Crop Analytics Orchestra

A coordinated, flexible constellation of R packages for crop analytics

Max Moldovan · Biometry Hub, Adelaide University · max.moldovan@adelaide.edu.au

AAGI Analytics for the Australian Grains Industry

Abstract

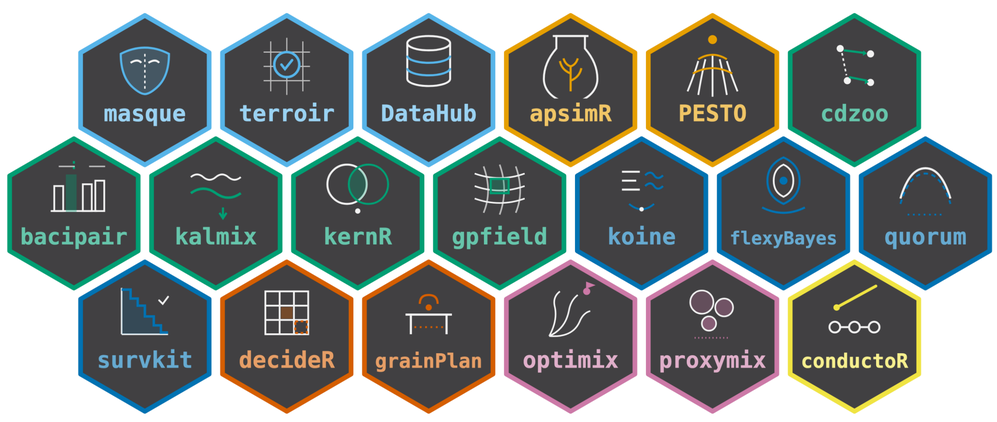
This work has been done towards GRDC-funded AAGI Research and Development project AAGI-AU-RD-CSL, *Causal learning for optimising industry relevant traits in Australian grain production systems*, with an intention to introduce a broader analytical paradigm to the grain industry, with a potential of AI integration. In particular, crop analytics mostly treat analytical methods either in isolation (e.g. linear modelling) or in small constellations (e.g. descriptives + modelling + economic evaluation). **The Crop Analytics Orchestra** is a larger constellation of 18 R packages and an interactive open-data catalogue called DataHub – all built to work together, also covering a wider open-ended range of methods. {conductoR} is a central analytical task dispatcher and outcome collector, and it is currently a pure R package, provisioned to be AI integrated. The Orchestra is practically illustrated with a basic causal analysis example performed on the open-source dataset. Each package, except {conductoR}, can function stand-alone, and each aims either to be entirely novel (e.g. {proxymix}, already on CRAN) or to perform among the best in its class.

DataHub: the data layer

101	27	55	8
open datasets held	of them Australian	more listed, awaiting access	kinds of question it can match

Each dataset is stored with its source and license, formatted, and its columns and coverage are recorded at the time of enrolment. We can ask, for example, which datasets compare several nitrogen rates, and DataHub would present candidates. The study on this poster utilises the open NVT results through DataHub.

The orchestra

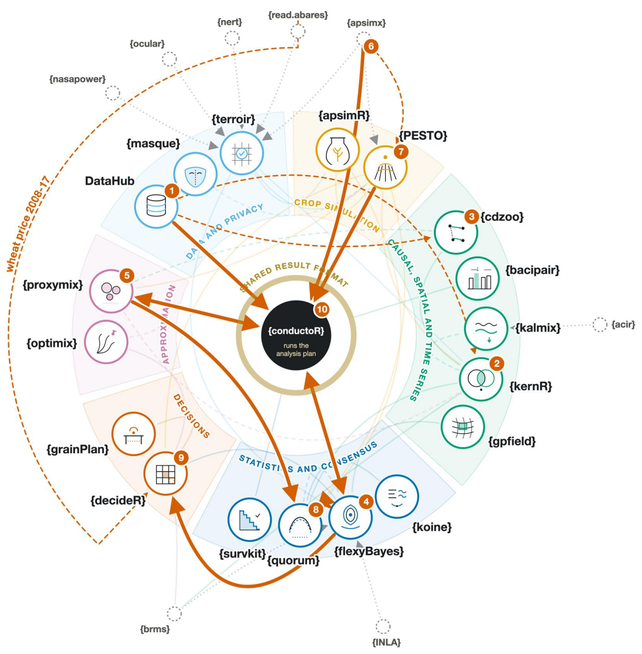


max578.github.io/orchestra

Each hexagon is one package, except DataHub. Its border colour shows its role, as listed below.

