Innovative procedures of data preparation to ensure data integrity for crop modelling

J. Liu, M. George, A.J. Michel, N. Arnold, I. Sorensen, H.E. Brown and R. Zyskowski

The New Zealand Institute for Plant & Food Research Limited, Private Bag 4704, Christchurch Mail Centre, Christchurch, 8140, New Zealand

jian.liu@plantandfood.co.nz

Abstract

Crop models require high-quality data to develop reliable functions that can represent the crop response to the surrounding environments. Data preparation, therefore, is a vital prerequisite for crop modelling exercises. Agricultural researchers spend a considerable amount of time collecting data from different sources and scales to ensure that reliable data is generated. In this paper, we discuss three common issues, 1) incomplete metadata, 2) missing values and 3) outliers during data integration for a 2-year project conducted at two Plant & Food Research sites. In the first section of the paper, we presented a programming approach to deal with these issues with data science tools, such as R and Python. Programming languages ensured the reproducibility of the process of data preparation. In section two, the application of a dashboard system was discussed to streamline sensor data for two purposes which were irrigation scheduling and data quality checking. The dashboard provided researcher the ability to monitor the soil water content in near-real-time, and prepare the sensor data ready for subsequent usage. Advanced data science tools can indeed mitigate the tedium of data preparation and increase the integrity of agricultural research data. However, it must be noted that the interactions between the data users and collectors are still vital to provide high-quality data from different sources.

Additional keywords: crop modelling status, data integrity, data preparation, data streamlining, field experiments

Introduction

The Crop models have helped scientists, policymakers and farmers make informed decisions on research directions, regulations and farm practices worldwide (Ewert *et al*., 2015; Holzworth *et al*., 2018; Holzworth *et al*., 2015; Jones *et al*., 2017). Further improvement of crop models is necessary to represent the temporal and spatial variations that occur in the real world. Robust datasets play a critical role in improving the applicability of crop models and minimising the uncertainty within such tools (Holzworth *et al*., 2018). Specialised teams have been organised to tackle data quality and shareability in the Agricultural Model Intercomparison and Improvement Project (AgMIP). High-level protocols for data quality evaluation are in place, yet data quality remains one of the major difficulties for crop modelling exercises (Seidel *et al*., 2018). Low quality data, which can be caused by unnoticed modification of samples or incorrect data manipulation, possess relatively high uncertainty (Gianni *et al*., 2010). The data uncertainty will propagate to the model output, which can reduce the model accuracy and usefulness (Mannschatz & Dietrich, 2017). The complexity of data preparation increases when there are various data sources over multiple experimental sites, which also requires considerable amounts of time to verify the integrity of the data for model development (Brown *et al.*, 2018).

Datasets obtained from different sources tend to vary in different ways (Wickham & Grolemund, 2017). Depending on individual experience, the preparation may involve multiple stages such as error checking, variable aggregation and exploratory data analysis. Data integrity can be jeopardised when there are several manual processes without a standardised protocol. Nevertheless, efforts have been made to generalise data preparation for statistical analysis. Zuur *et al.* (2010) reported a detailed checklist for data exploration to avoid common statistical issues in the ecological field. The checklist standardised visual tools that can aid data preparation, although these tools are orientated towards statistical science and could be challenging to use for junior researchers who have limited statistical background. A project-oriented procedure could shed some light on data preparation for modelling practices in respect to the FAIR (Findable, Accessible, Interoperable and Reusable) data principles (Wilkinson *et al.*, 2016).

Model development and validation generally requires data from three aspects: 1) local climate; 2) crop and management; 3) soil (Kersebaum *et al*., 2015). The complete dataset consists of these three aspects, therefore, can have uncertainties from two levels. First, each aspect has random errors stemmed from equipment, sampling protocols or other unnoticeable reasons (Gianni *et al*., 2010). Secondly, integrated data are incomplete or lost track of which data are up-to-date due to inconsistent practices of data manipulation during preparation, which can cause incorrect interpretation and model structure. We will focus on addressing the second level in this paper because it is the most cost-effective approach to evaluate and collate existing data for model development.

