Seamless Data Analysis, Visualizations and Sharing: Perspective from IRRI’s Rainfed Rice Breeding Program

Waseem Hussain (✉ waseem.hussain@irri.org)  
International Rice Research Institute  
https://orcid.org/0000-0002-6861-0193

Sankalp Bhosale  
International Rice Research Institute

Margaret Catolos  
International Rice Research Institute

Mahender Anumalla  
International Rice Research Institute

Apurva Khanna  
International Rice Research Institute

Ma Teresa Sta. Cruz  
International Rice Research Institute

Joie Ramos  
International Rice Research Institute

Short communication

Keywords: Rice Breeding, Analysis workflow, Interactive visualizations, Seamless sharing and Reproducibility.

DOI: https://doi.org/10.21203/rs.3.rs-49247/v1

License: Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Phenotypic data analysis is a key component in crop breeding to extract meaningful insights from data in making better breeding decisions. Each year the rainfed rice breeding (RRB) program at IRRI conducts trials in the national agricultural research and extension systems (NARES) network-partner sites across South Asia, Southeast Asia and Africa. Analyzing the data from the network trials and sharing the results with the partners in the best possible format is a daunting task. It is crucial to demystify data analysis to the NARES partners for making better breeding decisions. Here, we provide an overview of how RRB program at IRRI has leveraged R computational power with open-source resource tools like R Markdown, plotly, LaTeX and HTML to develop a unique data analysis workflow and redesigned it to a reproducible document for better interpretation, visualization and seamlessly sharing with partners. The generated report is the state-of-the-art implementation of analysis workflow and outputs either in text, tables or graphics in a unified way as one document. The analysis is highly reproducible and can be regenerated based at any time. The plots are built with enhanced dynamic and interactive visualizations to aid in better understanding and extract information with ease. Tables are highly interactive and manageable rendering liberty to be exported within the document in numerous formats. The source code and demo data set for download and use is available at https://github.com/whussain2/Analysis-pipeline.

Conclusively, the analysis workflow and document we presented is not limited to IRRI's RRB program but is applicable to any organization or institute with full-fledged breeding programs.

Background

The International Rice Research Institute (IRRI), established in the 1960s, is the world’s premier research organization dedicated to rice science. RRB at IRRI started since the establishment of the institute and is continuously committed to innovate and develop improved rice germplasm for improving livelihood of farmers encountering challenging climates (Dar et al. 2020). Currently, the ongoing rice breeding project, “Accelerated Genetic Gains in Rice Alliance” at IRRI funded by the Bill and Melinda Gates Foundation (BMGF) is mandated to modernize breeding strategies and framework to increase the current rates of genetic gains in close collaboration with NARES network-partner’s across South Asia (India, Bangladesh, and Nepal), East and Southern Africa (Kenya, Mozambique, Tanzania and Burundi).

Every year rainfed breeding programs at IRRI shares the breeding germplasm tolerant to drought, salt, heat and submergence with the regional partner’s for phenotypic evaluation and in return receives raw data from several trials at different locations. For instance, the rainfed breeding program during the year 2019 received data from approximately 20 trials from the NARES partner’s. It is crucial to demystify data analysis for regional partner’s to make better breeding decisions and present the results in an easy and understandable format. Explicit documentation will contribute to a crystal clear interpretation and understanding of results along with promoting collaborations. Furthermore, simultaneously analyzing and documenting the results has not been possible with readily available computational tools that require a ‘copy and paste’ system to document or report the results which in turn are highly error-prone. Thus, we believe an immediate up-gradation of data analysis workflow is crucial to be more effective and would
enhance reproducibility (Beaulieu-Jones and Greene 2017). The high-end improvement is necessary for conveniently documenting and sharing the reports.

Technology advances have made data management, analysis, interpretation, visualization and sharing more convenient. For example, R software (R Core Team 2018) packages *viz.*, ggplot2 (Wickham 2016), plotly ([https://plotly.com/](https://plotly.com/)), DT ([https://rstudio.github.io/DT/](https://rstudio.github.io/DT/)) has made the data mining manageable and visualizations interactive and dynamic. Similarly with R Markdown (Baumer and Udwin 2015), data analysis can be turned into high quality reproducible reports in which codes, text, tables, graphics and more are embedded in one unified document. Furthermore, the reports can be generated in a variety of formats including MS Word, PDF, HTML (Hyper-Text Markup Language) and more for seamless sharing ([https://rmarkdown.rstudio.com/](https://rmarkdown.rstudio.com/)).

Here, we provide an overview of how RRB program at IRRI has leveraged R computational power with open-source resource tools of R Markdown, plotly, LaTeX (Triantafyllidis and Papageorgiou 2018) ([https://www.latex-project.org/get/](https://www.latex-project.org/get/)) and HTML to develop a unique workflow of phenotypic data analysis and redesigning it to a reproducible document for better interpretations, visualization and easy sharing with collaborators. The analytical pipeline we report is unique because the data analysis workflow with modern methodology, description of scripts, outputs (either in text, tables or graphics) and interpretation of results are compiled as a single document or in simpler words ‘everything is at one place’.

