Bayesian joint-regression analysis of unbalanced series of on-farm trials.

Michel Turbet Delof, Pierre Rivière, Julie C Dawson, Arnaud Gauffreteau, Isabelle Goldringer, Gaëlle van Frank, Olivier David

Bayesian joint-regression analysis of unbalanced series of on-farm trials

Michel Turbet Delof[1], Pierre Rivière[2], Julie C. Dawson[3], Arnaud Gauffreteau[4], Isabelle Goldringer[1], Gaëlle van Frank[1] & Olivier David[5]

[1] UMR GQE-Le Moulon, Université Paris-Saclay - INRAE - CNRS - AgroParisTech – Gif-sur-Yvette, France
[2] Mètis – Prayssas, France
[3] Department of Horticulture, University of Wisconsin-Madison – WI 53706, USA
[4] UMR Agronomie, Université Paris-Saclay - AgroParisTech - INRAE – Palaiseau, France
[5] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

Correspondence: Michel.Turbet-Delof@inrae.fr

Abstract

Participatory plant breeding (PPB) is aimed at developing varieties adapted to agroecologically-based systems. In PPB, selection is decentralized in the target environments, and relies on collaboration between farmers, farmers' organisations and researchers. By doing so, evaluation of new genotypes takes genotype $\times$ environment ($G \times E$) interactions into account to select for specific adaptation. In many cases, there is little overlap among genotypes assessed from farm to farm because the farmers participating in a PPB project choose which ones to assess on their farm. In addition, on-farm trials can often generate more extreme observations than trials carried out on research stations. These features make the estimation of genotype, environment and interaction effects more difficult. This challenge is not unique to PPB, as many breeding programs use sparse testing or incomplete block designs to evaluate more genotypes, however in PPB genotypes are not assigned randomly to environments. To explore methods of overcoming these challenges, this article tests various data analysis scenarios using a Bayesian approach with different models and a real wheat PPB dataset over 11 years. Four morpho-agronomic traits were studied, representing over 1000 $G \times E$ combinations from 189 on-farm trials. This dataset was severely unbalanced with more than 90% of $G \times E$ combinations missing. We compared various Bayesian Finlay-Wilkinson models and found that placing hierarchical distributions on model parameters and modelling residuals using a Student's t distribution jointly improved the estimates of main effects and interactions. This statistical framework allowed us to estimate two indicators of genotype stability (one static and one dynamic) despite the high disequilibrium of the data. We found differences in mean and stability between genotype categories, with mixtures tending to be more stable. The methods developed could be used for evaluation and/or selection within networks of various stakeholders such as farmers, gardeners, plant breeders or managers of genetic resource centres.

Keywords: decentralized participatory plant breeding; bread wheat; $G \times E$ interaction; hierarchical model; Finlay-Wilkinson model; Student's t distribution; varietal stability

1 Introduction

Developing new varieties adapted to Organic Agriculture (OA), agroecological and low input systems is a major concern to achieve improvements in agricultural sustainability (Wolfe et al., 2008). In OA, the use of synthetic inputs (nitrogen, phytochemicals) is not allowed, therefore, cropping environments are not standardized by inputs and varieties grow in more diverse conditions from farm to farm (Dawson et al., 2008). These environments are dependent on pedoclimatic conditions, yearly weather, farmers' management practices and interactions between these factors (Desclaux et al., 2008).

In order to develop varieties adapted to such a diversity of environments two strategies can be used: (i) centralized and indirect selection, or (ii) decentralized and direct selection. The key difference between these approaches is the way they take genotype-by-environment ($G \times E$) interactions into account. These interactions are considered by plant breeders as the main factor limiting the efficiency of the response to selection in breeding programs (Ceccarelli et al., 2001). In centralized and indirect selection, breeding lines are evaluated and selected at a few research stations assumed to represent the target environments. This is efficient if there is a high additive genetic correlation between the trait measured on the station and the same trait measured in the target environment, and if the narrow sense heritability is high in the selection environment (Falconer, 1960).

Decentralized selection can take account of $G \times E$ interactions that are important in OA (Dawson et al., 2008; Murphy et al., 2007). In this approach, the selection and evaluation environments are very close to the target environments (the production environments of farms). Selection then maximizes the use of the reproducible part of G $\times$ E interactions to select for specific adaptations (Annicchiarico et al., 2010). This method is close to direct selection and has been shown to be effective (Annicchiarico et al., 2010; Ceccarelli et al., 2001; Murphy et al., 2007; Smith et al., 2001; Virk et al., 2005).

Many PPB programs have been carried out over the last 20 years targeting low-input farming systems in the Global South and also OA and agroecological systems in Europe and North America (Ceccarelli and Grando, 2020). A few programs tested different experimental designs and specific statistical methods to analyze data taking GxE into account (Mohammadi et al., 2011; Snapp and Silim, 2002). However, few had an extensive dataset with a large number of farms, years and genotypes to allow investigation of the best relevant methods in detail. One program with such data is a wheat PPB program that started in France in 2005, as a collaboration between INRAE GQE-Le Moulon and the Farmers' Seed Network (Réseau Semences Paysannes, RSP). This PPB program had three objectives: (i) develop population-varieties adapted to farmers' practices and needs (organic management, artisanal bread quality ...) using a participatory approach, (ii) develop strategies for preserving genetic diversity through on-farm dynamic management and breeding, and (iii) learn from and improve farmers' individual and collective breeding methods and diffuse successful methods broadly.

In this program, a large number of populations was evaluated over a large network of farms in the RSP (Dawson et al., 2011; Goldringer et al., 2020; Rivière et al., 2015a,b; van Frank et al., 2020). Because of the extensive trial network, the assessment of population performance within an environment could potentially be improved by taking account of the average performance of populations over the network and the stability of their performance, in particular temporal stability, as it determines agronomic and economic risks. Two types of stability assessment have been developed: static and dynamic stability (Becker and Leon, 1988; Lin et al., 1986).

The farmers involved in the program chose which populations to evaluate on their farm, based on prior information about the parents, and characteristics of interest. As very few populations were present in all the trials, the resulting series of trials was very unbalanced, so that the estimation of population average performances and stabilities was difficult. Joint regression is a robust method for estimating genotype main effects and stability with incomplete datasets (Finlay and Wilkinson, 1963; Pereira et al., 2007). It is based on the Finlay-Wilkinson (FW) model, which is parsimonious since the interaction effect between a genotype and an environment is modelled as the product of a genotype stability parameter, called sensitivity, and the

environment main effect. Various Finlay-Wilkinson models have been used in a frequentist framework, which include fixed-effect models and models with random environmental effects (Nabugoomu et al., 1999; Ng and Williams, 2001; Patterson and Silvey, 1980). In the latter models, environmental effects are assumed to come from a common distribution, thereby leading to shrunk estimates. FW models in which genotype main effects, environment main effects and genotype sensitivities are all random effects have recently been developed that can handle unbalanced data and take into account the similarity between some genotypes and the similarity between some environments. These have been implemented in a Bayesian framework and when they include random effects, these are called hierarchical models (Carlin and Louis, 2008; Robert, 2007). Thus far, these models have been used to analyze slightly unbalanced trials (Lian and de los Campos, 2016). Hierarchical joint regression has also been used to analyze very unbalanced simulated data (van Frank et al., 2019). This simulation study has shown that genotypes should be tested in sufficiently many trials in order to estimate their main effects and sensitivities reliably. However, this method had not been used to analyze real and very unbalanced trials. Thus, it was not clear if it could cope with the actual levels of unbalanced data seen in the French PPB on-farm trials and what insight it could give into the behavior of genotypes across environments.

Extreme data is an important issue in data analysis. In multi-environment trials (MET), they may come from either (1) errors between scoring and data formatting (measurement error, wrong labelling, etc.), or (2) environmental heterogeneity in the trial (weed infestation, soil fertility, etc.), or (3) particular environmental conditions that fall outside the normal range of environments under study (poor emergence, extreme weather, strong pest/disease pressure, etc.). In our PPB program, as cultivation environments are less controlled, extreme observations (types 2 and 3) could be more frequent than expected. This could bias estimates based on the normal distribution. Extreme observations could be removed from the dataset to solve this problem, but it is difficult to decide which observations to remove. If too many extreme observations are removed, then the variability of the data may be underestimated and the precision of the statistical analysis overestimated. Alternatively, statistical methods that are robust to extreme observations may be used (Hampel et al., 2011; Huber and Ronchetti, 1981). Various robust methods have been developed in a frequentist or a Bayesian framework, in particular methods consisting in replacing the normal distribution by a Student's t distribution in statistical models. This distribution is more robust to extreme observations than the normal distribution, because it has heavier tails (Carlin and Polson, 1991; Choy and Chan, 2008; Lange et al., 1989; Rosa et al., 2003). It has been used to handle the extreme observations of a single trial in a Bayesian framework (Besag and Higdon, 1999; Cao et al., 2022; Gianola et al., 2018). However, to our knowledge, it has not been used to analyze an unbalanced network of trials.

Recently, participatory variety trials using crowdsourcing has been used in several countries with great success (van Etten et al., 2019). These methods typically use an experimental design called a triadic comparison of technologies (tricot), followed by an analysis of variety ranks (Beza et al., 2017). In the tricot design, large numbers of farmers each compare three variety subsets from the complete set of entries, and provide direct comparison rankings among them for a few traits (i.e. best/middle/worst). By using ranking methods and structuring the entry distribution as an incomplete block design, this allows for comparisons of larger numbers of varieties without overburdening individual farmers. These design options enhance breeders' ability to engage farmers in trialing experimental lines, since on-farm trials are often limited by space and farmers' time. Trialing a few experimental lines, including a check line or variety that is replicated across sites is more realistic for farmers than implementing a fully replicated design. Triadic methods are very useful in many situations, but they are not applicable to more mature networks of farmer breeders such as we have in our PPB program for wheat. The farmers in this network have selected populations over time according to their own rationale, and the populations are not randomly assigned to farmers. In addition, farmers test different numbers of populations, with some only trialing a few and others trialing several dozen. Farmers also want access to quantitative data rather than simple ranks, and so a non-parametric ranking of varieties with no assumptions about distribution will not produce a satisfactory analysis for this purpose.

This article was aimed at improving the assessment of the population-varieties of our program by using

the information at the level of the network. As our dataset was very unbalanced and could include extreme observations, we compared several Finlay-Wilkinson models, in particular hierarchical models and models based on the t distribution. These models were developed in a Bayesian framework, since this framework is rigorous and since it facilitates the implementation of complex models (Carlin and Louis, 2008; Robert, 2007). Finally, the best Finlay-Wilkinson model we obtained was used to analyze our data and characterize the behaviour of our population-varieties across environments.

2 Material and methods

Notation	Meaning
PPB	Participatory plant breeding
OA	Organic agriculture
RSP	French farmers' seed network
MET	Multi-environment trial
$G \times E$	Genotype $\times$ environment interaction
FW	Finlay Wilkinson
MCMC	Markov chain Monte Carlo
LOO	Leave one out
α	Germplasm main effect
θ	Environment main effect
η	Germplasm sensitivity (FW coefficient)
S^2	Germplasm static stability
W	Germplasm ecovalence (a dynamic stability)
elpd_{loo}	LOO expected logarithmic predictive density

Table 1. Main notations.