This paper aims to: 1) document the procedure we adapted to integrate data obtained from different sources; 2) explore the solutions for fast-tracking the identification of sources of errors; 3) attempt to standardise the data preparation of field data for crop modelling exercises. We hope these efforts can help to improve the efficiency and accuracy of data preparation for crop modelling exercises carried out researchers. There are two sections after a brief description of the project in this paper. Section One describes the potential causes of three common issues in the dataset with practical solutions. Section Two introduces an example: using a dashboard to facilitate data preparation via advanced visualisation in near real-time for sensor data.

Brief description of the project

A two-year project (2018-2020) was conducted at the Plant and Food Research sites in Havelock North (39°39'13.4" S, 176°51'35.8" E) and Lincoln (43°37'28.5"S, 172°28'03.5"E) (Table 1). The project aim was to investigate the effect of agronomic practices on pea yield and pea protein composition. The main factors were cultivar, sowing date and fertiliser application rate. Two identical field experiments were conducted in the first year, one at each site. The same sampling protocol was followed to collect phenology and biomass data fortnightly. In the second year, two different field experiments were set up at each site. The experiment at the Havelock North site aimed to test the consistency of the findings from the first year with less frequent measurements. The experiment at the Lincoln site was set up in the rain-out shelter facility (Michel *et al.*, 2015). An auto logging system was set up in this facility to monitor the soil water content (SWC), with data recorded every 15 mins. The data were used to schedule the irrigation requirements of the different treatments and for subsequent water balance analysis during modelling activities.

Table 1: Key information about the experiments.

Experiment ID	Year	Location	Key Measurements	Frequency
1	2018/19	Havelock North	biomass phenology*	fortnightly and final yield assessment
2	2018/19	Lincoln	biomass phenology*	fortnightly and final yield assessment
3	2019/20	Havelock North	biomass phenology*	inter-season phenology development and final yield assessment
4	2019/20	Lincoln	biomass phenology* sensor data	Seven biomass and phenology measurements at key development stages 15 mins interval logging for soil water content (TDR); 10 mins interval logging for all the other sensors (soil temperature, canopy cover, canopy temperature, canopy reflectance)

**Phenology in this study consists of branch, node and leaf number, and flowering time.*

R (R Core Team, 2019) was the main software used to develop a procedure of data preparation, via a tool package called Rmarkdown (Allaire *et al.*, 2019). Python and Grafana were the tools used to develop the dashboard. Data were stored and transmitted by two database libraries SQLite and PostgreSQL.

Section One

The importance of metadata

The first challenge was the inconsistency of variable names within and across experimental sites. There were three specific issues identified with the variable names: 1) multiple variable names for the same plant trait; 2) a mismatch of variable names across different sites and; 3) the inconsistency of variable names presented in the data life cycle (from field data to downstream analysis). The first two issues were caused for various reasons. Sudden changes in plant development made it necessary to create new variables to record the changes. For example, pea grains sprouted while still in the mother plants, whereas we expected dormancy in the pea grains. The last issue was due to the naming conventions for crop models which usually differ from the variable names that researchers use for field data collection, further complicated by their individual preferences.

The reusability of the data increases by eliminating these issues. One possible approach adopted here was to communicate with team members and identify the causes of issues to form well-documented metadata. Generally, the metadata should contain sufficient information which can empower a second researcher to visualise the experiment and understand the meaning of variables. Table 2 provides an example of the tabular metadata used in this study. There were four critical parts including name, description, units and equation. Abbreviations were used for field data collection, while more informative variable names were used for the calculated variables. Each variable requires a minimum description of the name, meanings and possible methods if applicable. Units are for reporting the data on different scales. Equations are crucial for calculated variables since equations reflect the dependence of particular variables.

Table 2: An example of metadata for agricultural field experiments.

