve Relevance of Nutrient Management Research. *Agronomy journal*, *109*(6), 2441-2449. https://doi.org/10.2134/agronj2017.04.0215

ECMWF. (2019). *Python API to access the Copernicus Climate Data Store (CDS)* In (Version 0.5.1) ECMWF. https://github.com/ecmwf/cdsapi

Elias, A. A., Robbins, K. R., Doerge, R. W., & Tuinstra, M. R. (2016). Half a Century of Studying Genotype × Environment Interactions in Plant Breeding Experiments. *Crop Science*, *56*(5), 2090-2105. https://doi.org/10.2135/cropsci2015.01.0061

Ersoz, E. S., Martin, N. F., & Stapleton, A. E. (2020). On to the next chapter for crop breeding: Convergence with data science. *Crop Science*, *60*(2), 639-655. https://doi.org/10.1002/csc2.20054

Escoto, N. D. (2013). El Cultivo del Frijol. 44. http://www.agronegocioshonduras.org/wp-content/uploads/2014/06/el_cultivo_de_frijol_dicta.pdf

Eskridge, K. M. (1990). Selection of Stable Cultivars Using a Safety-First Rule. *Crop Science*, *30*(2), 369-374. https://doi.org/10.2135/cropsci1990.0011183X003000020025x

Eskridge, K. M. (1997). Evaluation of Corn Hybrids Using the Probability of Outperforming a Check Based on Strip-Test Data. *Journal of Agricultural, Biological, and Environmental Statistics*, *2*(3), 245-254. https://doi.org/10.2307/1400444

Eskridge, K. M., & Mumm, R. F. (1992). Choosing plant cultivars based on the probability of outperforming a check [journal article]. *Theoretical and Applied Genetics*, *84*(3), 494-500. https://doi.org/10.1007/bf00229512

Esposito, S., Carputo, D., Cardi, T., & Tripodi, P. (2020). Applications and Trends of Machine Learning in Genomics and Phenomics for Next-Generation Breeding. *Plants*, *9*(1), 34. https://www.mdpi.com/2223-7747/9/1/34

Eyland, D., Luchaire, N., Cabrera-Bosquet, L., Parent, B., Janssens, S. B., Swennen, R., . . . Carpentier, S. C. (2022). High-throughput phenotyping reveals differential transpiration behaviour within the banana wild relatives highlighting diversity in drought tolerance. *Plant, Cell & Environment*, *45*(6), 1647-1663. https://doi.org/10.1111/pce.14310

Fadda, C., & van Etten, J. (2019). Generating Farm-Validated Variety Recommendations for Climate Adaptation. In T. S. Rosenstock, A. Nowak, & E. Girvetz (Eds.), *The Climate-Smart Agriculture Papers: Investigating the Business of a Productive, Resilient and Low Emission Future* (pp. 127-138). Springer International Publishing. https://doi.org/10.1007/978-3-319-92798-5_11

Feed the Future Legume Innovation Lab, M. S. U., & USDA. (2018). *TEN-YEAR LEGACY REPORT of the Dry Grain Pulses CRSP (2007-2012) and the Feed the Future Innovation Lab for Collaborative Research on Grain Legumes (2013-2017)*. https://www.canr.msu.edu/legumelab/uploads/files/LIL%20Legacy%20Report%20FINAL.pdf

Feng, P., Wang, B., Liu, D. L., Waters, C., & Yu, Q. (2019). Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. *Agricultural and Forest Meteorology*, *275*, 100-113. https://doi.org/https://doi.org/10.1016/j.agrformet.2019.05.018

Fernández de Córdova, F., Gepts, P. L., & López Genes, M. (1986). *Etapas de desarrollo de la planta de fríjol común (Phaseolus vulgaris L.)*. Centro Internacional de Agricultura Tropical (CIAT). https://hdl.handle.net/10568/69648

Ferris, R. S. B., Ortiz, R., & Vuylsteke, D. (1999). Fruit quality evaluation of plantains, plantain hybrids, and cooking bananas. *Postharvest Biology and Technology*, *15*(1), 73-81. https://doi.org/10.1016/S0925-5214(98)00067-2

Ferrufino, A. (2014). *Catálogo de semillas de granos básicos: Variedades de arroz, frijol, maíz y sorgo liberadas por el INTA*. https://repositorio.iica.int/handle/11324/19830

Finlay, K., & Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, *14*(6), 742-754. https://doi.org/https://doi.org/10.1071/AR9630742

Firke, S. (2021). *janitor: Simple Tools for Examining and Cleaning Dirty Data*. In https://CRAN.R-project.org/package=janitor

Firth, D. (2020). *qvcalc: Quasi Variances for Factor Effects in Statistical Models*. In (Version 1.0.2) https://CRAN.R-project.org/package=qvcalc

Firth, D., & De Menezes, R. X. (2004). Quasi-variances. *Biometrika*, *91*(1), 65-80. https://doi.org/10.1093/biomet/91.1.65

Folta, K. M., & Klee, H. J. (2016). Sensory sacrifices when we mass-produce mass produce [Review Article]. *Horticulture Research*, *3*, 16032. https://doi.org/10.1038/hortres.2016.32

Franzel, S., & Coe, R. (2002). Participatory on-farm technology testing: the suitability of different types of trials for different objectives. In S. Franzel & R. Coe (Eds.), *Quantitative analysis of Data from participatory methods in plant breeding* (pp. 1-8). CIMMYT.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129-1164. https://doi.org/10.1002/spe.4380211102

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., . . . Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new

environmental record for monitoring extremes [Data Descriptor]. *Scientific Data*, *2*, 150066. https://doi.org/10.1038/sdata.2015.66

García-Solera, I., & Ramírez, P. (2012). Central America's seasonal climate outlook forum. *The Climate Services Partnership*, 8. https://climate-services.org/wp-content/uploads/2015/09/CRRH_Case_Study.pdf

Garnier, S., Ross, N., Rudis, R., Camargo, A. P., Sciaini, M., & Scherer, C. (2021). *Rvision - Colorblind-Friendly Color Maps for R*. In (Version 0.4.0) CRAN. https://sjmgarnier.github.io/viridis/

Gauch, H. G., Jr. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Elsevier.

Germeier, C. U., & Unger, S. (2019). Modeling Crop Genetic Resources Phenotyping Information Systems [Hypothesis and Theory]. *Frontiers in Plant Science*, *10*(728). https://doi.org/10.3389/fpls.2019.00728

Gourdji, S., Läderach, P., Valle, A. M., Martinez, C. Z., & Lobell, D. B. (2015). Historical climate trends, deforestation, and maize and bean yields in Nicaragua. *Agricultural and Forest Meteorology*, *200*, 270-281. https://doi.org/10.1016/j.agrformet.2014.10.002

Grambsch, P. M., & Therneau, T. M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika*, *81*(3), 515-526. https://doi.org/10.2307/2337123

Gross, Y., & Kigel, J. (1994). Differential sensitivity to high temperature of stages in the reproductive development of common bean (Phaseolus vulgaris L.). *Field Crops Research*, *36*(3), 201-212. https://doi.org/10.1016/0378-4290(94)90112-0

Guzman, M., Pérez Vicente, L., Carlier, J., Abadie, C., de Lapeyre de Bellaire, L., Carreel, F., . . . Pasberg-Gauhl, C. (2018). Black leaf streak. In D. R. Jones (Ed.), *Handbook of Diseases of Banana, Abacá and Enset*. CABI.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., . . . Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156-162. https://doi.org/10.1890/120103

Hansen, P., & Ombler, F. (2008). A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *Journal of Multi-Criteria Decision Analysis*, *15*(3-4), 87-107. https://doi.org/doi:10.1002/mcda.428

Harrell, F. E. (2015). Cox Proportional Hazards Regression Model. In *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (pp. 475-519). Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7_20

Harrell, F. E., Jr, Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the Yield of Medical Tests. *JAMA*, *247*(18), 2543-2546. https://doi.org/10.1001/jama.1982.03320430047030

