

final report

Project code: L.EQT.1624

Prepared by:

Garth Tarr and Kevin Wang Polkinghornes Pty Ltd

Date published:

9 June 2017

PUBLISHED BY Meat and Livestock Australia Limited PO Box 1961 NORTH SYDNEY NSW 2059

Analytics for Industry (Statistical) Consultancy Agreement

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

This publication is published by Meat & Livestock Australia Limited ABN 39 081 678 364 (MLA). Care is taken to ensure the accuracy of the information contained in this publication. However MLA cannot accept responsibility for the accuracy or completeness of the information or opinions contained in the publication. You should make your own enquiries before making decisions concerning your interests. Reproduction in whole or in part of this publication is prohibited without prior written consent of MLA.

Executive summary

The meat processing and pastoral industries are rapidly entering the digital age and are accumulating large and complex data bases from properties, feedlot and processing systems. The challenge is to integrate this data into the business using appropriate analyses to create value for the business and flow on benefits to industry.

Given that much of the data is sensitive it is most effective that analysis be done in-house, so the data handling skills of the relevant staff need to be advanced in order to handle the large volume of data which will be delivered back to the processing and production sectors.

Currently there are a number of courses available commercially however these are not targeted to agriculture and tend to be very expensive which deters companies from investing in these courses. Hence MSA/MLA saw a need to develop and run a course targeted to the data input and problems specific to the processing and pastoral sectors. It would provide a mechanism for current staff to upskill and better service company needs, with a longer-term outcome being that such courses will be able to be run in the future which will develop a pool of users with relevant skills who are focused on agricultural problems.

This project developed and delivered statistical training to company staff on a cost recovery basis to develop the data handling/analysis skills of the meat processing and pastoral industries. The course is aimed at introducing students to the possibilities that are available in data analysis and visualisation of data by providing them with basic R skills.

The initial training course has received positive feedback with requests to extend to follow up courses and intensive 'bootcamp' style retreats focussing on specific agriculture problems. With appropriate communications and advertising it is expected that future courses could be charged at a more commercial rate.

Table of contents

1	Background	4
2	Project objectives	4
3	Methodology	5
4	Results	5
5	Discussion	10
6	Conclusions/recommendations	11
7	Key messages	11
8	Bibliography	11
9	Appendix	12

1 Background

The meat processing and pastoral industries are rapidly entering the digital age and are accumulating large and complex data bases from properties, feedlot and processing systems.

These industries are not well prepared for this onslaught of data in terms of staff with the appropriate skills and tools to collate, prepare and analyse these complex databases.

To date the pastoral and meat processing companies have employed graduates with a rudimentary knowledge of data handling (largely excel, with some basic knowledge of other statistical packages) and a passion for agriculture. Data handling skills for these staff need to be advanced if they are to handle the large volume of satellite, pasture, animal performance and carcass data which will be delivered back to the processing and production sectors. The challenge is to integrate this data into the business using appropriate analyses to create value for the business.

In addition, the recent push for objective carcass measurement (OCM) will accelerate the need for data processing skills to be developed in processing companies to underpin development of value based training systems. Both the OCM and lamb supply chain projects will have a need for supply chain officers to have advanced data handling skills to help processing works undertake analyses inhouse.

Given that much of the data is sensitive it is critical that these analyses be done in-house where the

- objectives in the analyses be more clearly articulated
- there is continuity in the types of analyses being performed

It is much more cost efficient to undertake these analyses in-house than sub-contracting to external providers.

2 Project objectives

To deliver statistical training to company staff to develop the data handling/analysis skills of the meat processing and pastoral industries. The course is aimed at introducing students to the possibilities that are available in data analysis and visualisation of data by providing them with basic R skills.

Whilst the data problems being presented to staff are diverse, the course will achieve several things; firstly, it will show them how to address a statistical problem and break it up into logical steps. The R packages introduced to the students will then assist them to convert these steps into statistical code, running the code and then producing an output.

The longer-term advantage of running such courses is that it develops a pool of users that have focused on agricultural problems. This like-minded group will provide the bases for a user group that can meet regularly to share problems and solutions. It will upskill the group so that in addition to sharing problems they will begin to use the web resources to further refine their skills. This will not occur without delivering basic courses to introduce students to the packages and get them started.

The course would be run on a cost recovery basis and specifically focus on handling mapping, pasture, live animal, carcass quality and carcass yield data. There are a number of courses that are provided by the commercial sectors however these are not targeted to the agricultural applications and tend to be very expensive which deters companies from investing in these courses. The logic for MSA/MLA to run the course would be that it was targeted to the data input and problems which are specific to the processing and pastoral sectors. It would provide a mechanism for current staff to upskill and better service company needs.

The course frequency would be determined by the demand from industry. Obviously, there would be a call for more frequent courses initially and then perhaps one or at most two a year.

There are a number of software packages are capable of handling complex data sets. These include "R", "SAS", "SPSS", "GENSTAT" and "ASREML" etc. Whilst all these programs tend to have a similar capability in terms of the statistical analyses they undertake, they do differ greatly in both the visual quality of the outputs and the initial cost of the programs.

