Optimizing breeding schemes //

Manual

Estimating surrogates of genetic value

- This manual describes and compares different methods of estimating genetic value, with discussions of theory, strengths and weaknesses and practical examples.
- Published on 01/08/2020

excellenceinbreeding.org/toolbox/tools/eib-breeding-scheme-optimization-manuals

Excellence in Breeding PLATFORM

ExcellenceinBreeding.org

Estimating surrogates of genetic value

Authors

Ian Mackay / i.j.mackay@gmail.com

IMplant Consulting Ltd., Chelmsford, CM2 6HA, UK

Editors

Giovanny E. Covarrubias-Pazaran / g.covarrubias@cgiar.org

Breeding Optimization Lead, CGIAR Excellence in Breeding Platform (EiB)

Sam Storr / EiB

CGIAR Excellence in Breeding Platform

Contents

Introduction and definitions

Genetic value is the measure on which breeders most commonly rank and select among varieties, clones, families or individuals. This manual describes and compares different methods of estimating genetic value, with discussions of theory, strengths and weaknesses and practical examples.

Genetic value is typically estimated from one or more trait values (which could be marker scores), from which an estimate of the genetic value of a line, family or individual can be derived. The genetic value may be for the trait(s) initially scored, or could be for a different trait: predicting yield from markers is a good example.

In terms of classical genetics, the phenotype (P) is a linear function of genotype (G), the environment (E), and the interaction of these two. We know P and we wish to estimate G.

For the purposes of this manual, we shall refer to G as the "Genetic Value", which includes all sources of genetic determination in an individual: additive, dominance and epistatic.[1]

For simplicity in this document, unless otherwise stated, all other sources of variation are subsumed within E, including genotype x environment interactions.

G, the Genetic Value, is the estimated parameter on which we usually select.

[1] The definition of G as "Genetic Value" rather than "genotype", as it is usually interpreted, is to avoid confusion as genotype is often used to refer to genetic markers. Neither is G described as the "breeding value" since this has a precise meaning in terms of additive genetic effects and allele frequencies.

To describe different methods of estimating G, we start with the following model:

$$y_{ij} = \mu + g_i + e_{ij}$$

Where:

y_{ij} is the j^{th} observation on individual i.

μ is the mean.

g_i is the genetic value of individual i expressed as a deviation from the mean.

e_{ij} is the error of measurement, equivalently environmental noise, for observation i on individual j.

For brevity, this manual addresses the genetic value of an individual or a line. A *line* will be used to refer to any unit of selection other than an individual: such as an inbred line, a hybrid, family, clone etc. Unless specifically stated, there is no discrimination between these types.

A line, so defined, is made up of a set of individuals which need not be genetically identical, as long as we wish to estimate the genetic value of the line and not of the individuals contributing to it. For example, a line could be a full-sib family with observations collected on different family members. In this case the error of measurement – e_{ij} – incorporates the genetic deviation of each individual from the family mean; this does not usually affect the estimation of family genetic value in plant breeding as family sizes are usually large.

Taking the above into account, this manual will cover six different methods of estimating genetic value, with guidelines for when each method should be employed.

1. Estimation from a single observation

1.1 Method

$$y_{ij} = \mu + g_i + e_{ij}$$

The estimate of the genetic value from a single observation of the trait represents the simplest form of surrogate for genetic value.

1.2 Benefits

Estimation from a single observation is recommended for:

- Highly heritable traits.
- Traits scored pre-reproduction in outbreeding species.
- Traits scored post-reproduction in inbreeding species.
- High intensities of selection, when applied to single plants.

1.3 Constraints

Estimation from a single observation is not recommended for:

- Low heritability traits.
- Traits which show a lot of GxE sensitivity.
- Post-reproduction traits.

2. Estimation from a mean of several observations

2.1 Method

The arithmetic mean of a set of n observations is written as:

$$\mu+\hat{g}_i = \Sigma y_{ij} / n = \Sigma(\mu+g_i+e_{ij}) / n$$

Where:

'^' is commonly used to denote an estimate of the parameter.

μ is common to all individuals and can be ignored. In comparisons between individual means it cancels as:

$$y_{1j}/n - y_{2j}/n = \mu+ \hat{g}_1 - \mu+ \hat{g}_2 = \hat{g}_1 - \hat{g}_2$$

The contribution of the environment, or of error, to the estimate of breeding value is:

$$\Sigma e_{.j} / n$$

As the number of observations contributing to the mean increases, the precision of estimating is also increased. Means based on different numbers of observations will therefore differ in precision (**Figure 1**).

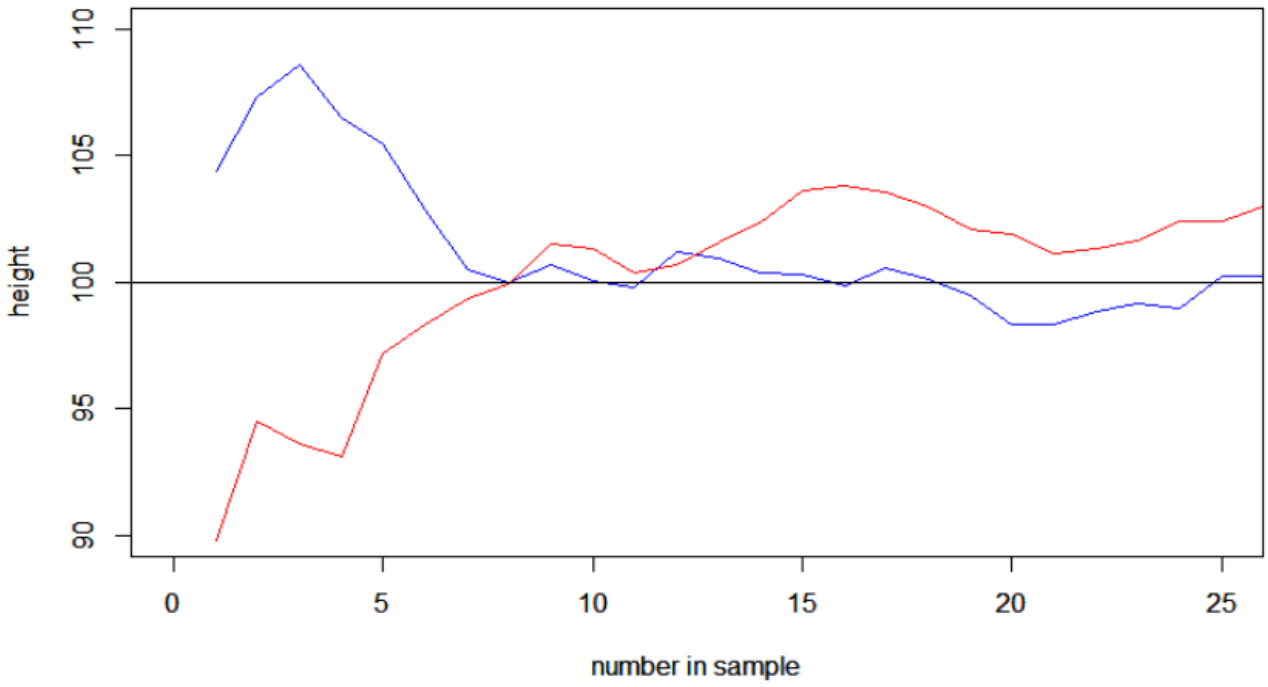


Figure 1. **Precision of means increases with sample size.** Simulated means from progressively adding one observation. Blue and red lines: true mean is 98 and 102 respectively, with error variance of 100.

2.2 Benefits

- Good for low heritability traits.
- Increased accuracy.

2.3 Constraints

- Not the best method to infer genetic value if means are based on different numbers of observations (unbalanced designs).
- Not the best method to infer genetic value if observations are made using different experimental protocols or methods.

3. Best linear unbiased estimate (BLUE)

3.1 Method

Observations on a line are commonly made on multiple replications of the line in different blocks within a trial and/or in multiple locations. Estimates of genetic value must take into account the varying contributions of these different nuisance factors to each observation.

To accommodate this, the notation for our simple model is extended from $y_{ij} = \mu + g_i + e_{ij}$ to the following:

$$y_{ijk} = \mu_i + g_j + e_{ijk}$$

There are now several different terms for μ. For example, in a trial with six blocks there would be six means. The subscripts have changed too: I for the i^{th} non-genetic effect, j for the j^{th} genetic value. Each individual observation has its own personal error e_{ijk}. The subscript k is required since multiple observations may occur for line j in environment i.

Parameters are estimated by least squares. The error sum of squares can be minimized as such:

$$\Sigma(y_{ijk} - \mu_i + g_j)^2 = \Sigma e_{ijk}^2$$

The values of u_i and g_j are called the least squares estimates, used to minimize the error, whereas g_j refers to the genetic values we require.

In designed experiments with no missing data, the least squares estimates of g_i can often still be obtained from the simple arithmetic means of the observations. This is true for a randomized complete block design for example. However, this does not apply if there are missing data or for an incomplete block design. In practice, a statistical package is commonly used to estimate BLUEs.

Matrix notation can also be used to express the BLUEs model. Matrix notation is used to help describe the more complex methods to estimate genetic value below.

We can write the above model as follows:

$$y_{ijk} = x_1\mu_1 + x_2\mu_2 + x_3\mu_3 + \ldots x_m\mu_m + x_{(m+1)}g_1 + x_{(m+2)}g_2 + x_{(m+3)}g_3 + \ldots + x_{(m+n)}g_n + e_{ijk}$$

Where:

x_i is an indicator variable. It takes the value 0 if a particular effect has no influence on an observation and a value of 1 if it does.

For example, with three blocks in a trial, the set of values {$x_1 = 0$, $x_2 = 1$, $x_3 = 0$} would indicate an observation taken on the second block ($x_2 = 1$) and not on the first or third ($x_1 = 0$, $x_3 = 0$).