Hartung, K., & Piepho, H. P. (2005). A Threshold Model for Multiyear Genebank Data Based on Different Rating Scales. *Crop science.*, *45*(3), 1045.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., . . . Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hernández Fonseca, J. (2009). *Manual de recomendaciones técnicas: cultivo de frijol (Phaseolus vulgaris)*. INTA. http://www.mag.go.cr/bibliotecavirtual/F01-9533.pdf

Hernández Núñez, R. (1995). *Selección in vitro e invernadero de clones de Musa sp. para la evaluación de su resistencia a la Sigatoka negra (Mycosphaerella fijiensis, Morelet)* CATIE]. https://repositorio.catie.ac.cr/handle/11554/3945

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., . . . Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. https://doi.org/10.1002/qj.3803

Hester, J., Wickham, H., & Csárdi, G. (2021). *fs: Cross-Platform File System Operations Based on 'libuv'*. In (Version 1.5.2) CRAN. https://CRAN.R-project.org/package=fs

Hijmans, R. J. (2010). *GADM: database of global administrative areas* https://www.gadm.org

Hijmans, R. J. (2021). *terra: Spatial Data Analysis. R package version*. In (Version 1.4-9) https://rspatial.org/terra/

Ho, M. D., Rosas, J. C., Brown, K. M., & Lynch, J. P. (2005). Root architectural tradeoffs for water and phosphorus acquisition. *Functional Plant Biology*, *32*(8), 737-748. https://doi.org/10.1071/FP05043

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674. https://doi.org/10.1198/106186006X133933

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, *16*(1), 3905-3909. https://www.jmlr.org/papers/v16/hothorn15a.html

Huehn, M. (1990). Nonparametric measures of phenotypic stability. Part 1: Theory [journal article]. *Euphytica*, *47*(3), 189-194. https://doi.org/10.1007/bf00024241

Hufkens, K., Stauffer, R., & Campitelli, E. (2019). The ecwmfr package: an interface to ECMWF API endpoints. https://doi.org/10.5281/zenodo.2647541

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, *32*(1), 384-406. https://doi.org/10.1214/aos/1079120141

Hyman, G., Espinosa, H., Camargo, P., Abreu, D., Devare, M., Arnaud, E., . . . Traore, S. (2017). Improving agricultural knowledge management: The AgTrials experience [version 2; peer review: 2 approved]. *F1000Research*, *6*(317). https://doi.org/10.12688/f1000research.11179.2

Hyman, G., Hodson, D., & Jones, P. (2013). Spatial analysis to support geographic targeting of genotypes to environments [Methods]. *Frontiers in Physiology*, *4*(40). https://doi.org/10.3389/fphys.2013.00040

IPCC. (2022). Summary for Policymakers. In H.-O. Pörtner, D. C. Roberts, E. S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, & A. Okem (Eds.), *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Irish, B. M., Goenaga, R., Montalvo-Katz, S., Chaves-Cordoba, B., & Van den Bergh, I. (2019). Host Response to Black Leaf Streak and Agronomic Performance of Banana Genotypes in Puerto Rico. *HortScience horts*, *54*(10), 1808. https://doi.org/10.21273/hortsci13876-19

Irish, B. M., Goenaga, R., Rios, C., Chavarria-Carvajal, J., & Ploetz, R. (2013). Evaluation of banana hybrids for tolerance to black leaf streak (Mycosphaerella fijiensis Morelet) in Puerto Rico. *Crop Protection*, *54*, 229-238. https://doi.org/10.1016/j.cropro.2013.09.003

Jacome, L. H., & Schuh, W. (1992). Effects of leaf wetness duration and temperature on development of Black Sigatoka disease on banana infected by Mycosphaerella fijiensis var. difformis. *Phytopathology*, *82*(5), 515-520.

Janssen, S. J. C., Porter, C. H., Moore, A. D., Athanasiadis, I. N., Foster, I., Jones, J. W., & Antle, J. M. (2017). Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology. *Agricultural Systems*, *155*, 200-212. https://doi.org/10.1016/j.agsy.2016.09.017

Jarvis, A., Reuter, H. I., Nelson, A., & Guevara, E. (2008). *Hole-filled seamless SRTM data V4*. International Centre for Tropical Agriculture (CIAT). http://srtm.csi.cgiar.org/

Jeuffroy, M.-H., Casadebaig, P., Debaeke, P., Loyce, C., & Meynard, J.-M. (2014). Agronomic model uses to predict cultivar performance in various environments and cropping systems. A review [journal article]. *Agronomy for Sustainable Development*, *34*(1), 121-137. https://doi.org/10.1007/s13593-013-0170-9

Jones, D., & Tézenas du Montcel, H. (1994). *Final Report for UNDP/World Bank on the Results of the International Musa Testing Programme (Phase I)*. INIBAP.

Jordan, C. A., Hoover, B., Dans, A. J., Schank, C., & Miller, J. A. (2019). The Impact of Hurricane Otto on Baird's Tapir Movement in Nicaragua's Indio Maíz Biological Reserve. In R. Reyna-Hurtado & C. A. Chapman (Eds.), *Movement Ecology of Neotropical Forest Mammals: Focus on Social Animals* (pp. 5-19). Springer International Publishing. https://doi.org/10.1007/978-3-030-03463-4_2

Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, *143*, 23-37. https://doi.org/https://doi.org/10.1016/j.compag.2017.09.037

Kang, M. S. (1997). Using Genotype-by-Environment Interaction for Crop Cultivar Development. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 62, pp. 199-252). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2113(08)60569-6

Kang, M. S., & Gorman, D. P. (1989). Genotype × Environment Interaction in Maize. *Agronomy journal*, *81*(4), 662-664. https://doi.org/10.2134/agronj1989.00021962008100040020x

Karamura, D., & Karamura, E. (1994). *A provisional checklist of banana cultivars in Uganda*. INIBAP.

Karamura, D., Karamura, E., & Tinzaara, W. (2012). Banana cultivar names, synonyms and their usage in East Africa. *Bioversity International, Uganda*.

Kassambara, A., Kosinski, M., & Biecek, P. (2021). *survminer: Drawing Survival Curves using 'ggplot2'*. In (Version 0.4.9) CRAN. https://CRAN.R-project.org/package=survminer

Kehel, Z., Crossa, J., & Reynolds, M. (2016). Identifying Climate Patterns during the Crop-Growing Cycle from 30 Years of CIMMYT Elite Spring Wheat International Yield Trials. In *Applied Mathematics and Omics to Assess Crop Genetic Resources for Climate Change Adaptive Traits* (pp. 151-174). CRC Press. https://doi.org/10.1201/9781315369006

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1-2), 81-93. https://doi.org/10.1093/biomet/30.1-2.81

Kendall, M. G., & Smith, B. B. (1939). The Problem of m Rankings. *The Annals of Mathematical Statistics*, *10*(3), 275-287. http://www.jstor.org.ezproxy.library.wur.nl/stable/2235668

Kholova, J., Urban, M. O., Cock, J., Arcos, J., Arnaud, E., Aytekin, D., . . . Xu, Y. (2021). In pursuit of a better world: crop improvement and the CGIAR. *Journal of Experimental Botany*. https://doi.org/10.1093/jxb/erab226

Kimunye, J., Jomanga, K., Tazuba, A. F., Were, E., Viljoen, A., Swennen, R., & Mahuku, G. (2021). Genotype X Environment Response of 'Matooke' Hybrids (Naritas) to Pseudocercospora fijiensis, the Cause of Black Sigatoka in Banana. *Agronomy*, *11*(6), 1145. https://doi.org/10.3390/agronomy11061145

Kimunye, J., Were, E., Swennen, R., Viljoen, A., & Mahuku, G. (2021). Sources of resistance to Pseudocercospora fijiensis, the cause of black Sigatoka in banana. *Plant Pathology*, *70*(7), 1651-1664. https://doi.org/10.1111/ppa.13408

