Predicting early bulking in potatoes

Study group Mathematics with Industry 2016

Fetsje Bijma, Alessandro Di Bucchianico, Eric Cator, Henk Don, Patrick Hafkenscheid, Jakub Nowotarski, Bijan Ranjbar-Sahraei

January 29, 2016
Company background

Supplies seed potatoes to various clients.
Company background

- Supplies seed potatoes to various clients.
- Large R&D department, analysing and developing new varieties of potatoes.
Company background

- Supplies seed potatoes to various clients.
- Large R&D department, analysing and developing new varieties of potatoes.
- Potatoes bred for various applications.
What is a tuber?
What is a tuber?

- Square size
- Weight
What is a tuber?

- Square size
- Weight
- Number of tubers

Study group Mathematics with Industry 2016
Predicting early bulking in potatoes
Early bulking

Study group Mathematics with Industry 2016
Predicting early bulking in potatoes
Early bulking

Study group Mathematics with Industry 2016
Predicting early bulking in potatoes
Why early bulking?

**More profitable** New harvest as early as possible.
Why early bulking?

- **More profitable** New harvest as early as possible.
- **More flexibility with scheduling** Takes less time until harvest.
Why early bulking?

- More profitable New harvest as early as possible.
- More flexibility with scheduling Takes less time until harvest.
- Climate factors Influences of rain, humidity etc.
Research questions

Question 1
How to model tuber growth and predict which varieties are more likely to bulk early?

Question 2
How to identify important genetic properties (SNPs) for early bulking?
In ’15 data for individual tubers:
Length, width, height, square size, weight and volume.
See table, each line is one tuber.

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>E</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VEL</td>
<td>RAS_KODE2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ras001</td>
<td>23.7</td>
<td>13.4</td>
</tr>
<tr>
<td>3</td>
<td>ras001</td>
<td>39.7</td>
<td>42.4</td>
</tr>
<tr>
<td>4</td>
<td>ras001</td>
<td>26.5</td>
<td>18.7</td>
</tr>
<tr>
<td>5</td>
<td>ras001</td>
<td>22.8</td>
<td>12.3</td>
</tr>
<tr>
<td>6</td>
<td>ras001</td>
<td>28.4</td>
<td>19.3</td>
</tr>
<tr>
<td>7</td>
<td>ras001</td>
<td>22.0</td>
<td>10.2</td>
</tr>
<tr>
<td>8</td>
<td>ras001</td>
<td>26.3</td>
<td>15.3</td>
</tr>
</tbody>
</table>
Dataset: tuber data

In ’15 data for individual tubers:
Length, width, height, square size, weight and volume.
See table, each line is one tuber.

For previous years (’11–’14) only summarized data:
Number of tubers and total weight per square size category.
Experimental design

100 varieties of tubers.
Experimental design

100 varieties of tubers.
4 Different harvest times.
**Experimental design**

- **100** varieties of tubers.
- **4** Different harvest times.
- **2** Replicates per harvest time.
### Experimental design

<table>
<thead>
<tr>
<th>Harvest time</th>
<th>Rep. 1</th>
<th>Rep. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1 = 61$</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>$t_2 = 80$</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>$t_3 = 101$</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>$t_4 = 133$</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
</tbody>
</table>

- **100** varieties of tubers.
- **4** Different harvest times.
- **2** Replicates per harvest time.

Study group Mathematics with Industry 2016
Predicting early bulking in potatoes
Dataset: Genetic information

SNP data available for 113 of 222 varieties.
SNP data available for 113 of 222 varieties.

For each variety assigns 0-4 to a certain SNP.
Dataset: Genetic information

- SNP data available for 113 of 222 varieties.
- For each variety assigns 0-4 to a certain SNP.
- Unfortunately overlap individual data 2015 and SNP data bad (12/100 varieties known).
Dataset: Genetic information

- SNP data available for 113 of 222 varieties.
- For each variety assigns 0-4 to a certain SNP.
- Unfortunately overlap individual data 2015 and SNP data bad (12/100 varieties known).
First impressions of the tuber data

**Figure**: Jitterplot for one variety. Horizontal time, vertical square size
First impressions of the tuber data (2)

Figure: Linear relation log square size and log weight

Study group Mathematics with Industry 2016
Predicting early bulking in potatoes
Approach for questions

**Approach tuber data**

1. Joint model for log-weight and log-square size as function of time
2. Predict yield as function of time

**Approach SNP data**

1. Apply elasticnet to preselect important SNPs
2. Fine-tune selection of SNPs
### Part 1: Model for variety \( v \)

<table>
<thead>
<tr>
<th>Problem outline</th>
<th>Available data</th>
<th>Tubers</th>
<th>SNPs</th>
<th>Conclusions</th>
</tr>
</thead>
</table>

#### Part 1: Model for variety \( v \)

\[
Y_{1v}(t) = \log \text{of square size of the tuber at a time } t \\
Y_{2v}(t) = \log \text{of weight of the tuber at a time } t \\
N_{v}(t) = \text{number of tubers.} \\
\]

\( Y_{1v}(t), Y_{2v}(t) \sim N((0, 0), \varSigma_v) \).

Maximum likelihood estimation for \( \mu_v, \mu_v, \) and \( \varSigma_v \).
Part 1: Model for variety $\nu$

- $Y_{1\nu}(t) = \log$ of **square size** of the tuber at a time $t$

- $N_{\nu}(t)$ number of tubers.
Part 1: Model for variety $\nu$

- $Y_1^{\nu}(t) = \log$ of **square size** of the tuber at a time $t$
- $Y_2^{\nu}(t) = \log$ of **weight** of the tuber at a time $t$
- $N^{\nu}(t)$ number of tubers.
- $(\varepsilon_1^{\nu}(t), \varepsilon_2^{\nu}(t)) \sim \mathcal{N}((0, 0), \Sigma^{\nu})$. 
Part 1: Model for variety \( \nu \)

- \( Y_1^\nu(t) = \log \text{of square size} \) of the tuber at a time \( t \)
- \( Y_2^\nu(t) = \log \text{of weight} \) of the tuber at a time \( t \)

Individual model per tuber

\[
(Y_1^\nu(t), Y_2^\nu(t)) = (1\ t\ t^2\ N^\nu(t)) \begin{pmatrix} \beta_{11}^\nu & \beta_{12}^\nu \\ \beta_{21}^\nu & \beta_{22}^\nu \\ \beta_{31}^\nu & \beta_{32}^\nu \\ \lambda_1^\nu & \lambda_2^\nu \end{pmatrix} + (\epsilon_1^\nu(t), \epsilon_2^\nu(t)).
\]

- \( N^\nu(t) \) number of tubers.
- \( (\epsilon_1^\nu(t), \epsilon_2^\nu(t)) \sim \mathcal{N}((0, 0), \Sigma^\nu). \)
- Maximum likelihood estimation for \( \beta, \lambda \) and \( \Sigma \).
# Part 1: Prediction

Plug $N^\gamma(t)$ into the model.
Part 1: Prediction

- Plug $N^v(t)$ into the model.
- Compute for each $t$: expected total weight of big tubers.
Part 1: Prediction

Plug $N^v(t)$ into the model.

Compute for each $t$: expected total weight of big tubers.

Total weight of potatoes with $d>45$, all varieties

Study group Mathematics with Industry 2016
Predicting early bulking in potatoes
Part 2: Model

Linear regression model with \((l_1, l_2)\)-penalty (elasticnet):

\[
\begin{pmatrix}
W_1 \\
\vdots \\
W_{113}
\end{pmatrix}
= \begin{pmatrix}
x_{1,1} & \ldots & x_{1,11673} \\
\vdots & : & \vdots \\
x_{113,1} & \ldots & x_{113,11673}
\end{pmatrix}
\begin{pmatrix}
\gamma_1 \\
\vdots \\
\gamma_{11673}
\end{pmatrix}
+ \begin{pmatrix}
\epsilon_1 \\
\vdots \\
\epsilon_{113}
\end{pmatrix}
\]

\(W_i\) is total weight of all tubers with square size 45+ of variety \(i\).
\(x_{i,j}\) is the SNP-data for variety \(i\).
Part 2: Model

Linear regression model with $(l_1, l_2)$-penalty (elasticnet):

$$
\begin{pmatrix}
W_1 \\
\vdots \\
W_{113}
\end{pmatrix} =
\begin{pmatrix}
x_{1,1} & \cdots & x_{1,11673} \\
\vdots & \ddots & \vdots \\
x_{113,1} & \cdots & x_{113,11673}
\end{pmatrix}
\begin{pmatrix}
\gamma_1 \\
\vdots \\
\gamma_{11673}
\end{pmatrix} +
\begin{pmatrix}
\varepsilon_1 \\
\vdots \\
\varepsilon_{113}
\end{pmatrix}.
$$

- $W_i$ is total weight of all tubers with square size 45+ of variety $i$.
- $x(i,:)$ is the SNP-data for variety $i$. 