Although this formula is more drawn-out, it is also more flexible. For example, x_4 and x_5 might indicate whether a trait was scored on one of two different dates. In this case, each observation would have two $x_i = 1$ values: one for a block and one for date. Each line also has a personal x value, indicating whether that variety was present (1) or missing (0) from that plot.

This model can now be written in matrix notation as follows:

$$\mathbf{y} = \mathbf{Xu} + \mathbf{e}$$

Where:

y is a vector of length n of the n observations

X is a matrix of 1s and 0s, of dimensions n observations x m parameters (i.e. m things to be estimated).

e is a vector of length n of the n error terms.

The error sum of squares in matrix form is expressed as:

$$(\mathbf{y} - \mathbf{Xu})'(\mathbf{y} - \mathbf{Xu}) = \mathbf{e}'\mathbf{e}$$

And solved as:

$$\hat{\mathbf{u}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

There is one additional complication: as it stands, **X'X** cannot be inverted. This is because the parameters, including the genetic values, are not independent.

For example, with 10 lines, for a particular observation, if we know x_i=0 for the first nine lines, then we know that x_{10} must equal 1. Similarly, with four replicates in a randomized complete block experiment, if x = 0 for three replicated, then we know that x must equal 1 for the fourth.

There are two main ways of accounting for this dependency:

1. Treat the breeding values to be estimated, and the replicate effects (say) as deviations from an overall mean. In this case there are m-1 breeding values to be estimated and n-1 replicate effects, along with an overall or experimental mean. The genetic value for the dropped variety is estimated as minus the sum of all the other varieties. Likewise, the genetic value for the dropped replicate is minus the sum of all the other varieties. This is the way GenStat works.
2. Treat the estimated effect for the first replicate and first line as a reference and measure all other replicate and line effects as deviations from the reference. This the way R (using lm) works.

These two methods will produce different mean effects, but the differences in genetic value between any pair of lines will be identical. For selection, therefore, it doesn't matter which of these two parameterizations you use.

This process is necessary whatever form of the model is used; it is not just a complication for matrix algebra. It may help to think in terms of degrees of freedom (df). For example, with n replicates and m lines, there are (nm-1) degrees of freedom. These can be partitioned into (m-1) df for (m-1) independent line effects, (n-1) for (n-1) replicate effects and (n-1)(m-1) degrees of freedom for error. Each df estimates one effect.

3.2 Example

As an example, consider the following five observations of yield on five wheat lines:

Yield (kg)	Variety code	μ	V1	V2
97.5	1	1	1	0
86.2	1	1	1	0
102.8	2	1	0	1
108.9	2	1	0	1
110.3	3	1	-1	-1

There are two entries each of the first two lines and only one for the third line. These have been coded for use in matrix notation as a column of 1's for a mean effect (μ) and two columns (V1 and V2) for the three variety effects. The third variety is indicated by -1 in both the V1 and V2 columns: i.e. V3 is indicated as not-V1 and not-V2.

Since there is no complication caused by the inclusion of blocks, the least squares estimates of the variety means are just the mean of the values for each variety:

Variety	Yield
V1	91.85
V2	105.85
V3	110.30
average	102.67

Note that the mean of the three variety estimates is not the same as the mean of all entries in the experiment (101.14). This is a consequence of the unequal replication of the varieties.

To estimate these effects using the matrix method:

X		
1	1	0
1	1	0
1	0	1
1	0	1
1	-1	-1

X′				
1	1	1	1	1
1	1	0	0	-1
0	0	1	1	-1

X′X		
5	1	1
1	3	1
1	1	3

$(X'X)^{-1}$		
0.22222	-0.05556	-0.05556
-0.05556	0.38889	-0.11111
-0.05556	-0.11111	0.38889

y	X′y
97.5	505.7
86.2	73.4
102.8	101.4
108.9	
110.3	

u	
102.67	= estimate of mean
-10.817	= BLUE for V1
3.183	= BLUE for V2

The BLUE for V3 is - (BLUE for V1 + BLUE for V2) = - (-10.817 + 3.183) = 7.633.

We have three BLUEs for varieties, which is all that is required for selection.

If desired, we can estimate the mean effect of each variety by adding to the BLUE the estimate of the mean:

V1	=	102.67 – 10.817	=	91.85
V2	=	102.67 + 3.183	=	105.85
V3	=	102.67 + 7.633	=	110.30

These are identical to the estimates obtained from the simple means, as they should be in this case. In more complex cases, for example with incomplete blocks, this is no longer the case, and the BLUEs are more accurate estimates of genetic value than simple means.

3.3 Benefits

Selecting on BLUEs is recommended for:

- Balanced data with equal replication
- Designed trials with experimental designs (i.e. incomplete block designs)
- Comparison of advanced-trial materials with commercial checks

3.4 Constraints

Selecting on BLUEs is not recommended for:

- Trials with variable replication (unbalanced) and precision among varieties
- Selection of early generation materials

4. Best linear unbiased prediction (BLUP)

4.1 Method

The model used previously for estimation of genetic value from BLUEs is:

$$y_{ijk} = x_1\mu_1 + x_2\mu_2 + x_3\mu_3 + \ldots x_m\mu_m + x_{(m+1)}g_1 + x_{(m+2)}g_2 + x_{(m+3)}g_3 + \ldots + x_{(m+n)}g_n + e_{ijk}$$

In matrix form:

$$\mathbf{y} = \mathbf{Xu} + \mathbf{e}$$

There was no discrimination in estimation between the genetic values (the g terms) and the other effects (the μ terms). BLUP treats the estimation of these effects differently. We partition the model as:

$$\mathbf{y} = \mathbf{Xu} + \mathbf{Zg} + \mathbf{e}$$

Where:

g	is a vector of the genetic values we wish to estimate.
Z	is a matrix of 0s and 1s describing on which line the observation has been made.
Xu + **Zg**	is simply a split of the matrix used for BLUEs (also called **Xu**) vertically into two parts.

With this partition, the estimation of BLUEs is as follows:

$$\begin{bmatrix} \hat{u} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ ZX' & Z'Z \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

The least squares estimates are as before, but now there is an explicit difference between the genetic values (**g**) and the other terms in the model (**u**).

Writing the model in this form allows us to modify the estimation of genetic values. We treat the lines as "random effects" and the other factors as "fixed effects". Random effects can often be regarded as samples from a population, for example doubled haploid (DH) lines from an F_2 population. Each DH line is one of a potentially infinite population of lines which could be produced.

The population has an associated genetic variance, which we shall denote as σ_g^2. Fixed effects are the other factors we must include in the model to get fair and accurate estimates of genetic values. These are treated as having a common error variance, here denoted as σ_e^2, though the terms themselves cannot be regarded as samples from a population. Different fertilizer treatments are one example in plant breeding. More detail on the difference between fixed and random effects is provided in **A2**.

The estimation is made as follows:

$$\begin{bmatrix} \hat{u} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ ZX' & Z'Z + I\gamma \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Note the modification to the bottom right corner of the matrix to be inverted.

Where:

I is a square matrix of dimensions n x n (n = the number of genetic values to be estimated) with value 1 down the leading diagonal and 0 elsewhere: it is an identity matrix.

γ $= \sigma_e^2/\sigma_g^2$

γ is closely related to heritability of a single observation (**Figure 2**).

h^2 $= \sigma_g^2 / (\sigma_e^2 + \sigma_g^2)$

γ $= \sigma_e^2 / \sigma_g^2 = (1-h^2)/h^2$

h^2 $= 1/(1+\gamma)$

The modification therefore has the effect of inflating the bottom right corner of the matrix by adding σ_e^2 / σ_g^2 to the diagonal. We shall see later that this is a special case of more general forms of BLUP.

The consequence here is that the estimated genetic effects are identical to the BLUEs multiplied by the heritability of the line mean. Since heritability is always less than one, we say that the BLUPs are *shrunk* estimates of the BLUEs:

The similarity between this relationship and the breeders' equation is not coincidence:

$$R = h^2S$$

Response to selection, R, is the *predicted* response from the current *estimate* of performance, S, shrunk by the heritability. BLUPs and BLUEs apply this relationship to single lines rather

than a selected group. The genetic value is the *predicted* future performance of a line on retesting. It is based on the current *estimate* of performance of the same line.

Recall that the heritability of a line mean is calculated from the variance components as follows:

$$h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2/n)$$

Where:

n is the number of observations.

This has three important consequences for estimation using BLUPs:

1. As the number of observations goes up, there is less shrinkage.
2. If the number of observations is the same for all lines, then shrinkage is a constant proportion for all lines; the top 10% of lines on BLUEs remain the top 10% on BLUPs. However, if selecting lines above a threshold, say those exceeding the mean of the controls, then fewer lines may be selected using BLUPs.
3. If the number of observations differs from line to line, then the heritability of each line mean will differ, and the degree of shrinkage will differ. This can change the ranking of lines and is therefore important. It can have a marked effect in p-rep trials in the early stages of testing, where γ can be high and the heritability of line means is low but can change markedly with replication.

4.2 Examples

We shall consider first a simple case with balanced data.

As an example, consider the following five observations of yield on three wheat lines:

Yield (kg)	Variety code	μ	V1	V2	V2
97.5	1	1	1	0	0
86.2	1	1	1	0	0
102.8	2	1	0	1	0
108.9	2	1	0	1	0
102.1	3	1	0	0	1
110.3	3	1	0	0	1

This is the same example used to illustrate BLUE but with one additional observation on V3 to give equal observations for all lines.