Kipp, S., Mistele, B., Baresel, P., & Schmidhalter, U. (2014). High-throughput phenotyping early plant vigour of winter wheat. *European Journal of Agronomy*, *52*, 271-278. https://doi.org/https://doi.org/10.1016/j.eja.2013.08.009

Kleinbaum, D. G., & Klein, M. (2012a). Evaluating the Proportional Hazards Assumption. In *Survival Analysis: A Self-Learning Text* (pp. 161-200). Springer New York. https://doi.org/10.1007/978-1-4419-6646-9_4

Kleinbaum, D. G., & Klein, M. (2012b). Introduction to Survival Analysis. In *Survival Analysis: A Self-Learning Text* (pp. 1-54). Springer New York. https://doi.org/10.1007/978-1-4419-6646-9_1

Krajewski, P., Chen, D., Ćwiek, H., van Dijk, A. D. J., Fiorani, F., Kersey, P., . . . Weise, S. (2015). Towards recommendations for metadata and data handling in plant phenotyping. *Journal of Experimental Botany*, *66*(18), 5417-5427. https://doi.org/10.1093/jxb/erv271

Krupnik, T. J., Andersson, J. A., Rusinamhodzi, L., Corbeels, M., Shennan, C., & GÉRard, B. (2019). Does size matter? A critical review of meta-analysis in agronomy. *Experimental Agriculture*, 1-30. https://doi.org/10.1017/S0014479719000012

Kuhn, M. (2022). *caret: Classification and Regression Training*. In (Version 6.0-91) CRAN. https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). Data Pre-processing. In *Applied Predictive Modeling* (pp. 27-59). Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_3

Lammerts van Bueren, E. T., Struik, P. C., van Eekeren, N., & Nuijten, E. (2018). Towards resilience through systems-based plant breeding. A review [journal article]. *Agronomy for Sustainable Development*, *38*(5), 42. https://doi.org/10.1007/s13593-018-0522-6

Lan, Y., Zhu, Y., Guo, J., Niu, S., & Cheng, X. (2014). Position-Aware ListMLE: A Sequential Learning Process for Ranking. UAI,

Laurent, A., Pelzer, E., Loyce, C., & Makowski, D. (2015). Ranking yields of energy crops: A meta-analysis using direct and indirect comparisons. *Renewable and Sustainable Energy Reviews*, *46*, 41-50. https://doi.org/https://doi.org/10.1016/j.rser.2015.02.023

Lecomte, C., Prost, L., Cerf, M., & Meynard, J.-M. (2010). Basis for designing a tool to evaluate new cultivars [journal article]. *Agronomy for Sustainable Development*, *30*(3), 667-677. https://doi.org/10.1051/agro/2009042

Leonelli, S., Davey, R. P., Arnaud, E., Parry, G., & Bastow, R. (2017). Data management and best practice for plant science [Comment]. *Nature Plants*, *3*, 17086. https://doi.org/10.1038/nplants.2017.86

Lesk, C., Rowhani, P., & Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. *Nature*, *529*, 84. https://doi.org/10.1038/nature16467

Liakos, K., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, *18*(8), 2674. http://www.mdpi.com/1424-8220/18/8/2674

Lim, J. (2011). Hedonic scaling: A review of methods and theory. *Food Quality and Preference*, *22*(8), 733-747. https://doi.org/10.1016/j.foodqual.2011.05.008

Lim, J., Wood, A., & Green, B. G. (2009). Derivation and Evaluation of a Labeled Hedonic Scale. *Chemical Senses*, *34*(9), 739-751. https://doi.org/10.1093/chemse/bjp054

Lin, B. B. (2011). Resilience in Agriculture through Crop Diversification: Adaptive Management for Environmental Change. *BioScience*, *61*(3), 183-193. https://doi.org/10.1525/bio.2011.61.3.4

Lin, S. (2010). Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(5), 555-570. https://doi.org/doi:10.1002/wics.111

Lin, S., & Ding, J. (2009). Integration of Ranked Lists via Cross Entropy Monte Carlo with Applications to mRNA and microRNA Studies. *Biometrics*, *65*(1), 9-18. https://doi.org/doi:10.1111/j.1541-0420.2008.01044.x

Linacre, J. M., & Wilson, M. (1992). Objective measurement of rank-ordered objects. In *Objective measurement: Theory into practice.* (pp. 195-209).

Llano, A., Molina, J., Urbina, R., Guzmán, M., & Cuadra, S. (2013). *Avances de la Biofortificaciónde Frijol en Nicaragua(2005-2013)* Nicaragua, INTA. http://lac.harvestplus.org/wp-content/uploads/2015/02/Avances-biof-frijolINTA_ALLano_Taller-Nica_Nov2013.pdf

Lobell, D. B., Bänziger, M., Magorokosho, C., & Vivek, B. (2011). Nonlinear heat effects on African maize as evidenced by historical yield trials. *Nature Climate Change*, *1*(1), 42-45. https://doi.org/http://dx.doi.org/10.1038/nclimate1043

Loomes, G., & Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, *92*(368), 805-824. https://doi.org/10.2307/2232669

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., . . . Halpern, B. S. (2017). Our path to better science in less time using open data science tools [Perspective]. *Nature Ecology &Amp; Evolution*, *1*, 0160. https://doi.org/10.1038/s41559-017-0160

Lu, G., & Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, *23*(20), 3105-3124. https://doi.org/doi:10.1002/sim.1875

Luce, R. D. (1959). *Individual choice behavior: a theoretical analysis*. Dover Publications. https://doi.org/10.1037/14396-000

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, *21*(16), 2313-2324. https://doi.org/doi:10.1002/sim.1201

Lysen, S. (2009). *Permuted inclusion criterion: a variable selection technique* (Publication Number 28) University of Pennsylvania]. Publicly accessible Penn Dissertations. https://repository.upenn.edu/edissertations/28

Madden, L. V., Piepho, H. P., & Paul, P. A. (2016). Statistical Models and Methods for Network Meta-Analysis. *Phytopathology*, *106*(8), 792-806. https://doi.org/10.1094/PHYTO-12-15-0342-RVW

Maia, A. H. N., & Meinke, H. (2010). Probabilistic methods for seasonal forecasting in a changing climate: Cox-type regression models. *International Journal of Climatology*, *30*(15), 2277-2288. https://doi.org/10.1002/joc.2042

Makowski, D. (2021). Editorial of the special issue “Evidence synthesis in agronomy”. *European Journal of Agronomy*, *122*, 126183. https://doi.org/10.1016/j.eja.2020.126183

Makowski, D., Piraux, F., & Brun, F. (2019). Basic Concepts in Meta-analysis. In *From Experimental Network to Meta-analysis: Methods and Applications with R for Agronomic and Environmental Sciences* (pp. 105-126). Springer Netherlands. https://doi.org/10.1007/978-94-024-1696-1_6

Malosetti, M., Ribaut, J.-M., & van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*, *4*, 44-44. https://doi.org/10.3389/fphys.2013.00044

Mark, H., Briony, F., Jon, T., Grace, B., Ross, W., Barend, M., . . . Igor, O. (2018). *The State of Open Data Report 2018* (Digital Science Report, Issue. https://digitalscience.figshare.com/articles/The_State_of_Open_Data_Report_2018/7195058

Matteis, L., Chibon, P. Y., Espinosa, H., Skofic, M., Finkers, R., Bruskiewich, R., . . . Arnaud, E. (2013). Crop ontology: Vocabulary for crop-related concepts. 1st International Workshop on Semantics for Biodiversity (S4BioDiv), Montpellier, France.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

McLaren, C. G., Bruskiewich, R. M., Portugal, A. M., & Cosico, A. B. (2005). The International Rice Information System. A Platform for Meta-Analysis of Rice Crop Data. *Plant Physiology*, *139*(2), 637-642. https://doi.org/10.1104/pp.105.063438

Meyer, H. (2018). CAST: 'caret' Applications for Spatial-Temporal Models. *0.5.1*. https://CRAN.R-project.org/package=CAST

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, *12*(9), 1620-1633. https://doi.org/10.1111/2041-210X.13650

