A novel method for the estimation of the relative importance of breeds in order to conserve the total genetic variance

Jörn BENNEWITZ[a*], Theo H.E. MEUWISSEN[b]

[a] Institute of Animal Breeding and Husbandry, Christian-Albrechts-Universität, 24098 Kiel, Germany
[b] Institute of Animal and Aquacultural Sciences, Agriculture University of Norway, Box 5052, 1432 Aas, Norway

Abstract – The need for conservation of farm animal genetic resources is widely accepted. A key question is the choice of breeds to be conserved. For this purpose, a core set of breeds was introduced in that the total genetic variance of a hypothetical quantitative trait was maximised (MVT core set). For each breed the relative contribution to the core set was estimated and the breeds were ranked for conservation priority according to their relative contribution. The method was based on average kinships between and within breeds and these can be estimated using genetic marker data. The method was compared to a recently published core set method that maximises the variance of a hypothetical population that could be obtained by interbreeding the conserved breeds (MVO core set). The results show that the MVT (MVO) core set favours breeds with a high (low) within breed kinship that are not related to other breeds. Following this, the MVT core set method suggests conserving breeds that show a large difference in the respective population mean of a hypothetical quantitative trait. This maximises the speed of achieving selection response for this hypothetical selection direction. Additionally, bootstrap based methods for the estimation of the breed's contribution to the core sets were introduced, substantially improving the accuracy of the contribution estimates.

genetic variance / conservation / kinship / livestock breed / bootstrap

1. INTRODUCTION

On a world-wide level, there are roughly 6000 breeds of 30 domestic mammalian and bird species. Around 35% of them are classified as having a high risk of extinction and every week two breeds permanently vanish [9] resulting in an irreversible loss of animal genetic resources. The need to conserve

---

* Corresponding author: jbennewitz@tierzucht.uni-kiel.de

these resources is widely accepted mainly because it can be seen as an insurance against future challenges and conditions but also for ecological and sociocultural reasons. Because funds to preserve animal genetic resources are limited, an optimal allocation of these funds is of central importance. Within this framework a key question is the choice of breeds for conservation programmes. Ruane [14] reported several criteria for this selection, such as specific adaptive features, particular traits of special interest and genetic uniqueness. This last feature is aimed at maintaining the genetic variance, an aspect upon which we will focus exclusively throughout this paper.

Pairwise genetic distances and their graphical representation in distance trees are common tools used to assess the genetic uniqueness of a particular breed within a set of breeds. