Statistical Indicators

E-49

Integration of genomic data into the national evaluation (DGV-PBLUP)

▪ Introduction

With the introduction of genomic evaluations a method of incorporating genomic data into the national evaluation became necessary. To incorporate genomic data into the Dutch/Flemish national genetic evaluation a method is used called DGV-PBLUP. It is a conventional pedigree BLUP analysis (PBLUP) in which *direct genomic values* or DGV are fitted as *prior means*, such that effects in the model are estimated as deviations from these (and not zero as is the case in a regular PBLUP).

DGV of genotyped animals are calculated from genotypes and allele substitution effects estimated in a single step SNP BLUP evaluation. These DGV are mathematically propagated to ungenotyped relatives of genotyped animals, such that potentially all animals in a evaluation benefit from inclusion of genomic data.

In the following paragraphs more details will be provided regarding the method of DGV-PBLUP. Additionally some more details are given about the method to derive genomic reliabilities for all animals in the evaluation. Finally, some information on the type of DGV allowed in the national evaluation is given.

▪ Principles of DGV-PBLUP

The equations of the model were derived from the single step SNPBLUP linear equations proposed by Liu et al. (2014). If we assume that estimates of SNP effects $\hat{\mathbf{g}}$ are known before performing a single-step genomic prediction, then the vector **d** with predicted DGV of genotyped animals can be computed as **d** = **Zg**, where **Z** is the genotyped matrix centred with observed allele frequencies, and we can assume the following prior multivariate normal (MVN) distribution for the genetic additive effects $\mathbf{u}$:

$$[\mathbf{u}|\hat{\boldsymbol{\mu}}, \mathbf{A}^*] \sim MVN(\hat{\boldsymbol{\mu}}, \mathbf{A}^* \sigma_u^2)$$

with

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1} \\ \mathbf{I} \end{bmatrix} \mathbf{d}$$

and

$$\mathbf{A}^{*-1} = \begin{bmatrix} \mathbf{A}^{nn} & \mathbf{A}^{ng} \\ \mathbf{A}^{gn} & \mathbf{A}^{gg} + \left(\frac{1}{w} - 1\right)\mathbf{A}_{gg}^{-1} \end{bmatrix},$$

where the subscripts n and g refer to ungenotyped and genotyped animals, respectively,

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{nn} & \mathbf{A}_{ng} \\ \mathbf{A}_{gn} & \mathbf{A}_{gg} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{nn} & \mathbf{A}^{ng} \\ \mathbf{A}^{gn} & \mathbf{A}^{gg} \end{bmatrix}$$

is the inverse of the pedigree relationship matrix partitioned between genotyped and ungenotyped animals, w is the proportion of additive genetic variance explained by the residual polygenic effects, σ_u^2 is the genetic variance, $\mathbf{d}$ is the vector with DGV of genotyped animals, and $\mathbf{I}$ is an identity matrix.

The system of equations associated with these assumptions, hereafter called DGV-PBLUP, is written as follows:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{*-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} + \mathbf{A}^{*-1}\sigma_u^{-2}\widehat{\boldsymbol{\mu}} \end{bmatrix}$$

where $\widehat{\boldsymbol{\beta}}$ is the vector of estimated fixed effects, $\mathbf{y}$ is the vector of records, $\mathbf{R}^{-1}$ is the inverse of the residual variance structure matrix, and $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating records to the fixed and additive genetic effects, respectively.

The system of equations of DGV-PBLUP is equivalent to a single-step genomic evaluation, provided that the SNP effects $\hat{\mathbf{g}}$ were estimated using the same phenotypic, genomic and pedigree information (Vandenplas et al, 2021).

An attractive feature of the DGV-PBLUP method is that no extra correlated traits have to be fitted to incorporate DGV information in a pedigree BLUP evaluation. Neither does it require a post-processing step to integrate DGV. The run times of DGV-PBLUP are comparable to the run times of conventional pedigree BLUP evaluations. The run times of routine single-step SNPBLUP evaluations on average are 2.4 times longer than either conventional or DGV-PBLUP evaluations.