Among R's many features is the capability to output high quality static plots and figures. Increasingly, there are many interactive visualisations that can also be created with R, including geospatial mapping applications and web-based interfaces so that end-users can query data sets and analysis without needing to know any R programming.

3 Methodology

One 5 day statistical workshop

The 5 day workshop was run from 15-17 May and 5-6 June in the boardroom at the MLA Brisbane office. There were 7 participants taught by Dr Garth Tarr (University of Newcastle) and Mr Kevin Wang (University of Sydney). A short biography for both presenters can be found in the appendix.

A course webpage (https://garthtarr.github.io/meatR) has been developed to host the course materials and provides a reference portal for past participants. It will continue to be added to and refined over time. The majority of course preparation work is in building, updating and adding to the course webpage.

One 2 day follow up workshop

The 2 day workshop will be held at the Rydges Fortitude Valley from 29-30 June.

Ongoing statistical support to MSA

- Upkeep of the GCheck grader analysis shiny app
- Assistance to Kevin McDermott in creating the MSA non-compliance mapping app
- Meeting with MSA staff in Brisbane on Wednesday 28 June to discuss reporting needs.

4 Results

At the end of the 5 day workshop, the participants each gave presentations an analyses they'd performed in R on their own data sets. Example analyses include:

- Interactive shiny app to explore MSA non-compliance spatiotemporally.
- Interactive shiny app to explore infra-red heat measurements taken on different animals and superimposing summary statistics to help inform the best way to summarise the data
- Merging data sets and building charts to enable comparison of eating quality scores between family members.
- Exploring many blood measurement variables graphically, identifying questionable observations and linking measurements to eating quality measurements.
- Wrangling walk-over-weighting data into a clean data frame; identifying temporal trends, such as calving; constructing growth curves for calves.



Image 1: Workshop participants

Feedback

It's essential to collect participant feedback to help improve the structure and delivery of the workshop over time. A survey was created to collect participant feedback and participants were strongly encouraged to provide their feedback on the final day. The feedback form can be found here

https://docs.google.com/forms/d/e/1FAIpQLScVkJ9yWGbLLG4_nCxR1f4IZAX156wcdJbNHL8oUsdfk4 YLLQ/viewform 4 out of 7 participants have responded so far, waiting on the final 3 (a reminder email has been sent). So far we have 100% agreement or strongly agree to each of the statements below. The free form comments that they provided have also been included.

About the presenters:

- The presenters communicated the information clearly
- The presenters made the subject matter compelling
- The presenters were able to answer questions
- The presenters were friendly and approachable

Comments:

They were patient and able to demonstrate examples with our own data which made the course so much easier to understand

amazing teachers, very supportive and fun - and sympathetic towards my minimal understanding of statistics!

Real teaching skills

About the material:

- The workshop was relevant to me
- The workshop was interesting
- The workshop met my purpose in attending
- I want to tell others about what was presented
- The website resources were useful
- The opportunity to work on my own data was useful

Comments:

Starting with basic code then continually building on it with my own data set was brilliant. Also so helpful that the participants had relevant research so my own (meat / animal based research) so learnt how to evaluate things differently.

Provided a clear, logical approach to analysing data. Gave me the confidence and skills to analyse my own data.

Lessons adjusted for a first step with R

About the workshop in general:

- The duration of the workshop was appropriate
- The workshop was well organised
- I would recommend this workshop to others

Comments:

Good amount of time each day. Online notes excellent.

You could honestly spend another month working through individual analyses - so this workshop gave me a starting point to work on my own, with the skills and understanding so I can ask for help if needed.

Not necessarily a criticism more of an observation: I found the first three days really, really full on, it was good though in that it got us to a place where we could work on our own projects after that with Garth input, but perhaps have this a little more spread out. I know this may not be possible though as people have other commitments. I found the pace of day 4/5 much more enjoyable.

Good time management between the lessons and the work on our data.

What were the best aspects of the workshop?

The class size was brilliant. Small enough to get lots of one on one attention yet the other researchers work was relevant so learning from peers also very helpful. Case study presentations was very helpful to get feedback and learn new code from peers.

Garth and Kevin

Working with our own data and the interaction with Garth between sessions. It also demonstrated things that I did not think were feasibly possible for me ever to do are now possible with regard to some of the interactive graphs and data analysis

The first steps in R with lessons, then the opportunity to work on our data and to show our use with the presentation.

What aspects of the workshop need improvement?

I would have preferred a month in-between the first and second weeks of the course to allow more time to play with R.

nil.

One suggestion may be to formally set up an R-buddy system?? This would require matching people who are at the same skill levels who can help each other out after the 5 days are done.

Please give a couple of concrete examples of how you will be able to apply what was learnt in the workshop to improve your organisation's decision making abilities and drive change.

Graphing potential of R will revolutionise our teams graphing. Currently slow via excel. This will also remove potential data errors when moving across to excel.

The html markdown is a very effective tool to share statistics with adjunct supervisors.