Name	Description	Units	Equation
Date	The date for trait measurement	NA	NA
Treatment	3 cultivars: A, B and C	NA	NA
	3 sowing dates: 5-Oct-18, 26-Oct-18 and 16-Nov-18		
HA	Harvest area for each measurement. Quadrats of 0.5 m^2 were used	m^2	NA
TFW	Total fresh weight of plant sample in the harvest area. Machinery cut to ground level.	g	NA
Total_Fresh Yield	The total fresh yield calculated from specific treatment plot	kg/ha	TFW/Ha*10
VWC	Volumetric water content in a layer (100 mm thickness) in the designated plot.	mm^3/mm^3	NA
SWC	Soil water content in a designated plot	mm	VWC*100

The metadata table can be extended incrementally to incorporate more variables when data reach a new analysis phase. For instance, yield-relevant variables for each plot are probably sufficient during the exploratory data analysis phase. Summarised variables that represent treatment effects will be necessary when a conclusion is needed. Soil water content (SWC) and climate variables are required for water balance calculations. Parameters of crop models can be included once a significant effect is identified.

An R (R Core Team, 2019) script-based approach was used with Rmarkdown (Allaire *et al.*, 2019) to tackle metadata management. At the beginning of data preparation, a great amount of effort had been made to clarify the meaning and sample methods for all variables via frequent communication with data collection teams. The metadata for different data sources, such as biomass and climate data, were then processed by R script with clear documentation of rationales about any modification on the metadata. The metadata were stacked and stored in a relational database (DB) powered by SQLite (a light version of the Structured Query Language database engine). Relations of metadata and data were described and established in the Rmarkdown file. This approach presents five advantages including 1) Automation of the process by programming languages; 2) Variables can be tracked efficiently because of the direct comparison from different sources; 3) R and SQLite are free for any purpose of use; 4) SQLite DB is stable and storage efficient; 5) high shareability and findability.

Are missing values (NA) truly missing?

Missing values (NA) are common in phenological data collected in field experiments due to the high likelihood of unexpected events. However, NA could be introduced to Excel worksheets during the data entry stage, especially when there are more columns and rows than can be comfortably displayed on the monitor. In this pea project, we observed that some measurements were mismatched with incorrect plot number or treatment. These mismatches are difficult to detect by direct observation of the raw data and will confuse subsequent data users.

Visual aids are particularly useful to identify NA. Using an R package inspectdf (Rushworth, 2019), Figure 1 shows an example of a ranking chart that was generated by counting the percentage of NA presented in each variable. The variable, Nitrogen content in Dead materials (N_Content_Dead), had 100% of NA and stood out immediately. Close attention should be paid to the difference between the bound variables. For example, the variable sub-sample fresh weight (SSFW) had 16% of NA, while the NA percentage increased to 23% in sub-sample dry weight (SSDW) after the samples were oven-dried. These two variables should have the same number of observations.

To further investigate the source of NA, key variables were selected to do preliminary exploratory data analysis. The dry matter percentage of plant samples is usually a constant number within one sampling event, regardless of sampling methods used. Hence, the dry matter percentage of sub-samples (SSDM_PC) and partitioning samples (partDM_PC) were plotted together (Figure 2), with the harvest events labelled on each data point. A positive relationship with less variation is expected between the two calculated values. However, the harvest event H3 in site 2 has SSDM_PC at 0 while partDM_DCs are greater than 0, which suggests that H3 could be the source of NA (Note: NA are denoted as 0s here since NA cannot be drawn as numeric values). The process can be iterated through other key variables to clarify the meaning of NA.