In our study, we will call a germplasm any biological entity whose individuals are derived from the same breeding process, including varieties registered in the official catalog, landraces, historic varieties, mixtures or populations stemming from crosses. An environment is the combination of a farm and a year.

2.1 Germplasm

We studied 206 germplasm covering different "germplasm types": 98 "cross" germplasm resulting from crosses made either on the farm or at the research station (Rivière et al., 2015b), 50 "landraces", i.e. population varieties grown before 1850, 30 "historic varieties", developed by professional breeding before 1950, 17 "mixtures", which were generally complex, with numerous genotypes from potentially all the other germplasm types. In addition, 11 "registered varieties" after 1950 and widely used in organic farming were included: Maitre Pierre (1954), Poncheau (1956), Renan (1990), Ataro (2004), Pollux (2004), Rubisco (2012), Hendrix (2012), Kampmann selected in Renan, and Hermes (1982), Alauda (2004) and Goldritter (2013), all three selected in Probus (1957).

2.2 Experimental designs and data

2.2.1 Experimental designs

Data were collected between 2008 and 2019. The wheat PPB program followed numerous experimental designs due to the different constraints of farmers, collectives and researchers. The designs have been grouped into 5 classes (Tab. 2).

The "regional farm" and "satellite farm" designs were co-designed to be adapted to the farmer-breeders' constraints and to be used in their agricultural routine. In these designs, the germplasm common to all

farms (control germplasm) were collectively chosen by farmers and researchers, while each farmer individually chose the additional germplasm to be cultivated in his farm. At the beginning the control was a selection in a landrace, and after 2014 it was a germplasm stemming from a cross. Both complete-block and incomplete-block designs were used to address specific research questions such as the study of the evolution of traits (Rivière et al., 2015b), local adaptation (van Frank et al., 2020) or the evaluation of agronomic performance (Goldringer et al., 2020). Some unreplicated trials corresponded to trials with replications but for which measurements could not be performed in some replications.

	Nb of blocks	Nb of repeated germplams	Nb of gemplasms by environment	Nb of environments
Complete blocks	2 to 3	6 to 45	7 to 45	24
Incomplete blocks	3 to 4	3 to 49	6 to 58	11
Regional farm	2	5 to 16	7 to 81	20
Statellite farm	1	1 to 22	5 to 79	102
Unreplicated	1	0	5 to 39	32

Table 2. Experimental designs of the 189 trials used in the statistical analysis. Nb: number, Environment: combination of a year and a location.

2.2.2 Data collected

Four traits were studied, plant height (60% of the data was the average height of 25 individuals and 40% was the overall height of the microplot, mm), spike weight (mean of 25 individual measures, g), protein content of the grain (on the microplot, measured with NIRS technology at INRAE Clermont-Ferrand France, %) and thousand kernel weight (TKW, measured on the microplot, g). These four traits were among those collectively chosen by farmers and researchers to be measured during the PPB program (Table 3). Plant height was measured in the field, while the other traits were measured after harvest at the research station on samples of spikes sent by farmers. The data analyzed were the adjusted means for block effects if these effects were significant, and the empirical means if otherwise. Obvious outliers were excluded.

van Frank et al. (2019) analyzed the sensitivity of the hierarchical FW model to different MET set-ups with simulated data. They found that, in contrast to the environmental effects, the germplasm effects and FW coefficients were difficult to estimate. This is why they recommended that a large number of environments be used and that the germplasm be repeated sufficiently. We have therefore made a selection of the data and kept the environments with at least five germplasm and the germplasm that were present in at least four environments. Thus, the data analyzed comprised 70 to 76% of the initial data, depending on the trait.

The multi-environment data were very unbalanced, with most of the germplasm occurring in a limited number of environments (the median number of replicates across environments was seven, and about 20% of the germplasm were replicated in four environments only). For each trait, the number of observations was between 1300 and 2000 and the measures were spread over more than nine years (Tab. 3).

Trait	Observations	Germplasm	Environments	Disequilibrium	Farms	Years
Plant height	1437	124	117	90	44	11
Spike weight	1804	172	148	93	52	10
Protein	1332	144	111	92	44	9
TKW	1982	177	165	93	58	11

Table 3. Description of the dataset. Disequilibrium: proportion of missing values in the Germplasm x Environment table in %.

2.3 Models

The phenotypic value $Y_{ij} \in \mathbb{R}$ for a given trait Y, germplasm i and environment j was assumed to be equal to

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij},$$

where $(i, j) \in \mathcal{C}$, $\mathcal{C}$ was the set of the germplasm x environment combinations occurring in the data set, $\mu_{ij} \in \mathbb{R}$ was an expectation term, and $\varepsilon_{ij} \in \mathbb{R}$ was a residual term. Five models were developed, which modelled the expectation term, the residual term and the prior distribution differently (Tab. 4).

	Expectation term	Residual term	Prior distribution
ADHn	Additive	Normal	Hierarchical
ADHs	Additive	Student	Hierarchical
FWHn	Finlay Wilkinson	Normal	Hierarchical
FWHs	Finlay Wilkinson	Student	Hierarchical
FWs	Finlay Wilkinson	Student	Weakly informative

Table 4. The five models fitted.

2.3.1 Expectation term

In models ADHs and ADHn, the expectation term was modelled as additive effects of both the germplasm and the environment without interaction:

$$\mu_{ij} = \alpha_i + \theta_j,$$

where $\alpha_i \in \mathbb{R}$ was the main effect of germplasm i, and $\theta_j \in \mathbb{R}$ was the main effect of environment j. Models FWHs, FWs and FWHn modelled genotype-environment interaction using the Finlay-Wilkinson, also called joint-regression, model (Finlay and Wilkinson, 1963). In these models, the expectation term was assumed to be equal to

$$\mu_{ij} = \alpha_i + \theta_j + \eta_i\theta_j,$$

where $\eta_i \in \mathbb{R}$ was the sensitivity of germplasm i to environments (linear regression coefficient, Perkins and Jinks, 1968). Finlay and Wilkinson (1963) defined their coefficient as $b_i = 1 + \eta_i$. As the average sensitivity is equal to 0, a germplasm with $\eta_i > 0$ (resp. $\eta_i < 0$) is more (resp. less) sensitive to environments than a germplasm with the average sensitivity (Nabugoomu et al., 1999). In these models, the interaction between germplasm i and environment j was modelled as a multiplicative term $\eta_i\theta_j$ contributing to the expectation term with the remaining part adding to the residual term. The Finlay-Wilkinson coefficient is considered as both a static and a dynamic indicator of stability (Becker and Leon, 1988; Lin et al., 1986). In this model, statically stable genotypes have a coefficient close to -1. Dynamically stable genotypes have a coefficient close to zero, but having a coefficient close to zero is not sufficient to determine dynamic stability, this also depends on the amount of $G \times E$ variation that remains unexplained by the model.

2.3.2 Residual term

In models ADHn and FWHn, the distribution of the residual term was assumed to be normal:

$$\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right),$$

where $\mathcal{N}\left(0, \sigma^2_\varepsilon\right)$ was the normal distribution with expectation 0 and variance σ^2_ε. However, to limit the influence of extreme values on the results of the analyses, we also developed models based on Student's t distributions. Thus, in models FWHs, FWs and ADHs, the distribution of the error term was assumed to be equal to

$$\varepsilon_{ij} \sim t\left(0, \sigma^2_\varepsilon, \nu\right),$$

where $t\left(0, \sigma_\varepsilon^2, \nu\right)$ was the Student's t distribution with dispersion parameter $\sigma_\varepsilon^2 > 0$ and $\nu > 2$ degrees of freedom. We assumed that $\nu > 2$ to ensure that the expectation and the variance of ε_{ij} were defined and finite. In models FWHs, FWs and ADHs, the variance of ε_{ij} was equal to $\nu\sigma_\varepsilon^2/(\nu-2)$. The normal distribution can be considered as a t distribution with ν tending to $+\infty$. For additive models, the residual combined the $G \times E$ contribution and errors, i.e. experimental errors and environmental heterogeneity in each trial, while for FW models, it combined the part of $G \times E$ not explained by η and errors.

2.3.3 Prior distribution

The statistical analysis was carried out in a Bayesian framework, so that a joint prior distribution was placed on model parameters. We placed weakly informative priors on σ_ε and ν (Cao et al., 2022; Gelman, 2006; Juárez and Steel, 2010):

$$\sigma_\varepsilon \sim \mathcal{N}^+(0, \lambda_\varepsilon^2), \quad \nu \sim \Gamma(2, 0.1),$$

with $\nu > 2$, and where λ_ε was a known prior value of the standard deviation of the trait, $\mathcal{N}^+(0, \lambda_\varepsilon^2)$ was the normal distribution restricted to positive values with parameters 0 and λ_ε^2, and $\Gamma(2, 0.1)$ was the gamma distribution with shape parameter 2 and rate parameter 0.1.

Given the high data disequilibrium and the large numbers of germplasm and environments, we decided to implement a hierarchical Bayesian approach. In all the models except the FWs model, α_i, θ_j and when present η_i were assumed to follow hierarchical distributions:

$$\alpha_i \sim \mathcal{N}\left(\mu_Y, \sigma_\alpha^2\right), \quad \eta_i \sim \mathcal{N}\left(0, \sigma_\eta^2\right), \quad \theta_j \sim \mathcal{N}\left(0, \sigma_\theta^2\right),$$

where μ_Y, σ_α, σ_η and σ_θ were unknown parameters. Then, we placed weakly informative prior distributions on the hyperparameters μ_Y, σ_α, σ_η and σ_θ:

$$\mu_Y \sim \mathcal{N}\left(\lambda_\mu, \lambda_\varepsilon^2\right), \quad \sigma_\alpha \sim \mathcal{N}^+(0, \lambda_\varepsilon^2), \quad \sigma_\theta \sim \mathcal{N}^+(0, \lambda_\varepsilon^2), \quad \sigma_\eta \sim \mathcal{N}^+(0, 0.75^2),$$

where λ_μ was a known prior value of the trait mean. Germplasm main effects, environment main effects, germplasm sensitivities and residuals were assumed to be independent given the hyperparameters, σ_ε and ν. In model FWs, the hierarchical distributions of α_i, η_i and θ_j were replaced by weakly informative prior distributions:

$$\alpha_i \sim \mathcal{N}\left(\mu_Y, \lambda_\varepsilon^2\right), \quad \eta_i \sim \mathcal{N}\left(0, 0.75^2\right), \quad \theta_j \sim \mathcal{N}\left(0, \lambda_\varepsilon^2\right).$$

The values chosen for λ_ε and λ_μ are in Appendix A.1.

2.3.4 Posterior distribution

Bayesian inference is based on the posterior distribution of the model parameters. This distribution was estimated using Markov chain and Monte Carlo (MCMC) methods. These methods simulate the values of the model parameters according to a Markov chain that converges to the posterior distribution of these parameters (Robert, 2007). The MCMC methods were implemented using `R` (R Core Team, 2014) and the package `rstan` (Stan Developpement Team, 2016), that performs Hamiltonian Monte Carlo (HMC) sampling. This method aims at reducing the correlation between successive sampled values by using a proposal distribution based on Hamiltonian dynamics (Neal, 2011). Four MCMC chains were run independently to test for convergence. The initial values of each chain were taken randomly. For each chain, the burn-in consisted of 200 iterations, then 5,000 iterations were performed for all models, except FWs where 8,000 iterations were required. The average calculation time (for a given trait and a given model) was 9 minutes and the maximum time was 22 minutes (with FWs), with a computer *intel CORE i7*©. Estimates of the Gelman-Rubin statistic were smaller than 1.02 and the effective sample size was greater than 400 for each parameter in all tested models.