Möhring, J., & Piepho, H.-P. (2009). Comparison of Weighting in Two-Stage Analysis of Plant Breeding Trials. *Crop Science*, *49*(6), 1977-1988. https://doi.org/10.2135/cropsci2009.02.0083

Moore, E. K., Kriesberg, A., Schroeder, S., Geil, K., Haugen, I., Barford, C., . . . Parr, C. (2021). Agricultural data management and sharing: Best practices and case study. *Agronomy journal*, *n/a*(n/a). https://doi.org/https://doi.org/10.1002/agj2.20639

Morales, N., Ogbonna, A. C., Ellerbrock, B. J., Bauchet, G. J., Tantikanjana, T., Tecle, I. Y., . . . Mueller, L. A. (2022). Breedbase: a digital ecosystem for modern plant breeding. *G3 Genes|Genomes|Genetics*. https://doi.org/10.1093/g3journal/jkac078

Moyo, M., Ssali, R., Namanda, S., Nakitto, M., Dery, E. K., Akansake, D., . . . Muzhingi, T. (2021). Consumer Preference Testing of Boiled Sweetpotato Using Crowdsourced Citizen Science in Ghana and Uganda [Original Research]. *Frontiers in Sustainable Food Systems*, *5*. https://doi.org/10.3389/fsufs.2021.620363

Murad, M. H., Wang, Z., Chu, H., & Lin, L. (2019). When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *BMJ (Clinical research ed.)*, *364*, k4817-k4817. https://doi.org/10.1136/bmj.k4817

Nalley, L. L., Barkley, A., Watkins, B., & Hignight, J. (2009). Enhancing Farm Profitability through Portfolio Analysis: The Case of Spatial Rice Variety Selection. *Journal of Agricultural and Applied Economics*, *41*(3), 641-652. https://doi.org/10.1017/S1074070800003126

Nalley, L. L., & Barkley, A. P. (2010). Using Portfolio Theory to Enhance Wheat Yield Stability in Low-Income Nations: An Application in the Yaqui Valley of

Northwestern Mexico. *Journal of Agricultural and Resource Economics*, *35*(2), 334-347.

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes*. In (Version 1.1-3) CRAN. https://CRAN.R-project.org/package=RColorBrewer

Neyhart, J. L., Gutiérrez, L., & Smith, K. P. (2021). Using environmental similarities to design training sets for genomewide selection. *Crop Science*, *61*(1), 396-409. https://doi.org/10.1002/csc2.20303

Ning, S.-P., Xu, L.-B., Lu, Y., Huang, B.-Z., & Ge, X.-J. (2007). Genome composition and genetic diversity of Musa germplasm from China revealed by PCR-RFLP and SSR markers. *Scientia Horticulturae*, *114*(4), 281-288. https://doi.org/10.1016/j.scienta.2007.07.002

Njukwe, E., Amah, D., Ndango, R., & Tenkouano, A. (2010). Fast-Track Release of Black Leaf Streak Resistant Banana and Plantain (Musa spp.) Hybrids and Related Technologies to Farmers in Cameroon: a Strategy to Enhance Food Security. *Acta Horticulturae*, *879*, 577-584. https://doi.org/10.17660/ActaHortic.2010.879.64

Noupadja, P., & Tomekpé, K. (2001). Agronomic performance of IITAs improved Musa germplasm at CRBP, Njombé, Cameroon. *Afr. Crop Sci. J*, *9*(3), 481-486. http://www.bioline.org.br/request?cs01067

Nyombi, K., Van Asten, P. J. A., Leffelaar, P. A., Corbeels, M., Kaizzi, C. K., & Giller, K. E. (2009). Allometric growth relationships of East Africa highland bananas (Musa AAA-EAHB) cv. Kisansa and Mbwazirume. *Annals of Applied Biology*, *155*(3), 403-418. https://doi.org/10.1111/j.1744-7348.2009.00353.x

Oostra, V., Gomes, L. G. L., & Nijman, V. (2008). Implications of deforestation for the abundance of restricted-range bird species in a Costa Rican cloud-forest. *Bird Conservation International*, *18*(1), 11-19. https://doi.org/10.1017/S0959270908000038

Orjeda, G. (1998). Evaluation of Musa germplasm resistance to Sigatoka diseases and Fusarium wilt. *INIBAP technical guidelines*.

Orjeda, G. (2000). *Evaluating Bananas: A Global Partnership: Results of IMTP Phase II*. Bioversity International.

Orlando Narváez, A. A. (2004). *Evaluación agronómica y de la resistencia a fitonemátodos y sigatoka negra de cultivares mejorados de banano y plátano en Ecuador* CATIE]. Turrialba (Costa Rica).

Ortiz-Crespo, B., Steinke, J., Quirós, C. F., van de Gevel, J., Daudi, H., Gaspar Mgimiloko, M., & van Etten, J. (2020). User-centred design of a digital advisory service: enhancing public agricultural extension for sustainable intensification in Tanzania. *International Journal of Agricultural Sustainability*, 1-17. https://doi.org/10.1080/14735903.2020.1720474

Ortiz, R., & Vuylsteke, D. (1998). `BITA-3': A Starchy Banana with Partial Resistance to Black Sigatoka and Tolerance to Streak Virus. *HortScience HortSci*, *33*(2), 358-359. https://doi.org/10.21273/hortsci.33.2.358

Papoutsoglou, E. A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I. N., Chaves, I., . . . Pommier, C. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytologist*, nph.16544. https://doi.org/10.1111/nph.16544

Parkes, B., Higginbottom, T. P., Hufkens, K., Ceballos, F., Kramer, B., & Foster, T. (2019). Weather dataset choice introduces uncertainty to estimates of crop yield responses to climate variability and change. *Environmental Research Letters*, *14*(12), 124089. https://doi.org/10.1088/1748-9326/ab5ebb

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R journal*, *10*(1), 439--446. https://doi.org/10.32614/RJ-2018-009

Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. In (Version 1.1.1) https://CRAN.R-project.org/package=patchwork

Pennisi, E. (2010). Armed and Dangerous. *Science*, *327*(5967), 804-805. https://doi.org/10.1126/science.327.5967.804

Perez-Vicente, L., Carreel, F., Roussel, V., Carlier, J., & Abadie, C. (2021). Guidelines for the evaluation of Pseudocercospora resistance to leaf spots of banana. In M. Dita (Ed.), *Practical guidelines for early screening and field evaluation of banana against Fusarium wilt, Pseudocercospora leaf spots and drought.* Bioversity International.

Peryam, D., & Girardot, N. (1952). QM Pins Food 'Likes" and 'Dislikes' with Advanced Taste-Test Method. *Food Engineering*, *24*(58).

Philibert, A., Loyce, C., & Makowski, D. (2012). Assessment of the quality of meta-analysis in agronomy. *Agriculture, Ecosystems & Environment*, *148*, 72-82. https://doi.org/10.1016/j.agee.2011.12.003

Philipp, M., Rusch, T., Hornik, K., & Strobl, C. (2018). Measuring the Stability of Results From Supervised Statistical Learning. *Journal of Computational and Graphical Statistics*, *27*(4), 685-700. https://doi.org/10.1080/10618600.2018.1473779

Philipp, M., Zeileis, A., & Strobl, C. (2016). *A toolkit for stability assessment of tree-based learners*. http://hdl.handle.net/10419/146128

Pickett, S. T. A., Kolasa, J., & Jones, C. G. (2007). Integration and Synthesis. In S. T. A. Pickett, J. Kolasa, & C. G. Jones (Eds.), *Ecological Understanding (Second Edition)* (pp. 146-167). Academic Press. https://doi.org/10.1016/B978-012554522-8.50009-1

Piepho, H.-P., Denis, J.-B., & van Eeuwijk, F. A. (1998). Predicting Cultivar Differences Using Covariates. *Journal of Agricultural, Biological, and Environmental Statistics*, *3*(2), 151-162. https://doi.org/10.2307/1400648