▪ Genomic reliabilities

The estimation of genomic reliabilities relies on the concept of *expected daughter contributions* or EDC. A reliability *r* can be transformed into EDC by:

$EDC_i = \frac{\alpha r_i}{1-r_i}$ where $\alpha = \frac{4-h^2}{h^2}$ is a multiplication factor accounting for the heritability (h^2) of a trait. Back transformation of EDC to a reliability is achieved by:

$$r_i = \frac{EDC_i}{\alpha + EDC_i}$$

The concept of EDC is a convenient way of making reliabilities additive.

Basic algorithm. The estimation of genomic reliabilities consists of the following steps:

1. Calculate conventional reliabilities, transform to EDC
2. Calculate SNP *genotype confidences*

    a. Set genotyped animals to rate of imputation
    b. Propagate genotype confidence
3. Calculate DGV reliabilities for all animals; transform to EDC
4. Sum conventional and genomic EDC to obtain final EDC; transform to GEBV reliabilities

Conventional reliabilities are calculated using standard algorithms such as the method for calculating multi-trait EDC (MT-EDC) and transforming these to reliabilities of traits in a multi-trait evaluation devised by Liu et al. (2001).

Direct genomic reliability. The reliability of a single genotype is derived from validation results in the following manner: GEBV validation is performed on data of a group of validation animals; typically candidate animals without progeny or observations, only a genotype. In the GEBV validation procedure GEBV for a trait from a full evaluation ($GEBV_{full}$) are used as the phenotype that the breeding values (BV) is predicting:

$$GEBV_{full} = b_0 + b_1 BV$$

The regression is performed twice: Once on conventional EBV, usually PA, and once on genomic breeding value (GBV) from a truncated run, where phenotypic information from daughters of bulls is omitted. The realized coefficient of determination R^2 of these regressions can be interpreted as reliabilities of PA en GBV in predicting the full GEBV. The difference in R^2 is a measure of the information added by the genotype. The exact amount of information added by the genotype can be calculated by transforming the R^2 into EDC, given the heritability of the trait under consideration:

$$GEBV_{full} = b_0 + b_1 PA \quad => \quad R^2_{PA} \quad => EDC_{PA}$$
$$GEBV_{full} = b_0 + b_1 GBV \quad => \quad R^2_{GBV} \quad => EDC_{GBV}$$

And

$$EDC_{GBV} - EDC_{PA} = EDC_{geno}$$

The variable EDC_{geno} is subsequently transformed back to a reliability R^2_{geno} which is the mean reliability of the direct genomic effect (DGV) for a particular trait. It is assumed that R^2_{geno} for a trait is approximately equal for all genotyped animals.

Genotype confidence. Since propagation is aimed at modelling the information on unknown genotypes of ungenotyped animals, given genotyped relatives, the concept of *genotype confidence* (R^2_{SNP}) is introduced. These genotype confidences are a number between 0 and 1 for each animal, indicating how well the genotype of that animal is known. A value of 0 indicates the genotype of that animal is not known at all, while a value of 1 indicates the genotype is fully known.

The genotype confidence of genotyped animals is set to the rate of imputation (typically nearly 1, but accounting for errors in the genotyping process). For ungenotyped animals genotype confidences are initially set to zero. Genotype confidences are transformed to EDC using an arbitrary heritability and propagation is performed.

The resulting vector is back transformed to genotype confidences R^2_{SNP}, which are (very nearly) 1.0 for genotyped animals (particularly those with many genotyped offspring) and < 1.0 for ungenotyped animals with few or no genotyped relatives. The reliability of the DGV for each animal *i* , genotyped or ungenotyped, for a trait *t* is calculated from the direct genomic reliability of trait *t* and the genotype confidence of animal *i*:

$$R^2_{DGV,it} = R^2_{SNP,i} \times R^2_{geno,t}$$

Transforming R^2_{DGV} to EDC for each animal and all traits gives EDC^*_{add} (propagated added EDC) which are added to the EDC from conventional information (EDC_{conv}):

$$EDC_{final} = EDC^*_{add} + EDC_{conv} \Rightarrow R^2_{final} \text{ (GEBV reliability)}$$

Implementation in MT-EDC
To correctly account for the multi-trait character of many evaluations the procedure described above was adjusted in the following manner, based on the MT-EDC method described in Liu et al. (2001).