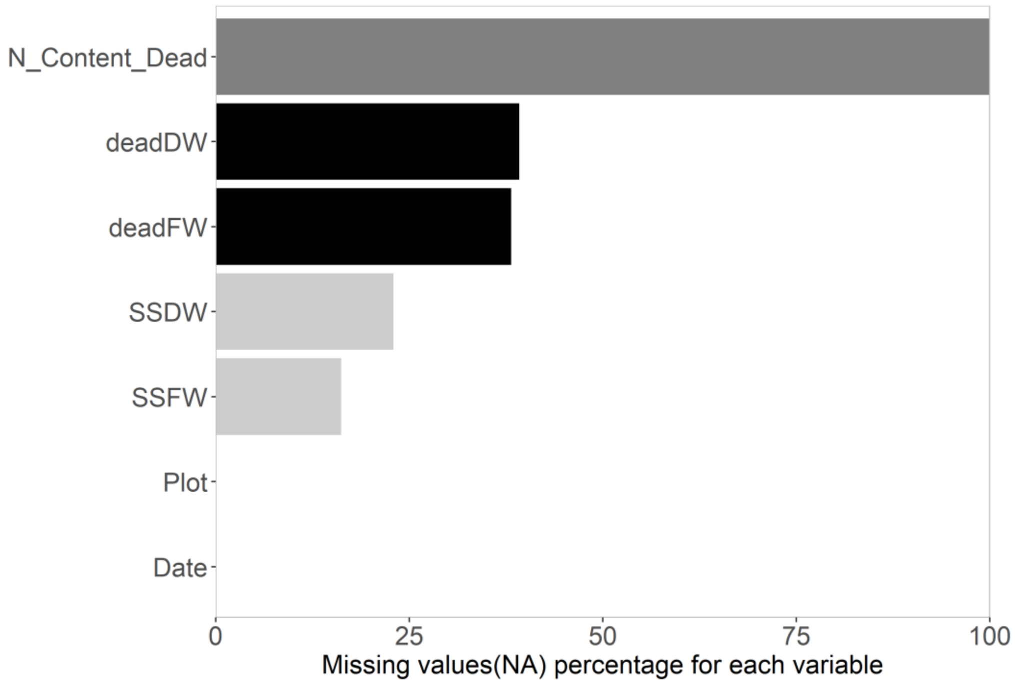

Figure 1: Percentage of missing values (NA) in each variable.

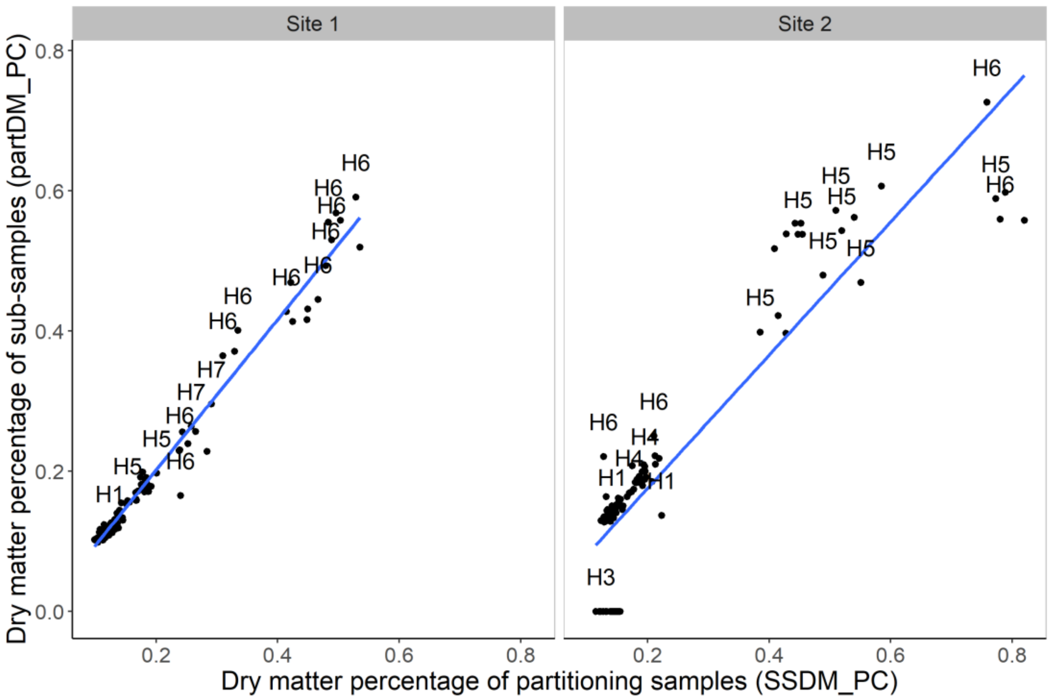

Figure 2: Exploratory data analysis for data quality inspection. The text label represented the harvest number.

Outliers or typographical errors (Typos)

The black dots on both sides of the line in Figure 2 illustrate the variability in the weight measures (SSDW and partDW). Variability could be attributed to various factors such as instrument errors, errors introduced during sample processing, typos during data entry or true outlier values. Standard processing for samples may consist of three steps: measurements of fresh samples, measurements of dried samples and data entry. These steps could be error-prone. For instance, decimals could be misplaced when manually transferring values from the hard copy datasheet to electronic ones. On occasions, we observed that the plant component dry weight was greater than its fresh weight because of misplaced decimals in the Excel sheet.

Scatter, histogram and boxplot are the three standard graphical tools to examine abnormal values. Figure 3 shows the boxplot for the number of leaves per stem per plant in six treatments over two sites. In the first treatment (Cultivar A, Early sown and zero Nitrogen; CA_E_0N), the leaf number has a group of values (triangles) exceed 40 leaves per stem and categorised as outliers for the cultivar A. It could be a group of true outliers because of the cultivar responding to the unique environment or errors caused by incorrect data recording. Further investigation, such as seeking expert knowledge and extra visualisation effort, is necessary to trace the reason for this particular data behaviour.

Typos can be mitigated by reducing the number of manual steps in the process. For instance, scales with the auto-recording system could minimise the handling errors during sample weighing. In contrast, phenological data measurements like the number of leaves and branches can be challenging to be automated and rely on individual experience. Specific training could reduce the possibilities of errors and using pre-setup data validation rules can help to prevent data entry errors.

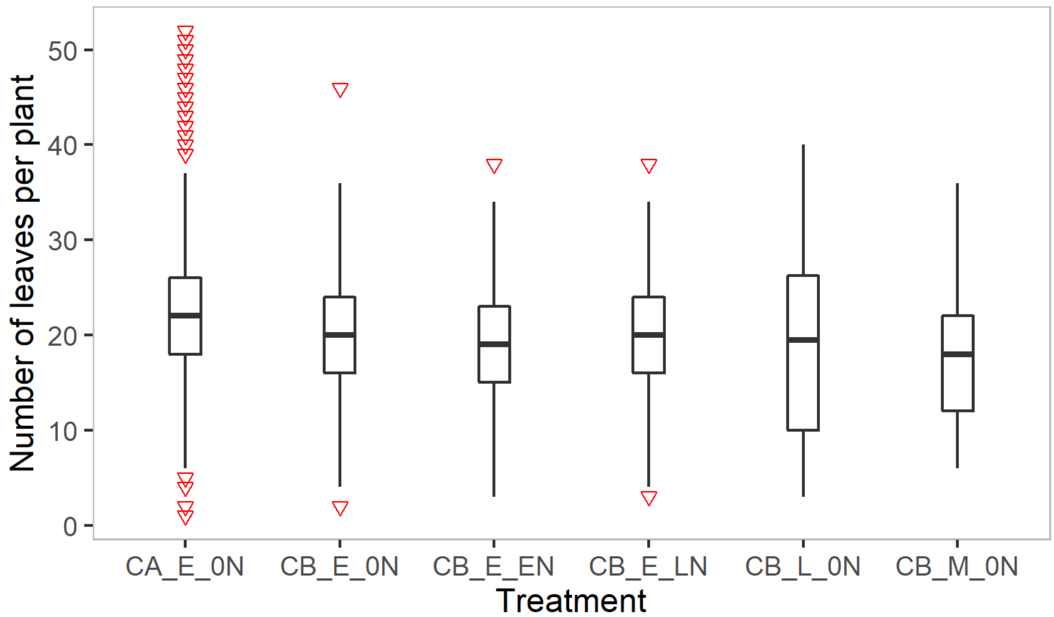

Figure 3: Boxplot of leaf number over six treatments for data quality check.

Section Two

Information from sensors can provide a dramatic improvement in data resolution. However, issues like the failure of electronic components in the sensor can cause missing data. Additionally, manual sensor data processing can be tedious. We implemented a workflow that could automate the process from data preparation and storage to dashboard reporting at near-real-time (15 to 30 mins delay). Figure 4 illustrates an overview of the workflow. The sensors recorded the soil moisture readings and stored the data in a .dat file by the Campbell Scientific Logger Net software via radio transmission. Python scripts were used to concatenate sensor data with the correct metadata (plot design and treatment) in a Jupyter notebook (An integrated development environment). We re-used Python scripts that were developed previously to calculate soil water content (SWC) and soil water deficit (SWD; deficit to the profile water content at the beginning of the experiment). The raw sensor data and processed values were then uploaded into a PostgreSQL database which connects with the dashboard tool, Grafana.

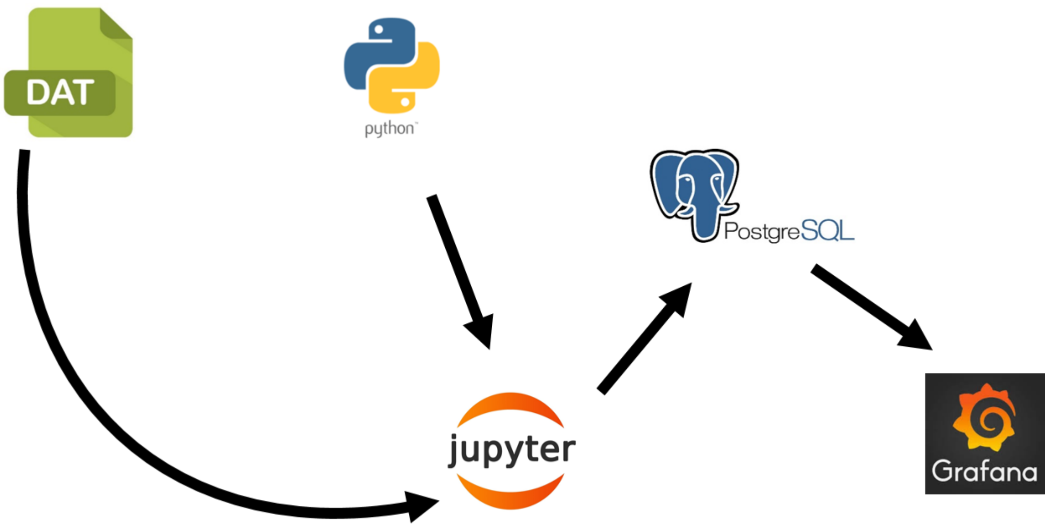

Figure 4: Overview of the workflow for near real-time data streaming.

To establish the workflow, the first step was to configure the database structure for the processed data. Therefore, an independent Jupyter notebook was developed as a configuration file. Python scripts in the notebook generated two critical outcomes, including a .npy file (a file created by Python to store array structured data) and a PostgreSQL database. The .npy file contained the information about the number of observations in the existing raw data, which was starting the point for importing subsequent data. The database was the structure used to store processed values. The similar Python script for data manipulation was packed into an iteration function within a second Jupyter notebook. The second notebook imported the new data and appended the information to the database at an interval of 15 mins.

Grafana automatically updated the dashboard by querying the database every 15 mins. Figure 5 shows a snapshot from the Grafana dashboard for soil water profile.

The interactive dashboard provided detailed information when users hovered on the panel. Multiply panels could be displayed in one dashboard to monitor the status of individual sensors and detailed water usage from each plot.