First, we estimate the BLUEs as before:

X		
1	1	0
1	1	0
1	0	1
1	0	1
1	-1	-1
1	-1	-1

X′					
1	1	1	1	1	1
1	1	0	0	-1	-1
0	0	1	1	-1	-1

X′X		
6	0	0
0	4	1
0	1	4

$(X'X)^{-1}$		
0.16667	0	0
0	0.33333	-0.16667
0	-0.166667	0.333330

y	X′y
97.5	607.8
86.2	-28.7
102.8	-0.7
108.9	
102.1	
110.3	

u	
101.3	= estimate of mean
-9.45	= BLUE for V1
4.55	= BLUE for V2
4.9	= BLUE for V3 (i.e. -V1-V2)

Next, we estimate the BLUPs. This adds a penalty to the diagonal of the entries to be shrunk in **X'X**, so that the original **X** is partitioned into a component for the fixed effects (**X**) and a component for the random effects (**Z**).

It is assumed that: $\sigma_g^2 = \sigma_e^2 = 1$

So that: $\gamma = \sigma_e^2 / \sigma_g^2 = 1$

Firstly, $\gamma = 1$ corresponds to a heritability of 0.5 for an individual observation and not for a mean. Secondly, once we treat the lines or individuals for which we wish to estimate genetic value as members of a population, there is no longer a requirement to add constraints of the form V3 = -V1 - V2. In essence, V1, V2, and V3 are treated as samples from a population with variance σ_g^2 and the need to impose a restraint so that they add to zero is removed.

y	μ	V1	V2	V3
97.5	1	1	0	0
86.2	1	1	0	0
102.8	1	0	1	0
108.9	1	0	1	0
110.3	1	0	0	1

X
1
1
1
1
1
1

Z		
1	0	0
1	0	0
0	1	0
0	1	0
0	0	1
0	0	1

(X Z)'					
1	1	1	1	1	1
1	1	0	0	0	0
0	0	1	1	0	0
0	0	0	0	1	1

	(X'X X'Z) (ZX' Z'Z)		
6	2	2	2
2	2	0	0
2	0	2	0
2	0	0	2

Exactly as before, prior to splitting the design matrix into fixed and random components.

Next, we penalize the random components **Z'Z** by the addition of γ (with a value of 1 in our example) to the diagonal.

In matrix terms:

(X'X X'Z) (ZX' Z'Z + I γ)			
6	2	2	2
2	3	0	0
2	0	3	0
2	0	0	3

(X'X X'Z)$^{-1}$ (ZX' Z'Z + I γ)			
0.5	-0.333	-0.333	-0.333
-0.333	0.556	0.222	0.222
-0.333	0.222	0.556	0.222
-0.333	0.222	0.222	0.556

The solution:

y	X'y Zy
97.5	607.8
86.2	183.7
102.8	211.7
108.9	212.4
102.1	
110.3	

$\begin{bmatrix} X'X & X'Z \\ ZX' & Z'Z + I\gamma \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$	
101.3	= estimate of mean
-6.3	= BLUP for V1
3.033	= BLUP for V2
3.267	= BLUP for V3

The mean is identical to the estimate from the BLUEs analysis. The BLUPs are identical to the BLUEs multiplied by the heritability of the variety means, 0.67, and not the heritability of a single observation. which is 0.5. See Table 1. If the heritabilities are identical for each variety mean, there is no difference in ranking. However, the shrunk estimates may be more realistic, particularly in single replicate trials with low heritability.

Table 1. Shrinkage of estimates

$\sigma_g^2 = \sigma_e^2 = 1$

$h^2_{(2\ reps)} = 0.5/(0.5+0.5/2) = 0.6667$

BLUE	BLUP	BLUP/BLUE*
-9.45	-6.3	0.67
4.55	3.03	0.67
4.90	3.27	0.67

*Note that h^2=BLUP/BLUE as suggested by the method of Walsh and Lynch referred in the heritability manual of EiB.

4.3 Variance components

To estimate the BLUPs, we require a population mean and variance. These can come from prior knowledge or other experiments. However, more commonly they are not known, and are estimated from the data together with other fixed effects. This is described in more detail in **A1**.

The simplest case of BLUP described here can be extended to include multiple traits and environments. The more common uses of BLUP are described in the Annex. BLUP is increasingly used in preference to BLUE in trials.

4.4 Benefits

- Good for unbalanced data – lines with variable replication, especially p-rep designs.
- Good for designed trials – including incomplete block designs.
- Lines can be grouped into exchangeable sets.

- Good for selection of lines which exceed a fixed threshold, as opposed to selecting a proportion of lines.

4.5 Constraints

- Unnecessary for simple cases.
- Unnecessary for uniform trials with equal replication.
- Not best used for very heterogeneous sets of varieties (population structure). Consider fitting >1 random effect (see Annex).
- Not best used for selection among lines with very variable genetic relationships.
- Lack of understanding of the methodology.
- Shrinkage of high-yielding lines can be unpopular with breeding program management.

5. Pedigree BLUP (pBLUP)

5.1 Method

Imagine testing a set of clonal lines from two different crosses, but that the four parents are unrelated. Thus, clones within a cross are related as full-sib individuals and clones from different crosses are unrelated. You are given a new line from one of the crosses. If it had no trait data, an obvious first estimate of its genetic value would be the mean of the other clones in the same cross. If the new line had extensive trait data, you would likely judge it directly on its own merits and ignore the data on siblings. When heritabilities of line means are lower, the ideal is to weight the two sources of information (cross mean and individual phenotype) to give a more accurate estimate of genetic value. This is the essence of pedigree BLUP (pBLUP).

However, it is possible to take into account information from all relatives, not only siblings but also half-sibs, parents, progeny, second degree relatives and so on. The weighting of information from all relatives will vary with the degree of relationship: data on parents is more important than data on great-grandparents, for example. A key point to note is that now we are able to estimate the genetic value of a line even though it has no trait data.

There are two ways to estimate genetic values incorporating information from relatives. The first is entirely empirical and is possible if individuals are grouped into families of the same type (full-sibs, half-sibs F_2s etc.).

We can partition the genetic value of a line into two parts:

$$g_{ij} = g_{bi} + g_{wj}$$

Where:

g_{bi} is the genetic value of the b^{th} family

g_{wi} is the genetic value of the deviation of the j^{th} individual from the i^{th} family.

These two components of the genetic value of an individual are independent and their values can be shrunk independently by their respective heritabilities, to give a BLUP for the individual in the following manner:

$$g_{ij} = h^2_f p_{bi} + h^2_w p_{wj}$$

Where:

h^2_f and h^2_w are between and within family heritabilities.

p_{bi} is the deviation of the i^{th} family mean from the overall mean.

p_{wj} is the deviation of the j^{th} individual from the i^{th} family mean

There is a slight complication depending on whether the phenotype of the individual under consideration also contributes to the family mean. If the family size is large, this makes little difference. If family size is small, it can be taken into account. An advantage of this approach is that it requires no genetics: the estimates of $\mathbf{h}^2_f$ and $\mathbf{h}^2_w$ can come from the data and require no genetic assumptions about relationships among individuals or the genetic composition of the trait or of the population from which the families were sampled. The

disadvantage of this approach is that it is hard to apply to more complex and variable pedigree relationships. To take these into account we must be explicit about an assumed starting population and also modify the mixed model equations again.

The model remains unchanged however:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e}$$

but now the effects are estimated as

$$\begin{matrix} \hat{\mathbf{u}} \\ \hat{\mathbf{g}} \end{matrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{b} \\ \mathbf{Z}\mathbf{R}^{-1}\mathbf{X}' & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-1} \begin{matrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{matrix}$$

The definition of **y**, **u**, **g**, **X**, and **Z** is unchanged.

R is new. It is a square matrix, of dimension n (the number of observations). It is the matrix of error variances and covariances associated with the **e** terms. **R** must be included if fitting a spatial model to a variety trial (most commonly using AR1 x AR1 or two-dimensional splines). We shall not discuss these further here. If errors are treated as independent (which is always valid in randomized trials but not necessarily optimal), then $\mathbf{R}^{-1}$ is a diagonal matrix with values $1/\sigma_e^2$. In this case, all **R** terms cancel, and the solution is simplified. This is why we have omitted them previously.

G is also new and is a square matrix of dimensions equal to the number of genetic values to be estimated. Its terms account for the genetic correlations or relationships among lines. For a simple BLUP, lines are treated as unrelated and **G** reduces to a diagonal matrix with element σ^2_g. $1/\sigma_g^2$ is then added to the diagonal of $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}$. If errors are also treated as independent and **R** is dropped from the solution, we must add σ_e^2 / σ_g^2 (or γ) to $\mathbf{Z}'\mathbf{Z}$ to obtain the same solution as before.

In the form given here, and ignoring the fixed effects, the BLUPs can be regarded as BLUEs multiplied the multi-line analogue of heritability for line means: **G** / (**G**+**E**) compared to σ_g^2 / (σ_e^2 + σ_g^2) for a single individual.

5.2 Composition of the G, genetic variance/covariance matrix

In pBLUP, **G** is estimated from σ_g^2 in an ancestral or base population, in which all individuals are assumed to be unrelated and none-inbred. In this reference population, the covariance between individuals is 0 and the genetic variance of an individual is σ_g^2. Founders of the pedigree are treated as a sample of this population. Variances and covariances change among descendants of the founders and estimates of the changes are provided by the pedigree.

G can be written as:

$$\mathbf{G} = \mathbf{K}\sigma_g^2$$

K is a matrix of relationships. Most commonly, this describes relationships resulting only from additive genetic variation. pBLUP can be extended to incorporate dominance and epistasis by including additional matrices for these effects but this is not described here. Common practice is to consider only additive variation in pBLUP and this is generally adequate since dominance and epistatic interactions are not inherited. The elements of **K** are the coefficients of the additive genetic (co)variance between the individuals in the dataset. In animal breeding, the matrix generally includes all founder individuals and ancestors of those in the dataset. Founders are assumed to be non-inbred. This makes the prediction of animal breeding values robust to the effects of selection within the pedigree.

