Peer Community In
Mathematical &
Computational
Biology

Collecting, assembling and sharing data in crop sciences

Eric Tannier *based on peer reviews by* ***Christine Dillmann*** *and 2 anonymous reviewers*

Rémi Mahmoud, Pierre Casadebaig, Nadine Hilgert, Noémie Gaudio (2024) A workflow for processing global datasets: application to intercropping. HAL, ver. 2, peer-reviewed and recommended by Peer Community in Mathematical and Computational Biology. https://hal.science/hal-04145269

Submitted: 04 July 2023, Recommended: 26 February 2024

Cite this recommendation as:
Tannier, E. (2024) Collecting, assembling and sharing data in crop sciences. *Peer Community in Mathematical and Computational Biology*, 100197. 10.24072/pci.mcb.100197

Published: 26 February 2024

It is often the case that scientific knowledge exists but is scattered across numerous experimental studies. Because of this dispersion in different formats, it remains difficult to access, extract, reproduce, confirm or generalise. This is the case in crop science, where Mahmoud et al [1] propose to collect and assemble data from numerous field experiments on intercropping.

It happens that the construction of the global dataset requires a lot of time, attention and a well thought-out method, inspired by the literature on data science [2] and adapted to the specificities of crop science. This activity also leads to new possibilities that were not available in individual datasets, such as the detection of full factorial designs using graph theory tools developed on top of the global dataset.

The study by Mahmoud et al [1] has thus multiple dimensions:

- The description of the solutions given to this data assembly challenge.
- The illustration of the usefulness of such procedure in a case study of 37 field experiments on cereal-legume associations. The dataset is publicly available [3], while some results obtained from it have been independently published elsewhere [e.g. 4].
- The description of an algorithm able to detect complete factorial designs.
- An informed discussion of the merits of global datasets compared to alternatives, in particular meta-analyses
- A documented reflection on scientific practices in the era of big data, guided by the principles of open science.

I was particularly interested in the promotion of the FAIR principles, perhaps used a little too uncritically in my view, as an obvious solution to data sharing. On the one hand, I am admiring and grateful for the availability of these data, some of which have never been published, nor associated with published results. This approach is likely to unearth buried treasures. On the other hand, I can understand the reluctance of some data producers to commit to total, definitive sharing, facilitating automatic reading, without having thought about a certain reciprocity on the part of users and use by artificial intelligence. Reciprocity in terms of recognition, as is discussed by Mahmoud et al [1], but also in terms of contribution to the commons [5] or reading conditions for machine learning.
But this is another subject, to be dealt with in the years to come, and for which, perhaps, the contribution recommended here will be enlightening.

References:

[1] Mahmoud R., Casadebaig P., Hilgert N., Gaudio N. A workflow for processing global datasets: application to intercropping. 2024. □hal-04145269v2□ ver. 2 peer-reviewed and recommended by Peer Community in Mathematical and Computational Biology. https://hal.science/hal-04145269

[2] Wickham, H. 2014. Tidy data. Journal of Statistical Software 59(10) https://doi.org/10.18637/jss.v059.i10

[3] Gaudio, N., R. Mahmoud, L. Bedoussac, E. Justes, E.-P. Journet, et al. 2023. A global dataset gathering 37 field experiments involving cereal-legume intercrops and their corresponding sole crops. https://doi.org/10.5281/zenodo.8081577

[4] Mahmoud, R., Casadebaig, P., Hilgert, N. et al. Species choice and N fertilization influence yield gains through complementarity and selection effects in cereal-legume intercrops. Agron. Sustain. Dev. 42, 12 (2022). https://doi.org/10.1007/s13593-022-00754-y

[5] Bernault, C. « Licences réciproques » et droit d'auteur : l'économie collaborative au service des biens communs ?. Mélanges en l'honneur de François Collart Dutilleul, Dalloz, pp.91-102, 2017, 978-2-247-17057-9. https://shs.hal.science/halshs-01562241

Reviews

Evaluation round #1

DOI or URL of the preprint: https://hal.science/hal-04145269
Version of the preprint: 1

Authors' reply, 12 February 2024

Download author's reply
Download tracked changes file

Decision by Eric Tannier, posted 07 December 2023, validated 07 December 2023

Minor revision

The first draft of the manuscript was reviewed by three experts. All three praised the originality and usefulness of the approach, and are in favor of the recommendation by PCI mathematical and computational biology.