It was necessary to host the second notebook and a Grafana instance on an active computer or server in order to keep the data streamlined. We used two systems, a Windows desktop computer and a Linux server, to host notebook and the Grafana instance respectively due to internal network restrictions.

The workflow created a template for data preparation of other sensors such as pyranometer and temperature sensors. We could monitor the sensor status and check data quality simultaneously. Processed data in PostgreSQL DB is ready for any subsequent analysis exercises.

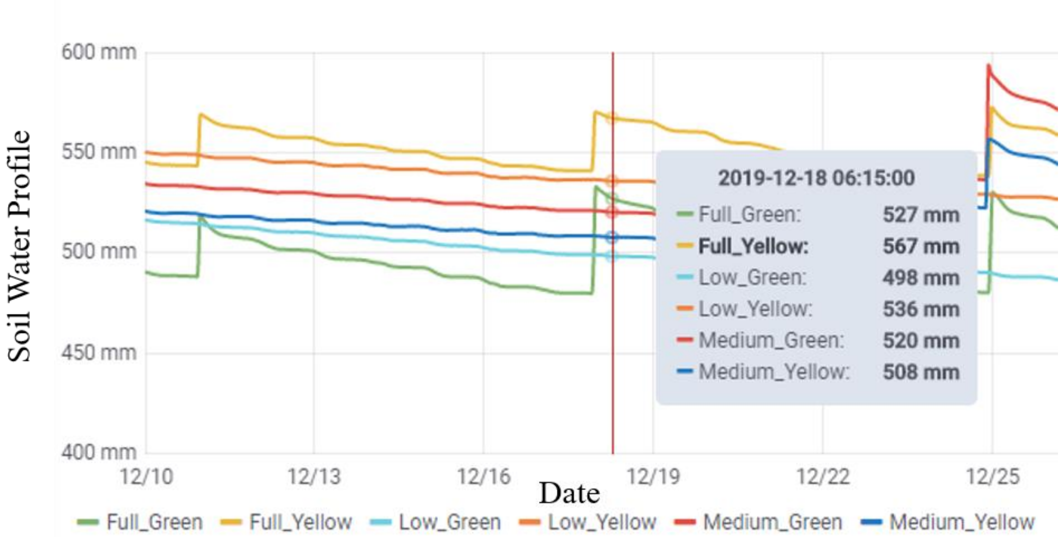

Figure 5: A Snapshot of the dashboard for irrigation scheduling.

Conclusion

We applied a combination of data science tools to attempt to achieve error-free datasets. We found that the programming language R can efficiently identify and track potential errors in the data using its well-documented metadata and visualisation tools and that the dashboard system is useful for processing sensor data and irrigation scheduling. This integration of data science tools provides agricultural researchers with a new avenue to improve the efficiency of data preparation and ensure data integrity for future usage. However, researchers who have limited experience in programming languages and databases may find it challenging to adapt the applications in this paper. It must be noted that an understanding of the data is more important than the tools one chooses to process the data and maintain data integrity. Poor data quality commonly stems from lack of understanding and communication between the data users and the collection team. Wrong assumptions are possible if researchers who process samples misunderstand the objectives of the sampling protocol or have insufficient information about the sources of missing values or outliers. Metadata developed by both sides to capture details of experiments, therefore, are critical to maximise the value of agricultural data.

Acknowledgements

The authors would like to thank the colleagues from Plant & Food Research who contributed their time to the project. Two programmes, Discovery Science and Sustainable AgroEcosystems, supported the completion of this project. Both programmes were funded from the Strategic Science Investment Fund by Plant & Food Research.

References

Allaire, J.J.; Xie, Y.; McPherson, J.; Luraschi, J.; Ushey, K.; Atkins, A.; Wickham, H.; Cheng, J.; Chang, W.; Iannone, R. 2019. *Rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown.

Brown, H.; Neil, H.; Holzworth, D. 2018. Crop Model Improvement in Apsim: Using Wheat as a Case Study. *European Journal of Agronomy 100*: 141–50. https://doi.org/https://doi.org/10.1016/j.eja.2018.02.002.