There are two key operations in the mathematics of MT-EDC:

$$\mathbf{Y} = Y\_from\_R(\mathbf{R}) = 4((\mathbf{I}-\mathbf{R})^{-1} -\mathbf{I})*\mathbf{G}^{-1} \quad \text{; MT-EDC from reliability matrix}$$
$$\mathbf{R} = R\_from\_Y(\mathbf{Y}) = \mathbf{I} - (¼\ \mathbf{YG} + \mathbf{I})^{-1} \quad \text{; reliability matrix from MT-EDC}$$

Where **Y** is a *n* x *n* matrix of MT-EDC of n traits in a multi-trait evaluation for an animal. The matrix **R** is the corresponding matrix containing reliability values of multiple trait EBV for an animal. **G** is the *n* x *n* genetic covariance matrix and **I** is an *n* x *n* identity matrix.

In the conventional reliability estimation $\mathbf{Y}_{conv}$ is stored in memory for each animal in the evaluation.

For the genomic reliability approximation the vector **r** containing the R^2_{DGV} of an animal for all traits in the evaluation is transformed to a diagonal matrix $\mathbf{R}_{add}$. The MT_EDC matrix for this reliability matrix is obtained through:

$$\mathbf{Y}_{add} = Y\_from\_R(\mathbf{R}_{add})$$

The matrix $\mathbf{R}_{final}$ is then obtained summing the two MT-EDC matrices:

$$\mathbf{R}_{final} = R\_from\_Y(\mathbf{Y}_{conv} + \mathbf{Y}_{add})$$

Reliability of traits and indices
The MT-EDC algorithm allows for reliability estimation of extra traits, defined as indices of traits present in the evaluation. To this end a matrix **Θ** of size *m* x *n* is defined, where *m* is the number of output traits. Each line of **Θ** contains index weights relating the index trait to evaluation traits (if reliabilities are only needed for evaluation traits, **Θ** simply is an identity matrix).

A reference variance matrix $\mathbf{G}_v$ is calculated:

$$\mathbf{G}_v = \mathbf{\Theta}\ \mathbf{G}\ \mathbf{\Theta}'$$

This results in a *m* x *m* genetic covariance matrix of traits and/or indices.

The variance matrix of an animal *i* with reliability matrix $\mathbf{R}_{final}$ is calculated as:

$$\mathbf{G}_i = \mathbf{\Theta}\ (\mathbf{G}\ \mathbf{R}_{final}\ )\ \mathbf{\Theta}'$$

For each trait *j* of *m* defined traits in **Θ** the approximate (genomic) reliability for animal *i* is then calculated as:

$$r^2_{ij} = \mathbf{G}_i(j,j)/\mathbf{G}_v(j,j)$$

▪ Selection of data

The direct genomic values or DGV are calculated from genotypes and allele substitution effects estimated in a single step genomic evaluation. Animals with DGV will have their DGV included in the national evaluation if:

- The animal is a genotyped bull and
    - owned by an AI organization participating in the genomic evaluation.
    - the animals is a Eurogenomics bull.
    - a fee has been paid.
    - is not an AI bull and has been culled.
- Or when the animal is a female.

Currently, AI organizations participating in the Dutch national genomic evaluations are CRV and KI Kampen.

Eurogenomics is a European network of AI companies performing genomic evaluations. To more accurately estimate DGV and widen the scope of genomic evaluations, the participants in Eurogenomics have agreed to exchange genotype information of their sires for use in genomic evaluations in the participants respective countries. In practice this means that a Eurogenomics bull with a genotype will have a Dutch/Flemish DGV used in the DGV-PBLUP system.

▪ References

Liu Z., Reinhardt F., and Reents R. (2001), The effective daughter contribution concept appliedto multiple trait models for approximating reliability of estimated breeding values. Interbull Bulletin 27: 41-47.

Liu Z., M. E. Goddard, F. Reinhardt, and R. Reents (2014), A single-step genomic model with direct estimation of marker effects, J. Dairy Sci. 97 :5833–5850. DOI: 10.3168/jds.2014-7924

Liu, Z. et al. (2017), Approximating genomic reliabilities for national genomic evaluation. Interbull Bulletin 27: 75-85.

Vandenplas J., Eding H. and Calus M. (2021), Interim genomic prediction considering newly acquired genotypes and phenotypes, Interbull Bulletin 56. https://journal.interbull.org/index.php/ib/article/view/87

VanRaden P.M., and Wiggans G.R. (1991), Derivation, calculation, and use of national animal model information, J. Dairy Sci. 74(8):2737-2146, DOI: 10.3168/jds.s0022-0302(91)78453-1