The diagonals of **K** are the coefficients of additive genetic variance for the individuals themselves. The off-diagonal entries in the table are coefficients of relationship, or twice the coefficients of kinship. The coefficient of kinship between two individuals is the probability than an allele picked at random from one individual is identical by descent (ibd) to an allele

picked at random from the other, or p(ibd). With no inbreeding, these coefficients are 1/4 for full-sibs and 1/8 for half-sibs. Other common relationships are shown in **Table 2**. For a population with no inbreeding, the diagonal entries of **K** are also coefficients of relationships or twice the coefficients of kinship. The p(ibd) of a non-inbred individual with itself is a half (the inbreeding coefficient of its selfed progeny), so the coefficient of relationship of an outbred individual with itself is double this, or one.

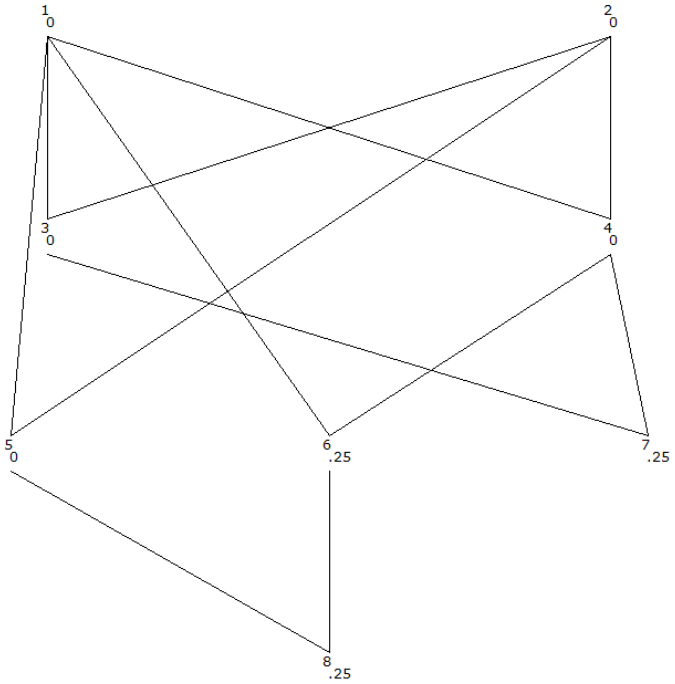
Table 2. Coefficient of relationship among commonly encountered relatives of outbred individuals

Relationship	Coefficient of relationship
Itself	1
Parent	0.5
Full sib	0.5
Half sib	0.25
Grandparent	0.25
Aunt or uncle	0.25
Great-grandparent	0.125
Unrelated individual	0

In the absence of inbreeding, the complete relationship matrix required in the mixed model equations is twice the kinship matrix. In animals and in plant species which do not self, inbreeding only occurs if relatives mate.

In this case, elements of **G** are still genetic relationships or twice the coefficients of kinship. The diagonal elements, however, may be better viewed as 1+F; the coefficient of the additive genetic variance for an individual with inbreeding coefficient F. The diagonals will thus have a maximum value of 2 and a minimum of 1. Software used for estimation of pBLUPs will also compute the **K** matrix, though stand-alone packages also exist. The process for estimating **K** (or $\mathbf{K}^{-1}$) uses some simple recursive tricks.

A small example pedigree and its corresponding relationship matrix is shown in **Figure 2**.



ID	1	2	3	4	5	6	7	8
1	1							
2	0	1						
3	0.5	0.5	1					
4	0.5	0.5	0.5	1				
5	0.5	0.5	0.5	0.5	1			
6	0.75	0.25	0.5	0.75	0.5	1.25		
7	0.5	0.5	0.75	0.75	0.5	0.625	1.25	
8	0.625	0.375	0.5	0.625	0.75	0.875	0.5625	1.25

Figure 2. **Example pedigree and corresponding relationship matrix.** Top: Pedigree for eight individuals showing inbreeding coefficients of each individual. Bottom: Relationship matrix for the eight individuals. Matrix is symmetrical, only the lower half is shown. Diagonals are (1+F) where F is the inbreeding coefficient of the individual. Entries are coefficients of σ_g^2 for pedigree BLUP.

5.3 Example

We shall add some pedigree information to the example with balanced data used previously for basic BLUP:

Yield (kg)	Variety code	μ	V1	V2	V3
97.5	1	1	1	0	0
86.2	1	1	1	0	0
102.8	2	1	0	1	0
108.9	2	1	0	1	0
102.1	3	1	0	0	1
110.3	3	1	0	0	1

We now assume that individuals V1 and V2 are members of the same full-sib family and that the parents of V1 and V2 are unrelated. V3 is unrelated to V1 or V2. For simplicity, we shall also treat $\sigma^2_g = \sigma^2_e$, so that the matrix $\mathbf{R} = \mathbf{R}^{-1} = \mathbf{I}\sigma_e^2$ and can be ignored, and $\mathbf{G} = \mathbf{K}\sigma^2_g = \mathbf{K}$.

The genetic variance/covariance matrix is therefore:

G	V1	V2	V3
V1	1	0.5	0
V2	0.5	1	0
V3	1	0	1

With inverse

G^{-1}	V1	V2	V3
V1	1.333	-0.667	0
V2	-0.667	1.333	0
V3	0	0	1

	(X'X (ZX'	X'Z) Z'Z)	
6	2	2	2
2	2	0	0
2	0	2	0
2	0	0	2

Identical to the previous balanced example for BLUP

Adding $\mathbf{G^{-1}}$ to **Z'Z**

(X'X (ZX'	X'Z) Z'Z+G^{-1})		
6	2	2	2
2	3.333	-0.667	0
2	-0.667	3.333	0
2	0	0	4

Solution:

y	X'y Z'y
97.5	607.8
86.2	183.7
102.8	211.7
108.9	212.3
102.1	
110.3	

(X'X (ZX'	$X'Z)^{-1}$ $Z'Z + I\gamma$)	(X'y) (Z'y)	
	101.79		= estimate of mean
	-5.705		= BLUP for V1
	1.295		= BLUP for V2
	2.940		= BLUP for V3

5.4 Benefits

- BLUPs can be estimated for individuals with no trait data.
- Historical deep pedigrees may be available.
- Good for outbreeding species with validated pedigrees.
- It accounts for the population structure issues found with regular BLUP.

5.5 Drawbacks

- Pedigree errors and inconsistencies will affect estimates.
- Assumption that all founders are unrelated outcrossed members of the same population are rarely true and dealing with selfing species is problematic.
- Cannot estimate genetic value of any individual with missing pedigree information.
- Cannot discriminate between individuals from the same cross.

5.6 Comparison of BLUE, BLUP and pBLUP methods

The genetic values obtained for the example data using BLUE, BLUP and pBLUP are presented in **Table 3**.

Table 3. Comparison of three genetic value estimation methods (BLUE, BLUP and pBLUP) using example data.

	BLUE	BLUP	pBLUP
V1	-9.45	-6.3	-5.705
V2	4.55	3.03	1.295
V3	4.90	3.27	2.940

Using pedigree BLUP, the full-sib relationship between V1 and V2 has moved their estimated genetic values towards each other compared with basic BLUP. Running the example above with increased or decreased estimates of relationship between V1 and V2, and varying σ^2_g (i.e. scaling $\mathbf{G} = \mathbf{K}\sigma^2_g$ up or down) confirms the intuition that:

If relationships are strong: pBLUPs are **similar**.

If relationships are distant: pBLUP and basic BLUP are **equivalent**.

If heritability is high: pBLUPs **approach** BLUEs.

If heritability is low: pBLUPs **approach** zero.