**Main Features Of Analysis Report**

The phenotypic data analysis workflow we developed is very unique, having an analysis workflow, description of scripts results with detailed description embedded as one unified document. A sample document is given in Additional file 1. The main features of the document are:

1. The report shows a complete data analysis workflow and a modern way of analyzing the data (Fig. 1). Data modelling includes advanced mixed-models to account both for experimental designs and spatial variations (Isik et al. 2017). Five-mixed-models are tested and the best model is selected for downstream analysis and generation of results. The description of the models is detailed in Additional file 1 under sections 4.1 and 4.2.

2. The analysis pipeline and document is highly reproducible and the same report and analysis pipeline can be generated if needed. The sample source codes of the analytical pipeline and demo data set can be directly downloaded from the GitHub repository ([https://github.com/whussain2/Analysis-pipeline](https://github.com/whussain2/Analysis-pipeline)). The instructions on how to run the analysis pipeline on a local computer is given on the GitHub repository page.

3. Any new data and or editing/corrections to the existing pipeline can be done by simply re-knit the R markdown `.Rmd` document ([https://rmarkdown.rstudio.com/articles_intro.html](https://rmarkdown.rstudio.com/articles_intro.html)). This analytical pipeline avoids manually updating or generating reports or PowerPoint slides which are otherwise highly prone to errors and time-consuming.
4. The document includes metadata (information about the field trial, data collection, experimental design, and more) in the beginning for quick identification, location and association of data and analysis at any given time (Fig. 2a).

5. The document is well structured and organized. For example, the document is divided into sections with headings and subheadings to increase accessibility and cognition. The table of contents is always visible in the document making it faster and easier to navigate within a document (Fig. 2a). Additionally, readers have the flexibility to hide the sections for better readability and accessibility.

6. The document is currently generated in HTML which upon download can be easily opened in any browser without requiring any access to the internet. Further, HTML files can be shared easily and/or hosted on websites for easy sharing and future use.

7. The graphics in the document are highly dynamic and interactive. Simply hovering a cursor on the plot will display the additional and hidden information, which is not possible in static plots. For example, the box plots and heatmaps of field experimental design to visualize spatial trends are highly dynamic and interactive (Fig. 2b and 2c). Additionally, plots can be easily exported to the local drive.

8. The output generated in the form of tables is highly dynamic and interactive. Tables generated can be easily managed, searched, and sorted like a mini excel sheet (Fig 2d). Interestingly, tables can also be exported in various formats or printed directly within the document. The tables and result outputs being in the same file completely avoids the option of saving the files on computers and digging into them to extract the useful information in making presentations or in undertaking breeding decisions.

9. Complete description and details of scripts, procedures and methods used for analysis are elaborated in the same document. Results generated in the document in the form of plain plots and or tables have been thoroughly described to aid in the interpretation and better understanding. Hyperlinks have been embedded in the required sections to help in understanding the concepts and add knowledge to the users. For example, web sources on how to interpret the box plots; methods used to calculate heritability with complex models; spatial analysis modelling and much more has been hyperlinked in the document (Fig. 2e and 2f).

Conclusion

We report unique implementation of data analyses workflow and methodology, and document it for better understanding and easy sharing with partners. The enhanced interactive visualization of plots and tables makes it easier for users to extract the more information available at fingertips with ease. To fully harness the benefits of this approach basic knowledge of R programming and R Markdown is required. We believe this is a great initiative to modernize the data analysis of IRRI’s RRB program and in the future can be further improved. Conclusively, the analysis pipeline will be of great use to the crop breeding communities having full-fledged breeding programs.

List Of Abbreviations
Declarations

Ethical Approval and Consent to Participate

Not applicable

Consent for Publication

Not applicable.

Availability of supporting data

The R source codes and demo data is available on the GitHub repository https://github.com/whussain2/Analysis-pipeline. The detailed instructions how to use the source codes and run the analysis pipeline on a local computer is given on the GitHub repository page.

Competing interests

The authors declare no potential competing interests.

Funding

This work was supported by the Transforming Rice Breeding project from the Bill and Melinda Gates Foundation (Grant No. OPP1076488).

Authors contributions

WH designed and conceived the concept and study. SB supervised and directed the study. MC, MA and AP were involved in the data analysis. SMT and JR helped in the data collections and preparations. All authors read and approved the manuscript.

Acknowledgements
The rainfed breeding team expresses gratitude to the IRRI’s irrigated breeding team led by Joshua N. Cobb to initiate the idea of the data analysis workflow. Authors are also thankful to the IRRI’s internal reviewers Shalabh Dixit and Jerome Bartholome for a constructive feedback and comments for improving the manuscript.