2.3.5 Model comparison

We compared the predictive ability of models using leave-one-out cross-validation, which seems more appropriate than Bayes factors for selecting models that approximate the process generating the data (Lartillot, 2023). We estimated the expected logarithmic predictive density using the R package LOO (Vehtari et al., 2017). This criterion was equal to

$$\text{elpd}_{\text{loo}} = \sum_{(i,j)\in\mathcal{C}} \ln(p(Y_{ij}|Y_{-ij})),$$

where Y_{-ij} was the dataset without observation Y_{ij}, and $p(Y_{ij}|Y_{-ij})$ was the leave-one-out posterior density of Y_{ij}. The larger this criterion, the better the agreement between the model and the data. The elpd_{loo} criterion was also used to identify extreme observations. The quantity $\ln(p(Y_{ij}|Y_{-ij}))$ can be understood as the contribution of observation Y_{ij} to elpd_{loo}. Observations with low contributions are unlikely and can be considered extreme observations.

For main effects and sensitivities, we estimated the average standard deviation of estimates, which allowed us to have an estimate of the precision of these effects. To be able to compare the precision between traits, for α and θ we estimated the average coefficient of variation by dividing this standard deviation by the general average μ_Y.

2.4 Data analysis

Model parameters were studied using the best model as determined by the methods described above.

2.4.1 Variance decomposition

In order to quantify the influence of model terms on observations, the variance of an observation was decomposed. Since α_i, θ_j, η_i and ϵ_{ij} were assumed to be conditionally independent, the variance of an observation given the hyperparameters, σ^2_ε and ν was equal to

$$Var(Y_{ij}) = Var(\alpha_i + \theta_j + \eta_i\theta_j + \varepsilon_{ij}) = \sigma^2_\alpha + \sigma^2_\theta + \sigma^2_\eta\sigma^2_\theta + Var(\varepsilon_{ij}).$$

As the best model involved the t distribution, $Var(\varepsilon_{ij})$ was equal to $\nu\sigma^2_\epsilon/(\nu-2)$. The proportions of variance explained by the germplasm main effect, the environment main effect and the interaction effect were equal to

$$\pi(\alpha) = \frac{\sigma^2_\alpha}{Var(Y_{ij})}, \quad \pi(\theta) = \frac{\sigma^2_\theta}{Var(Y_{ij})}, \quad \pi(\eta\theta) = \frac{\sigma^2_\eta\sigma^2_\theta}{Var(Y_{ij})}.$$

$\pi(\alpha)$ is also called broad-sense heritability. The proportion of variance explained by the model (coefficient of determination) was equal to

$$R^2 = \pi(\alpha) + \pi(\theta) + \pi(\eta\theta) = \frac{\sigma^2_\alpha + \sigma^2_\theta + \sigma^2_\eta\sigma^2_\theta}{Var(Y_{ij})}.$$

This definition of R^2 ensured that $R^2 \leq 1$ (Gelman et al., 2019). We also estimated the proportion of the variance of $G \times E$ interactions and experimental errors that was explained by the $\eta_i\theta_j$ term, defined by

$$\rho = \frac{Var(\eta_i\theta_j)}{Var(\eta_i\theta_j + \varepsilon_{ij})} = \frac{\sigma^2_\eta\sigma^2_\theta}{\sigma^2_\eta\sigma^2_\theta + Var(\varepsilon_{ij})}.$$

2.4.2 Characterization of germplasm

Germplasm main effects and sensitivities were estimated. In addition, we estimated two stability indicators, the static stability S^2_i (Becker and Leon, 1988) and the ecovalence W_i (Wricke, 1962) which is an indicator of dynamic stability. Static stability describes the response of a genotype that maintains a constant performance across environments, while dynamic stability describes the response of a genotype showing a constant difference with the average response of all genotypes tested in each environment (Annicchiarico, 2002). Due to

data imbalance, the empirical estimates of these indicators were biased. Thus, we defined stability indicators by means of theoretical variances from the Bayesian model described above (Cotes et al., 2006; Piepho, 1999). Using the independence assumptions of the model, we obtained for germplasm i,

$$\begin{aligned} W_i &= Var(\eta_i\theta_j + \varepsilon_{ij}) = \eta_i^2\sigma_\theta^2 + Var(\varepsilon_{ij}), \\ S_i^2 &= Var(\theta_j + \eta_i\theta_j + \varepsilon_{ij}) = (1+\eta_i)^2\sigma_\theta^2 + Var(\varepsilon_{ij}) = (1+2\eta_i)\sigma_\theta^2 + W_i. \end{aligned}$$

The larger these indicators, the less stable the germplasm. Becker (1981) applied the same decomposition with the empirical variances. These stability indicators are approximations of the static stability and ecovalance in a balanced framework.

Moreover, we tested whether the "type" of germplasm (cross, landrace, registered variety, mixture of germplasm and historic variety) had an influence on germplasm parameters (α_i, η_i, S_i^2 and W_i) by running a one-way ANOVA and Tukey–Kramer HSD test with germplasm type as factor.

3 Results

3.1 Model comparison

3.1.1 Predictive capacity of models

According to the elpd_{loo} criterion, the non-hierarchical FWs model was less predictive than the hierarchical FWHs model for all the traits (Fig. 1). Using the latter model shrank the estimates of η and sometimes α (Fig. 2). With the non-hierarchical model (FWs), some estimates (α_i and η_i) seemed to be unreliable, in particular some germplasm means were extreme and some FW coefficients were larger than 1 or smaller than -1.

The hierarchical models with a Student likelihood (FWHs, ADHs) were more predictive than the models with a normal distribution (FWHn, ADHn), all the more as ν was low (Tab. 5). For protein content, ν was equal to 20, so the t distribution was close to a normal distribution. The t distribution reduced the shrinkage of FW coefficients (Fig. 3). Moreover, t models better accounted for extreme data than normal models (Fig. 4). These extreme data mainly came from germplasm that were not replicated in the trials.

The Finlay-Wilkinson models (FWHs, FWHn) were slightly more predictive than the simple additive models (ADHs, ADHn), except for protein content, where the difference was not significant (Fig. 1). This difference was smaller than the differences due to the distribution of residuals and the hierarchization of parameters.

The elpd_{loo} criterion was estimated using Pareto smoothed importance sampling (Vehtari et al., 2017). This method tends to be less precise for models that do not fit the data well. Thus, as expected, estimates of elpd_{loo} were more reliable for the two hierarchical models with a t likelihood (FWHs and ADHs) than for the other models, in particular model FWs (Supplementary Tab. B.1).

3.1.2 Precision of estimates and distribution of residuals

For the models with a t distribution, the estimate of the number of degrees of freedom (ν) varied between 3.8 and 28.2 (close to a normal distribution) (Tab. 5). Thus, the shape of the distribution of residuals depended on the trait. This result confirmed that the number of extreme observations was not negligible in our data, and that models with a t distribution were more appropriate. In the latter case, the variation ranges of residuals were wider but with more residual values close to 0 for the t distribution than the normal distribution (Fig.3). Models had similar estimate precision, except for model FWs, which had less precise estimates. This result confirmed that a basic joint regression, i.e. non-hierarchical model, was not suited to our unbalanced data. Parameters α and θ were estimated more precisely (difference in coefficient of variation between -0.1 and 0.4, Tab. 5) for t models (ADHs and FWHs) than for normal models (ADHn and FWHn). This result was consistent with Fig. 4, where extreme data were better predicted by FWHs than by FWHn, except for protein content.

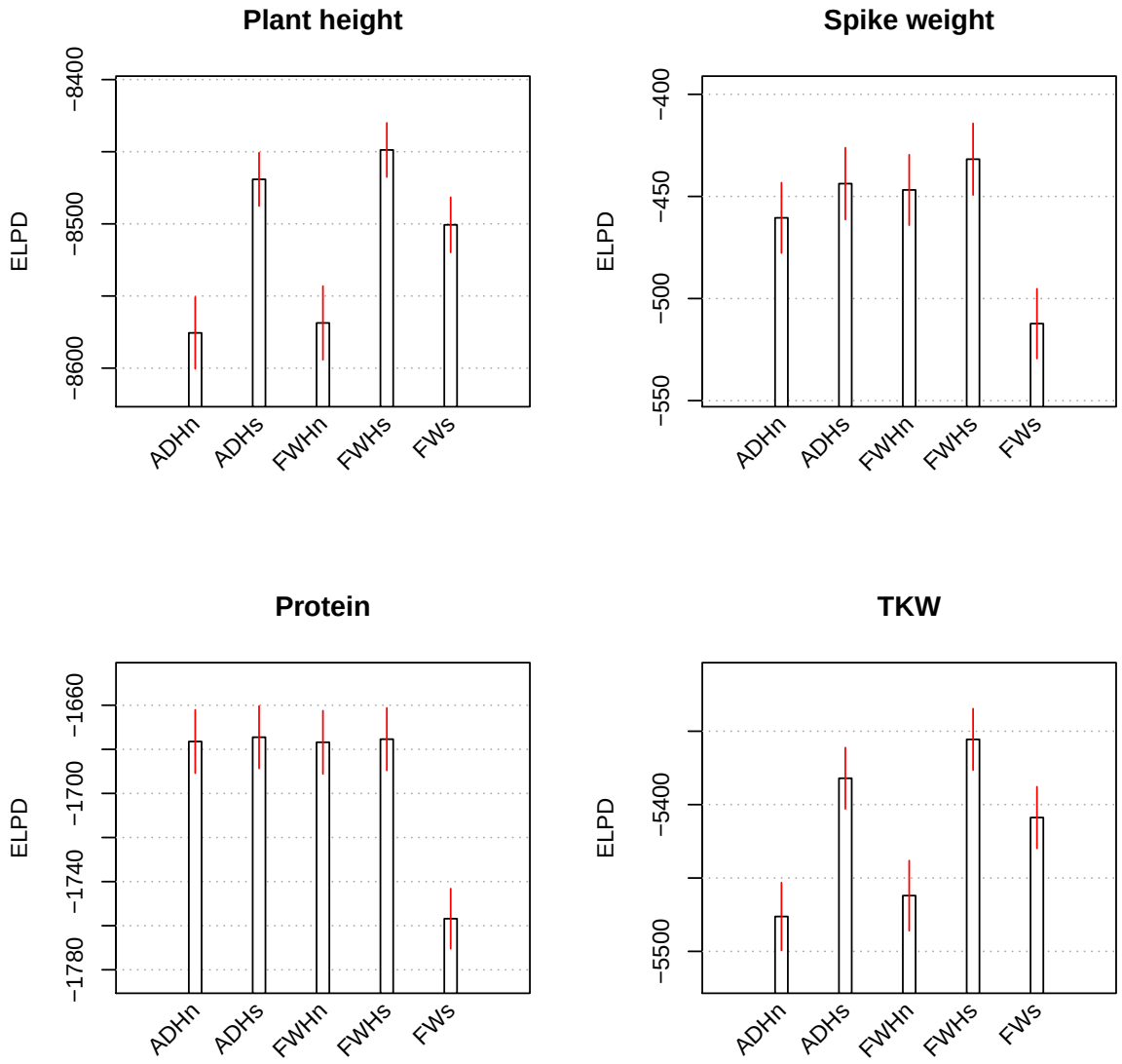

Figure 1. Predictive capacity of models. elpd_{loo} and its associated standard error for the four studied traits.