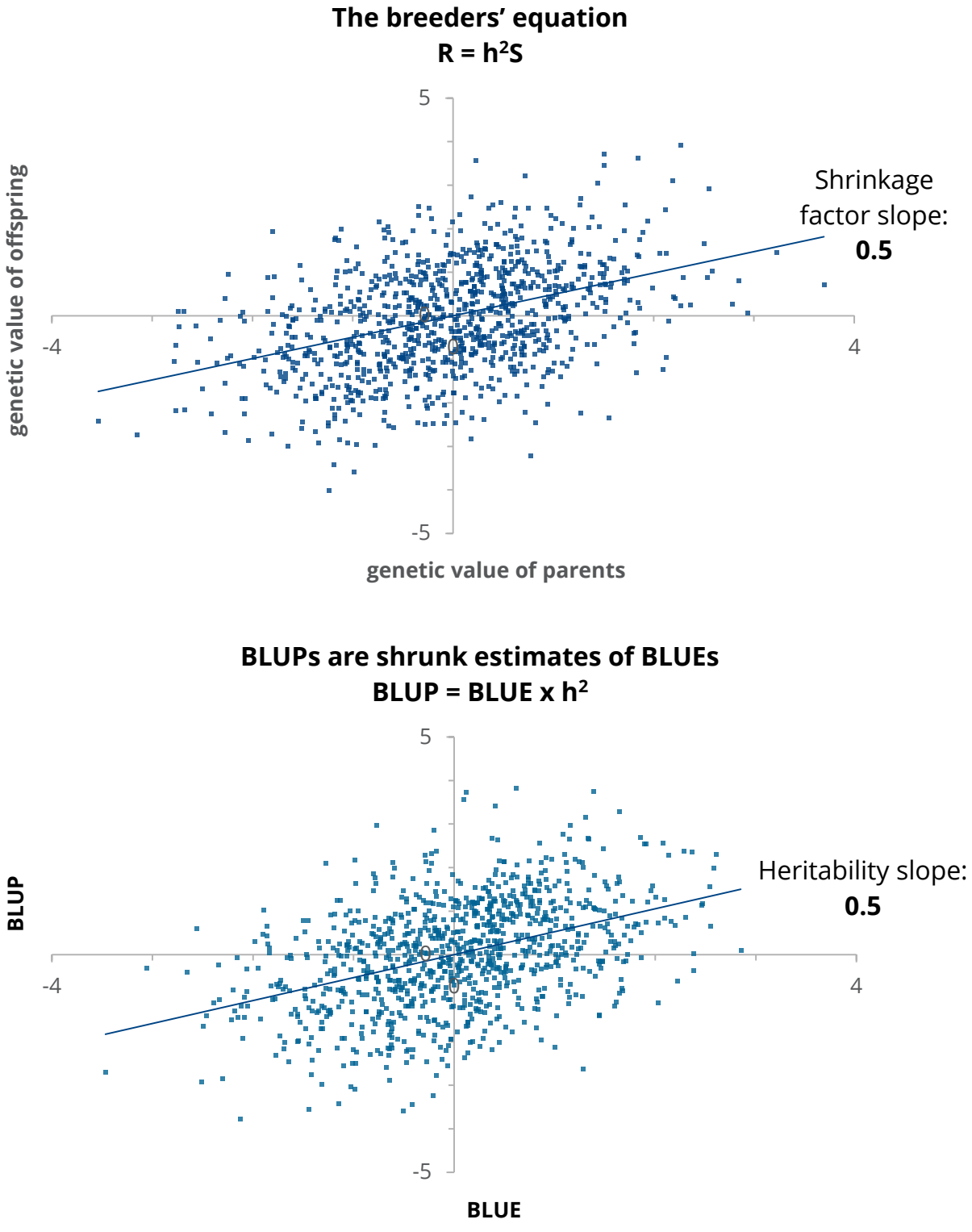


Figure 3. Shrinkage of BLUPs from BLUEs and the breeder equation are equivalent

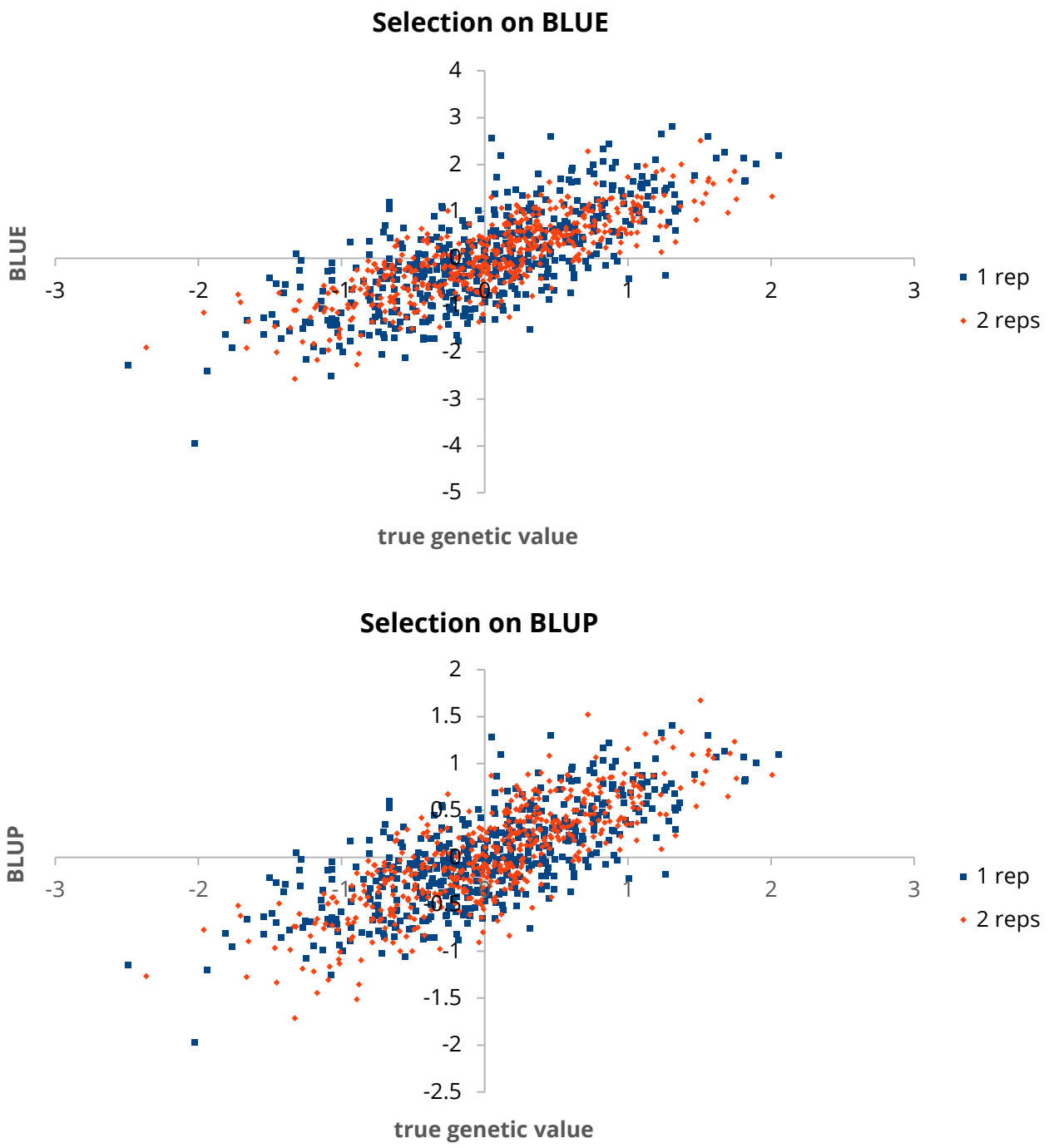


Figure 4. **Selection on BLUPs is more accurate than selection on BLUEs when lines are tested with unequal replication.** Selection on BLUE (top) selects disproportionately many lines tested in one replicate with low true genetic values (X axis). Using the same data, selecting on BLUP (bottom) selects more entries tested in two replicates with higher true genetic values. 1,000 lines simulated, half tested in one replicate (blue) and half in two (red). Genetic variance = error variance = 0.5.

6. Genomic BLUP and ridge regression BLUP

Genomic BLUP (gBLUP) and ridge regression BLUP (rrBLUP) are both closely related methods and so are considered together here.

6.1 gBLUP method

gBLUP is conceptually similar to pedigree pBLUP, but overcomes many of the problems associated with it.

The model for genetic value in gBLUP is exactly the same as pBLUP:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{Zg} + \mathbf{e}$$

Effects are estimated in the same way for gBLUP as in pBLUP:

$$\begin{matrix} \hat{\mathbf{u}} = & (\mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} &)^{-1} & \mathbf{X'R^{-1}y} \\ \hat{\mathbf{g}} & (\mathbf{ZR^{-1}X'} & \mathbf{Z'R^{-1}Z+G^{-1}}) & & \mathbf{Z'R^{-1}y} \end{matrix}$$

The difference between gBLUP and pBLUP is that the matrix **G** expressing genetic variances and covariance among lines or individuals. However, just as for pBLUP, **G** is written in terms of relationships among individuals as:

$$\mathbf{G} = \mathbf{K}\sigma_g^2$$

In this case, **K** is now estimated from a genome-wide set of genetic markers. Since markers segregate within crosses, marker estimated relationships can also vary within crosses. This is the major advantage of gBLUP over all the previous methods described: genetic values among individuals within a cross can now be predicted by exploiting these relationships.

Methods of estimating **K** are described in **A5**.

6.2 rrBLUP method

A simple way of predicting genetic values from traits is to create a marker index and use least squares estimates from multiple regression of a trait on a set of markers. The regression equation can then be used to predict missing trait values from marker scores. However, this will only work if there are more individuals in the dataset than there are markers: with n individuals there are n-1 degrees of freedom in a regression. Each SNP requires 1 df to estimate its regression coefficient, which restricts the number of biallelic markers to n-1.

Returning to our initial least squares equations:

$$\mathbf{y} = \mathbf{Xu} + \mathbf{e}$$

y	is a vector of length n of the n observations
u	is a vector of fixed effects: here a mean and the m marker effects
X	is now a matrix of marker scores, of dimensions n by (m+1).
e	is a vector of length n of the n error terms.

This is solved as:

$$\hat{\mathbf{u}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

This is only possible by least squares if there are fewer markers than observations. However, we now switch from treating the markers as fixed effects to treating them as random effects.

Shrunk regression coefficients are estimated as follows:

$$\begin{matrix} \hat{\mathbf{u}} & = & (\mathbf{X'X} & \mathbf{X'Z})^{-1} & \mathbf{X'y} \\ \hat{\mathbf{s}} & & (\mathbf{ZX'} & \mathbf{Z'Z} + \mathbf{I}\lambda) & \mathbf{Z'y} \end{matrix}$$

λ is a penalty which is added to the diagonal of **Z'Z** with the consequence that the estimates of each marker effect, **ŝ**, are shrunk. In effect, we are applying BLUP to the marker effects. It is remarkable that this allows any number of marker effects to be estimated. Geometric explanations can be found in the **Recommended literature** on ridge regression.

The solution is analogous to the use of a penalty $\gamma = \sigma_e^2 / \sigma_g^2$ in simple BLUP. Suppose our genome wide set of markers was adequate to capture all the genetic variation for a trait. Suppose further, that each marker captured an equal amount of the available genetic variation. In this case, the expected variation captured per marker is σ_g^2/m for m markers and we can apply a penalty:

$$\lambda = \sigma_e^2 / (\sigma_g^2/m) = m\sigma_e^2 / \sigma_g^2 = m\gamma$$

Where trait heritability is 0.5, which is often a reasonable first approximation in yield trials, λ refers to the number of markers.

When $\lambda = \sigma_e^2 / (\sigma_g^2/m)$, this specific form of ridge regression is called rrBLUP. Provided markers are coded and standardized in the same manner, it can be shown that the solutions to rrBLUP should be identical to gBLUP and will give the same predicted genetic values.

Since gBLUP is numerically easier to solve than rrBLUP (since the matrix to be inverted is of dimension n x n individuals [usually hundreds], rather than of m x m markers [often thousands]) it is easier to work with gBLUP equations, then transpose the solution to provide the (shrunk) marker effects if these are also required.