Study group Mathematics with Industry 2016
Predicting early bulking in potatoes
Part 2: Model

Linear regression model with \((\ell_1, \ell_2)\)-penalty (elasticnet):

\[
\begin{pmatrix}
W_1 \\
\vdots \\
W_{113}
\end{pmatrix}
= 
\begin{pmatrix}
x_{1,1} & \ldots & x_{1,11673} \\
\vdots & & \vdots \\
x_{113,1} & \ldots & x_{113,11673}
\end{pmatrix}
\begin{pmatrix}
\gamma_1 \\
\vdots \\
\gamma_{11673}
\end{pmatrix}
+ 
\begin{pmatrix}
\varepsilon_1 \\
\vdots \\
\varepsilon_{113}
\end{pmatrix}.
\]

\(W_i\) is total weight of all tubers with square size 45+ of variety \(i\).

\(x(i, :)\) is the SNP-data for variety \(i\).

Penalty forces selection of \(\gamma\)’s.
Part 2: Selected SNPs, spurious effects

HZPC_SNP_02291

HZPC_SNP_07998

Study group Mathematics with Industry 2016
Predicting early bulking in potatoes
Part 2: Selected SNPs, positive and negative effects

Ad hoc criterion to select $\gamma$’s
<table>
<thead>
<tr>
<th>Questions</th>
<th>Insights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Question 1</strong></td>
<td><strong>Insights</strong></td>
</tr>
</tbody>
</table>
| How to model tuber growth and predict which varieties are more likely to bulk early? | 1. Linear relation between log-weight and log-square size  
2. Variance stabilization through log plots  
3. Moderate plot effect  
4. Number of tubers stabilizes after second harvest time. |
Key insights 1

Question 1
How to model tuber growth and predict which varieties are more likely to bulk early?

Insights
1. Linear relation between log-weight and log-square size
Question 1
How to model tuber growth and predict which varieties are more likely to bulk early?

Insights
1. Linear relation between log-weight and log-square size
2. Variance stabilization through log – log plots
Question 1
How to model tuber growth and predict which varieties are more likely to bulk early?

Insights

1. Linear relation between log-weight and log-square size
2. Variance stabilization through log–log plots
3. Moderate plot effect
Question 1

How to model tuber growth and predict which varieties are more likely to bulk early?

Insights

1. Linear relation between log-weight and log-square size
2. Variance stabilization through log–log plots
3. Moderate plot effect
4. Number of tubers stabilizes after second harvest time.
Key insights 2

Question 2

How to identify important genetic properties (SNPs) for early bulking?

Insights
**Key insights 2**

**Question 2**

How to identify important genetic properties (**SNPs**) for early bulking?

**Insights**

1. Do not include **SNPs** that are almost constant for all varieties.
Key insights 2

Question 2

How to identify important genetic properties (SNPs) for early bulking?

Insights

1. Do not include SNPs that are almost constant for all varieties.
2. About 1% of SNPs show an effect.
3. Both positive and negative effects occur.
### Future research

- **Investigate model accuracy**
Future research

- Investigate model accuracy
- Perform sensitivity analysis for harvest times
Future research

- Investigate model accuracy
- Perform sensitivity analysis for harvest times
- Extend the model to include plot effect
Future research

- Investigate model accuracy
- Perform sensitivity analysis for harvest times
- Extend the model to include plot effect
- Use different shapes of time-profiles (e.g. $\sqrt{t}$)
Future research

- Investigate model accuracy
- Perform sensitivity analysis for harvest times
- Extend the model to include plot effect
- Use different shapes of time-profiles (e.g. $\sqrt{t}$)
- Extend the model to capture group effects of SNPs
Future research

- Investigate model accuracy
- Perform sensitivity analysis for harvest times
- Extend the model to include plot effect
- Use different shapes of time-profiles (e.g. $\sqrt{t}$)
- Extend the model to capture group effects of SNPs
- Collect more data on the Genotype side (increase the number of varieties)
Future research

- Investigate model accuracy
- Perform sensitivity analysis for harvest times
- Extend the model to include plot effect
- Use different shapes of time-profiles (e.g. $\sqrt{t}$)
- Extend the model to capture group effects of SNPs
- Collect more data on the Genotype side (increase the number of varieties)
- Explore ways to avoid normality problems, e.g. other transformations (Box-Cox) or modeling of the joint distribution