CAN WE FRAME AND UNDERSTAND CROSS-VALIDATION RESULTS IN ANIMAL BREEDING?

Conference Paper · July 2017

CITATIONS
13

READS
746

2 authors, including:

Andres Legarra
French National Institute for Agriculture, Food, and Environment (INRAE)
295 PUBLICATIONS **8,803** CITATIONS

SEE PROFILE

CAN WE FRAME AND UNDERSTAND CROSS-VALIDATION RESULTS IN ANIMAL BREEDING?

A. Legarra[1], A. Reverter[2]

[1] UMR 1388 GenPhySE, INRA, Castanet Tolosan, France
[2] CSIRO Agriculture and Food, 306 Carmody Rd., St. Lucia, QLD 4067, Australia

SUMMARY

Performance of genomic selection is typically evaluated by cross-validation. In this work we review and point out some problems and features of the cross-validation metrics. Then we propose a semiparametric alternative using statistics derived from the "Method R".

INTRODUCTION

Genomic prediction of breeding values via genomic BLUP (GBLUP) is expensive and requires initial and continuous investments in genotyping. State of the art theory so far does not yield convincing *a priori* estimates of the increased accuracy of genomic prediction vs. pedigree-based predictions. Thus, cross-validation has been extensively used (e.g. Legarra *et al.* 2008; VanRaden *et al.* 2009; Mantysaari *et al.* 2010; Christensen *et al.* 2012). The theory of cross-validation is poorly understood in the context of heavily related and selected data (but see (Gianola and Schön, 2016)). For instance, how to evaluate accuracy for maternal traits is very unclear. Here we provide a brief review of this topic and suggest some options.

CROSS-VALIDATION BIAS AND ACCURACY

What cross-validation? Forecasters such as pedigree-BLUP and GBLUP may behave differently according to what the "forecasted" target is. Breeders have a difficult task, namely, to forecast the best reproducers in order to select them. In this, they are different from *machine learners*, whose objective is (from our perspective) to forecast present phenomena. Thus, it is rather obvious that for breeders the best method is such that allows taking the best selection decisions, that it is, the method that best predicts future performance of an individual knowing its genetic background.

We will call this *forward cross-validation*. Its features are three-fold: (1) It needs the definition of a cut-off date; (2) It needs the construction of "Full" and "Reduced" data sets (Mantysaari *et al.* 2010; Olson *et al.* 2011); and (3) In its crudest form, it does not provide any form of randomisation and therefore a point estimate of goodness of prediction is obtained, without any associated measure of uncertainty.

In contrast, the classical *random folding k*-fold cross-validation in its most classic form splits randomly the data into *k* distinct sets and predicts one set from the remaining *k-1* sets. Its key features include: (1) Extremely simple to implement; (2) Provides estimates of standard error of metrics of cross-validation; (3) Not realistic in an animal breeding setting and the ranking of methods is not suitable for practical purposes; and (4) Tends to overfit (case of leave-one-out)

Some more esoteric forms of cross-validation exist. Legarra *et al.* (2008) *split folds "across" or "within" families*, obtaining very different results. But this is undoable (and little useful) for regular animal breeding data. The *k-means for cross-validation* (Saatchi *et al.* 2011) separates individuals into "most distinct" folds, and the i-th fold is predicted from the remaining k-1 folds. This does not answer the breeder's question, which most often wants to predict from *close*, not from *far* animals.

Which metrics? To assess the *predictive ability* of the different forecasters, animal breeders are highly formatted by Henderson's BLUP, which in turn was highly dependent upon dairy cattle

genetic improvement. Metrics commonly used come from linear regression, named in this paper *predictive abilities*, are:

$$\text{Bias: } b_0 = E(u - \hat{u}); \qquad \text{Slope: } b_1 = \frac{Cov(u,\hat{u})}{Var(\hat{u})}; \qquad \text{Accuracy: } r = \frac{Cov(u,\hat{u})}{\sqrt{Var(u)Var(\hat{u})}}$$

Sometimes mean squared error is used ($MSE = b_0^2 + \sigma_u^2(1 + r^2/b_1^2 - 2r^2/b_1)$). Properties of BLUP in absence of selection are no bias, slope of 1, and maximum accuracy. Henderson defined this at the individual level on a frequentist basis (over conceptual repetitions). Bias=0 and slope=1 ensure fair comparisons across old and young animals. This is important if the scheme mixes proven and young animals, like dairy cattle. It seems less relevant in schemes were reproducers are culled quickly (pigs, chicken) with beef species falling someone in the middle, we believe. Deviations may exist if there is selection, because bias and slope are related to genetic gain and dispersion (see Figure 1).