6.3 Example of gBLUP

We shall use the same data on three individuals as in the previous examples, but substitute the kinship matrix **K** calculated from six markers as described in **A5.3** for the pedigree-based relationship matrix used previously.

Yield (kg)	Variety code	μ	V1	V2	V3
97.5	1	1	1	0	0
86.2	1	1	1	0	0
102.8	2	1	0	1	0
108.9	2	1	0	1	0
102.1	3	1	0	0	1
110.3	3	1	0	0	1

Previously, V1 and V2 were treated as members of the same full-sib family. This is unnecessary when using realized genomic relationships, though knowledge of this can help in interpreting results. For simplicity, and consistency with the pBLUP example, we shall treat $\sigma^2_g = \sigma^2_e$ so that the matrix $\mathbf{R} = \mathbf{R}^{-1} = \mathbf{I}$ and can be ignored and $\mathbf{G} = \mathbf{K}\sigma^2_g = \mathbf{K}$.

The genetic variance/covariance matrix, estimated from the six markers (**A5.3**) is therefore:

KK' = G	V1	V2	V3
V1	4.744	5.093	-0.488
V2	5.093	6.605	-0.721
V3	-0.488	-0.721	0.674

with inverse

G^{-1}	V1	V2	V3
V1	1.236	-0.968	-0.140
V2	-0.968	0.930	0.293
V3	-0.140	0.29	1.695

	(X'X	X'Z)	
	(ZX'	Z'Z)	
6	2	2	2
2	2	0	0
2	0	2	0
2	0	0	2

Identical to the previous balanced example for BLUP

Adding $\mathbf{G^{-1}}$ to **Z'Z**

	(X'X	X'Z)	
	(ZX'	$Z'Z+G^{-1}$)	
6	2	2	2
2	6.74	5.093	-0.488
2	5.093	8.605	-0.721
2	-0.488	-0.721	2.674

With inverse

	(X'X	$X'Z)^{-1}$	
	(ZX'	$Z'Z+G^{-1}$)	
0.779	-0.717	-0.730	-0.391
-0.717	1.003	0.784	0.364
-0.730	0.784	1.065	0.340
-0.391	0.364	0.340	0.469

And solution	y	X'y Z'y	$(X'X \quad X'Z)^{-1} \quad (X'y)$ $(ZX' \quad Z'Z + G^{-1}) \quad (Z'y)$	
	97.5	607.8	104.377	= estimate of mean
	86.2	183.7	-8.246	= BLUP for V1
	102.8	211.7	-1.801	= BLUP for V2
	108.9	212.4	0.817	= BLUP for V3
	102.1			
	110.3			

Table 5 compares estimates for BLUE, BLUP, pBLUP and gBLUP, before and after rescaling the means of the BLUPs to zero for ease of interpretation.

6.4 Benefits

- Prediction within crosses.
- Increased accuracy of field trials.
- Single cross prediction (i.e. hybrid prediction).
- Optimal contribution methods.

6.5 Drawbacks

- Low marker density with complex family structures.
- Datasets in which some individuals cannot be genotyped: e.g. ancestral lines for which no seed is available, but for which there may be pedigree and trait data.

Recommended literature

Books

Galwey, N.W., 2014. Introduction to mixed modelling: beyond regression and analysis of variance. John Wiley & Sons.

An easy to read and full account of BLUE and BLUP with little algebra. Very good on the difference between fixed and random effects. Examples are given with code in GenStat, R and SAS. Focus is on basic error structures with no covariance terms but pedigree BLUP and AR1 xAR1 models are described.

Lynch, M. and Walsh, B., 1998. Genetics and analysis of quantitative traits (Vol. 1, pp. 535-557). Sunderland, MA: Sinauer.

Comprehensive, more mathematical treatment. Includes pedigree BLUP but not genomic BLUP

Falconer, D.S. and Mackay, T.F.C., 1996. Introduction to quantitative genetics 4th edition. Harlow, UK: Longmans.

Classic textbook. Little on the mixed model but clear explanations of kinship relationship and breeding value.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning with Applications in R (Vol. 112, pp. 3-7). New York: Springer.

Excellent introductory account of statistical methods such as ridge regression and the lasso, which can be applied to the estimation of genetic values. Free to download.

Papers

BLUE, BLUP and Pedigree BLUP

Piepho, H.P., Möhring, J., Melchinger, A.E. and Büchse, A., 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica, 161(1-2), pp.209-228.

Kinship Calculation

Amadeu, R.R., Cellon, C., Olmstead, J.W., Garcia, A.A., Resende, M.F. and Muñoz, P.R., 2016. AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example. The plant genome, 9(3).

Speed, D. and Balding, D.J., 2015. Relatedness in the post-genomic era: is it still useful?. Nature Reviews Genetics, 16(1), pp.33-44.

> Review of method of estimating kinship from markers – focuses on human (i.e. outbreeding diploids).

Goudet, J., Kay, T. and Weir, B.S., 2018. How to estimate kinship. Molecular ecology 27(20), pp.4121-4135.

> Comments extensively on the difference between pedigree and genomic estimates of relationship.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. Journal of dairy science, 91(11), pp.4414-4423.

Genomic BLUP

Endelman, J.B., 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome, 4(3), pp.250-255.

> Outline of theory together with a description of the methods available in on of the most commonly used packages for genomic BLUP.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y. and Dreisigacker, S., 2017. Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science, 22(11), pp.961-975.

Xavier, A., Muir, W.M., Craig, B. and Rainey, K.M., 2016. Walking through the statistical black boxes of plant breeding. Theoretical and applied genetics, 129(10), pp.1933-1949.

More mathematical review of methods for genomic prediction.

Annexes

A1. Residual Estimation by Maximum Likelihood (REML) and estimation of variance components.

BLUPs require estimates of population means and variance components (minimally $\sigma_e^2 + \sigma_g^2$). Generally, the software we use to estimate BLUPs will also estimate these from the data. REML (Residual Estimation by Maximum Likelihood) is the default method in most packages. It has the advantage that it gives identical estimates of variance components, and therefore identical estimates of BLUEs and BLUPs, to those obtained by least squares estimates of variance components. This gives us confidence in using REML in cases where least squares estimates cannot be made.

For some of the examples used in this document, the BLUPs are not simply the BLUEs x h^2. This is because the estimates of the mean and the random effects are correlated. In datasets of the size encountered in practice in breeding, this is not a problem, but it can make a difference in very small datasets.

A2. Distinguishing fixed and random effects.

With the availability of good computer statistical packages, there is no requirement to be able to write down the analytical models and solve by matrix algebra as illustrated in our examples. It is important, however, for the user to be able to describe the model, even in longhand, and to understand which effects are treated as random effects and which fixed.

It is not necessary for the lines under testing to be members of a well-defined genetic population. A less stringent requirement is that the lines can be regarded as *exchangeable*. This means that the outcome or interpretation of the experiment is not affected by switching their coding. If line 20, say, was no longer treated as line 20 but as line 135, would this matter?

If not, then the lines are exchangeable and can be regarded as members of a common group for which BLUPs will be estimated.

For example, suppose lines 20 and 135 were clones being tested for the first time in a preliminary yield trial. They are new and unknown, and swapping their labels would have no outcome on how you judge their genetic values. Such a collection of lines could be tested in a p-rep design for example, saving space and money or allowing more lines to be tested, without worrying that the breeding value of one particular line was shrunk more than another. However, suppose line 135 is being tested for the first time, but line 20 is being retested after selection in the previous season. In this case, you are likely to want to treat new lines and retested lines differently; they are not exchangeable.

Equivalently, in analyzing a collection of lines of different type, say if inbred and hybrid lines were in the same trial, or if the lines under test were a mix of full-sib and half-sib families, then these would be expected to have different values of σ_g^2. This can be accommodated by having separate random effects for each type.

A3. Additional random effects: the mixed model in variety trials

The estimation of breeding values described in our examples has treated individuals as random effects and the mean as the only fixed effect. Other factors included in an experiment may be treated as either fixed or random. It is common practice in variety trials for the varieties to be treated as fixed, with BLUEs to be estimated, and the blocks, whether complete replications or incomplete blocks, to be treated as random.

Such a model could be written as follows:

$$\mathbf{y = Xg + Z\beta + e}$$

Where:

β is a vector of block effects. BLUPs for these can be estimated, but we are not usually that interested in them.

Z is the design matrix – allocating blocks to observations.

g is the vector of genetic values we wish to analysed, commonly estimated as BLUEs.

X is the design matrix allocating varieties to observations.

e is the usual vector of error effect.

Estimation can proceed as before. In this case, the variance component for block effects is always estimated from the experiment itself. The advantage of treating blocks as random rather than fixed is that there is some information of differences between varieties which is locked into estimates of differences between blocks, and this information can be released and incorporated into the estimates of variety effects to improve their precision. This process used to be referred to as "recovery of inter-block information" and predates current methods of estimation and terminology. The information recovered is greatest when block effects are of intermediate variability.

If block effects are large (typically in a bad trial), the blocks variance is also large so that block effects are hardly shrunk at all, and the BLUEs for variety effects are little changed in value or precision. If block effects are small, the blocks variance approaches zero (and blocks could be dropped from the model) and the variety BLUEs approach simple arithmetic means. With modest block effects however – as is often the case – there is information to be recovered and the precision of the BLUEs is improved.

There has been a near-philosophical discussion about whether blocks should be treated as fixed or random. The exchangeability argument above helps to establish them as random. The process of randomization of varieties over blocks also validates this choice: in general it is not important which variety gets allocated to a particular block.

This model (varieties fixed, blocks random) remains the most commonly used method of estimating variety performance globally, though it can frequently be improved. Variety effects, if exchangeable, can be treated as random and error variances may be better modeled, as described in **A4**.