Authors’ information

Rice Breeding Platform, International Rice Research Institute, Los Banos, Laguna, Philippines

Waseem Hussain*, Sankalp Bhosale, Margaret Catolos, Mahender Anumalla, Apurva Khanna, Ma. Teresa Sta. Cruz, and Joie Ramos

Description Of Supplementary Materials

Additional file 1: The file is a sample report of data analysis workflow and documentation. It is in HTML format.

References

Baumer B, Udwin D (2015) R Markdown. Wiley Interdiscip. Rev. Comput. Stat. 7: 167–177.

Beaulieu-Jones B, Greene C (2017) Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol 35: 342–346. https://doi.org/10.1038/nbt.3780

Dar MH, Waza SA, Shukla S, et al (2020) Drought Tolerant Rice for Ensuring Food Security in Eastern India. Sustainability 12:2214. https://doi.org/10.3390/su12062214

Isik F, Holland J, Maltecca C (2017) Spatial Analysis. In: Isik F, Holland J, Maltecca C (eds) Genetic Data Analysis for Plant and Animal Breeding. Springer International Publishing, Cham, pp 203–226

R Core Team 2018 R: A language and environment for statistical computing.e. R Foundation for Statistical Computing, Vienna, Austria

Triantafyllidis CP, Papageorgiou LG (2018) An integrated platform for intuitive mathematical programming modeling using LaTeX. PeerJ Comput Sci 4:e161. https://doi.org/10.7717/peerj-cs.161

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis, 2nd edn. Springer International Publishing

Figures
Figure 1

Schematic representation of data analysis workflow adapted in the current R Markdown generated report. The four main steps involved in the workflow process are: a) data importing, b) data pre-processing, c) data modelling, and d) generation of results. a) Data is imported and general information (metadata) on data is recorded in the beginning of the document. b) In pre-processing, data is checked for missing values. Descriptive statistics (mean, mode, coefficient of variation, standard deviation and more) are generated to get a general idea about the data. Interactive heatmaps of field experimental design are plotted to check the spatial trends in data. Data is visualized using histograms, box plots and QQ plots to check the data distribution, data normality and correlated errors among data points. Lastly, in pre-processing step outliers are identified and filtered using a univariate approach. c) In data modelling, a mixed model approach is used to correct for experimental design factors and spatial variations. The detailed description and type of the mixed-models used for analysis is given in the sample report. Modelling of trials is done separately and also through the combined approach if more than one trial is available. d) After modelling various results are extracted, which includes best unbiased linear predictors (BLUPs) or best linear unbiased estimates (BLUEs) depending upon whether genotypes were treated random or fixed. Variances and heritability for the variable are also estimated. The genotypes based on BLUPs or BLUEs are ranked to make breeding decisions. Finally, general remarks on the overall data analysis is noted at the end of the document.
The four main steps involved in the workflow process are: a) data importing, b) data pre-processing, c) data modelling, and d) generation of results. 

a) Data is imported and general information (metadata) on data is recorded in the beginning of the document. 

b) In pre-processing, data is checked for missing values. Descriptive statistics (mean, mode, coefficient of variation, standard deviation and more) are generated to get a general idea about the data. Interactive heatmaps of field experimental design are plotted to check the spatial trends in data. Data is visualized using histograms, box plots and QQ plots to check the data distribution, data normality and correlated errors among data points. Lastly, in pre-processing step outliers are identified and filtered using a univariate approach. 

c) In data modelling, a mixed model approach is used to correct for experimental design factors and spatial variations. The detailed description and type of the mixed-models used for analysis is given in the sample report. Modelling of trials is done separately and also through the combined approach if more than one trial is available. 

d) After modelling various results are extracted, which includes best unbiased linear predictors (BLUPs) or best linear unbiased estimates (BLUEs) depending upon whether genotypes were treated random or fixed. Variances and heritability for the variable are also estimated. The genotypes based on BLUPs or BLUEs are ranked to make breeding decisions. Finally, general remarks on the overall data analysis is noted at the end of the document.
Panels showing the screenshots of some of the features in the generated document. 2(a) Shows the general information and table of content to easily navigate the document. 2(b) Shows the example of an interactive heatmap to visualize the field experimental design and check for spatial variations in field. 2(c) Shows interactive boxplot. 2(d) Shows the example of interactive tables which can be managed and exported in various formats. 2(e) Shows the example of description of the content to aid in understanding the data analysis, code and procedure. 2(f) Shows the detailed description of one of the spatial models used for the analysis of data.
Figure 2

Panels showing the screenshots of some of the features in the generated document. 2(a) Shows the general information and table of content to easily navigate the document. 2(b) Shows the example of an interactive heatmap to visualize the field experimental design and check for spatial variations in field. (2c) Shows interactive boxplot. 2(d) Shows the example of interactive tables which can be managed and exported in various formats. 2(e) Shows the example of description of the content to aid in understanding the data analysis, code and procedure. 2(f) Shows the detailed description of one of the spatial models used for the analysis of data.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.html
- Additionalfile1.html