Piepho, H. P. (1997). Analyzing Genotype-Environment Data by Mixed Models with Multiplicative Terms. *Biometrics*, *53*(2), 761-766. https://doi.org/10.2307/2533976

Piepho, H. P. (2000). Exact Confidence Limits for Covariate-Dependent Risk in Cultivar Trials. *Journal of Agricultural, Biological, and Environmental Statistics*, *5*(2), 202-213. https://doi.org/10.2307/1400531

Piepho, H. P., Möhring, J., Schulz-Streeck, T., & Ogutu, J. O. (2012). A stage-wise approach for the analysis of multi-environment trials [Article]. *Biometrical Journal*, *54*(6), 844-860. https://doi.org/10.1002/bimj.201100219

Piepho, H. P., Williams, E. R., & Madden, L. V. (2012). The Use of Two-Way Linear Mixed Models in Multitreatment Meta-Analysis. *Biometrics*, *68*(4), 1269-1277. https://doi.org/doi:10.1111/j.1541-0420.2012.01786.x

Pillemer, D., & Light, R. (1980). Synthesizing Outcomes: How to Use Research Evidence from Many Studies. *Harvard Educational Review*, *50*(2), 176-195. https://doi.org/10.17763/haer.50.2.v755316522jkup33

Plackett, R. L. (1975). The Analysis of Permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *24*(2), 193-202. https://doi.org/10.2307/2346567

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, *7*(1), 217-240. https://doi.org/10.5194/soil-7-217-2021

Priyadarshan, P. M. (2019). Maintenance Breeding and Variety Release. In *PLANT BREEDING: Classical to Modern* (pp. 561-570). Springer Singapore. https://doi.org/10.1007/978-981-13-7095-3_25

PRR-FIPAH. (2019). Catálogo de Variedades de Frijol y Maíz Desarrolladas en Honduras, bajo Enfoque de Fitomejoramiento Participativo. In (pp. 15). Honduras.

R Core Team. (2022a). *R: A language and environment for statistical computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2022b). *Writing R Extensions*. R Foundation for Statistical Computing. https://cran.r-project.org/doc/manuals/r-release/R-exts.html

Ramirez-Villegas, J., & Challinor, A. (2012). Assessing relevant climate data for agricultural applications. *Agricultural and Forest Meteorology*, *161*, 26-45. https://doi.org/10.1016/j.agrformet.2012.03.015

Rangarajan, A. (2002). Vigor or Rigor? The Competing Goals of Variety Trials. *12*(4), 562. https://doi.org/10.21273/horttech.12.4.562

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Educational Tests*. The Danish Institute for Education Research.

Reyes, B. A., Maredia, M. K., Bernsten, R. H., & Rosas, J. C. (2016). Opportunities seized, opportunities missed: Differences in the economic impact of bean research in five Latin American countries. *MSU International Development Working Paper*(151), 42. https://ageconsearch.umn.edu/record/251850/files/idwp151.pdf

Reynolds, M., Chapman, S., Crespo-Herrera, L., Molero, G., Mondal, S., Pequeno, D. N. L., . . . Sukumaran, S. (2020). Breeder friendly phenotyping. *Plant Science*, 110396. https://doi.org/https://doi.org/10.1016/j.plantsci.2019.110396

Rijgersberg, H., & Top, J. L. (2000). Exchanging crop trials information: Standardization by means of data model templates [Article]. *Computers and Electronics in Agriculture*, *25*(3), 221-231. https://doi.org/10.1016/S0168-1699(99)00070-8

Ritchie, J. T. (1995). International consortium for agricultural systems applications (ICASA): Establishment and purpose. *Agricultural Systems*, *49*(4), 329-335. https://doi.org/https://doi.org/10.1016/0308-521X(95)00028-4

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., . . . Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913-929. https://doi.org/10.1111/ecog.02881

Rosas, J. C., Araya, R., & Ortega, I. (2009). Variedades de frijol rojo, obtenidas por Fitomejoramiento Participativo en Honduras y Nicaragua. In (pp. 40). Estación Experimental “Fabio Baudrit”, Alajuela, Costa Rica: Programa Colaborativo de Fitomejoramiento Participativo en Mesoamérica (FPMA).

Rosas, J. C., Beaver, J. S., Beebe, S., & Viana, A. (2004). Nomenclatura de variedades de frijol común liberadas en Centro América y El Caribe. *Agronomía Mesoamericana*, *15*(2), 221. https://doi.org/10.15517/am.v15i2.11917

Rosas, J. C., Beaver, J. S., Escoto, D., Perez, C. A., Llano, A., Hernandez, J. C., & Araya, R. (2004). Registration of ‘Amadeus 77’ Small Red Common Bean. *Crop Science*, *44*(5), 1867-1868. https://doi.org/10.2135/cropsci2004.1867

Rosas, J. C., & Escoto, D. (2011). Validación comercial de las variedades de frijol de grano rojo tipo criollo. In DICTA-SAG & Programa de Investigaciones en Frijol - Escuela Agrícola Panamericana Zamorano (Eds.). Honduras.

Rosas, J. C., Escoto, D., & Meza, N. (2016). *Propuesta de liberación de la variedad de frijol biofortificado de grano rojo claro brillante "Honduras Nutritivo"*. Tegucigalpa, Honduras: DICTA Retrieved from http://dicta.gob.hn/files/Frijol-2016-012.pdf

Rosas, J. C., Rodriguez, I. Y., Escoto, N. D., & Meza, N. (2019). Liberación de las variedades de frijol "Rojo Chortí" y "Tolupan Rojo" con adaptación a estreses abióticos y resistencia a enfermedades en honduras. 64 reunión Anual del Programa Cooperativo Centroamericano para el Mejoramiento de Cultivos y Animales (PCCMCA), Honduras.

Rowe, P., & Rosales, F. (1993). Diploid breeding at FHIA and the development of Goldfinger (FHIA-01). *InfoMusa*, *2*(2), 9-11.

Ruas, M., Guignon, V., Sempere, G., Sardos, J., Hueber, Y., Duvergey, H., . . . Rouard, M. (2017). MGIS: managing banana (Musa spp.) genetic resources information and high-throughput genotyping data. *Database*, *2017*. https://doi.org/10.1093/database/bax046

Sabaghnia, N. (2016). Nonparametric statistical methods for analysis of genotype × environment interactions in plant pathology [journal article]. *Australasian Plant Pathology*, *45*(6), 571-580. https://doi.org/10.1007/s13313-016-0453-0

Sakyi-Dawson, E., Asamoah-Bonti, P., & Amponsah Annor, G. (2008). Biochemical changes in new plantain and cooking banana hybrids at various stages of ripening. *Journal of the Science of Food and Agriculture*, *88*(15), 2724-2729. https://doi.org/10.1002/jsfa.3399

Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*, *3*(2), 80-97. https://doi.org/doi:10.1002/jrsm.1037

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., . . . Crowley, J. (2021). *GGally: Extension to 'ggplot2'*. In https://CRAN.R-project.org/package=GGally

Schutz, H. G., & Cardello, A. V. (2001). A labeled affective magnitude (LAM) scale for assessing food liking/disliking. *Journal of Sensory Studies*, *16*(2), 117-159. https://doi.org/10.1111/j.1745-459X.2001.tb00293.x

Shorter, R., Lawn, R. J., & Hammer, G. L. (1991). Improving Genotypic Adaptation in Crops – a Role for Breeders, Physiologists and Modellers. *Experimental Agriculture*, *27*(2), 155-175. https://doi.org/10.1017/S0014479700018810

Shrestha, R., Sanchez, H., Ayala, C., Wenzl, P., & Arnaud, E. (2010). Ontology-driven International Maize Information System (IMIS) for Phenotypic and Genotypic Data Exchange. *Nature Precedings*. https://doi.org/10.1038/npre.2010.5029.1

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., & Arppe, A. (2021). DescTools: Tools for descriptive statistics. *R package version 0.99.43.*, *28*, 17. https://cran.r-project.org/web/packages/DescTools/index.html