A4. Multiple random effects

In the previous discussion, both lines and blocks can be treated as random effect. Multiple fixed effects could also be included, but first we shall consider only the mean. The model can be written as:

$$\mathbf{y} = \mathbf{X\mu} + \mathbf{Z}_1\boldsymbol{\beta}\mathbf{g}_1 + \mathbf{Z}_2\mathbf{g}_2 + \mathbf{e}$$

$\boldsymbol{\mu}$	Is a vector of fixed effects.
$\mathbf{X}$, $\mathbf{Z}_1$, $\mathbf{Z}_2$	Are matrices allocating the fixed effects
$\boldsymbol{\mu}$	Random block effects
$\mathbf{g}_1$ and $\mathbf{g}_2$	The genetic values to each observation.

If just the mean is included in the model, **X** is a vector of 1s.

The effects are estimated as:

$$\begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{g}}_1 \\ \hat{\mathbf{g}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X'X} & \mathbf{X'W} & \mathbf{X'Z} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{I}\gamma_1 & \mathbf{W'Z} \\ \mathbf{Z'X} & \mathbf{Z'W} & \mathbf{Z'Z} + \mathbf{I}\gamma_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X'y} \\ \mathbf{W'y} \\ \mathbf{Z'y} \end{pmatrix}$$

γ_1 is σ_e^2/σ_b^2

(Where σ_b^2 is the blocks component of variation, estimated from the data)

γ_2 is σ_e^2/σ_g^2

(As before)

Different penalties are therefore added to the diagonals of **W'W** and **Z'Z** depending on the relative magnitudes of the variance components for blocks and lines.

We can extend the model by more complex relationships among the residual errors terms (such as correlations between adjacent plots in AR1 x AR1 designs, and correlations among the random effects (such as genetic relationships among varieties or individuals) in which case BLUPs and BLUEs are estimated as:

$$\begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{g}}_1 \\ \hat{\mathbf{g}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X'R}^{-1}\mathbf{X} & \mathbf{X'R}^{-1}\mathbf{W} & \mathbf{X'R}^{-1}\mathbf{Z} \\ \mathbf{W'R}^{-1}\mathbf{X} & \mathbf{W'R}^{-1}\mathbf{W} + \mathbf{G}_1^{-1} & \mathbf{W'R}^{-1}\mathbf{Z} \\ \mathbf{Z'R}^{-1}\mathbf{X} & \mathbf{Z'R}^{-1}\mathbf{W} & \mathbf{Z'Z} + \mathbf{G}_2^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X'R}^{-1}\mathbf{y} \\ \mathbf{W'R}^{-1}\mathbf{y} \\ \mathbf{Z'R}^{-1}\mathbf{y} \end{pmatrix}$$

$\mathbf{G}_1$ is the variance covariance matrix among the $\mathbf{g}_1$ random effects.

$\mathbf{G}_2$ is the variance covariance matrix among the $\mathbf{g}_2$ random effects.

There is no requirement to stop here: additional random terms could be added, each with separate variance components. For example, in variety trials, if one set of lines were F_1 hybrids and another inbred lines, BLUPs for each could be estimated independently. This seems complicated, but at heart, we are simply adding a matrix of penalty effects to the bottom right the variance/covariance matrix (anything involving **W** and **Z**). The penalty effects vary from method to method and from experiment to experiment, but the basic principle is the same: penalize the bottom right and leave the fixed effects in the first set of rows and columns untouched. If each observation can be treated as independent, with uncorrelated errors (always valid with randomization), then **R** can be ignored too.

Fortunately, provided that we can specify the model in terms of fixed and random effects, together with the variance/covariance structure of the random effects, there is no requirement to write down the mixed model solutions: the software will take care of the estimation for us. With very complex models however, the software sometimes struggles too.

A5. Estimation of relationships from markers

Given its central importance in GBLUP, the estimation of **K** has received much attention. Here we describe the two most common methods.

A5.1. Identity by state

If two alleles are identical (e.g. both are the same nucleotide for a SNP or both are the same repeat length for a microsatellite) they are called identical by state (IBS). This terminology is to distinguish IBS from identity by descent (IBD) which will not be described here. Consider two diploid individuals. At a single locus they may carry 0, 1 or 2 alleles in common. If we chose an allele at random from each individual, we can assign a probability that the two alleles are identical. **Table 4** gives examples.

Table 4. **Example p(IBS) between pairs of individuals**

Individual 1	Individual 2	p(IBS)
$\mathbf{A_1A_1}$	A_1A_1	1
$\mathbf{A_1A_1}$	A_1A_2	0.5
$\mathbf{A_1A_2}$	A_1A_2	0.5
$\mathbf{A_1A_2}$	A_1A_3	0.25
$\mathbf{A_1A_2}$	A_3A_4	0
$\mathbf{A_1A_1}$	A_2A_2	0

Subscripts represent four alleles. For a SNP there are only two; for a microsatellite there could be many. Over multiple loci, the average p(IBS) is an estimate of the relationship of a pair of individuals.

A feature of p(IBS) is that the relationship of an individual with itself is 1 for and inbred line and 0.5 for a completely heterozygous individual, while the relationship between two completely different individuals is zero: in line with expectations from pedigree relationships. For use in the mixed model equations, relationships estimated by p(IBS) would be doubled for a diploid, so that the diagonal of **K** would be 2 for an inbred line and 1 for an outbred individual. If this is not done however, it simply means that the estimate of σ_g^2 will be doubled to compensate and the estimates of genetic value will still be correct: that is to say, the BLUPs will still be correctly shrunk.

Advantages of p(IBS) are that it is simple to understand, is easily calculated for all ploidy levels and is easily applied to multi-allelic loci.

A5.2. van Raden's method.

IBS relationship matrices are not the favored method, as they treat all alleles and loci equally, yet a match between two rare alleles is more indicative of close relationship than a match

between two common alleles. Biallelic markers are usually called by the numbers of copies of the reference allele an individual carries: 0, 1, 2 for a diploid or 0, 1, 2, 3, 4 for an autotetraploid. The common practice is to standardize such marker scores to a mean of zero by subtracting twice the reference allele frequency for a diploid, four times the reference allele frequency for a tetraploid and so on. After standardizing, the relationship between a pair of individuals is estimated as the average of the cross product of these standardized variables. Since carriers of rare alleles have a greater deviation from zero, a match of rare alleles now indicates a closer relationship than a match of common alleles. Writing the matrix of standardized variables as **W** with rows equal to the number of individuals, and columns equal to m, the number of markers, the relationship matrix is:

$$\mathbf{K} = \mathbf{WW}'/2\Sigma\, p_k q_k$$

or

$$r_{ij} = \Sigma\,[(w_{ik}-2p_k)(w_{jk}-2p_k)] \;/\; 2\Sigma\, p_k q_k$$

Many modifications and alterations to this method have been published and discussed, but this has withstood the test of time and is the default method in many packages.

For autopolyploids, this becomes:

$$\mathbf{G} = \mathbf{WW}'/[\text{trace}(\mathbf{WW}')/n]$$

Where:

n is the number of individuals (not the number of markers)

For diploids, the diagonal elements (the relationship of an individual with itself) is an estimate of 1+F.

The individual values of **K** can be less than zero. This is in contrast to pedigree estimates, which must lie between 0 and 2. The negative values of the genomic relationship matrix must not be set to zero. A relationship of zero should be regarded as an average relationship among the lines contributing to the dataset. Pairs of lines may be less related than the average and will therefore have estimates of relationship below 1. Similarly, diagonal elements which are greater than 2 should be left alone.

A5.3. Example calculation of relationship matrix (K) from three individuals using six markers.

ID	M1	M2	M3	M4	M5	M6
1	2	0	1	2	1	2
2	2	0	2	2	1	1
3	1	2	0	0	0	2

Markers are coded as the number of reference alleles carried by a diploid individuals.

Assume these individuals come from a larger population with allele frequencies:

M1	M2	M3	M4	M5	M6
0.5	0.9	0.1	0.1	0.5	0.9

Standardize the marker scores to a (population) mean of zero by subtracting 2x the allele frequency.

	M1	M2	M3	M4	M5	M6
p(A)	0.5	0.9	0.1	0.1	0.5	0.9
2pq	0.25	0.18	0.25	0.18	0.5	0.18
Σ(2pq)	1.72					

ID	M1	M2	M3	M4	M5	M6
1	1	-1.8	0.8	1.8	0	0.2
2	1	-1.8	1.8	1.8	0	-0.8
3	0	0.2	-0.2	-0.2	-1	0.2

Strictly, the allele frequencies used to adjust the marker scored should be those of the founder or ancestral population. In practice, the sample allele frequencies are commonly used and are acceptable provided the sample size is not too small (as here). Ignoring the ID column this is the relationship matrix, **K** (before scaling).

K'		
1	1	0
-1.8	-1.8	0.2
0.8	1.8	-0.2
1.8	1.8	-0.2
0	0	-1
0.2	-0.8	0.2

KK'		
8.16	8.76	-0.84
8.76	11.36	-1.24
-0.84	-1.24	1.16

KK' / Σ(2pq)		
4.744	5.093	-0.488
5.093	6.605	-0.721
-0.488	-0.721	0.674

Although the values appear improbably high (a consequence of using only six markers), there are several points to note:

1. The diagonals are generally larger than the off-diagonals: a relationship of an individual to itself is generally expected to be higher than its the relationship to other individuals.
2. The off-diagonals can be negative.
3. Individuals 1 and 2 appear to be closely related: high off-diagonal relationship: looking at the marker data they are identically homozygous at four out of the six loci.

A5.4. Missing data

Missing marker data causes problems in the estimation of **K**. Ideally, the missing data should be imputed. For most SNP data sets, after quality control, including removal of poor markers, the problem is slight and simply inserting the average genotype score is acceptable (which is zero after standardizing). For genotyping by sequencing, the missing data problem is extreme and one of the several methods to impute missing data must be used.