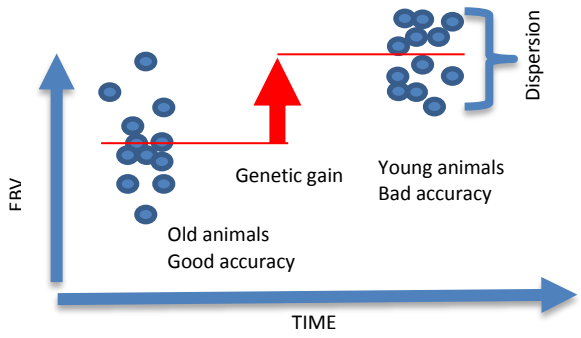

Figure 1 Typical scenario for retrospective analysis.

What is it meant by classical bias? Animal breeders probably agree to Henderson's (1973) sentence "*most users would, I think, be reluctant deliberately to bias comparisons between different groups, for example to underevaluate young sires as compared to older ones*". Here we have an operational definition of bias. In formal terms this implies that at a given point in time:

$$\begin{aligned} b_0^{[Henderson]} &= \left(\mathbf{1}'\hat{\boldsymbol{u}}_{group1} - \mathbf{1}'\hat{\boldsymbol{u}}_{group2}\right) - \left(\mathbf{1}'\boldsymbol{u}_{group1} - \mathbf{1}'\boldsymbol{u}_{group2}\right) \\ &= \left(\mathbf{1}'\hat{\boldsymbol{u}}_{group1} - \mathbf{1}'\boldsymbol{u}_{group1}\right) - \left(\mathbf{1}'\hat{\boldsymbol{u}}_{group2} - \mathbf{1}'\boldsymbol{u}_{group2}\right) \end{aligned}$$

This definition has practical implications: if the candidates are chosen *across* groups, selection decisions are optimal if there is no bias. Thus, it is expected that $b_0^{[Henderson]} = 0$. There may be several definitions of groups: (1) Different conditions (grazing vs. indoor fed cattle). This case should be addressed by the model used for evaluation; (2) Within country, different amounts of information that cumulate in time (progeny-tested vs. genomic bulls). This case is strongly affected by within-country genetic trend (see below); (3) Same amount of information, but different origins (US vs. FR). This case is most affected by wrong estimates of the difference in genetic level across countries (Bonaiti *et al.* 1993; Powell and Wiggans 1994).

The Interbull definition. Interbull uses retrospective tests (Boichard *et al.* 1995; Mantysaari *et al.* 2010) that compare EBV's *before* and *after* progeny testing.

$$b_0^{[Interbull]} = \mathbf{1}'\hat{\boldsymbol{u}}_t - \mathbf{1}'\hat{\boldsymbol{u}}_{t-1}$$

If progeny testing gives exact EBVs, then $\hat{\boldsymbol{u}}_t = \boldsymbol{u}_t$ and $b_0^{[Interbull]} = \mathbf{1}'\boldsymbol{u} - \mathbf{1}'\hat{\boldsymbol{u}}_{t-1}$.Note that $b_0^{[Henderson]} \neq b_0^{[Interbull]}$, but if group1 is "very old" proven bulls and $\hat{\boldsymbol{u}}_t = \boldsymbol{u}_t$ and group2 is genomic bulls (then becoming proven bulls) then $b_0^{[Henderson]} = b_0^{[Interbull]}$. This may be rather obvious, but it only holds for progeny testing data.

