PCI Comp. Biol.
A workflow for processing global datasets : application to intercropping, hal-04145269

christine.dillmann@universite-paris-saclay.fr

December 3, 2023

The paper proposes the concept of global datasets, as opposition to meta-analyses. Global data sets are curated aggregation of experimental data sets, far richer than summary datas that can be extracted from the litterature. The authors provide guidelines and methods to create and exploit those global datasets to answer scientific questions in agricultural research. Indeed, field experiments are costly but provide with valuable data that are often underexploited and used to answer a specific question. Aggregating the raw observations from numerous experiments into global dataset allows to study diverse phenotypic observations from varying soils and climates and may enable reliable generalizations of local findings. With the generalization of public data repositories that can handle data from field experiments, there is a real need for methodological developments like the ones that are proposed here. The paper is organized into three main parts.

- In the first part, a global workflow is presented for gathering, tidying and distributing datasets. A tidy dataset is a dataset where every column is a variable, every row an observation, and every sigle cell is a single value. The authors nicely review the general recomendations to end-up with FAIR open data. *To me, this part lacks a section about the progresses that have been achieved during the last decade on ontologies and, in particular, plant phenotyping ontologies (see e.g. Krajewski et al, 2015, doi:10.1093/jxb/erv271 or the https://www.miappe.org/ project).*

- The second part relates a case-study and describes the creation of a global dataset gathering 37 field experiments involving cereal-legume intercrops and their corresponding sole crops. The global dataset is publicly available on Zenodo. The creation of this global dataset is a remarquable result that is insufficiently described, even in the data_report.pdf file on Zenodo.

  In particular, on page 8 of the manuscript, the method used to redistribute the variables into four categories : trials, management, traits and climate should be better described. For example, trait BBCH uses the decimal code proposed for cereals in 1974. I gess this was not the code used in all experiments. When another coding system was used, was it translated or noted as NA ? Which method was used to end-up with the consensual trait names that figure in the global dataset ? How many traits were left-aside from the original data sets ? Which traits are reliably informed in the original datasets (see figure below) ? What is the "management" trait with 65 levels ? It seems that M1 and M2 are Sole-Cropping only (M1 cereals and M2 legumes ?). Figure 1 is not very informative (8 trial sites in Europe with the three french sites overrepresented) and could be avantageously replaced by a figure showing the organization of the global data-set into four csv files and a metadata file. Similarly, Table 1 would gain in being better commented. What do the

colour codes stand for ? I gess that spatial arrangement stands for the mixing pattern (within_row or alternate_row). Could the species mixtures be described ? Why were there only two Nitrogen fertilization status instead of 3 as in Mahmoud et al, 2022 ?

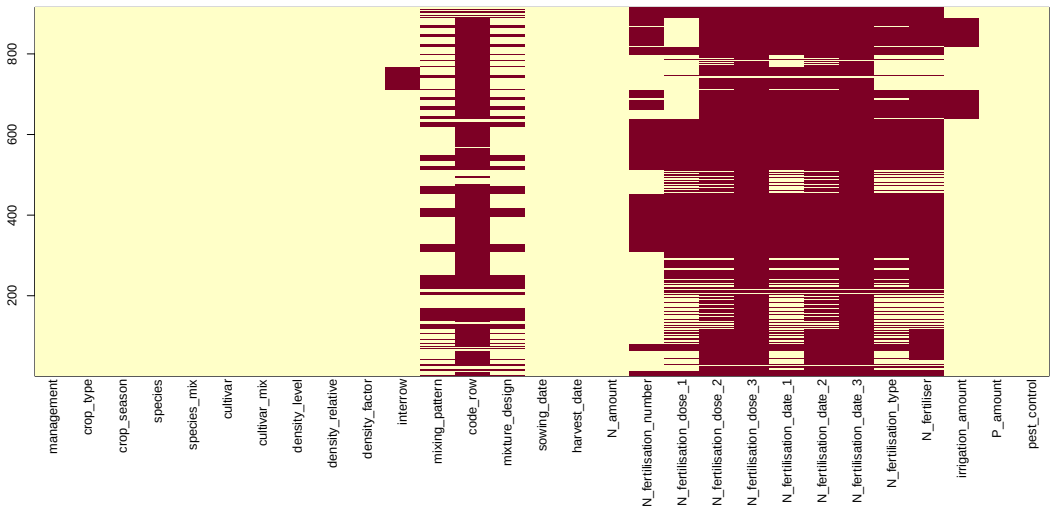

Figure 1: **Description of the data concerning the management system.** The 37 experiments were expanded into 960 experimental trials. The management.csv data file contains informations about the management system. In red, missing datas.

- The last part proposes a methods from Graph Theory to identify subsets associated with complete factorial designs in the global dataset, allowing for statistical analyses. It is associated with the production of a very nice R package available on github that allows to visualize the structure of the global dataset and to enumerate the maximal k-cliques present in the graph, each k-clique representing a factorial design. *If I understand correcly, the global dataset graph has a special structure, with each experiment being a k-partite graph, uniquely described by the levels of the k factors taken into account. This could have been stated explicitely in the manuscript.*

  A didactic application of the method is proposed with a fictive global dataset that helps to understant de concepts. *Could the edges in Figure 3 have different width, depending on the number of replicate experiments with the same k-plets ? The authors claim that the method was applied to the intercropping global dataset to identify 2-factors factorial designs (field location and nitrogen fertilization) that contain two levels of N-fertilization. But no results were provided. They refer to a published analysis (Mahmoud et al 2022) but the paper cited does not refer to the method used to select the experiments. An additional figure with the selection results would be nice.*

Finaly, I really enjoyed this paper and would recomend it to colleagues. It raises highly relevant issues concerning the processes of data production and data analysis in agricultural sciences along with the question of opening research. *I missed the step further in the application concerning the intercropping global dataset. It would have been very nice to compare the global analysis to single location/single mixture analyses for a basic agronomical trait like overyielding for example. Is the global-dataset really more powerfull ? Because several confounding factors are necessarily aggregated in the global-dataset analysis, how much larger is the residual variance ?*