Two reviewers have produced significant work to suggest improvements to the text, and I think their reports deserve to be processed in depth by the submission authors in order to consider a revision.

Most of the suggestions require only minor rewrites, in order to facilitate the reader's understanding, the completeness of the bibliography, and the clarity and completeness of the results. Some suggestions may require major changes or additional calculations. These should be considered by the authors as optional suggestions, even if some of them could be useful. For example, I also found that some of the theoretical aspects presented in the first part only made sense when their application to an example was read later. But again it will be up to the authors to appreciate the ratio of improvement over amount of needed work.

I am also sensitive to the question raised by a reviewer about the possible ethical reasons (confidentiality, environmental costs..) for sometimes departing from the FAIR principles. This may not be relevant in this case, it will be up to the authors to judge, but this could be useful to the global dataset/meta-analysis discussion.

PCI Mathematical and Computational Biology would be pleased to receive a corrected version, together with replies to the reviewers. And we apologize to the authors for the lengthy editing process.

Reviewed by anonymous reviewer 1, 14 October 2023

The manuscript by Mahmoud et al. describes a workflow for processing global datasets. The authors have collected datasets from different laboratories to study the effect of different variables related to intercropping (e.g. species composition, environmental conditions and cultural practices). They describe how to reconstruct the most complete experimental design possible from partial datasets combined into a single one. Four papers have been published based on subsets of this global dataset, focusing on different scientific questions. Overall, the manuscript is well written and I have no specific requests. I just wonder how close it is to the scope of computational biology. One improvement might be to include the actual figure from the four papers mentioned in the text in addition to Figure 3.

Reviewed by Christine Dillmann, 03 December 2023

Download the review

Reviewed by anonymous reviewer 2, 06 December 2023

The manuscript « A workflow for processing global datasets : application to intercropping » tackles the problem of collecting raw crop data from heterogeneous sources, in order to analyze them jointly. To this aim, the authors present a 3 step workflow that is illustrated on a concrete example (already published in a previous article). They also provide some methodological contribution by suggesting the representation of the overall experimental design as a connected graph. This representation allows one to reformulate the problem of extracting a complete factorial design subset from the global dataset design into a problem of finding cliques in the corresponding graph.

The main focus of the article (gathering and tidying datasets in order to make them amenable to a joint analysis by data analysts) is of real interest for the community, as it offers opportunities to i) increase the power when testing hypotheses compared to considering each dataset separately and ii) query the initial datasets in new ways that were not initially accessible. It is also important to emphasize that the community should support initiatives that make data reusable as they require a significant amount of time and work to deliver ready-to-use augmented datasets to the largest audience.

A first concern it the exact status of the manuscript : in its current state, it mixes different aspects, going from feedbacks from a previous experiment (the one of Gaudio, 2021, and related references), to general guidelines for global dataset construction, through methodological contribution. This combination makes it hard to get the real significant contribution of the paper. Some rewriting of the Introduction section could help here.

Here are some additional elements of discussion of the article.

Identification of complete factorial subsets

The idea of proposing a way to automatically extract a subset of the complete global dataset that corresponds to a complete factorial design on a restricted number of levels is really interesting, for further analysis obviously, but also as a way to describe the global dataset. But the authors must provide a more detailed discussion about the use and limitations of the proposed procedure. Here are some points that should be discussed / explained more thoroughly :