What happens under selection? Assume that we want to compare selection candidates with "proven" animals. If there is *no* selection, then $\mathbf{1}'\boldsymbol{u}_{group1} = \mathbf{1}'\boldsymbol{u}_{group2}$ and there is actually no need to make the test. Alas, if *there is* selection, then

$$b_0^{[Henderson]} = \left(\mathbf{1}'\hat{\boldsymbol{u}}_{group1} - \mathbf{1}'\hat{\boldsymbol{u}}_{group2}\right) - \left(\mathbf{1}'\boldsymbol{u}_{group1} - \mathbf{1}'\boldsymbol{u}_{group2}\right) = n\left(\hat{\Delta} - \Delta\right)$$

in other words, unbiasedness requires a correct (unbiased!) estimate of the realized genetic trend.

What is overdispersion, a.k.a {Interbull, genomic} bias? Is it affected by selection?

Dairy cattle breeders are much concerned by overdispersion of genomic proofs. If there is too much dispersion of $\hat{\boldsymbol{u}}_{genomic}$, the retained candidates will have unfairly high $\hat{\boldsymbol{u}}_{genomic}$. This could be staten more formally as "the mean of the EBVs of the selected candidates should be equal to the mean of the TBVs". If selection is by truncation and under multivariate normality, the true mean *after* selection is $\mu_T = (\mathbf{1}'\boldsymbol{u})/n + ir\sigma_u$, but this mean is (implicitly) predicted before selection as $\mu_E = (\mathbf{1}'\hat{\boldsymbol{u}})/n + i\sigma_{\hat{u}}$.

For $\mu_T = \mu_E$ to hold, we need the first unbiasedness condition (b_0 above), plus a second condition, $\sigma_{\hat{u}} = r\sigma_u$. But this condition *only* holds if $Cov(u, \hat{u}) = Var(\hat{u})$, which amounts to the regression coefficient to be 1:

$$b_1 = \frac{Cov(u, \hat{u})}{Var(\hat{u})}$$

This is the Interbull official, and most put forward, test of unbiasedness and nowadays more often called as "bias". It is easy to see why $b_1 = 1$ may not hold, namely, because selection modifies variances in rather unpredictable manners. The expected $Cov(u, \hat{u}) = Var(\hat{u})$ holds under quite restrictive conditions (Henderson 1982).

Evaluations can easily be biased. Unbiasedness of current genetic evaluations is more wishful thinking than an established fact. Unbiasedness exist only if several conditions hold:

- The model is correct (linear model, effects, heritabilities…)
- The selection process is described by the data
- Multivariate normality

Thus, there are many reasons why there is wrong estimate of the genetic trend and thus there will be bias:

- Collinearity of contemporary groups and genetic trend (this is the usual case)
- Genetic groups in the model
- Heritability is wrong (or changes with time)
- Analysis are single trait whereas selection is multiple trait
- Selection decisions not based on data.

In addition, genetic gain can be estimated one generation forward (but no more) unless an explicit selection model is included. In other words, retrospective analysis cannot be done deleting two generations of records. This would need explicit introduction of the selection process.

Why some species/traits seem biased where others do not? Basically, if there is *no* selection then *automatically* $b_0 = 0$ holds (i.e., all possible sets of candidates have 0 average value), and most likely $b_1 = 1$ holds, because selection does not change variances, and if a decent estimator of genetic variance is used, then genetic parameters are such that $b_1 = \frac{Cov(u,\hat{u})}{Var(\hat{u})} = 1$ by construction, in particular in a BLUP context. So, bias is expected to increase more with higher genetic gains.

An example is *pigs*. Christensen *et al.* (Christensen *et al.* 2012) found slopes below 1 (~0.9) for a heritable, selected trait (daily gain), whereas Xiang *et al.* (Xiang *et al.* 2016) found regressions nearly one for hard-to-select trait litter size.

In Lacaune dairy *sheep* (Baloche *et al.* 2014), we can put together the following. Figure 1 shows the regression slopes vs. the expected genetic gain or the expected loss of genetic variance based on Robertson (1977) . In theory, the reduction in variance is accounted for by genetic evaluation (Bijma 2012). In practice, this does not seem to be the case. A possible solution may be to reestimate this variance in each cycle of selection.