Initial analyses of raw data (structured approach) that is more efficient and lovely to look at.

During the course a Heat Map was developed which will greatly aid in identifying problem areas meaning workshops/open days/comms can be effectively targeted

I can now work on automating some processes that were previously manual which will mean a saving of staff hours and also giving a constant, reliable result

Try to replace Excel/JASP use for data analysis with R

Do you have any suggestions about the best way to continue to support and grow your analytics skills in R?

An online group and annual bootcamps (small groups) would be great.

bootcamps, call outs for targeted workshops (e.g. if there is a research team working on a dataset - gets everyone on the same page)

I do like the idea of a 6 monthly bootcamp for 2/3 days. Not sure how well the skype would work.

An online group to share our difficulties and help each other

Any other comments?

Thank you. It was brilliant. It has revolutionised my understanding of how I should approach data for analysis. I would love a follow up course on model selection and model design.

This was one of the best workshops I have ever attended. Thank you so much!

This course has definitely improved what I can offer my company as well as increasing the enjoyment I get from working with data and data visualisation.

Thank you, this workshop was really helpful for me!!

5 Discussion

The demographics of the 5 day workshop was highly skewed towards young academics and research students, many of whom already have strong ties to MSA through their supervisors. This meant that the workshop went further down the modelling/statistics path than the previous workshop.

Participants were charged \$2500 for the 5 days and had to cover their own transport and accommodation costs. This may seem expensive, but is at the cheaper end of the spectrum for R courses.

Private training companies such as Presciient charge between \$2000 and \$2600 (plus GST) for twoday introduction to R courses.¹ The University of Technology Sydney charges \$800 for a one-day introduction to R course; \$900 for a one-day intermediate R course and \$1000 for a one-day advanced R course - total of \$2,700 for three days.² UNSW charges \$500/day for its R courses.³ For the Presciient, UTS and UNSW courses there would be many more than 10 participants, which means less personalised attention and less opportunity for participants to make progress on their data analysis work.

¹ Source: http://presciient.com/current-courses/

² Source: https://www.uts.edu.au/research-and-teaching/our-research/advanced-analytics-institute/education-and-research-opportuniti-3

³ Source: http://www.analytical.unsw.edu.au/facilities/stats-central-2

6 Conclusions/recommendations

Recommendation 1

Continue with the 5 day format for initial Analytics for Industry training workshops. The dates should be finalised months in advance and advertised widely. The recommended number of participants remains 7-10. If the course is held in Brisbane again, the boardroom at the MLA offices was a perfect venue. While we're perfecting the course delivery, the price can remain unchanged, however, there is scope to increase the fees in the future. For example, \$3000 for 5 days is not at all unreasonable.

Recommendation 2

There seems to be demand for annual boot camps or refresher courses to build on and consolidate the techniques learnt in the initial workshop. A possible model for this type of course is the "writing retreat" style. Under this "data retreat" model, participants bring one or more analysis projects with them to a location that is removed from their usual environment and the usual distractions so they make significant progress with their analyses.

During the retreat, participants work on a specific writing task in a conducive working environment, where the facilitator runs a series of R workshops, intermingled with ample time for participants to practice the skills learnt and implement similar approaches on their own data set.

Retreat locations are typically a short distance outside major cities. We would require:

- Enough bedrooms to house all participants
- Good internet access
- A dining room table (or similar) large enough for everyone to work at
- A white wall near the table (for projection)
- A data projector

The aim of the retreat would be to improve research outcomes, increase R coding confidence, build better interdisciplinary networks and understanding and facilitate collegiality.

7 Key messages

- The 5 day workshop was a clear success. I have already been contacted about hosting more in the future.
- Fix course dates months before hand and advertise widely.
- Move towards a "data retreat" model for annual refresher courses/boot camps.

8 Bibliography

Tarr (2017). Analytics for Industry. Web resource. <u>https://garthtarr.github.io/meatR/</u>

9 Appendix

Short biographies of the presenters

Dr Garth Tarr (University of Newcastle)

Garth received his PhD in Mathematical Statistics from the University of Sydney and has held postdoctoral positions at the University of Sydney and the Australian National University. Garth is an expert R user, he has developed three R packages and contributed to a number of others. He is currently a lecturer at the University of Newcastle. His diverse interests include data visualisation, meat science, robust statistics, model selection, econometric modelling (including value based marketing), educational research and biostatistics. He has received citations and awards for his teaching and was awarded first prize for the best oral presentation by a student at the Australian Young Statisticians Conference (2013) as well as the EJG Pitman Prize for the most outstanding talk presented by a young statistician at the Australian Statistical Conference (2012).

Kevin Wang (University of Sydney)

Kevin is currently a PhD candidate and Postgraduate Teaching Fellow in the School of Mathematics and Statistics at the University of Sydney. Kevin's main research area is in statistical bioinformatics and is developing novel methods brought forward by high dimensional biomedical data. A central focus of his current research focuses on the increasingly popular boutique array platform and its application both as a validation platform for biomarkers for patients in melanoma studies. His other interests include data visualisation, statistical computing, model selection and biostatistics.