- **Data and privacy.** {terroir}: climate, soil and farm data with their sources · {masque}: synthetic copies of private data · DataHub: open-data catalogue
- **Causal, spatial and time series.** {kernR}: tests whole distributions · {cdzoo}: causal discovery · {bacipair}: before and after comparisons · {gpfield}: maps a field · {kalmix}: series over time
- **Approximation and optimisation.** {proxymix}: simplifies distributions · {optimix}: chooses an optimiser
- **At the centre.** {conductoR}: runs an analysis plan
- **Crop simulation.** {apsimR}: runs APSIM from R · {PESTO}: fits a crop simulator to field data
- **Statistics and consensus.** {flexyBayes}: Bayesian models for field trials · {koine}: an independent second analysis · {quorum}: combines estimates · {survkit}: time to an event
- **Decisions.** {decideR}: decisions under uncertainty · {grainPlan}: plans for growers

One study through the orchestra



1. **DataHub** provides the pairs of trials from the open-source dataset.
2. **{kernR}** checks that Early and Main trials matched before sowing.
3. **{cdzoo}** looks for causes of sowing date without the graph.
4. **{flexyBayes}** fits the main model and gives the posterior the decision uses.
5. **{proxymix}** checks the answer on the pair differences, and agrees.
6. **{apsimx}** writes each site's weather and simulation files and runs APSIM at both sowing dates.
7. **{PESTO}** calibrates APSIM to each pair's site.
8. **{quorum}** compares the three analyses.
9. **{decideR}** uses the Bayesian posterior and turns the yield gain and the wheat price of the study seasons into a decision.
10. **{conductoR}** runs the study as one plan, and logs each dispatch and collect.

Colours are as declared in the list above. Dotted circles are packages from outside the Orchestra, used by Orchestra members. Bold arrows are the handoffs {conductoR} recorded in the presented study, and dashed arrows ran beside it. {asreml} serves as the benchmark and is not drawn.

How the packages work together

- **One shared philosophy**. The Orchestra works together to take a grain-production question from open data to a decision a grower can act on.
- **One plan.** In its present form, {conductoR} runs an analysis algorithmically, reading from a YAML file, dispatching analytical tasks and re-distributing results. The study on this poster ran this way.
- **One data and result format.** Each dataset is accepted by {conductoR} and dispatched further in a unified format. Each result contains its estimate, its uncertainty range and its data source, so the next package can check it and use it either directly or through {conductoR}.
- **One install.** The sixteen public packages install with one call:

```
install.packages(c("apsimR", "bacipair", "cdzoo", "conductoR",
  "decideR", "gpfield", "grainPlan", "kalmix", "kernR", "koine",
  "masque", "optimix", "PESTO", "proxymix", "survkit", "terroir"),
  repos = c("https://max578.r-universe.dev",
            "https://cloud.r-project.org"))
```

Does wheat sown earlier yield more?

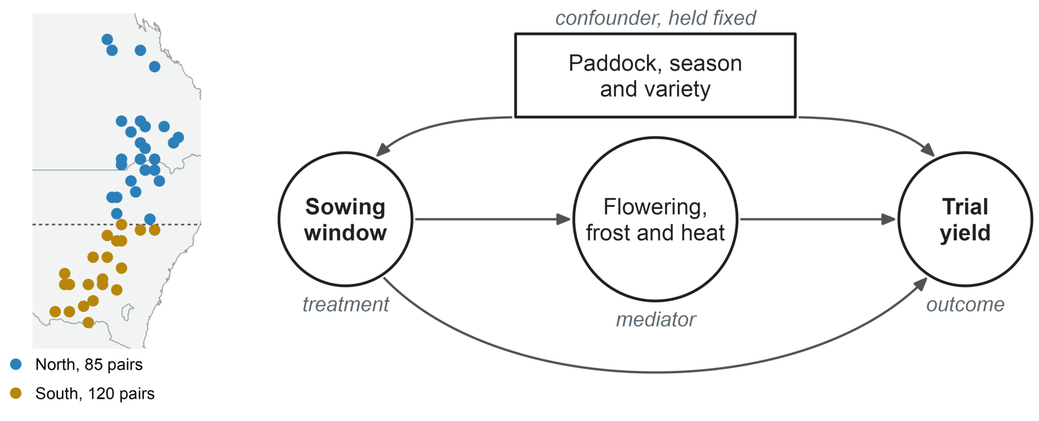
For the dataset, some sites sow wheat twice in a season, side by side: an Early series ahead of the usual sowing window and the Main series within it. The study uses these same-paddock pairs in NSW and Queensland, 2008 to 2017.