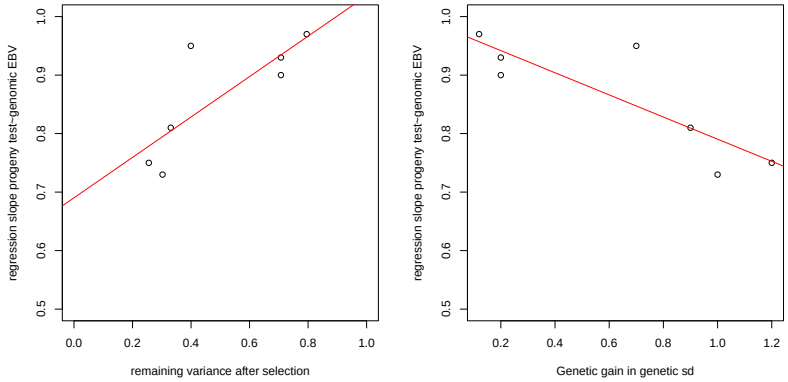

Figure 2 Slope b_1 vs. expected reduction in genetic variance (left) or genetic gain (right) by trait in Lacaune dairy sheep.

Vitezica *et al.* (2011) compared by *simulation* several predictors in selected populations in a SSGBLUP context. Statistic b_1 generally indicated bias, that was higher with less heritability. High heritability increases the selection differential and reduces variances, but it also gives more information. Interestingly, the only method which provided unbiased $b_1 = 0.99$ resulted in strong bias $b_0 = 1.38\sigma_u$. Thus, *both* bias should be checked.

What do we mean by accuracy? In animal breeding textbooks, accuracy (r, with reliability r^2) is presented twice: first, as a component of $\Delta_G = ir\sigma_u$ (so, a populational parameter) and, second, as a measure of uncertainty of $\hat{u}$ (an individual parameter). However, when selecting from real populations, EBVs are correlated across individuals, so the individual accuracies may be meaningless. In other words: it is pointless to obtain $r_i = 0.70$ and $r_j = 0.70$ if $r(\hat{u}_i, \hat{u}_j) = 0.69$.

Cross-validation accuracies are computed as correlations $r^2 = \frac{Cov(u,\hat{u})}{Var(u)Var(\hat{u})}$. They indicate our ability to rank individuals *within* a cohort. The fact that these accuracies are computed regardless of the correlated structure of both u and $\hat{u}$ has unclear implications. In fact, it can be shown that, if Hendersonian conditions hold, $E(r)^2 = 1 - \frac{(\overline{diag(\boldsymbol{C}^{22})} - \overline{\boldsymbol{C}^{22}})}{(\overline{diag(\boldsymbol{G})} - \overline{\boldsymbol{G}})}$ is the expectation of the observed reliability. This reliability takes into account the "classical" reliability contained in the diagonal terms but also the relationships a priori (in $\boldsymbol{G}$) and a posteriori (in $\boldsymbol{C^{22}}$) across individuals. If the evaluation method cannot rank correctly *within* the validation sample, then diagonal and off-diagonal values of $\boldsymbol{C}^{22}$ are similar and reliability drops down. This is a desirable behaviour.

Selection also affects observed cross-validation accuracy (Edel et al., 2012; Bijma 2012). If the cross-validation test uses elite animals, accuracies are underestimated. In other words, it is easy to rank all animals, but more difficult to rank elite animals. The reduction is such that

$$r^2_{selected} = 1 - (1 - r^2_{unselected}) \frac{\sigma^2_{u_{unselected}}}{\sigma^2_{u_{selected}}} \quad .$$

ISSUES OF CROSS-VALIDATION METRICS

The accuracy of cross-validation metrics. After an experiment has been carried out, the breeder wants to know if the genomic accuracy is really different from the parents average accuracy. A

simple method is to use the theoretical standard error of the estimates; for b_0 and b_1 these are from classical regression theory. For the correlation, this is a bit more convoluted, but an option is to use Fisher's z-transform: $z = \frac{1}{2} ln \frac{1+r}{1-r}$ has approximate s.e. $1/\sqrt{n-3}$ where n is the number of data points used. From this a confidence interval can be worked out. For instance, in the Basco-Bearnaise breed genomic predictions of 87 rams were 0.06 more accurate than parent averages (Legarra *et al.* 2014); this implies a rather symmetric 95% confidence interval of $[-0.15,0.27]$.