Simko, I., Hayes, R. J., & Kramer, M. (2012). Computing Integrated Ratings from Heterogeneous Phenotypic Assessments: A Case Study of Lettuce Postharvest Quality and Downy Mildew Resistance. *Crop Science*, *52*(5), 2131. https://doi.org/10.2135/cropsci2012.02.0111

Simko, I., & Linacre, J. M. (2010). Combining partially ranked data in plant breeding and biology: II. Analysis with Rasch model. *Communications in Biometry & Crop Science*, *5*(1). http://agrobiol.sggw.waw.pl/~cbcs/articles/CBCS_5_1_8.pdf

Simko, I., & Pechenick, D. A. (2010). Combining partially ranked data in plant breeding and biology: I. Rank aggregating methods. *Communications in Biometry and Crop Science*, *5*(1), 41-55. http://agrobiol.sggw.waw.pl/~cbcs/articles/CBCS_5_1_7.pdf

Simko, I., & Piepho, H.-P. (2011). Combining phenotypic data from ordinal rating scales in multiple plant experiments. *Trends in Plant Science*, *16*(5), 235-237. https://doi.org/10.1016/j.tplants.2011.02.001

Singh, Y. P., Nayak, A. K., Sharma, D. K., Gautam, R. K., Singh, R. K., Singh, R., . . . Ismail, A. M. (2014). Farmers' Participatory Varietal Selection: A Sustainable Crop Improvement Approach for the 21st Century. *Agroecology and Sustainable Food Systems*, *38*(4), 427-444. https://doi.org/10.1080/21683565.2013.870101

Smith, A. B., Cullis, B. R., & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches [Review]. *Journal of Agricultural Science*, *143*(6), 449-462. https://doi.org/10.1017/S0021859605005587

Smith, K. F., & Fennessy, P. F. (2011). The use of conjoint analysis to determine the relative importance of specific traits as selection criteria for the improvement of perennial pasture species in Australia. *Crop and Pasture Science*, *62*(4), 355-365. https://doi.org/https://doi.org/10.1071/CP10320

Snapp, S. (2002). Quantifying farmer evaluation of technologies: The mother and baby trial design. Quantitative analysis of data from participatory methods in plant breeding. In M. R. Bellon & J. Reeves (Eds.), *Quantitative analysis of Data from participatory methods in plant breeding* (pp. 9-17). CIMMYT.

Sparks, A. H. (2018). nasapower: a NASA POWER global meteorology, surface solar energy and climatology data client for R.

Spyns, P., Meersman, R., & Jarrar, M. (2002). Data modelling versus ontology engineering. *SIGMOD Rec.*, *31*(4), 12-17. https://doi.org/10.1145/637411.637413

Steinke, J., Achieng, J. O., Hammond, J., Kebede, S. S., Mengistu, D. K., Mgimiloko, M. G., . . . van Etten, J. (2019). Household-specific targeting of agricultural advice via mobile phones: Feasibility of a minimum data approach for smallholder context. *Computers and Electronics in Agriculture*, *162*, 991-1000. https://doi.org/https://doi.org/10.1016/j.compag.2019.05.026

Steinke, J., & van Etten, J. (2017). Gamification of farmer-participatory priority setting in plant breeding: Design and validation of "AgroDuos". *Journal of Crop Improvement*, *31*(3), 356-378. https://doi.org/10.1080/15427528.2017.1303801

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677-680. https://doi.org/10.1126/science.103.2684.677

Stone, H., Bleibaum, R. N., & Thomas, H. A. (2012). Chapter 1 - Introduction to Sensory Evaluation. In H. Stone, R. N. Bleibaum, & H. A. Thomas (Eds.), *Sensory Evaluation Practices (Fourth Edition)* (pp. 1-21). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-382086-0.00001-7

Streck, E. A., de Magalhaes Jr., A. M., Aguiar, G. A., Henrique Facchinello, P. K., Reis Fagundes, P. R., Franco, D. F., . . . de Oliveira, A. C. (2018). Genetic Progress in 45 Years of Irrigated Rice Breeding in Southern Brazil. *Crop Science*, *58*(3), 1094-1105. https://doi.org/10.2135/cropsci2017.06.0383

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological methods*, *14*(4), 323-348. https://doi.org/10.1037/a0016973

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning. *Journal of Educational and Behavioral Statistics*, *36*(2), 135-153.

Sukcharoen, K., & Leatham, D. (2016). Mean-variance versus mean-expected shortfall models: an application to wheat variety selection. *Journal of Agricultural and Applied Economics*, *48*(2), 148-172. https://doi.org/10.1017/aae.2016.8

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., & Bennett, M. (2017). Plant Phenomics, From Sensors to Knowledge. *Current Biology*, *27*(15), R770-R783. https://doi.org/https://doi.org/10.1016/j.cub.2017.05.055

Tenkouano, A., Lamien, N., Agogbua, J., Amah, D., Swennen, R., Traoré, S., . . . Ortiz, R. (2019). Promising High-Yielding Tetraploid Plantain-Bred Hybrids in West Africa. *International Journal of Agronomy*, *2019*, 3873198. https://doi.org/10.1155/2019/3873198

Tenkouano, A., Ortiz, R., & Nokoe, S. (2012). Repeatability and optimum trial configuration for field-testing of banana and plantain. *Scientia Horticulturae*, *140*, 39-44. https://doi.org/10.1016/j.scienta.2012.03.023

Tesfaye, K., Sonder, K., Cairns, J., Magorokosho, C., Tarekegn, A., Kassie, G. T., . . . Erenstein, O. (2016). Targeting drought-tolerant maize varieties in Southern Africa: A geospatial crop modeling approach using big data [Article]. *International Food and Agribusiness Management Review*, *19*(A), 75-92. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85021318536&partnerID=40&md5=3a713687763079c7d57edfaa0df369eb

Theobald, C. M., Talbot, M., & Nabugoomu, F. (2002). A Bayesian approach to regional and local-area prediction from crop variety trials [Article]. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*(3), 403-419. https://doi.org/10.1198/108571102230

Therneau, T. M. (2021). *A Package for Survival Analysis in R*. In (Version 3.2-13) https://CRAN.R-project.org/package=survival

Therneau, T. M., & Grambsch, P. M. (2000). The Cox Model. In *Modeling Survival Data: Extending the Cox Model* (pp. 39-77). Springer New York. https://doi.org/10.1007/978-1-4757-3294-8_3

Thessen, A. E. (2016). Adoption of Machine Learning Techniques in Ecology and Earth Science. *One Ecosystem*, *1*. https://doi.org/10.3897/oneeco.1.e8621

Thomas, D. S., Turner, D. W., & Eamus, D. (1998). Independent effects of the environment on the leaf gas exchange of three banana (Musa sp.) cultivars of different genomic constitution. *Scientia Horticulturae*, *75*(1), 41-57. https://doi.org/10.1016/S0304-4238(98)00114-9

Tomlins, K., Rwiza, E., Nyango, A., Amour, R., Ngendello, T., Kapinga, R., . . . Jolliffe, F. (2004). The use of sensory evaluation and consumer preference for the selection of sweetpotato cultivars in East Africa. *Journal of the Science of Food and Agriculture*, *84*(8), 791-799. https://doi.org/doi:10.1002/jsfa.1712

Tonin, F. S., Rotta, I., Mendes, A. M., & Pontarolo, R. (2017). Network meta-analysis: a technique to gather evidence from direct and indirect comparisons. *Pharmacy practice*, *15*(1), 943-943. https://doi.org/10.18549/PharmPract.2017.01.943

Trabucco, A., & Zomer, R. J. (2019). *Global High-Resolution Soil-Water Balance* DOI.org (Datacite). https://figshare.com/articles/dataset/Global_High-Resolution_Soil-Water_Balance/7707605

Turner, D. W. (1995). The response of the plant to the environment. In S. Gowen (Ed.), *Bananas and Plantains* (pp. 206-229). Springer Netherlands. https://doi.org/10.1007/978-94-011-0737-2_9