Ewert, F.; Rötter, R.P.; Bindi, M.; Webber, H.; Trnka, M.; Kersebaum, K.C.; Asseng, S. 2015. Crop modelling for integrated assessment of risk to food production from climate change. *Environmental Modelling & Software 72*: 287-303. doi:https://doi.org/10.1016/j.envsoft.2014.12.003

Gianni, B.; Mike, R.; Marcello, D.; Keith, M. 2010. Validation of biophysical models: issues and methodologies. A review. *Agronomy for Sustainable Development 30*: 109–130. doi: 10.1051/agro/2009001

Holzworth, D.; Huth, N.I.; Fainges, J.; Brown, H.; Zurcher, E.; Cichota, R.; Verrall, S.; Herrmann, N. I.; Zheng, B.; Snow, V. 2018. APSIM Next Generation: Overcoming challenges in modernising a farming systems model. *Environmental Modelling & Software 103*: 43-51. doi:https://doi.org/10.1016/j.envsoft.2018.02.002

Holzworth, D.; Snow, V.; Janssen, S.; Athanasiadis, I.; Donatelli, M.; Hoogenboom, G.; White, J.; Thorburn, P. 2015. Agricultural Production Systems Modelling and Software: Current Status and Future Prospects. *Environmental Modelling & Software 72*: 276-286. https://doi.org/10.1016/j.envsoft.2014.12.013

Jones, J.W.; Antle, J.M.; Basso, B.; Boote, K.J.; Conant, R.T.; Foster, I.; Wheeler, T.R. 2017. Brief history of agricultural systems modeling. *Agricultural Systems 155(C)*: 240-254. doi:10.1016/j.agsy.2016.05.014

Kersebaum, K.C.; Boote, K.J.; Jorgenson, J.S.; Nendel, C.; Bindi, M.; Frühauf, C.; Gaiser, T.; Hoogenboom, G.; Kollas, C.; Olesen, J.E.; Rötter, R.P.; Ruget, F.; Thorburn, P.J.; Trnka, M.; Wegehenkel, M. 2015. Analysis and classification of data sets for calibration and validation of agro-ecosystem models. *Environmental Modelling and Software 72*: 402–417. https://doi.org/10.1016/j.envsoft.2015.05.009

Michel, A.J; Brown, H.E; Gillespie, R.N; George, MJ.; Meenken, E.D. 2015. Automated Measurement of Crop Water Balances Under a Mobile Rain-Exclusion Facility. *Proceedings Agronomy Society of New Zealand 45*: 39–46.

Mannschatz, T.; Dietrich, P. 2017. Model input data uncertainty and its potential impact on soil properties. pp. 25–52. *In*: Sensitivity Analysis in Earth Observation Modelling. Eds G. P. Petropoulos & P. K. B. T.-S. A. in E. O. M. Srivastava. Elsevier. https://doi.org/10.1016/B978-0-12-803011-0.00002-1

R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rushworth, A. 2019. *Inspectdf: Inspection, Comparison and Visualisation of Data Frames*. https://CRAN.R-project.org/package=inspectdf.

Seidel, S.J.; Palosuo, T.; Thorburn, P.; Wallach, D. 2018. Towards Improved Calibration of Crop Models – Where Are We Now and Where Should We Go? *European Journal of Agronomy* 94: 25–35. https://doi.org/https://doi.org/10.1016/j.eja.2018.01.006.

Wickham, Hadley; Grolemund, Garrett. 2017. R for Data Science: Import, Tidy, Transform, Visualise, and Model Data. 1st ed. O'Reilly Media, Inc.

Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, A. 2016. The Fair Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 15.

Zuur, A.F.; Ieno, E.N.; Elphick, C.S. 2010. A Protocol for Data Exploration to Avoid Common Statistical Problems. *Methods in Ecology and Evolution 1*: 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x.