3.2 Data analysis

In the following, we used the FWHs model which proved to be the best model in terms of prediction and accuracy of the estimated parameters.

3.2.1 Variance decomposition

The proportion of variance explained by each term of the model depended on the trait (Tab. 6). For all four traits, the environment effect was highly explanatory. For height and TKW, a relatively large part of the total variance was explained by the germplasm effect (resp. 23.8% and 16.1%), whereas this part was much smaller for spike weight and protein content (10.9% and 5.6%). The proportion of variance explained by the sensitivity effect η was not significantly different from 0 for protein content and low for the three other traits. It explained 6.5%, 4.8% and 6.9% of the variance of $G \times E$ interactions and experimental errors (ρ parameter) for plant height, spike weight and TKW, respectively.

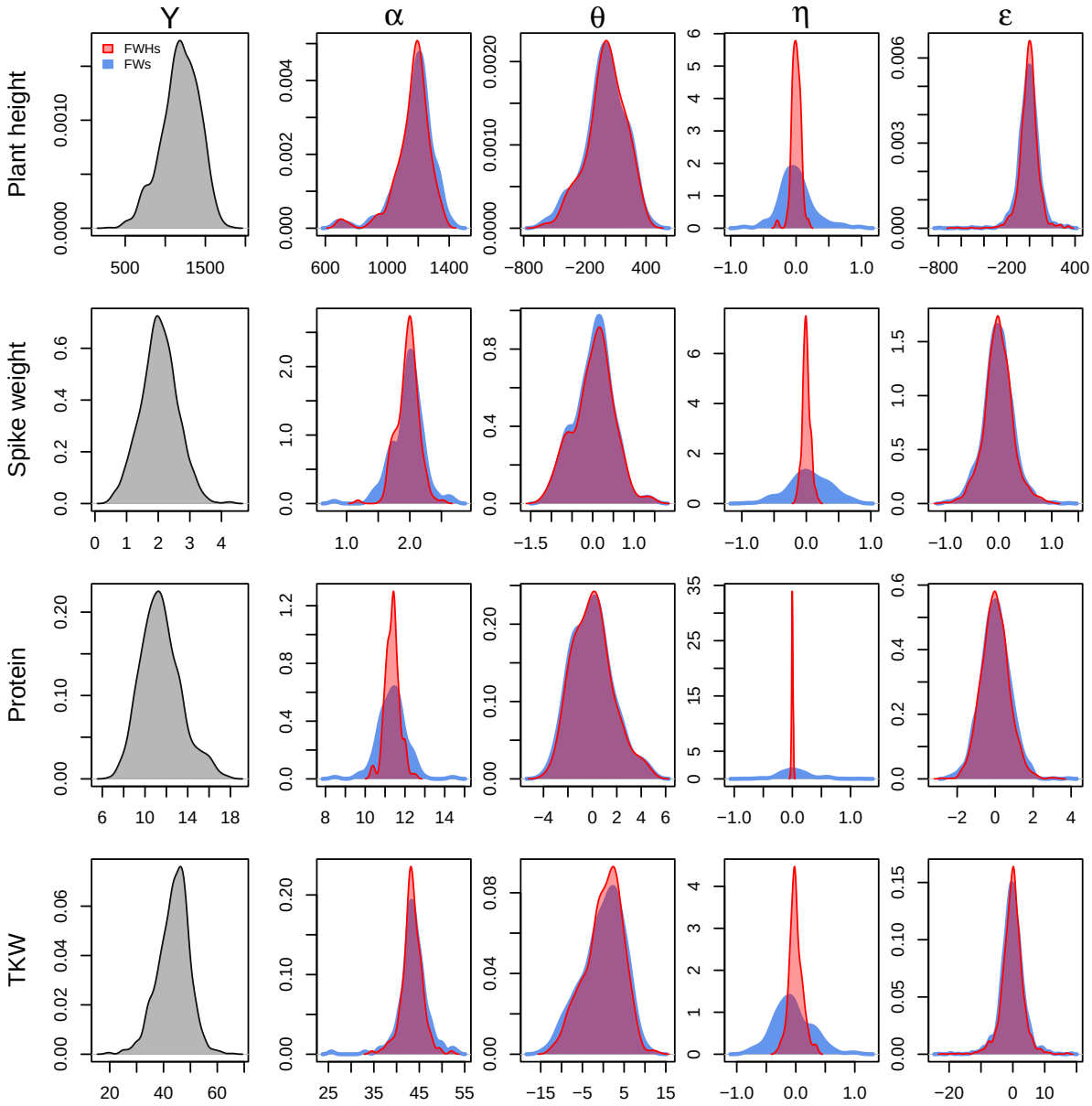

Figure 2. The first column presents the distribution of the trait to be explained (in grey). The last four columns compare the hierarchical (red) and non hierarchical (blue) versions of the FW model with a Student law for the residuals, and show the smoothed histograms of main effects and FW coefficients.

3.2.2 Characterization of germplasm

The correlation between germplasm sensitivity (η_i) and static stability (S_i^2) was very close to 1 for all traits while germplasm sensitivity was poorly correlated to W_i (Tab. 7). The main effect α_i had a low correlation with η_i and S_i^2, except for plant height and spike weight. Correlations between W_i and α_i were low and in most cases not significant.

Plant height was found to depend on the type of germplasm, landraces being taller than historic varieties, which were themselves taller than registered varieties. Registered varieties were significantly more stable (static stability and FW coefficient) than everything other than mixtures, but less stable dynamically (ecovalance). No germplasm parameters were significantly dependent on germplasm type for protein content and spike weight. TKW germplasm main effects did not depend on the type of germplasm, but for this trait landraces and mixtures appeared statically more stable than historic varieties.

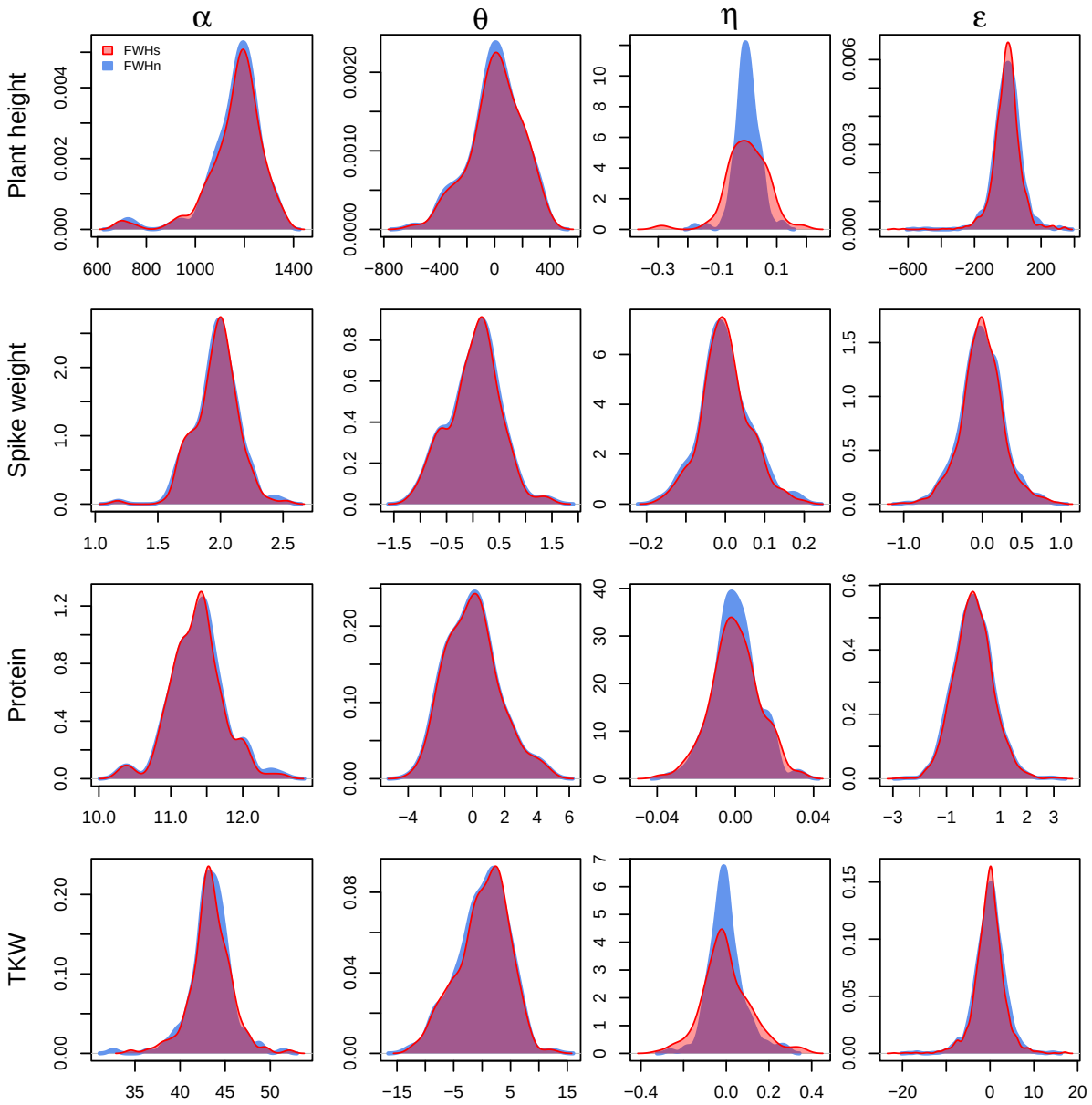

Figure 3. Comparison of hierarchical FW models with different residual laws, the Student (red) and the normal (blue). These graphics show the smoothed histograms of main effects and FW coefficients.

4 Discussion

To fit the characteristics of PPB trials, i.e., few inter-farm replicates and possible extreme data, we developed several models and we found that the hierarchical Finlay-Wilkinson model with t residuals was the best for prediction and parameter precision. Then we compared the performance and stability of different germplasm types.