Turner, D. W., Fortescue, J. A., & Thomas, D. S. (2007). Environmental physiology of the bananas (Musa spp.). *Brazilian Journal of Plant Physiology*, *19*, 463-484. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1677-04202007000400013&nrm=iso

Turner, H., & Firth, D. (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. *2012*, *48*(9), 21. https://doi.org/10.18637/jss.v048.i09

Turner, H. L., van Etten, J., Firth, D., & Kosmidis, I. (2020). Modelling rankings in R: the PlackettLuce package. *Computational Statistics*. https://doi.org/10.1007/s00180-020-00959-3

Tushemereirwe, W., Batte, M., Nyine, M., Tumuhimbise, R., Barekye, A., Tendo, S., . . . Uwimana, B. (2015). *Performance of NARITA banana hybrids in the preliminary yield trial for three cycles in Uganda*. I. NARO. http://www.iita.org/c/document_library/get_file?uuid=1500c39e-7c05-4219-b649-85cdb1bd73f6&groupId=25357

Ushey, K., Allaire, J., & Tang, Y. (2022). *reticulate: Interface to 'Python'*. In (Version 1.24) https://cran.r-project.org/web/packages/reticulate/index.html

van Eeuwijk, F., Denis, J., & Kang, M. (1996). Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In *Genotype-by-environment interaction* (pp. 15-50). CRC Press.

van Eeuwijk, F. A., Bustos-Korts, D. V., & Malosetti, M. (2016). What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype × Environment Interactions? *Crop Science*, *56*(5), 2119-2140. https://doi.org/10.2135/cropsci2015.06.0375

van Eeuwijk, F. A., Malosetti, M., Yin, X., Struik, P. C., & Stam, P. (2005). Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. *Australian Journal of Agricultural Research*, *56*(9), 883-894. https://doi.org/https://doi.org/10.1071/AR05153

van Etten, J., Abidin, E., Arnaud, D., Brown, E., Carey, E., Laporte, M.-A., . . . Valle-Soto, J. (2020). *The tricot citizen science approach applied to on-farm variety evaluation: methodological progress and perspectives*. CGIAR Research Program on Roots, Tubers and Bananas (RTB). https://doi.org/10.4160/23096586RTBWP20212

van Etten, J., Beza, E., Calderer, L., van Duijvendijk, K., Fadda, C., Fantahun, B., . . . Zimmerer, K. S. (2019). First experiences with a novel farmer citizen science approach: crowdsourcing participatory variety selection through on-farm triadic comparisons of technologies (tricot). *Experimental Agriculture*, *55*(S1), 275-296. https://doi.org/10.1017/S0014479716000739

van Etten, J., de Sousa, K., Aguilar, A., Barrios, M., Coto, A., Dell'Acqua, M., . . . Steinke, J. (2019). Crop variety management for climate adaptation supported by citizen science. *Proceedings of the National Academy of Sciences*, *116*(10), 4194-4199. https://doi.org/10.1073/pnas.1813720116

van Etten, J., Quirós, C. F., Madriz, B., & de Sousa, K. (2017). Climmob, digital platform to support triadic comparisons of technologies (tricot). In: CGIAR Gender & Breeding Initiative, Nairobi, Kenya.

van Etten, J., Steinke, J., & van Wijk, M. (2017). How can the Data Revolution contribute to climate action in smallholder agriculture? *Agriculture for Development*, 44-48.

van Tilborg, E. (2018). *Spatial Analysis and Visualization of Triadic Crop Variety Trials in Central America* Wageningen University and Research Centre]. Wageningen, The Netherlands.

van Wesemael, J., Kissel, E., Eyland, D., Lawson, T., Swennen, R., & Carpentier, S. (2019). Using Growth and Transpiration Phenotyping Under Controlled Conditions to Select Water Efficient Banana Genotypes [Original Research]. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00352

Vanhove, A.-C., Vermaelen, W., Panis, B., Swennen, R., & Carpentier, S. (2012). Screening the banana biodiversity for drought tolerance: can an in vitro growth model and proteomics be used as a tool to discover tolerant varieties and understand homeostasis [Original Research]. *Frontiers in Plant Science*, *3*. https://doi.org/10.3389/fpls.2012.00176

Vargas, A., Watler, W., Morales, M., & Vignola, R. (2018). Prácticas efectivas para la reducción de impactos por eventos climáticos en Costa Rica - Cultivo de Frijol. 79. Retrieved 2018, from http://www.mag.go.cr/bibliotecavirtual/reduccion-impacto-por-eventos-climaticos/Informe-final-frijol.pdf

Vargas, A. G., Rosas, J. C., Pavon, M., & Mejia, P. (2011). Campechano JR: Variedad de frijol rojo desarrollada por la asociación de CIALES del Lago de Yojoa (ASOCIALAYO). In Programa de Investigaciones en Frijol, Escuela Agrícola Panamericana Zamorano & Programa de Reconstrucción Rural (PRR) (Eds.). Honduras.

Vargas, M., Crossa, J., Reynolds, M. P., Dhungana, P., & Eskridge, K. M. (2007). Structural equation modelling for studying genotype×environment interactions of physiological traits affecting yield in wheat. *The Journal of Agricultural Science*, *145*(2), 151-161. https://doi.org/10.1017/S0021859607006806

Vargas, M., Crossa, J., van Eeuwijk, F. A., Ramírez, M. E., & Sayre, K. (1999). Using Partial Least Squares Regression, Factorial Regression, and AMMI Models for Interpreting Genotype × Environment Interaction. *Crop Science*, *39*(4), 955-967. https://doi.org/10.2135/cropsci1999.0011183X003900040002x

Virk, D. S., Pandit, D. B., Sufian, M. A., Ahmed, F., Siddique, M. A. B., Samad, M. A., . . . Witcombe, J. R. (2009). REML is an effective analysis for mixed modelling of unbalanced on-farm varietal trials. *Experimental Agriculture*, *45*(1), 77-91. https://doi.org/10.1017/S0014479708007047

Waldman, K. B., Kerr, J. M., & Isaacs, K. B. (2014). Combining participatory crop trials and experimental auctions to estimate farmer preferences for improved common bean in Rwanda. *Food Policy*, *46*, 183-192. https://doi.org/https://doi.org/10.1016/j.foodpol.2014.03.015

Wan, Z., Hook, S., & Hulley, G. (2015). *MYD11A1 MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MODIS/MYD11A1.006

Weltzien, E., & Christinck, A. (2017). Participatory breeding: developing improved and relevant crop varieties with farmers. In *Agricultural Systems (Second Edition)* (pp. 259-301). Elsevier.

White, J. W., Hunt, L. A., Boote, K. J., Jones, J. W., Koo, J., Kim, S., . . . Hoogenboom, G. (2013). Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. *Computers and Electronics in Agriculture*, *96*, 1-12. https://doi.org/https://doi.org/10.1016/j.compag.2013.04.003

White, J. W., & van Evert, F. K. (2008). Publishing agronomic data. *Agronomy journal*, *100*(5), 1396-1400.