There is a source of bias and two sources of randomness in cross-validation metrics. The source of bias is that individuals are related both at the stage of prediction (parent average and genomic) and later, at the stage of validation (moment at which they have data; except for the case of progeny-tested animals for which proofs can be assumed uncorrelated). This has been discussed above. The two sources of randomness are: (1) Sampling of the reference population, (2) Sampling of the validation population. Fisher's z-transform and Hotelling-Williams test include both. However, they do not consider that individuals are related, and therefore the accuracy is likely to be overestimated. Again, a theoretical equation can be worked out to estimate $Var(r)$.

(Re)Sampling of the validation population. A more practical approach involves using (re)sampling techniques. In k-fold cross-validation this is immediate but, as discussed before, the setting is not realistic. In (Mäntysaari and Koivula 2012; Legarra *et al.* 2014; Cuyabano *et al.* 2015), sampling of the validation population was addressed by *bootstrapping*, i.e. sampling *n* individuals with replacement from the original *n* individuals in the validation data set. This method main virtue is that it avoids strong influence of outliers in the validation data set. It also allows formal comparisons of accuracies. Its main drawback is that it does not addresses the sampling of the reference population.

(Re)sampling of the reference population. Recently, (Mikshowsky *et al.* 2016) bootstrapped, not the validation, but the *reference* population. This also provides distribution of metrics. However, it may be argued that, in a dairy cattle reference population, including a sire twice (what the bootstrapping actually does) is like including it once, because the accuracy of the sire pseudo-phenotype is close to 1 in dairy cattle. Thus, including it twice will not change much the solution for the sire – or the contribution of the sire to SNPs solutions. Therefore, randomness comes from *removing* sires more than by *overrepresenting* sires. In that sense, Mikshowsky *et al.* (2016) bootstrap corresponds to Tukey's jackknife with more than one data point removed.

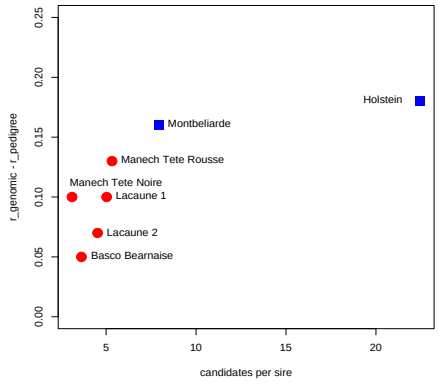

Figure 3 Genomic accuracy and family size.

Superiority of genomic on pedigree predictions is a function of family structure of the validation data set. Consider a set of two generations, a generation of parents and one of descendants: *n* full-sib families with *k* offspring each. Parents have information (say, own weight) but there is not information for the offspring. We can ask: is it worth doing genomic prediction?

Families can be easily ranked based on parent average, but there is not possibility to rank within families with pedigree information. However, genomic information *can* rank *within* family as well as *across* families. Thus, the observed benefit of GBLUP by retrospective analysis will be larger in a

set composed of *few* families with a large number of candidates *within* families. In the limit, if there is one big family, pedigree prediction has 0 accuracy, whereas if there are n families with 1 offspring each, pedigree and genomic predictions should behave similarly.

This is supported by Figure 3 in which we plot the genomic vs pedigree accuracy for milk yield for five dairy sheep and two dairy cattle breeds in France, as a function of family size. Clearly, the larger the family size, the larger the benefit because genomic selection allows distinguishing sibs. This raises several questions: (1) Do comparisons reflect "genetic architecture" or merely data structure in the validation? (2) Do selection schemes that select across families get less benefit from genomic selection? (3) Is Holstein gaining a lot from genomic selection because it has higher LD than other breeds or just as an artefact of its family structure?