4.1 Handling the data from a highly unbalanced series of trials

As the farmers of the program chose the germplasm they assessed, the data obtained from the series of trials were very unbalanced, with more than 90% of the $G \times E$ combinations missing. This made the estimation of germplasm main effects and sensitivities difficult. Although the Finlay-Wilkinson model was parsimonious, a basic joint regression with weakly-informative prior distributions (model FWs) was not able to cope with this level of disequilibrium. According to the elpd_{loo} criterion, model FWs was not the best model (Fig. 1). In

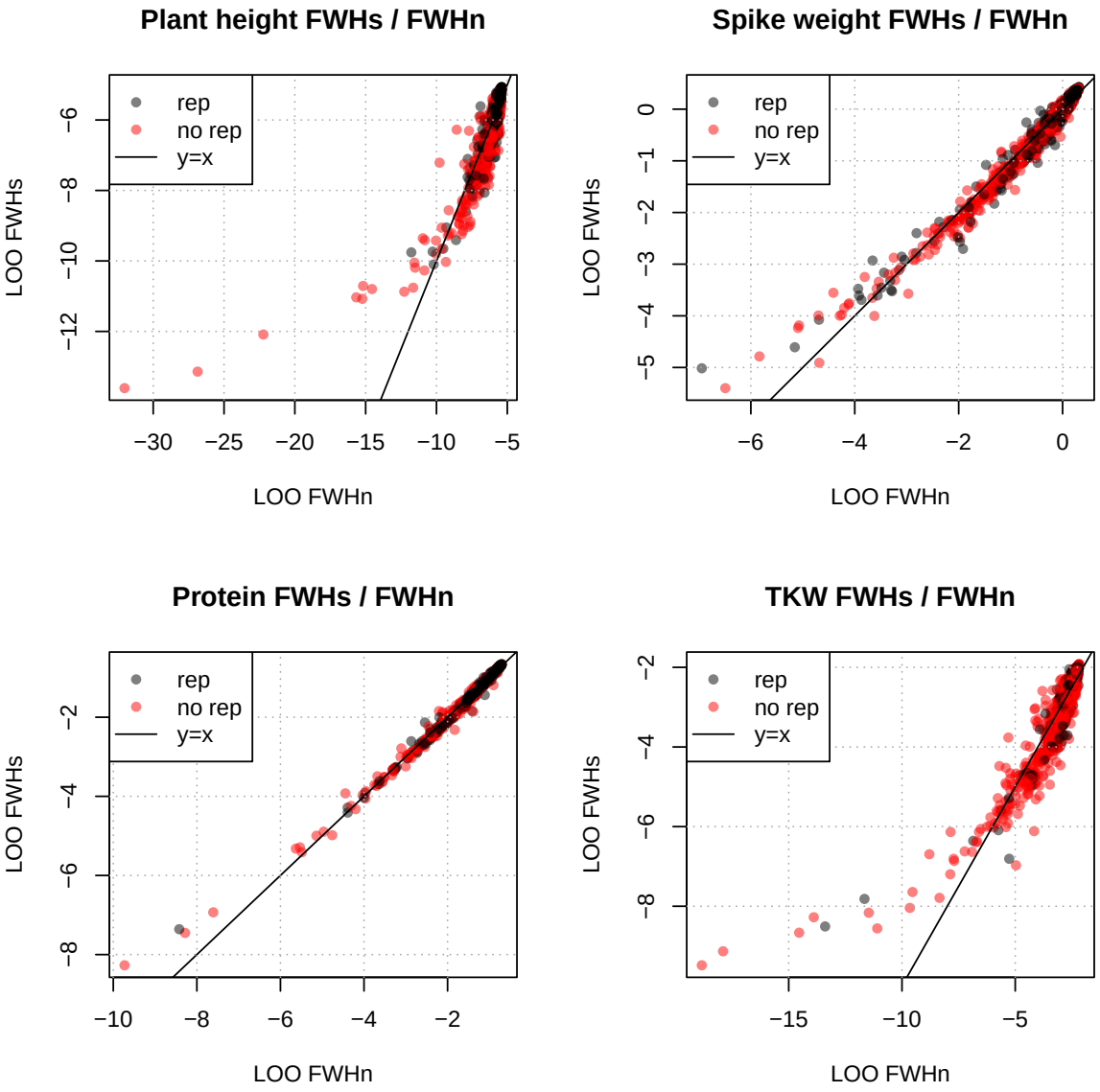

Figure 4. Comparison of t and normal models (FWHs vs FWHn) in terms of the contributions of observations to the elpd_{loo} criterion. Black (resp. red) dots correspond to observations that were measured on germplasm that were replicated (resp. not replicated) within trials.

addition, its estimates had poor precision and it led to extreme sensitivity estimates, with values close to 1 or -1 (Fig.2).

In contrast, hierarchical joint regression appeared more suited to our data structure. Model FWHs had the largest elpd_{loo} values for three traits out of four. Placing a hierarchical distribution on sensitivities constrained estimates and brought them closer to 0. This led to more satisfactory sensitivity estimates, since they were well below 1 in absolute value.

Three strategies have previously been used to manage incomplete $G \times E$ data: i) subset the total dataset to obtain an almost balanced subset for the analysis (Ceccarelli and Grando, 2007), ii) predict missing data with a more or less complex model and use these predictions in the analysis (Kumar et al., 2012; Woyann et al., 2017), and iii) use a model more robust to unbalanced data, provided it complies with model validation conditions (Assis et al., 2018; van Frank et al., 2019). We used the last strategy to maximise the amount of information from the data (less data excluded than in the first strategy) with a one-step process (unlike the second strategy).

Trait	Model	ν	cv(α)	cv(θ)	sd(η)
	ADHn		3.0	2.8	
	ADHs	3.8 (0.5)	2.8	2.6	
Plant height	FWHn		3.1	2.9	0.08
	FWHs	3.5 (0.4)	2.8	2.7	0.09
	FWs	3.3 (0.4)	4.7	4.5	0.20
	ADHn		5.3	5.2	
	ADHs	8.1 (2.2)	5.2	5.0	
Spike weight	FWHn		5.3	5.2	0.12
	FWHs	8 (2.2)	5.3	5.1	0.11
	FWs	10.2 (4.1)	6.8	5.7	0.29
	ADHn		2.7	2.7	
	ADHs	19.9 (9.6)	2.6	2.7	
Protein	FWHn		2.6	2.7	0.05
	FWHs	19.6 (9.6)	2.7	2.7	0.05
	FWs	28.2 (13.4)	4.8	4.1	0.25
	ADHn		2.8	2.8	
	ADHs	4.2 (0.5)	2.6	2.5	
TKW	FWHn		2.8	2.8	0.15
	FWHs	4 (0.5)	2.7	2.5	0.17
	FWs	3.8 (0.4)	3.5	3.2	0.33

Table 5. Number of degrees of freedom and precision of estimates.
ν: posterior means, with posterior standard deviations in parentheses, of the number of degrees of freedom of the t distribution; cv(α), cv(θ): average posterior coefficients of variation of germplasm and environment main effects; sd(η): average posterior standard deviation of germplasm sensitivities (FW coefficients).

	Plant height		Spike weight		Protein		TKW	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
R^2	87.1	82.9∼90.2	78.2	73.8∼82.1	83	79.1∼86.7	69.7	64.1∼74.8
$\pi(\alpha)$	23.8	17.9∼30.7	10.9	7.7∼14.7	5.6	3.7∼8.2	16.1	12∼20.8
$\pi(\theta)$	62.4	54.5∼69.8	66.2	60.1∼71.9	77.2	72∼82	51.4	44.6∼58.1
$\pi(\eta\theta)$	0.9	0.4∼1.6	1.1	0.4∼2.1	0.2	0∼0.8	2.2	1∼3.8
ρ	6.5	2.8∼12	4.8	1.7∼9.3	1.2	0∼4.7	6.9	3.1∼12

Table 6. Variance decomposition.
Estimates were made using model FWHs and are given in %. Mean: mean of the posterior distribution; 95% CI: 95% credible intervals. R^2 is the coefficient of determination. $\pi(\alpha)$, $\pi(\theta)$ and $\pi(\eta\theta)$ are respectively the proportion of variance explained by α, θ and $\eta\theta$. ρ is the proportion of the variance of $G \times E$ and errors explained by $\eta\theta$.

van Frank et al. (2019) investigated the influence of MET design on the evaluation of germplasm using model FWHn and simulations. They found that when data were highly unbalanced, this evaluation was more reliable with at least 100 environments and when some germplasm were replicated over at least 5 environments per germplasm. Therefore, we restricted the analysis to a subset of data that met these conditions. Having 100 environments seems difficult to achieve, but it should be noted that this number combines the number of trials per year and the number of years, which may be achieved by a network of many farmers experimenting over several years.

Cotes et al. (2006) used a Bayesian approach to estimate FW coefficients in a MET study in order to take prior information on germplasm coming from other studies into account. A similar approach was used by Couto et al. (2015), Foucteau and Denis (2001), and Nascimento et al. (2020) and was found to greatly improve the results. Here, we used little prior information. But in the future, previous evaluation studies may provide

Trait	Pearson correlation between					
	$\alpha_i \vert \eta_i$	$\alpha_i \vert S_i^2$	$\alpha_i \vert W_i$	$\eta_i \vert S_i^2$	$\eta_i \vert W_i$	$S_i^2 \vert W_i$
Plant height	0.43***	0.4***	-0.41	0.997***	-0.34	-0.26
Spike weight	0.34***	0.35***	0.24*	0.999***	0.17	0.21
Protein	0.13	0.14*	-0.01	0.999***	-0.01	0
TKW	0.23*	0.23*	0.12	0.996***	0.27	0.36

Table 7. Correlation between germplasm parameters.
*, **, *** : significant at $P = 0.05, P = 0.01, P = 0.001$ respectively.
α_i: germplasm effect, η_i: germplasm sensitivity (FW coefficient), S_i^2: static stability, W_i: ecovalence.

Trait		Registered		Historic		Landrace		Cross		Mixture		P-value
Plant height	α_i		862^{c}		1136^{b}		1221^{a}		1175^{ab}		1189^{ab}	**<0.001**
	η_i	n=6	-0.11^{b}	n=16	-0.01^{a}	n=21	0^{a}	n=74	0.01^{a}	n=7	-0.01^{ab}	**0.001**
	S_i^2		37566^{b}		44170^{a}		44616^{a}		45556^{a}		43911^{ab}	**0.005**
	W_i		8835^{a}		7680^{b}		7678^{b}		7716^{b}		7707^{b}	**<0.001**
Spike weight	α_i		2.03		1.99		1.94		1.96		2.02	0.439
	η_i	n=8	0.02	n=20	0.02	n=40	-0.01	n=89	0	n=15	-0.01	0.297
	S_i^2		0.34		0.34		0.32		0.33		0.32	0.303
	W_i		0.08		0.08		0.08		0.08		0.08	0.712
Protein	α_i		11.1		11.4		11.4		11.4		11.5	0.125
	η_i	n=9	0.001	n=27	0.001	n=34	-0.001	n=62	0	n=12	0.001	0.935
	S_i^2		3.41		3.41		3.39		3.4		3.41	0.951
	W_i		0.61		0.61		0.61		0.61		0.61	0.641
TKW	α_i		43.6		43.9		43.1		43.4		43.9	0.681
	η_i	n=8	-0.03^{ab}	n=20	0.08^{a}	n=42	-0.03^{b}	n=92	0.01^{ab}	n=15	-0.05^{b}	**0.003**
	S_i^2		32.9^{ab}		37.4^{a}		32.7^{b}		34.3^{ab}		31.7^{b}	**0.002**
	W_i		12.7		13		12.8		12.7		12.6	0.081

Table 8. Performance and stability of types of germplasm.
P-values in bold indicate significant differences (P<0.05) between germplasm types for a given trait and different letters indicate significant differences (P<0.05) based on Tukey–Kramer HSD tests.
n: number of germplasm, α_i: mean germplasm effect, η_i:mean sensitivity (FW coefficient), S_i^2: mean static stability and W_i: mean ecovalence.

stronger prior information on germplasm behaviour.