- Usefulness of a complete design : while a complete design prevents the complete confounding between factors there is a huge literature on balanced incomplete design and strategies to organize the confounding to keep small order interaction distinguishable. On the other hand, note that completeness does not prevent partial confounding as the number of samples in each cell of the selected factorial crossing may be quite imbalanced, a feature that is not accounted for in the proposed approach. I would like to see some more discussion on this aspect in the manuscript.
- Is the procedure amenable to extensions ? For instance, can one investigate the graph representation to look for icomplete but connected subparts of the global dataset ? How can the proposed procedure account for additional constraints (i.e. a minimum number of species/varieties ? Require some combinations of levels to be present) ? It seems that if the procedure explicitly enumerate all possible maximum cliques then an a posteriori filtering is always possible, but can the constraints reduce the computational burden, making the procedure amenable to larger global datasets ?
- It is mentioned that the procedure has an NP hard complexity, then that in the example the solutions can be found « quickly ». There is no clear quantification, so one has no idea about the size of designs that can be handled in practice. I recommend the authors to provide e.g. a table, displaying for different combinations of numbers of factors and number of levels per factor the computational time.
- The illustration of the method on a synthetic example is very clear, but the application to the real dataset is quite vague (l224-230). More precisely it is unclear whether the application is trivial or not : if one has a 2 partite graph where the second set of vertices (N fertilization) has only 2 nodes and one looks for the maximal 2-clique with the constraint that the 2 levels of N fertilization must be present, it seems that the problem boils down to looking at Table 1 and select Experiments (i.e. rows) for which the Hitrogen fertilization column is full (13 experiments satisfy this criterion). Is this what was done or am I missing something ? If this is the result I guess this is not the best way to illustrate the usefulness of the procedure.

About raw data

The authors aim at providing a sounded way to collect and tidy different datasets in view of their joint analysis, which is a useful initiative. However the procedure advocated by the authors is to provide the raw data, without any normalization. This point is roughly not commented in the mnuscript, except in l265-269 where it is mentioned that researchers may be willing to access the data at different levels (e.g. plant or crop level). While I understand that such accesses requires the data to be « as raw as possible », it is important to mention that i) many data scientists experienced the frustrating case where one is unable to reproduce the results of a publication dur to the impossibility to rebuild the normalized dataset from the raw data, and ii) the initial data producers are the ones aware of the experimental specificities, and consequently the ones that can suggest a sounded way to normalize the data (for e.g. spatial field effects, experimenter effects, etc). So it would be really nice to have both the raw data and the codes to rebuild the dataset as preprocessed in the inital publication on side, as an option to be used. This should be feasible as the authors mention that collecting the data requires a strong interaction with the data providers anyway. One could also think about future authors contributing to the global dataset by adding additional data but also alternative normalization codes, corresponding e.g. to new ways to analyze the data.

Scope of the paper

In many places the use of a global dataset is compared to the meta-analysis approach. Different aspects are discussed, going from the size of the dataset one can expect to collect in the two cases, to the working time these two types of data require for being processed. While the comparison makes sense in the crop science context, one can notice that similar initiatives (i.e. development of methods/ressources for meta-analysis or global datasets) are developed in other fields (e.g. quantitative genetics) where a same discussion would possibly lead to different conclusions. As an example, meta-analysis has now become a popular practice for genome-wide association studies in human genetics, where (among others convenient features) it provides a way to share results without sharing individual data that may be protected for ethics considerations. Meta-analysis can also be a way to avoid the modeling/fitting of complex correlation patterns between traits/panels/environments. It is consequently quite important for the author to give the precise scope of their study in terms of field application where their recommendations apply. Also referring to the previous point note that in between global datasets and meta-analysis data there is the case of collected pre-normalized datasets that should be discussed.

Minors

1/ It is a little bit awkward to read a paper about reusability of datasets that does not provide any link for the code associated to the procedure they present. Maybe the code can be found in the Gaudio article, but I would prefer to have this mentioned and the weblink available if any.

2/ I found the following sentence to be a little bit misleading : « the resulting overall design did not allow an intermediate statistical analysis... » (l185). What does that mean ? One can perform an ANOVA on this dataset, including main effects and maybe some low order interactions, just as we can with any imbalanced or non-complete dataset. One just needs to be aware about the consequence of the partial confounding when interpreting the results. I emphasize here again that completeness does not amount to balance, so the complete subdesigns that are extracted will also require some caution when it comes to their interpretation.

3/ The authors chose to distinguish between theory and practice by first having a section introducing the main concepts of global dataset constitution, then illustrating these concepts through the case study. When reading the conceptual part one may not understand the implied consequences of the different guidelines (the fact e.g. that one will possibly have to deal with different programing languages to process the different datasets), so I was wondering if an alternative presentation where each concept is directly illustrated through the case study would be more sensible. This is not a strong recommendation as I'm aware it would require some significant rewriting of the paper, and both organizations (the one chosen by the authors, the alternative one I'm suggesting) make sense, just a suggestion for consideration.