Peer Community In Mathematical & Computational Biology

Handling Data Imbalance and G×E Interactions in On-Farm Trials Using Bayesian Hierarchical Models

Sophie Donnet based on peer reviews by **David Makowski** and **Pierre Druilhet**

Michel Turbet Delof , Pierre Rivière , Julie C Dawson, Arnaud Gauffreteau , Isabelle Goldringer , Gaëlle van Frank , Olivier David (2024) Bayesian joint-regression analysis of unbalanced series of on-farm trials. HAL, ver. 2, peer-reviewed and recommended by Peer Community in Mathematical and Computational Biology.
https://hal.science/hal-04380787

Submitted: 11 January 2024, Recommended: 08 November 2024

Cite this recommendation as:
Donnet, S. (2024) Handling Data Imbalance and G×E Interactions in On-Farm Trials Using Bayesian Hierarchical Models. *Peer Community in Mathematical and Computational Biology*, 100272. 10.24072/pci.mcb.100272

Published: 08 November 2024

The article, "Bayesian Joint-Regression Analysis of Unbalanced Series of On-Farm Trials," presents a Bayesian statistical framework tailored for analyzing highly unbalanced datasets from participatory plant breeding (PPB) trials, specifically wheat trials. The key goal of this research is to address the challenges of genotype-environment (G×E) interactions in on-farm trials, which often have limited replication and varied testing conditions across farms.

The study applies a hierarchical Bayesian model with Finlay-Wilkinson regression, which improves the estimation of G×E effects despite substantial data imbalance. By incorporating a Student's t-distribution for residuals, the model is more robust to extreme values, which are common in on-farm trials due to variable environments. Note that the model allows a detailed breakdown of variance, identifying environment effects as the most significant contributors, thus highlighting areas for future breeding focus. Using Hamiltonian Monte Carlo methods, the study achieves reasonable computation times, even for large datasets.

Obviously, the limitation of the methods comes from the level of data balance and replication. The method requires a minimum level of data balance and replication, which can be a challenge in very decentralized breeding networks Moreover, the Bayesian framework, though computationally feasible, may still be complex for widespread adoption without computational resources or statistical expertise.

The paper presents a sophisticated Bayesian framework specifically designed to tackle the challenges of unbalanced data in participatory plant breeding (PPB). It showcases a novel way to manage the variability in on-farm trials, a common issue in decentralized breeding programs.

This study's methods accommodate the inconsistencies inherent in on-farm trials, such as extreme values and minimal replication. By using a hierarchical Bayesian approach with a Student's t-distribution for robustness, it provides a model that maintains precision despite these real-world challenges. This makes it especially relevant for those working in unpredictable agricultural settings or decentralized trials. From a more general perspective, this paper's findings support breeding methods that prioritize specific adaptation to local conditions. It is particularly useful for researchers and practitioners interested in breeding for agroecological or organic farming systems, where G×E interactions are critical but hard to capture in traditional trial setups.

Beyond agriculture, the paper serves as an excellent example of advanced statistical modeling in highly variable datasets. Its applications extend to any field where data is incomplete or irregular, offering a clear case for hierarchical Bayesian methods to achieve reliable results.

Finally, although begin quite methodological, the paper provides practical insights into how breeders and researchers can work with farmers to achieve meaningful varietal evaluations.

References:

Michel Turbet Delof , Pierre Rivière , Julie C Dawson, Arnaud Gauffreteau , Isabelle Goldringer , Gaëlle van Frank , Olivier David (2024) Bayesian joint-regression analysis of unbalanced series of on-farm trials. HAL, ver.2 peer-reviewed and recommended by PCI Math Comp Biol https://hal.science/hal-04380787

Reviews

Evaluation round #2

Reviewed by Pierre Druilhet, 04 November 2024

The revised version of the document clearly and satisfactorily addresses the points raised in the first submission (I agree with the variance calculation). I therefore have no objection to its publication.

Reviewed by David Makowski, 17 October 2024

The authors have satisfactorily addressed my comments.

Evaluation round #1

DOI or URL of the preprint: https://hal.science/hal-04380787
Version of the preprint: 1

Authors' reply, 03 October 2024

Download author's reply

Decision by Sophie Donnet, posted 20 June 2024, validated 21 June 2024

Revision

This paper aims to show that advanced modeling and Bayesian inference can be used to estimate the joint effects of environment and genotype from participatory plant breeding data. These programs, which involve volunteer farmers, make it possible to acquire a large volume of data, but according to unbalanced and uncontrolled experimental designs. Among other things, the authors propose to move away from the linear

Gaussian model by using residuals distributed according to a Student's law.
The two reviewers underline the interest of the subject and the quality of the work, but raise a number of questions that may require considerable work.

Consequently, I consider this preprint worthy of revision before recommendation.

Reviewed by Pierre Druilhet, 21 May 2024

The authors compare several models to estimate main effects and stability of genotype effects. Data are collected over a 12-year period and on several OA farms. This is a non-randomized trial since the farmer chose the germoplasms to be tested in they own farm. A control gerplams is common to all the farm.

Overall, the article is well-written and interesting. Here are a few comments:

1) l.160. It is not clear to the reader which variable is used as the blocking variable: the farm or some spatial or temporal variables or other blocking variables.

2) l. 216 : Is there any point in using a truncated normal distribution for ϵ rather than a usual Gamma or inverse-Gamma distribution?