4.2 Extreme observations

Extreme observations were more frequent in our dataset than expected under the normal distribution for three traits out of four (Fig. 4). For these traits, using a t distribution increased elpd_{loo} values, and the estimate of the number of degrees of freedom of this distribution was smaller than 10 (Tab. 5). These extreme observations could occur in our dataset for several reasons: because most of the populations were not replicated within the trials, because cultivation environments were less controlled, or because a non-negligible part of the GxE interaction was not captured by the multiplicative term of the FW model. The normal distribution was appropriate for the trait protein content. It is difficult to explain why this trait had fewer extreme observations. A possible explanation could be that the measurement of protein content is more standardized than other trait measurements. For plant height, extreme values occurred only for non-replicated micro-plots with a global measurement and never with data from the average of 25 plants (Sect. 2.2.2), suggesting that this measurement is less accurate. For TKW, the kernel count could be affected by broken kernels due to over-drying or incorrect threshing settings leading to an overestimation of the number of kernels in the sample. Another possible explanation is that protein content is less variable under different conditions than plant height and

spike weight (Kazakou et al., 2014).

Using a t distribution did not affect the estimates of germplasm and environment main effects. On the contrary, it improved the estimates of sensitivities. It reduced their shrinkage and allowed the multiplicative term of the FW model to better capture $G \times E$ interactions (Fig.3).

The Student distribution is expected to take better account of extreme data and to yield more robust estimates (Besag and Higdon, 1999; Lange et al., 1989; Rosa et al., 2003). Extreme data are more likely to occur when varieties are not replicated within trials, which is frequent in this dataset (Fig. 4). Rosa et al. (2003) found that a normal likelihood underestimated a sex effect compared to a t likelihood. This effect was estimated less precisely with a normal distribution, which is consistent with our results for plant height, spike weight and TKW. A Student likelihood appears to be a good solution for dealing with extreme data, in particular in stability analyses, where extreme observations are sometimes removed (this is justified when they are extreme because of experimental errors, but not when they are due to natural variability). While this distribution has recently been used to implement robust alternatives to BLUP (Gianola et al., 2018) or to handle environmental heterogeneity in a single trial (Cao et al., 2022), to our knowledge, it has not already been used in MET studies.

4.3 Computing time

Series of trials often include many genotypes and environments, leading to large data sets. Thus, their analysis using mixed or hierarchical models is generally computationally demanding (Smith et al., 2005). The computational load can be reduced by using approximate estimation methods (Nabugoomu et al., 1999) or efficient algorithms, such as algorithms based on sparse matrix operations (Gilmour et al., 1995; Thompson et al., 2003). Hierarchical joint regression has already been implemented using Gibbs sampling or Jags (Lian and de los Campos, 2016; van Frank et al., 2020). Our implementation based on Hamiltonian Monte Carlo and Stan was more efficient since it required fewer iterations (Gelman, 2005). It allowed us to analyze large datasets in about 10 minutes.

To reduce computing time, the analyses were carried out in two steps. First, germplasm means were estimated using within-trial analyses. Then, these estimates were gathered and analyzed using a between-trial analysis. Thus, this two-stage approach analyzes $G \times E$ means without taking account of their standard error. Rivière et al. (2015a) developed a flexible method for estimating the experimental variance of trials with low intra-farm replication (farm design presented in Tab. 2). An easy way to integrate the variability estimated in the first stage would be for instance to use the same method as Couto et al. (2015).

4.4 Main effects

For the four traits studied, we found that the environmental part of the variance was large (from 51% for TKW to 77% for protein content, Tab. 6), which is consistent with the diversity of the cropping environments encountered (soil, climate, cropping practices...). Nevertheless, heritability was still significant with plant height $>$ TKW $>$ spike weight $>$ protein. Rivière et al. (2015b) found (with data included in our study) a similar ranking in heritability: plant height $>$ TKW $=$ protein $>$ spike weight. Plant height is known to be quite heritable due to a relatively simple genetic architecture with a few major genes, such as the well known Green Revolution Rht1 and Rht2 genes (Peng et al., 1999). In our study, the presence of both quite recent registered varieties and varieties dating from before the second World War, very likely led to varieties containing different alleles for these loci and increased variability for height. The decrease in plant height from landraces to historic varieties and registered varieties appears very clearly (Tab. 8) as also found in several studies (Bektas et al., 2016; Cantarel et al., 2021).

4.5 Germplasm stabilities

FW coefficients explained a low proportion of the total variance (between 0.2% and 2.2%) and a low proportion of the variance of $G \times E$ interactions and errors (between 1.2% and 6.9%, Tab. 6). We can presume that

the explanation of the interaction by the FW parameter is weaker the greater the number of environments for example, 29% for 12 studies with less than 10 environments, and 12% for studies with more than 10 environments (Brancourt-Hulmel et al., 1997). Other classical models, such as AMMI (additive main effect and multiplicative interaction) or GGE ($G + G \times E$) models, might explain a larger part of GxE interactions.

Missing data estimation methods allow these models to be used when the data is highly unbalanced, with up to from 40% unbalanced data for a MET with less than 20 environments to 60% unbalanced data for MET with at least 40 environments (Woyann et al., 2017; Yan, 2013). However, these datasets are more balanced than ours, and, as found by Rodrigues et al. (2011), FW is more robust than AMMI when the data are highly unbalanced (75%). In our study, most germplasm occurred in a limited number of environments, so that a parsimonious and very simple modelling of $G \times E$ interactions had to be used. An alternative approach would be to better characterize the environments and thus explain the environmental effects and part of the $G \times E$ interaction using environmental variables (Piepho, 2022).

Although sensitivities explained a rather low proportion of variance, HFWs model had larger elpd_{loo} values than additive models for three traits out of four. In addition, for these traits, some sensitivity estimates were not negligible, with values close to 0.2 or 0.3. Interaction effects then represented 20% or 30% of environmental effects. Additive models were appropriate for the protein content trait. It was found that the multiplicative term of the FW model was not significant for protein content, both in a balanced network of 15 environments in Serbia (Hristov et al., 2010) and in 12 environments in Swiss organic trials (Knapp et al., 2017). On the contrary, Mut et al. (2010) found significant FW coefficients for a balanced network of 7 environments in Turkey. These contrasting results could be explained by differences between numbers of environments or between genetic diversities.

For plant height, we found that registered varieties were more statically stable but less dynamically stable (Tab.8). This can be explained by the fact that there are only a few registered varieties in the trials, therefore they have little influence on the average height, which can fluctuate greatly between trials, and therefore the deviation from this average will be greater for this type.

Static and dynamic stabilities were difficult to estimate since our series of trials was very unbalanced. In particular, raw estimates of these stabilities were not reliable, since they were much influenced by the unbalanced nature of the data. By using theoretical variances, the FW model allowed us to calculate simple indicators of static and dynamic stability in the wheat PPB dataset. To our knowledge, the FW model has never been used for this purpose before.

Dependence between stability and mean is widespread (Reckling et al., 2021), but in our case, the correlation was low, which simplified interpretation of the stability analysis. Several studies for different traits and with balanced MET found a very strong correlation between FW coefficient and the static stability (Becker, 1981; Fasahat et al., 2015; Reckling et al., 2021). However, in our case, this relationship was even stronger (Tab. 7), probably because of the assumption that the variance of residuals did not depend on the genotype. As in many other studies, the residual variance was assumed to be independent of germplasm throughout our study. Allowing the residual variance to depend on the genotype could improve the estimates of stability indicators (Cotes et al., 2006; Couto et al., 2015). In particular, the dynamic indicator would be similar to the Shukla Stability Variance, i.e, the varietal variance of $G \times E$ interactions (Cotes et al., 2006). However, estimating a residual variance and a FW coefficient for each germplasm could be difficult in our study, as most of the germplasm appeared in only a few environments.

When relationships were significant, mixtures were always in a more stable (statically and dynamically) statistical group (Tab. 7). This result supports the fact that within-plot diversity stabilizes performances (Döring et al., 2015; Kiær et al., 2012).

In the wheat PPB program, the populations tested were heterogeneous and their genotypic composition could vary over years and farms (David et al., 2020). In this analysis, such variations were considered as part of the response of a population to a given environment for the sake of simplicity. Therefore the $G \times E$ interactions could be overestimated (resp. underestimated) if populations underwent diversifying (resp.

stabilizing) selection pressures within farms.

4.6 Role of statistical methods in the wheat PPB project

This work was developed following the co-construction of an experimental set-up suitable for decentralized on-farm evaluation and selection, and research into the best methods for analyzing the resulting data. It was part of the methodology we set up in a wheat PPB program, which was based on a collaboration between farmers, associations of farmers and researchers (Dawson et al., 2011). The farmers could freely choose the populations they wanted to test, so that a wide genetic diversity could be evaluated in a wide range of environments. There were on average more than 130 environments resulting from the combination of years and farms. The number of genotypes evaluated was large compared to other studies, but it was smaller than in CIMMYT's MET, which involved between 500 and 800 genotypes tested in 12 MET between 1945 and 1986 (Braun et al., 1997).

One aim of the project was to provide farmers with information to help them select new germplasm for testing on their farm. The statistical tools we developed sought to cope with the large degree to which this series of trials was unbalanced. Their objectives were the same as in other MET analyses : (i) estimate and predict germplasm' values for traits of interest for breeding, (ii) study the stability of germplasm over several environments, (iii) select new germplasm to be tested in new locations (Cotes et al., 2006). MET are usually carried out to find stable germplasm that perform well on average over many locations, or to detect special local adaptations to certain environments (Annicchiarico et al., 2005; Gauch et al., 2008). Here, while farmers were mostly interested in selecting the best germplasm adapted to their local pedo-climatic conditions, farming practices and marketing objectives, information retrieved from the farmers' network on new varieties to introduce in their trials could also be useful.

5 Conclusion

The proposed hierarchical model was aimed at improving the estimates of the parameters of the FW model from unbalanced datasets. This model was complex and was easier to implement in a Bayesian framework. Placing hierarchical distributions on model parameters and modelling residuals using a t distribution improved the estimates of main and interaction effects. This model allowed us to estimate static and dynamic stability indicators despite the high level of unbalanced data. Main effects and stability indicators provide information on the behaviour of genotypes in different environments, which farmers could use in their selection process.

Participatory research raises new research questions and contributes to the development of new methods for societal action (Kastenhofer et al., 2011). In PPB programs, all the methodology is based on collective and collaborative work and action between farmers, associations of farmers and researchers (Brac de la Perrière et al., 2011). New statistical methods can contribute to a better use of such complex multi-environment data in the selection process, and more generally to the effectiveness of participatory research (Martin and Sherington, 1997).

6 Acknowledgements

We thank Estelle Serpolet, Nathalie Galic and Sophie Pin for their great help in the measurements on participating farms and research stations. We thank all the farmers and facilitators participating in the project. We thank Gérard Branlard from INRAE Clermont-Ferrand for his time when carrying out NIRS analysis and Jean-Marc Le Goff for this feedback.

7 Fundings

M. Turbet Delof was funded by Program PPR-CPA MoBiDiv (2021–2026) under grant agreement ANR-20-PCPA-0006 and INRAE (program on organic scaling METABIO and *Biologie et Amélioration des Plantes* department).

8 Conflict of interest disclosure

The authors of this article declare that they have no financial conflict of interest with the content of this article. XXX and XXX are recommenders for PCI XXX.

9 Data, script and code availability

Data and script for models are available online: DOI 10.57745/SUTZ9U https://doi.org/10.57745/SUTZ9U .

Supplementary information

A Models

Tab. A.1 provides supplementary information on the prior distribution of model parameters.

	λ_μ	λ_ε	μ_{emp}	σ_{emp}
Plant height	1200	500	1188	234
Spike weight	2.00	0.80	2.03	0.58
Protein	12.0	4.0	11.5	1.9
TKW	45.0	10.0	43.7	5.8

Table A.1. Known values of the parameters of the prior distribution, empirical mean and standard deviation of traits.