1. Pair. An Early and a Main trial of the same season on the same paddock form a pair, fixed before any yield was read: 205 pairs at 46 sites, sown 7 to 52 days apart (median 19).

2. Compare. Each variety grown in both trials gives one difference, Early yield minus Main yield. Aggregated across the whole study, this leads to the Average Treatment Effect (ATE) estimation. The same differences, aggregated separately for northern and southern paddocks, lead to Conditional ATE (CATE).

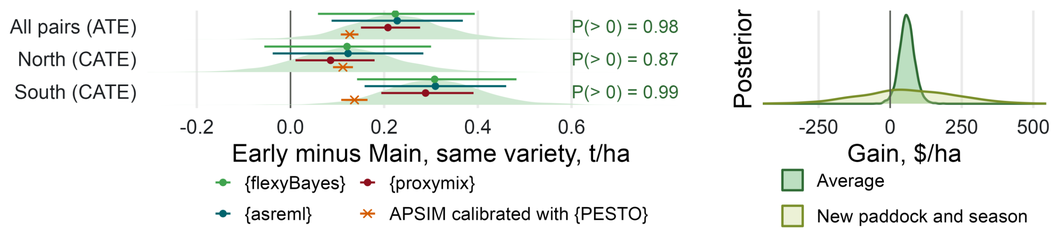
3. Check the match. {kernR}'s permutation test found no evidence of a difference between Early and Main trials in nitrogen applied by sowing.

4. **Negative control.** 82 pairs sown under a week apart differed by 0.05 t/ha (-0.03 to 0.13), so part of the gap may come from causes other than time spread.



Left: sites north and south of 31.5° S. ***Right:*** the Directed Acyclic Graph (DAG, Hernán and Robins, 2020), simplified for expositional purposes. Paddock, season and variety normally affect both **Sowing window** (the treatment) and **Trial yield** (the outcome) and are classical confounders. For the purposes of estimating causal effect of Early versus Main time of sowing on yield, these factors have been fixed through the matched pairs design, hence eliminating their confounding influence. Flowering, frost and heat themselves depend on the date and are not adjusted for. The lower arrow in the DAG transmits the date's other effects: establishment, nitrogen, and rain before sowing.

On the same paddock and in the same season, the same variety is expected to yield 0.22 t/ha (6% of the 3.8 t/ha Main-window mean) more when sown in the Early window than in the Main window. Further, a gain of 0.31 t/ha is expected in southern NSW, but only 0.12 t/ha in the northern paddocks. For the latter, the 90% interval covers zero (-0.06 to 0.30), so a gain there is not established. Utilising the Bayesian posterior, {flexyBayes} puts the chance of a gain there at 87%.



Bars and ranges are 90% intervals; shaded curves are {flexyBayes} posteriors.

- **The verdict.** Early sowing gained 0.22 t/ha, worth $58/ha ($15 to $103) at $261/t (ABARES wheat gross unit value, study seasons), provided early sowing does not cost extra. The average gain is positive with 98% probability. For a new paddock and season, which add their own variation, it is 65%.
- **Four estimates.** {flexyBayes}, the main analysis: 0.22 t/ha (0.06 to 0.39). {asreml} agrees: 0.23 (0.09 to 0.37). {proxymix}, without the normality assumption, agrees: 0.21 (0.15 to 0.28). APSIM, calibrated with {PESTO}, expected 0.13 t/ha (0.11 to 0.15).

Get it and explore

Sixteen packages are open source (MIT) on max578.r-universe.dev; {flexyBayes} and {quorum} are in development. Every package, with its role and links: max578.github.io/orchestra/members.

References

ABARES (2026). *Agricultural commodities: September quarter 2026*, statistical tables, Table 6. Australian Bureau of Agricultural and Resource Economics and Sciences.

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

Sparks, A. (2026). *read.abares: Read data from ABARES and ABS* (R package version 3.0.0). https://docs.ropensci.org/read.abares/

The VSNi Team (2026). *asreml: Fits linear mixed models using REML* (R package version 4.2.0.482). VSN International.

GRDC Grains Research & Development Corporation · Adelaide University · Curtin University · The University of Queensland Australia

This work was done under AAGI project AAGI-AU-RD-CSL, Causal learning for optimising industry relevant traits in Australian grain production systems. The support of the Grains Research and Development Corporation (GRDC) provided through the Analytics for the Australian Grains Industry (AAGI) project is acknowledged.