A big data approach to predicting grain crop yield

Presented by:
Liana Johnson

Other team members:
- Edward Jones
- Thomas Bishop
- Niranjan Acharige
- Sanjeewani Dewage
- Patrick Filippi
- Sabastine Ugbaje
- Thomas Jephcott
- Stacey Paterson
- Brett Whelan
A problem

- Farms across Australia have large amounts of unused data
- This data may be difficult to utilise to make management decisions
  - Different formats and located in a variety of repositories
  - Different spatial and temporal resolutions
How can we transform all of these disparate data streams into something useful, and then inform management decisions?
An opportunity

- Hackathon run by CSIRO and Lawson Grains
- Provided us with an abundance of spatial agricultural data (~15,000 ha)
  - Yield
  - EM and gamma surveys
  - Management data
  - Soil tests

- There is a also a lot of publicly available spatio-temporal environmental data
  - Rainfall, soil map of Australia, remote sensing etc.
Available spatial and temporal data

Provided farmer data:
- Yield – 10 m (space & time)
- Radiometrics – 10 m (space)
- EM surveys – 10 m (space)
- Soil test results (space & time)

Project-created data
- Clay and sand soil maps – 10 m (space)

Publicly-available data:
- TERN – soil maps ~ 90m (space)
- NDVI – 250 m & 16 day (space & time)
- Rainfall forecasts – monthly (time)
- Rainfall received – 5km & daily (space & time)
Approach: Our predictive model

- Using this farmer data and publicly-available datasets, we created a model to predict the yield for these three crops in the production rotation:
  - Wheat
  - Barley
  - Canola

- Modelling method: Machine learning (Random Forest)
  - Data-driven rather than mechanistic

- The idea is to use the data from all fields and years to predict yield within each individual field for a farm
Modelling for decision support

- 3 different predictive yield models for 3 important time points to inform key management decisions:

1. APRIL MODEL
   » To provide suggestions for variable sowing N rates

2. JULY MODEL
   » To provide suggestions for variable top-dress N rates

3. SEPTEMBER MODEL
   » To determine final yield prediction

- More data becomes available as the season progresses
Results – Map example

10 m resolution  Hakea - July 2015

5,500 Ha
Results – paddock resolution model assessment

1. Predict yield within a paddock, all years of previous yield data excluded

2. Predict yield within a paddock, previous yield data included

<table>
<thead>
<tr>
<th>TIME</th>
<th>APRIL</th>
<th>JULY</th>
<th>SEPTEMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV PADDOCK PREDICTIONS</td>
<td>RMSE (t/ha)</td>
<td>LCCC</td>
<td>RMSE (t/ha)</td>
</tr>
<tr>
<td>1) Without previous yield</td>
<td>0.64</td>
<td>0.19</td>
<td>0.63</td>
</tr>
<tr>
<td>2) With previous yield</td>
<td>0.42</td>
<td><strong>0.89</strong></td>
<td>0.39</td>
</tr>
</tbody>
</table>

LCCC of 1 characterises a perfect fit

Including previous data from the prediction paddock results in a better prediction

Models are very good
We have a model that predicts yield, but how can we make this **useful** and **user-friendly** for growers and consultants to inform management decisions?
USYD AgData Challenge

Select a paddock
Aggregate:
Oldfield
Paddock:
Doyle_Plats

Select and run a model
Model:
Topdress (Jul)

Commodity price ($/tonne):

150 240 600

Answer: Our user interface
Conclusions

- We used large amounts of agricultural and environmental data to:
  - accurately predict wheat, barley and canola yield across a collection of farms
  - developed a user-friendly application for farmers to aid key management decisions

What next?

- More data, more accurate predictions- model will improve over time
  ⇒ potential to integrate fine spatial resolution remote sensing (drones, Landsat, Sentinel etc.)
- With more consistent data collection it will be useful for:
  - Identifying yield gaps
  - Sowing seeding rate
  - Variable rate application – Lime, P, K & S
  - Futures contracts and market speculation

For any cropping system