B Model comparison

Tab. B.1 provides supplementary information on the estimation of the elpd_{loo} criterion. Fig. B.1 provides supplementary information on the comparison of models FWHs and ADHs. Fig. B.2 provides supplementary information on the comparison of models ADHs and ADHn.

References

Annicchiarico P (2002). *Genotype X Environment Interactions: Challenges and Opportunities for Plant Breeding and Cultivar Recommendations*. Food & Agriculture Org.

Annicchiarico P, F Bellah, and T Chiari (2005). Defining Subregions and Estimating Benefits for a Specific-Adaptation Strategy by Breeding Programs: A Case Study. *Crop Science* 45, 1741–1749. https://doi.org/10.2135/cropsci2004.0524.

Annicchiarico P, E Chiapparino, and M Perenzin (2010). Response of Common Wheat Varieties to Organic and Conventional Production Systems across Italian Locations, and Implications for Selection. *Field Crops Research* 116, 230–238.

Assis TOG de, CTdS Dias, and PC Rodrigues (2018). A Weighted AMMI Algorithm for Nonreplicated Data. *Pesquisa Agropecuária Brasileira* 53, 557–565.

Trait	Model	$k < 0.5$	$0.5 < k < 0.7$	$0.7 < k < 1$	$k > 1$
Spike weight	ADHn	1787	16	1	0
	ADHs	1804	0	0	0
	FWHn	1760	43	1	0
	FWHs	1803	1	0	0
	FWs	1658	129	16	1
Plant height	ADHn	1399	34	2	2
	ADHs	1437	0	0	0
	FWHn	1384	45	7	1
	FWHs	1436	1	0	0
	FWs	1405	22	8	2
TKW	ADHn	1960	21	1	0
	ADHs	1982	0	0	0
	FWHn	1918	56	8	0
	FWHs	1981	1	0	0
	FWs	1928	47	7	0
Protein	ADHn	1312	19	1	0
	ADHs	1331	1	0	0
	FWHn	1308	23	1	0
	FWHs	1331	1	0	0
	FWs	1099	164	67	2

Table B.1. Estimates of tail shape parameters (k) used to estimate elpd_{loo}. The contribution of each observation to elpd_{loo}, i.e., $\ln(p(Y_{ij}|Y_{-ij}))$, was estimated using Pareto smoothed importance sampling (Vehtari et al., 2017). For each observation, the largest importance weights of the importance sampling were smoothed using a generalized Pareto distribution with shape parameter k. Estimates of pointwise contributions with $k > 0.7$ are less reliable.

Becker HC (1981). Correlations among Some Statistical Measures of Phenotypic Stability. *Euphytica* 30, 835–840.

Becker HC and J Leon (1988). Stability Analysis in Plant Breeding. *Plant breeding* 101, 1–23.

Bektas H, CE Hohn, and JG Waines (2016). Root and Shoot Traits of Bread Wheat (Triticum Aestivum L.) Landraces and Cultivars. *Euphytica* 212, 297–311.

Besag J and D Higdon (1999). Bayesian Analysis of Agricultural Field Experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61, 691–746. https://doi.org/10.1111/1467-9868.00201.

Beza E, J Steinke, J Van Etten, P Reidsma, C Fadda, S Mittra, P Mathur, and L Kooistra (2017). What are the prospects for citizen science in agriculture? Evidence from three continents on motivation and mobile telephone use of resource-poor farmers. *PloS one* 12, e0175700.

Brac de la Perrière RA, P De Kochko, C Neubauer, and B Storup (2011). Visions Paysannes de La Recherche Dans Le Contexte de La Sélection Participative.

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

Braun HJ, S Rajaram, and M Ginkel (1997). CIMMYT's Approach to Breeding for Wide Adaptation. In: *Adaptation in Plant Breeding: Selected Papers from the XIV EUCARPIA Congress on Adaptation in Plant Breeding Held at Jyväskylä, Sweden from July 31 to August 4, 1995*. Ed. by Tigerstedt PMA. Developments in Plant Breeding. Dordrecht: Springer Netherlands, pp. 197–205. https://doi.org/10.1007/978-94-015-8806-5_25.

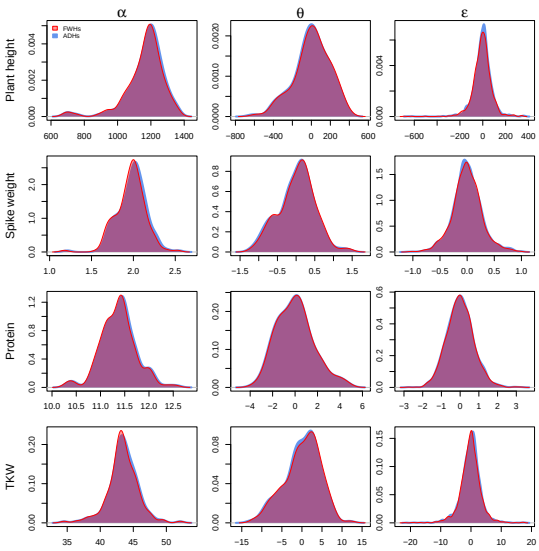

Figure B.1. Comparison of models FWHs and ADHs for the distribution of germplasm main effects (α), environment main effects (θ), FW coefficients (η) and residuals (ε) for each trait. Red: model FWHs; Blue: model ADHs.

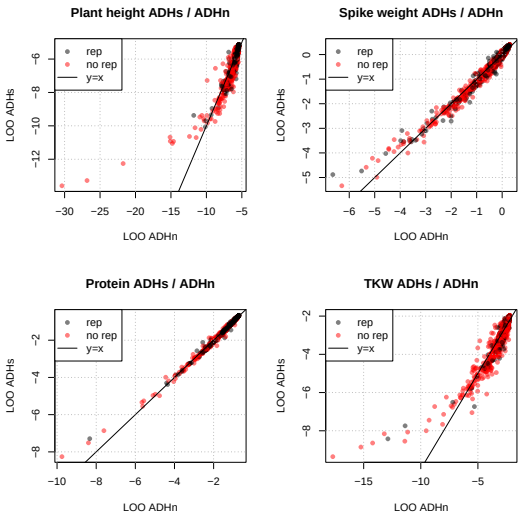

Figure B.2. Comparison of models ADHs and ADHn in terms of the contributions of observations to the elpd_{loo} criterion. Black (resp. red) dots correspond to observations that were measured on germplasm that were repeated (resp. not repeated) within trials

Cantarel AA, V Allard, B Andrieu, S Barot, J Enjalbert, J Gervaix, I Goldringer, T Pommier, S Saint-Jean, and X Le Roux (2021). Plant Functional Trait Variability and Trait Syndromes among Wheat Varieties: The Footprint of Artificial Selection. *Journal of Experimental Botany* 72, 1166–1180.

Cao Z, K Stefanova, M Gibberd, and S Rakshit (2022). Bayesian Inference of Spatially Correlated Random Parameters for On-Farm Experiment. *Field Crops Research* 281, 108477.

Carlin BP and NG Polson (1991). Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler. *Canadian Journal of statistics* 19, 399–405.

Carlin BP and TA Louis (2008). *Bayesian Methods for Data Analysis*. CRC Press.

Ceccarelli S, S Grando, E Bailey, A Amri, M El-Felah, F Nassif, S Rezgui, and A Yahyaoui (2001). Farmer Participation in Barley Breeding in Syria, Morocco and Tunisia. *Euphytica* 122, 521–536. https://doi.org/10.1023/A:1017570702689.

Ceccarelli S and S Grando (2007). Decentralized-Participatory Plant Breeding: An Example of Demand Driven Research. *Euphytica* 155, 349–360. https://doi.org/10.1007/s10681-006-9336-8.

– (2020). Participatory Plant Breeding: Who Did It, Who Does It and Where? *Experimental Agriculture* 56, 1–11.

Choy BST and JSK Chan (2008). Scale Mixtures Distributions in Statistical Modelling. *Australian & New Zealand Journal of Statistics* 50, 135–146.

Cotes JM, J Crossa, A Sanches, and PL Cornelius (2006). A Bayesian Approach for Assessing the Stability of Genotypes. *Crop Science* 46, 2654–2665. https://doi.org/10.2135/cropsci2006.04.0227.

Couto MF, M Nascimento, AT do Amaral Jr, FF e Silva, AP Viana, and M Vivas (2015). Eberhart and Russel's Bayesian Method in the Selection of Popcorn Cultivars. *Crop Science* 55, 571–577.

David O, G van Frank, I Goldringer, P Rivière, and M Turbet Delof (2020). Bayesian Inference of Natural Selection from Spatiotemporal Phenotypic Data. *Theoretical Population Biology* 131, 100–109. https://doi.org/10.1016/j.tpb.2019.11.007.

Dawson J, KM Murphy, and SS Jones (2008). Decentralized Selection and Participatory Approaches in Plant Breeding for Low-Input Systems. *Euphytica* 160, 143–154. https://doi.org/10.1007/s10681-007-9533-0.

Dawson JC, P Rivière, JF Berthellot, F Mercier, P de Kochko, N Galic, S Pin, E Serpolay, M Thomas, S Giuliano, and I Goldringer (2011). Collaborative Plant Breeding for Organic Agricultural Systems in Developed Countries. *Sustainability* 3, 1206–1223. https://doi.org/10.3390/su3081206.

Desclaux D, JM Nolot, Y Chiffoleau, E Gozé, and C Leclerc (2008). Changes in the Concept of Genotype × Environment Interactions to Fit Agriculture Diversification and Decentralized Participatory Plant Breeding: Pluridisciplinary Point of View. *Euphytica* 163, 533–546. https://doi.org/10.1007/s10681-008-9717-2.

Döring TF, P Annicchiarico, S Clarke, Z Haigh, HE Jones, H Pearce, J Snape, J Zhan, and MS Wolfe (2015). Comparative Analysis of Performance and Stability among Composite Cross Populations, Variety Mixtures and Pure Lines of Winter Wheat in Organic and Conventional Cropping Systems. *Field Crops Research* 183, 235–245.

Falconer DS (1960). *Introduction to Quantitative Genetics.* Edinburgh/London: Olivier & Boyd.

Fasahat P, A Rajabi, SB Mahmoudi, MA Noghabi, and JM Rad (2015). An Overview on the Use of Stability Parameters in Plant Breeding. *Biometrics & Biostatistics International Journal* 2, 00043.

Finlay KW and GN Wilkinson (1963). The Analysis of Adaptation in a Plant-Breeding Programme. *Australian Journal of Agricultural Research* 14, 742–754. https://doi.org/10.1071/AR9630742.

Foucteau V and JB Denis (2001). Statistical Analysis of Successive Series of Experiments in Plant Breeding: A Bayesian Approach. *Quantitative genetics and breeding methods: the way ahead. Proceedings of the Eleventh Meeting of the EUCARPIA Section Biometrics in Plant Breeding, Paris, France, 30/31 August - 1 September, 2000*, 49–56.

Gauch HG, HP Piepho, and P Annicchiarico (2008). Statistical Analysis of Yield Trials by AMMI and GGE: Further Considerations. *Crop Science* 48, 866–889. https://doi.org/10.2135/cropsci2007.09.0513.

Gelman A (2005). Analysis of Variance—Why It Is More Important than Ever.