Whitley, E., & Ball, J. (2002). Statistics review 6: Nonparametric methods. *Critical Care*, *6*(6), 509. https://doi.org/10.1186/cc1820

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* Springer-Verlag. https://ggplot2.tidyverse.org

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. In (Version 1.4.0) CRAN. https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. In (Version 1.0.8) https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Bryan, J. (2022). *readr: Read Rectangular Text Data*. In (Version 2.1.2) https://CRAN.R-project.org/package=readr

Wickham, H., Hester, J., Chang, W., & Bryan, J. (2021). *devtools: Tools to Make Developing R Packages Easier*. In (Version 2.4.3) CRAN. https://CRAN.R-project.org/package=devtools

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship [Comment]. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Williams, S. C. (2012). Data Practices in the Crop Sciences: A Review of Selected Faculty Publications. *Journal of Agricultural & Food Information*, *13*(4), 308-325. https://doi.org/10.1080/10496505.2012.717846

Wu, Y.-C., Shih, M.-C., & Tu, Y.-K. (2021). Using Normalized Entropy to Measure Uncertainty of Rankings for Network Meta-analyses. *Medical Decision Making*, *41*(6), 706-713. https://doi.org/10.1177/0272989x21999023

Wyborn, C., Louder, E., Harrison, J., Montambault, J., Montana, J., Ryan, M., . . . Hutton, J. (2018). Understanding the Impacts of Research Synthesis. *Environmental Science & Policy*, *86*, 72-84. https://doi.org/https://doi.org/10.1016/j.envsci.2018.04.013

Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants [journal article]. *Theoretical and Applied Genetics*, *129*(4), 653-673. https://doi.org/10.1007/s00122-016-2691-5

Yan, W. (2014a). An Overview of Variety Trial Data and Analyses. *Crop Variety Trials*, 23-30. https://doi.org/doi:10.1002/9781118688571.ch2

10.1002/9781118688571.ch2 (Wiley Online Books)

Yan, W. (2014b). Theoretical Framework for Crop Variety Trials. In *Crop Variety Trials: Data Management and Analysis*. https://doi.org/10.1002/9781118688571.ch1

Yan, W., Hunt, L. A., Sheng, Q., & Szlavnics, Z. (2000). Cultivar Evaluation and Mega-Environment Investigation Based on the GGE Biplot. *Crop Science*, *40*(3), 597-605. https://doi.org/10.2135/cropsci2000.403597x

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, *28*(4), 556-580. https://doi.org/10.1017/S0021859600050978

Yıldız, İ., Dy, J., Erdoğmuş, D., Kalpathy-Cramer, J., Ostmo, S., Campbell, J. P., . . . Ioannidis, S. (2020). *Fast and Accurate Ranking Regression* Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v108/yildiz20a.html

Yonow, T., Ramirez-Villegas, J., Abadie, C., Darnell, R. E., Ota, N., & Kriticos, D. J. (2019). Black Sigatoka in bananas: Ecoclimatic suitability and disease pressure assessments. *PLoS ONE*, *14*(8), e0220601. https://doi.org/10.1371/journal.pone.0220601

Yu, P. L. H., Gu, J., & Xu, H. (2019). Analysis of ranking data. *WIREs Computational Statistics*, *11*(6), e1483. https://doi.org/10.1002/wics.1483

Zeileis, A., Hornik, K., & Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, *53*(9), 3259-3270. https://doi.org/10.1016/j.csda.2008.11.033

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514. https://doi.org/10.1198/106186008X319331

Zeileis, A., Hothorn, T., & Hornik, K. (2010). party with the mob: Model-Based Recursive Partitioning in R. *R package version 0.9-9999*.

Zhang, T. (2018). Data Synthesis. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 1-3). Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_503-1

Ziliak, S. T. (2019). How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little "p" Is Not Enough. *The American statistician*, *73*(sup1), 281-290. https://doi.org/10.1080/00031305.2018.1514325

Acknowledgements

I would like to thank my supervisors Arnold Bregt, Sytze de Bruin and Jacob van Etten for all their help, support, patience, and guidance during my PhD.

Arnold kindly agreed to be my promotor despite his full agenda as Dean of Education. His always positive attitude and thoughtful suggestions made our short interactions very fruitful.

Sytze has been a great supervisor, who generously embraced the challenge of supervising an external PhD candidate, teaching me many things, above all the importance of pragmatism.

Jacob has been my mentor for the past four years, always motivating me to believe in myself and to go one step further. Thank you, Jacob, for everything you have done to support me in completing my PhD.

I also thank Kauê de Sousa for all his help, especially with R related things and for always supporting my idea of developing the *ag5Tools* package. Juan Carlos Rosas and Juan Carlos Hernandez, for taking the time to share their knowledge with me about common bean in Central America. Catherine Abadie, for sharing her knowledge about black leaf streak disease. Lewis Machida, for sharing the breeder's perspective in crop variety evaluation and for spotting those small but important details in the manuscripts. Sebastien Carpentier, for being supportive about the idea of developing Chapter 4 and helping to make it happen. To all the co-authors, thank you for your great contribution to making all the papers a reality, but especially for allowing me to learn from each of you.

I am also grateful to Inge Van den Bergh for her support at the beginning of my PhD. Claudius vande Vijver and everyone at PE&RC, for always making an effort to provide great courses and activities. To Truus van de Hoef and Antoinette Stoffers GRS, for always being willing to help me. Thank you also to Olga Spellman, Vincent Johnson and Harri Washington at the Alliance of Bioversity International and CIAT for the English editing of my manuscripts. To my paranymphs, Ximena Tagle and Hugo Dorado, for all their help.

To my wife Carol, thank you for your support, suggestions and especially for always understanding how important it was for me to obtain my PhD. To my daughter Paula, thank you for your patience and understanding, but especially for all your questions. Thank you also to my parents, for always showing me the importance of education.

This has been an amazing journey, thank you to all for making it possible!

Short biography

David Brown Fuentes was born in 1982 in Turrialba, Costa Rica. In 2006, he graduated in Business Informatics at the University of Costa Rica. In 2015, David completed an MSc in Geographical Information Science and Systems at the University of Salzburg, Austria, through the UNIGIS program. In 2018, he started a PhD program at Wageningen University, at the Laboratory of Geo-information Science and Remote Sensing.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (6 ECTS)

Geographic analysis and meta-analysis of experimental data in crop improvement

Writing of Project proposal (4.5 ECTS)

Which variety for which farm? Creating location-specific information for banana variety release decisions and banana variety choice recommendations in Uganda and Tanzania

Post-graduate courses (5.8 ECTS)

- Hands-on digital soil mapping; ISRIC (2019)
- Linear models; PE&RC, WIMEK (2020)
- Generalized linear models; PE&RC, WIMEK (2020)
- Mixed linear models; PE&RC, WIMEK (2020)
- Meta-analysis; PE&RC, WIMEK (2020)
- Advances in intercropping; PE&RC (2021)

Deficiency, refresh, brush-up courses (2.45 ECTS)

- Data science: R basics; edX, HarvardX (2020)
- Building R packages; Coursera, Johns Hopkins University (2021)
- Python for data science, AI & development; Coursera, IBM (2021)
- Survival analysis in R for public health; Coursera, Imperial College London (2021)
- Recommender systems: evaluation and metrics; Coursera, University of Minnesota (2021)

Laboratory training and working visits (2.1 ECTS)

- Landrace spatial gap analysis workshop: Global Crop Diversity Trust, Bonn, Germany (2018)
- Breeding better bananas project, annual planning meeting and working package 4 workshop on the on-farm evaluation of NARITA hybrids in Tanzania and Uganda: International Institute of Tropical Agriculture (IITA), Nelson Mandela African Institution of Science and Technology (NM-AIST), Tanzania (2018)

Invited review of (unpublished) journal manuscripts (1 ECTS)

- Crop Science: crop evaluation (2021)

Competence strengthening / skills courses (2.65 ECTS)

- Scientific publishing; WGS (2018)
- Project and time management; WGS (2021)
- Reviewing a scientific manuscript; WGS (2021)
- Research data management; WGS (2021)

Scientific integrity / ethics in science activity (0.6 ECTS)

- Scientific integrity; WGS (2019)

PE&RC Annual meetings, seminars or weekends (1.65 ECTS)

- PE&RC First years weekend (2018)
- PE&RC Day (2018)
- PE&RC Afternoon event (2020)
- PE&RC Day (2021)

Discussion groups / local seminars or scientific meetings (4.5 ECTS)

- Bioversity International Science Seminar Series (2018-2021)

International symposia, workshops and conferences (4.4 ECTS)

- ASA-CSSA-SSSA International Annual Meeting; poster presentation; online (2021)
- 2nd International Agrobiodiversity Congress; poster presentation; online (2021)

The research described in this thesis was financially supported by the Alliance of Bioversity International and CIAT.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by David Brown

Map: David Brown

Image: Bioversity International

Printed by ProefschriftMaken