Which variables to use on the metrics? In the dairy industry, sires do not have phenotypes, so that comparisons are between (G)EBV's and the "true" progeny proofs or deregressed proofs. In other species, it is more common to compare (G)EBV's to "true" phenotypes, say $\boldsymbol{y}$, using an approximation $r = Corr(GEBV, y)/h$ where h^2 is the heritability (Legarra *et al.* 2008). This is unsatisfactory, for conceptual and practical reasons:

- The equation above for *r* assumes uncorrelated individuals and GEBV's
- Records $\boldsymbol{y}$ are typically pre-corrected to $\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{b}}$, and the results are sensitive to precorrection. It is unclear what happens if there are contemporary groups in $\boldsymbol{b}$ that are not present in the training data.
- If the whole data set is used for precorrection, then a relationship structure is fit (e.g. pedigree relationships) as $\boldsymbol{y}^* = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'(\boldsymbol{ZAZ}\sigma_u^2 + \boldsymbol{I}\sigma_e^2)^{-1}\boldsymbol{X})^-)\boldsymbol{y}$ where $\boldsymbol{A}\sigma_u^2$ is assumed to be "correct". If the assumed relationship is biased or incorrect, so will be $\widehat{\boldsymbol{b}}$ and $\boldsymbol{y}^*$, and the bias will be toward the assumed relationship. This may explain some puzzling results, e.g. poor performance of genomic prediction in low heritable traits such as fertility (Hayes *et al.* 2009).
- Even after precorrection, there will be a remaining covariance structure across pre-corrected $\boldsymbol{y}^*$. This structure is notoriously hard to model (and rarely modelled). This may explain phenomena such as $\frac{Corr(GEBV,y^*)}{h} > 1$.
- Some precorrected $\boldsymbol{y}^*$ are too clumsy (Ricard *et al.* 2013) to be believed or computed in practice, for instance maternal effects.

CROSS-VALIDATION ACCURACIES FROM METHOD R

Description of the method. We propose to use the properties of method R to construct metrics of cross-validation. Reverter *et al.* (1994) observed that the regression of EBVs obtained with "whole" (w) data on EBVs estimated with "partial" (p) data, $b_{w,p} = \frac{Cov(\hat{u}_w, \hat{u}_p)}{Var(\hat{u}_p)}$ is 1, and this checks bias (in the sense b_1 before). The correlation of partial on whole (eq. 7-9 in their paper) $\rho_{p,w} = \frac{Cov(\hat{u}_p, \hat{u}_w)}{\sqrt{Var(\hat{u}_w)Var(\hat{u}_p)}}$ is a function of respective accuracies. Invoking exchangeability, both equations can be extended to multivariate forms, and expectations can be taken in both the numerator and the denominator, resulting in:

$$b_{w,p} = \hat{\boldsymbol{u}}_w'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p / \hat{\boldsymbol{u}}_p'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p$$

where $\boldsymbol{K}$ is a matrix of relationships, $b_{p,w}$ with an expected value of 1, and

$$\rho_{w,p} = \hat{\boldsymbol{u}}_p'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_w / \sqrt{\hat{\boldsymbol{u}}_p'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p \hat{\boldsymbol{u}}_w'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_w}$$

with an expected value $E(\rho_{w,p}) = \sqrt{\frac{\mu_{acc_p^2}}{\mu_{acc_w^2}}}$ that is, proportional to the relative increase in average reliabilities. As more data cumulates, $\widehat{\boldsymbol{u}}$ tends towards the true breeding values, thus $\widehat{\boldsymbol{u}}_w$ is more accurate than $\widehat{\boldsymbol{u}}_p$. The empirical covariance $\widehat{\boldsymbol{u}}_w'\boldsymbol{K}^{-1}\widehat{\boldsymbol{u}}_p$ measures the strength of the association between the two, whereas $\widehat{\boldsymbol{u}}_p'\boldsymbol{K}^{-1}\widehat{\boldsymbol{u}}_p$ measures the extent of shrinkage due to lack of information. In other words, the theoretical prediction error covariances are replaced by empirical ones (Thompson 2001). By combining cross-validation and theory from mixed models, we hope to retain the best of both worlds: a measure of accuracy that corresponds to reality and that is little affected by the existence of related, unbalanced data. Therefore, an algorithm to estimate accuracy of (say) PBLUP and GBLUP is:

1. Compute EBV's with all data ("whole") using, say, GBLUP (which method should not be critical if all animals have data or progeny)
2. Choose cutoff date
3. Create "partial" data: Set values after cutoff date to missing
4. Compute EBVs based on "partial" and GBLUP
5. Compute statistic $b_{w,p}^{GBLUP} = \frac{\widehat{\mathbf{u}}_p'\mathbf{K}^{-1}\widehat{\mathbf{u}}_w}{\widehat{\mathbf{u}}_p'\mathbf{K}^{-1}\widehat{\mathbf{u}}_p}$
6. Compute statistic $\rho_{p,w}^{GBLUP} = \frac{\widehat{\mathbf{u}}_p'\mathbf{K}^{-1}\widehat{\mathbf{u}}_w}{\sqrt{\widehat{\mathbf{u}}_w'\mathbf{K}^{-1}\widehat{\mathbf{u}}_w\widehat{\mathbf{u}}_p'\mathbf{K}^{-1}\widehat{\mathbf{u}}_p}}$
7. Compute EBVs based on "partial" and PBLUP
8. Compute statistic $b_{w,p}^{PBLUP} = \frac{\widehat{\mathbf{u}}_p'\mathbf{K}^{-1}\widehat{\mathbf{u}}_w}{\widehat{\mathbf{u}}_p'\mathbf{K}^{-1}\widehat{\mathbf{u}}_p}$
9. Compute statistic $\rho_{p,w}^{PBLUP} = \frac{\widehat{\mathbf{u}}_p'\mathbf{K}^{-1}\widehat{\mathbf{u}}_w}{\sqrt{\widehat{\mathbf{u}}_w'\mathbf{K}^{-1}\widehat{\mathbf{u}}_w\widehat{\mathbf{u}}_p'\mathbf{K}^{-1}\widehat{\mathbf{u}}_p}}$

For forward cross-validation, the statistics should be computed for the focal individuals (i.e., candidates to selection). On exit, $b_{w,p}^{GBLUP}$ should be 1 (unbiased method) and is equivalent to b_1 and $\rho_{p,w}^{GBLUP}$ and $\rho_{p,w}^{PBLUP}$ describes the respective accuracies of GBLUP and PBLUP. An extra statistic is bias $\mu_{wp} = b_0 = (\mathbf{1}'\boldsymbol{K}^{-1}\widehat{\boldsymbol{u}}_w - \mathbf{1}'\boldsymbol{K}^{-1}\widehat{\boldsymbol{u}}_p)/n$. Matrix **K** should be the "true" relationship matrix across individuals but there should be no great difference in using either genomic or pedigree relationships as far as they are correct. The procedure has several advantages: is completely general (it can be used e.g. for maternal traits or random regression), it is semi-automatic, and can, at least potentially, provide estimates of the accuracy of the cross-validation metric. There are though many points that need to be addressed: robustness to misspecification, the role of selection (and how to avoid biases in the estimates of the different $b's$), how to sample efficiently, etc.

TEST WITH REAL LIFE DATA SETS

In beef cattle, we used genetic and phenotypic resources from Brahman cows (N = 995) and bulls (N = 1,116) outlined in (Porto-Neto *et al.* 2015). The phenotype was yearling body weight. A procedure "method R" as above was introduced to assess accuracy of GBLUP, and random (1000 replicates) splits of the data set in training and validation was used, as animals are quite unrelated and belong to a single generation. We only present very briefly the results. The statistic $b_{w,p} = 0.96 \pm 0.08$ (in the whole population) showed that evaluation was nearly unbiased, whereas $\rho_{p,w} = 0.67 \pm 0.02$ has a correlation of 0.81 with conventional cross-validation accuracy

estimated as $\frac{Corr(GEBV,\boldsymbol{y}^*)}{h}$.