– (2006). Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper). *Bayesian Analysis* 1, 515–534. https://doi.org/10.1214/06-BA117A.

Gelman A, B Goodrich, J Gabry, and A Vehtari (2019). R-Squared for Bayesian Regression Models. *The American Statistician*.

Gianola D, A Cecchinato, H Naya, and CC Schön (2018). Prediction of complex traits: robust alternatives to best linear unbiased prediction. *Frontiers in genetics* 9, 195.

Gilmour AR, R Thompson, and BR Cullis (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, 1440–1450.

Goldringer I, G van Frank, C Bouvier d'Yvoire, E Forst, N Galic, M Garnault, J Locqueville, S Pin, J Bailly, R Baltassat, JF Berthellot, F Caizergues, C Dalmasso, P de Kochko, JS Gascuel, A Hyacinthe, J Lacanette, F Mercier, H Montaz, B Ronot, and P Rivière (2020). Agronomic Evaluation of Bread Wheat Varieties from Participatory Breeding: A Combination of Performance and Robustness. *Sustainability* 12, 128. https://doi.org/10.3390/su12010128.

Hampel FR, EM Ronchetti, PJ Rousseeuw, and WA Stahel (2011). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

Hristov N, N Mladenov, V Djuric, A Kondic-Spika, A Marjanovic-Jeromela, and D Simic (2010). Genotype by Environment Interactions in Wheat Quality Breeding Programs in Southeast Europe. *Euphytica* 174, 315–324.

Huber PJ and EM Ronchetti (1981). *Robust Statistics, Wiley: New York*. USA.

Juárez MA and MFJ Steel (2010). Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions. *Journal of Business & Economic Statistics* 28, 52–66. https://doi.org/10.1198/jbes.2009.07145.

Kastenhofer K, U Bechtold, and H Wilfing (2011). Sustaining Sustainability Science: The Role of Established Inter-Disciplines. *Ecological Economics* 70, 835–843. https://doi.org/10.1016/j.ecolecon.2010.12.008.

Kazakou E, C Violle, C Roumet, ML Navas, D Vile, J Kattge, and E Garnier (2014). Are Trait-based Species Rankings Consistent across Data Sets and Spatial Scales? *Journal of Vegetation Science* 25, 235–247.

Kiær LP, IM Skovgaard, and H Østergård (2012). Effects of Inter-Varietal Diversity, Biotic Stresses and Environmental Productivity on Grain Yield of Spring Barley Variety Mixtures. *Euphytica* 185, 123–138.

Knapp S, C Brabant, M Oberforster, H Grausgruber, and J Hiltbrunner (2017). Quality Traits in Winter Wheat: Comparison of Stability Parameters and Correlations between Traits Regarding Their Stability. *Journal of Cereal Science* 77, 186–193.

Kumar A, SB Verulkar, NP Mandal, M Variar, VD Shukla, JL Dwivedi, BN Singh, ON Singh, P Swain, and AK Mall (2012). High-Yielding, Drought-Tolerant, Stable Rice Genotypes for the Shallow Rainfed Lowland Drought-Prone Ecosystem. *Field crops research* 133, 37–47.

Lange KL, RJ Little, and JM Taylor (1989). Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association* 84, 881–896.

Lartillot N (2023). Identifying the best approximating model in Bayesian phylogenetics: Bayes factors, cross-validation or wAIC? *Systematic Biology* 72, 616–638.

Lian L and G de los Campos (2016). FW: An R Package for Finlay–Wilkinson Regression That Incorporates Genomic/Pedigree Information and Covariance Structures Between Environments. *G3 Genes|Genomes|Genetics* 6, 589–597. https://doi.org/10.1534/g3.115.026328.

Lin CS, MR Binns, and LP Lefkovitch (1986). Stability Analysis: Where Do We Stand? *Crop science* 26, 894–900.

Martin A and J Sherington (1997). Participatory Research Methods—Implementation, Effectiveness and Institutional Context. *Agricultural systems* 55, 195–216.

Mohammadi R, KN Mahmoodi, R Haghparast, S Grando, M Rahmanian, and S Ceccarelli (2011). Identifying Superior Rainfed Barley Genotypes in Farmers' Fields Using Participatory Varietal Selection. *Journal of Crop Science and Biotechnology* 14, 281–288.

Murphy KM, KG Campbell, SR Lyon, and SS Jones (2007). Evidence of Varietal Adaptation to Organic Farming Systems. *Field Crops Research* 102, 172–177. https://doi.org/10.1016/j.fcr.2007.03.011.

Mut Z, N Aydin, H Orhan Bayramoglu, and H Ozcan (2010). Stability of Some Quality Traits in Bread Wheat (Triticum Aestivum) Genotypes. *Journal of Environmental Biology* 31, 489.

Nabugoomu F, RA Kempton, and M Talbot (1999). Analysis of Series of Trials Where Varieties Differ in Sensitivity to Locations. *Journal of Agricultural, Biological, and Environmental Statistics* 4, 310–325. https://doi.org/10.2307/1400388. JSTOR: 1400388.

Nascimento M, ACC Nascimento, FF e Silva, PE Teodoro, CF Azevedo, TRA de Oliveira, AT do Amaral Junior, CD Cruz, FJC Farias, and LP de Carvalho (2020). Bayesian Segmented Regression Model for Adaptability and Stability Evaluation of Cotton Genotypes. *Euphytica* 216, 30. https://doi.org/10.1007/s10681-020-2564-5.

Neal RM (2011). MCMC Using Hamiltonian Dynamics. In: *Handbook of Markov Chain Monte Carlo*. Ed. by Brooks S, Gelman A, Jones G, and Meng XL. Chapman and Hall/CRC, pp. 113–162.

Ng M and E Williams (2001). Joint-Regression Analysis for Incomplete Two-way Tables. *Australian & New Zealand Journal of Statistics* 43, 201–206. https://doi.org/10.1111/1467-842X.00165.

Patterson HD and V Silvey (1980). Statutory and Recommended List Trials of Crop Varieties in the United Kingdom. *Journal of the Royal Statistical Society: Series A (General)* 143, 219–240.

Peng J, DE Richards, NM Hartley, GP Murphy, KM Devos, JE Flintham, J Beales, LJ Fish, AJ Worland, and F Pelica (1999). 'Green Revolution'Genes Encode Mutant Gibberellin Response Modulators. *Nature* 400, 256–261.

Pereira DG, JT Mexia, and PC Rodrigues (2007). Robustness of Joint Regression Analysis. *Biometrical Letters* 44, 105–128.

Perkins JM and JL Jinks (1968). Environmental and Genotype-Environmental Components of Variability. *Heredity* 23, 339–356.

Piepho HP (1999). Stability Analysis Using the SAS System. *Agronomy Journal* 91, 154–160.

– (2022). Extending Finlay-Wilkinson Regression with Covariates. *bioRxiv*, 2022.12. 14.520390.

R Core Team (2014). R: A Language and Environment for Statistical Computing.

Reckling M, H Ahrends, TW Chen, W Eugster, S Hadasch, S Knapp, F Laidig, A Linstädter, J Macholdt, and HP Piepho (2021). Methods of Yield Stability Analysis in Long-Term Field Experiments. A Review. *Agronomy for Sustainable Development* 41, 1–28.

Rivière P, JC Dawson, I Goldringer, and O David (2015a). Hierarchical Bayesian Modeling for Flexible Experiments in Decentralized Participatory Plant Breeding. *Crop Science* 55, 1053–1067. https://doi.org/10.2135/cropsci2014.07.0497.

Rivière P, I Goldringer, JF Berthellot, N Galic, S Pin, P De Kochko, and JC Dawson (2015b). Response to Farmer Mass Selection in Early Generation Progeny of Bread Wheat Landrace Crosses. *Renewable Agriculture and Food Systems* 30, 190–201. https://doi.org/10.1017/S1742170513000343.

Robert C (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Science & Business Media.

Rodrigues PC, DGS Pereira, and JT Mexia (2011). A Comparison between Joint Regression Analysis and the Additive Main and Multiplicative Interaction Model: The Robustness with Increasing Amounts of Missing Data. *Scientia Agricola* 68, 679–686.

Rosa , CR Padovani, and D Gianola (2003). Robust Linear Mixed Models with Normal/Independent Distributions and Bayesian MCMC Implementation. *Biometrical Journal* 45, 573–590.

Smith , BR Cullis, and R Thompson (2005). The Analysis of Crop Cultivar Breeding and Evaluation Trials: An Overview of Current Mixed Model Approaches. *The Journal of Agricultural Science* 143, 449–462.

Smith ME, FG Castillo, and F Gómez (2001). Participatory Plant Breeding with Maize in Mexico and Honduras. *Euphytica* 122, 551–563. https://doi.org/10.1023/A:1017510529440.

Snapp SS and SN Silim (2002). Farmer Preferences and Legume Intensification for Low Nutrient Environments. *Plant and soil* 245, 181–192.

Stan Developpement Team (2016). RStan: The R Interface to Stan.

Thompson R, B Cullis, A Smith, and A Gilmour (2003). A Sparse Implementation of the Average Information Algorithm for Factor Analytic and Reduced Rank Variance Models. *Australian & New Zealand Journal of Statistics* 45, 445–459.

van Etten J, K de Sousa, A Aguilar, M Barrios, A Coto, M Dell'Acqua, C Fadda, Y Gebrehawaryat, J van de Gevel, A Gupta, et al. (2019). Crop variety management for climate adaptation supported by citizen science. *Proceedings of the National Academy of Sciences* 116, 4194–4199.

van Frank G, I Goldringer, P Rivière, and O David (2019). Influence of Experimental Design on Decentralized, on-Farm Evaluation of Populations: A Simulation Study. *Euphytica* 215, 126. https://doi.org/10.1007/s10681-019-2447-9.

van Frank G, P Rivière, S Pin, R Baltassat, JF Berthellot, F Caizergues, C Dalmasso, JS Gascuel, A Hyacinthe, F Mercier, H Montaz, B Ronot, and I Goldringer (2020). Genetic Diversity and Stability of Performance of Wheat Population Varieties Developed by Participatory Breeding. *Sustainability* 12, 384. https://doi.org/10.3390/su12010384.

Vehtari A, A Gelman, and J Gabry (2017). Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing* 27, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

Virk DS, M Chakraborty, J Ghosh, SC Prasad, and JR Witcombe (2005). Increasing the Client Orientation of Maize Breeding Using Farmer Participation in Eastern India. *Experimental Agriculture* 41, 413–426. https://doi.org/10.1017/S001447970500270X.

Wolfe MS, JP Baresel, D Desclaux, I Goldringer, S Hoad, G Kovacs, F Löschenberger, T Miedaner, H Østergård, and ET Lammerts van Bueren (2008). Developments in Breeding Cereals for Organic Agriculture. *Euphytica* 163, 323. https://doi.org/10.1007/s10681-008-9690-9.

Woyann LG, G Benin, L Storck, DM Trevizan, C Meneguzzi, VS Marchioro, M Tonatto, and A Madureira (2017). Estimation of Missing Values Affects Important Aspects of GGE Biplot Analysis. *Crop Science* 57, 40–52.

Wricke G (1962). Uber Eine Methode Zur Erfassung Der Okologischen Streubreite in Feldverzuchen. *Z. Pflanzenzuchtg* 47, 92–96.

Yan W (2013). Biplot Analysis of Incomplete Two-way Data. *Crop Science* 53, 48–57.