A5.5. A comment on the number of markers

An assumption of GBLUP is that a high density of markers is used to estimate genomic relationships. The precise number will vary depending on the history of the population with which you are working. This can be tested by empirically by cross-validation.

Trait data for a proportion of individuals, say 1/10th are removed from the dataset, and their genetic values predicted from their genomic relationships with the remaining 9/10^{ths} of individuals. The accuracy of the prediction is assessed by correlation of observed and predicted traits. This is repeated for other subdivisions of the data and also repeated with varying numbers of markers. The relationship between prediction accuracy and marker number can therefore be quantified.

For most plant breeding applications, as a rule of thumb, thousands but not tens of thousands of markers are required. In some very narrowly based populations, for example progeny from a single cross, many fewer markers, around a hundred, may give adequate prediction accuracy.

Table 5. Comparison of BLUE, BLUP, pBLUP and gBLUP from a small balanced dataset

	BLUE	BLUP	pBLUP	gBLUP
V1	-9.45	-6.3	-5.705	-8.246
V2	4.55	3.03	1.295	1.801
V3	4.90	3.27	2.940	0.817
Average	0	0	-0.490	1.493
V1	-9.45	-6.3	-5.125	-5.169
V2	4.55	3.03	1.785	1.276
V3	4.90	3.27	3.430	3.894
Average	0	0	0	0

All BLUPs are shrunk compared to the BLUEs. In this example, the pBLUPs and gBLUPs are shrunk to a similar amount. Just as in pBLUP, multiplying the relationship matrix **K** by larger values of σ_g^2 increases the influence of the genetic variance/covariance matrix **G** on estimates of genetic value and BLUPs shrink less from the BLUEs. Reducing σ_g^2 towards zero causes the BLUPs to shrink towards zero.

A6. Further methods and developments

The method of gBLUP described here to estimate genetic value from markers is more than adequate for most breeding purposes. However, methods for trait prediction continue to attract research. Typically, newer methods or developments offer some improvements in prediction accuracy in some circumstances. However, compared to rrBLUP and GBLUP, the improvements are usually slight and not large enough yet to warrant a switch from these standards. In addition, the alternatives are often computationally more intensive, harder to understand, and software is less accessible. We list some below, with limited explanation or comment:

1. "The Bayesian alphabet" is a set of methods – Bayes A, Bayes B ..., which approach prediction using Bayesian statistics rather than through the mixed model approaches described here.
2. Use of multiple sets of random effects, each with its own relationship matrix. For example, a separate genomic relationship matrix can be estimated for dominance effects, epistatic effects, or even for individual chromosomes or different marker classes.
3. Machine learning methods.
4. Feature selection methods. These select subsets of markers which appear, on their own, to give the best prediction accuracy. The easiest to understand of these is the LASSO, which is closely related to ridge regression but selects only a subset of markers whereas ridge regression includes all markers in the prediction equation.

5. Combined methods: pedigree BLUP and GBLUP can be combined. The lasso and ridge regression can be combined (called the elastic net). As a very simple example, BLUP and BLUE can be combined (some markers, tagging known QTL for example, could be treated as fixed effects and other markers as random effects).

A6.1. Relationship matrix in autopolyploid species.

Estimation of relationships for autopolyploid species differs from that for diploids. However, software is available, for example AGHmatrix. A ploidy-specific matrix should be substituted for the usual default diploid matrix used in most packages.

A6.2. Relationship matrix in selfing species

There are at least three major problems:

Firstly, for fully inbred lines or doubled haploids, if there are no crosses between related lines, **K** is simply twice the equivalent matrix for an outbred population and standard pBLUP software can be used. Using a relationship matrix half of its actual value will be compensated for by estimation of σ_g^2 which is twice its actual value (recalling that σ_g^2 is the estimate for the outbred ancestral population of unrelated lines). However, this is rarely the case; published pedigrees for inbreeding species are always complex, for example in Figure 5, and the relationship matrix **K** will be incorrect.

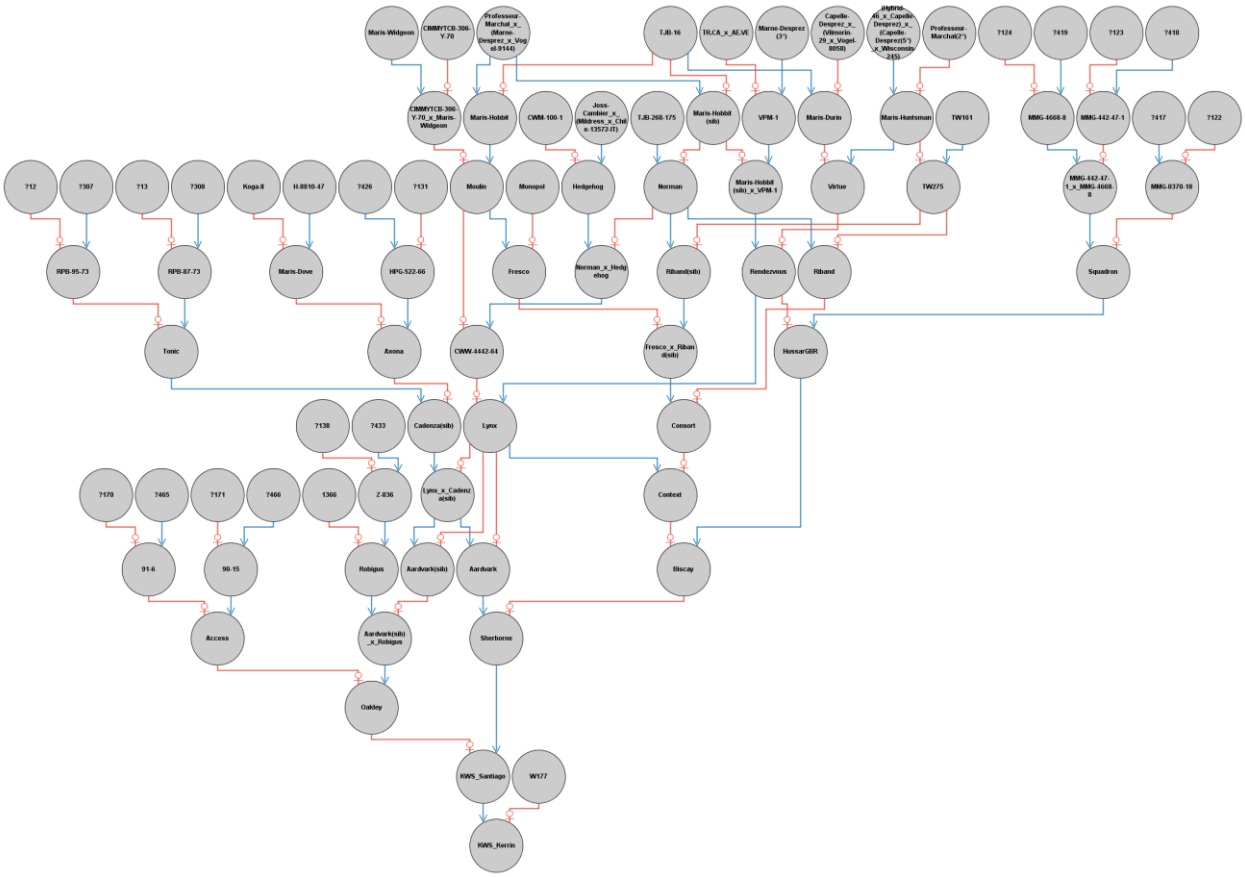


Figure 5. **An example plant pedigree.** Part of the UK wheat pedigree showing ancestors of the variety KWS Kerring. Derived from Fradgley et al. (2019) A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. PLoS biology, 17(2), p.e3000071.

This is easy to see: for inbred lines the relationship of an individual to itself is always 1, but in an outbreeding species this can vary between ½ and 1 depending on its inbreeding coefficient. At the moment, no pBLUP software exists which is explicit about the treatment of inbreeding species and that will ignore the distinction. The adverse consequences of this have not been described.

Secondly, the way pedigrees are recorded in crops like wheat and barley is a shorthand approximation. Line A x line B→ line C usually implies a number of generations between the AxB F_1 and line C. Even for doubled haploids, the F_1 is implicit rather than recorded in the pedigree. Kinship estimation can take this into account, but it is not routine.

Finally, the variety released and phenotyped often differs genetically from that used as a parent, even though they are recorded as identical. If selected lines are cycled quickly within the breeding program (good practice), then the parent could be an F_3 of F_4 individual to which the variety name is still attached. The released variety could be an $F_{5:7}$ family, for example, and the F_5 individual may not even be a direct descendant of the F_4 individual used in crosses. This introduces errors, with unknown effect, into the estimate of **K** for selfing species. Two 'smoking guns' for such problems are the seemingly very rapid cycling of lines in a conventional breeding program rather than the oft-quoted 10 years to create a variety in conventional breeding and miss-inheritance of genetic markers over and above the level expected from genotype errors.

A6.3. Prediction of lines with no trait data

pBLUP enables the prediction of genetic value for lines without a phenotype. It has a long history of use in estimating the genetic value of bulls for milk yield and cockerels for egg production. Use in crops has been more limited, though there are dioecious species (e.g. hemp, hops) where production is based on female plants but selection on the genetic value of males would increase response to selection. To estimate the genetic value of un-phenotyped individuals, they are included as extra columns in the **Z** matrix, with entries of zero, and extra rows and columns in the **G** matrix with entries equal to $k_{ij}\sigma_g^2$, where the k_{ij} are relationships between the phenotyped and unphenotyped individuals. Corresponding estimates of genetic value are returned in **g**.

Note however, that individuals or lines in the same cross or full-sib family will be predicted to have the same genetic value. Pedigree relationships cannot distinguish between individuals sharing the same parents. Trait information from more distant relatives will still be incorporated into the estimate, but Mendelian sampling variation within a family cannot be accessed by pedigree information alone.