3) l.216-217 : A Gamma distribution is chosen for \eta with the constraint \eta>2. Such a constraint generally reduces the efficiency of the MCMC used for inference. An alternative might be to use the parameter \eta=2+\eta' where \eta' is Gamma-distributed.

4) l.267 : I don't get the same variance decomposition since the second and third terms are not independent (given the hyperpriors). They may change the interpretation of the results.

Reviewed by David Makowski, 16 June 2024

The paper addresses an interesting topic concerning participatory plant breeding. It aims at comparing various Bayesian models for analyzing data collected in a highly unbalanced designs with a high share of missing data. I found the paper interesting but I had a lot of difficulties to understand what were the precise objectives of the on-farm trials analyzed by the authors. They didn't explain clearly enough what were the objectives of the genotype selection they aimed to perform with their on-farm trials. In addition, the description of the data is quite unclear and it is difficult to understand the different types of designs included in the datasets, and how this diversity could be properly handled. Moreover, some aspects of the modelling framework are not fully justified. In particular, I don't understand why your Bayesian models would be more suitable to analyze unbalanced data obtained in non-randomized experiments than other types of models. I don't see why your models would solve the issues related to the high level of heterogeneity of your datasets.

L17-20: The results presented in the abstract are qualitative. More quantitative results would be useful. In particular, it is unclear whether the results of the proposed model were accurate enough to be used in practice.

L17-20: Because of the use of an unbalanced design and of the lack of randomization, there is a risk of bias that may lead to errors in the ranking of the cultivars. I would be useful to reflect on this aspect in the abstract.

L20: « mixtures ». Do you mean « genotype mixtures » ? Risk of confusion with "mixtures of probability distributions ».

L30: Unclear (environments always depend on pedoclimatic conditions, weathers etc.). I guess you mean that crops may be impacted by a greater diversity of limiting factors in OA than in conventional agriculture.

L41-42: this is true only if we are able to test a large number of G in a large number of E. Otherwise, it does not allow you to estimate G x E, or for a very limited of G only. In addition, randomized trials are often more difficult to conduct on farms. The benefit is thus uncertain.

L54, 58, 65: the term « populations » is not clear. Maybe you mean « non-hybrid genotype » but I am not sure. This needs to be clarified in order to allow non-specialists to undertand.
L66: « as very few populations were present »: this appears to be in contradiction with the sentence « a large number of populations was evaluated » (L58). Moreover, for a non-specialist, it is not fully clear what is the difference between a genotype and a germ plasm.
L67-68: « genotype main effect and stability ». Unclear. Need to be explained. In fact, the concrete objective of the network of trials is unclear.
How do you want to use the data to select genotypes?
What are the criteria relevant for the assessment of the genotypes?
Do you want to select one genotype in each farm or select one cultivar for a group of farms sharing similar E?
L72-77: in this description, it is unclear what are the fixed parameters. Although, only some of the random parameters are listed, making you text hard to follow. In particular, it is not clear how the genotype sensitivities are defined. Do they correspond to genotype-specific random environmental effects (i.e., random GxE interactions). Definitions are provided much later in the paper, but this is not an optimal way to organize your paper.
Section 2.2.1: I don't understand what designs were used in the categories Regional, Statellite and unreplicated shown in Table 2.
Table 2: what is the difference between "a farm" and "an environment"?
161: « obvious outliers were excluded ». Define « obvious ».
The models are presented in section 2.3. The models look quite standard, with main effects and interactions. It is difficult to understand why these models are able to analyze on-farm trials with unbalanced and missing data.
In 2.3.1, based on the equations, « static stable » corresponds to absence of environmental effect, while « dynamic stable » corresponds to absence of interactions, but this is not how the authors presented these two types of stability in the text.
189: "Finlay and Wilkinson (1963) defined their coefficient as ». Which coefficients?
The model FWs is not based on a hierarchical structure. I don't understand what assumptions are made on the parameters in this model. Did you assume a specific parameter value for each germplasm and environment? What were the priors? This needs to be clarify.
2.3.2. Residual terms. The different types of designs included in your dataset are so different that it is unlikely that they could share the same residual variance. However, this is the assumption made in your models; you used a unique residual variance for all designs.
2.3.5. The use of leave-one-out cross validation is not fully justified as the data are not independent, in particular the data collected in the same environments are correlated.
2.3.6. Variance decomposition. I guess this decomposition is not valid for the model FWs. Please clarify.
L223: Here as well, because of the heterogenetity of the designs included in your dataset, it is unlikely that the between-germplasm and between-trial variances are the same for all types of trials.
281-283: Here, you define static and dynamic stability. But these definitions should be provided much earlier, the first time you are using these concepts.
290: why are they approximations? What would be their exact expressions?
L292-294: It is unconsistent to use frequentist statistical tests while your analysis is based on Bayesian models. You could use your Bayesian inference to derive credibility intervals for all quantities of interest.
3.1.1. This is interesting but, as mentioned above, the LOO CV does not seem well adapted to the clustering nature of your dataset.
3.2. The title is very general and could be applied to any part of the paper.
I found the results on the student distributions interesting, especially because they seem able to handle extreme data, but I am wondering why the authors only considered student and gaussian distributions and not others.
It was difficult to see how the models could be used in practice and how their use would change the practical

recommandations made to farmers.