In dairy sheep, we used a large data set (Manech Tete Rousse) of 1,700,000 milk yield performances, 500,000 animals in pedigree and 2,111 sires with 50K genotypes. Data was split at 2011 in training and validation. For *all* individuals, unbiasedness of (SSG)BLUP was checked with results $\mu_{w,p} = b_0 = 0.2\sigma_g = 5$ (liters), $b_{w,p} = b_1 = 0.996$, so genetic evaluation is virtually unbiased for b_1 (slope) but not for b_0 (genetic trend), which is unsurprising because the model includes Unknown Parent Groups. Later, candidates to selection were compared, with $\rho_{w,p}^{SSGBLUP} = 0.55$ vs. $\rho_{w,p}^{BLUP} = 0.39$, and both evaluations where notoriously biased ($b_1^{SSGBLUP} = 0.77$, $b_1^{BLUP} = 0.70$), possibly due to selection not well accounted for. All these results agree well with previous analysis (Legarra *et al.* 2014).

ACKNOWLEDGEMENTS

Toni Reverter and Andrés Legarra benefit funding from the INRA-CSIRO linkage proposals 2016/2017. AL also financed by grants GENOMIA (Poctefa, Feder) and GenoPyr (Feder).

REFERENCES

Baloche G., Legarra A., Sallé G., Larroque H., Astruc J. M., Robert-Granié C., Barillet F. (2014) *J. Dairy Sci.* **97**: 1107–1116.

Bijma P. (2012) *J. Anim. Breed. Genet.* **129**: 345–358.

Boichard D., Bonaiti B., Barbat A., Mattalia S. (1995) *J. Dairy Sci.* **78**: 431–437.

Bonaiti B., Boichard D., Barbat A., Mattalia S. (1993) *Interbull Bull.* **8**

Christensen O., Madsen P., Nielsen B., Ostersen T., Su G. (2012) *Animal* **6**: 1565–1571.

Cuyabano B. C. D., Su G., Rosa G. J. M., Lund M. S., Gianola D. (2015) *J. Dairy Sci.* **98**: 7351–7363.

Edel, C., Neuner, S., Emmerling, R. and Goetz, K.U., 2012. *Interbull Bull.* **46**

Gianola D., Schön C.-C. (2016) *G3 GenesGenomesGenetics* **6**: 3107–3128.

Hayes B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. (2009) *J Dairy Sci* **92**: 433–443.

Henderson C. R. (1973) *J Anim Sci* (**Symposium**) 10-41

Legarra A., Robert-Granié C., Manfredi E., Elsen J.-M. (2008) *Genetics* **180**: 611–618.

Legarra A., Baloche G., Barillet F., Astruc J., Soulas C., Aguerre X., Arrese F., Mintegi L., Lasarte M., Maeztu F. (2014*) J. Dairy Sci.* **97**: 3200–3212.

Mantysaari E., Liu Z., VanRaden P. (2010) *Interbull Bull* **41**.

Mäntysaari E. A., Koivula M. (2012) *Interbull Bull.* **46**

Mikshowsky A. A., Gianola D., Weigel K. A. (2016) *J. Dairy Sci.* **99**: 3632–3645.

Olson K., VanRaden P., Tooker M., Cooper T. (2011) *J. Dairy Sci.* **94**: 2613–2620.

Porto-Neto L. R., Barendse W., Henshall J. M., McWilliam S. M., Lehnert S. A., Reverter A. (2015) *Genet. Sel. Evol.* **47**: 84.

Powell R. L., Wiggans G. R. (1994) *Interbull Bull.* **10**.

Reverter A., Golden B. L., Bourdon R. M., Brinks J. S. (1994*) J. Anim. Sci.* **72**: 34–37.

Ricard A., Danvy S., Legarra A. (2013) *J. Anim. Sci.* **91**: 1076–1085.

Robertson A. (1977) Z. Für Tierz. Zücht. **94**: 131–135.

Saatchi M., McClure M. C., McKay S. D., Rolf M. M., Kim J., et al. (2011) *Genet. Sel. Evol.* **43**: 40.

Thompson R. (2001) *Livest. Prod. Sci.* **72**: 129–134.

VanRaden P. M., Tassell C. P. V., Wiggans G. R., Sonstegard T. S., Schnabel R. D., Taylor J. F., Schenkel F. S. (2009) *J Dairy Sci* **92**: 16–24.

Vitezica Z., Aguilar I., Misztal I., Legarra A. (2011) *Genet. Res.* **93**: 357–366.

Xiang T., Nielsen B., Su G., Legarra A., Christensen O. F. (2016) *J. Anim. Sci.* **94**: 936–948.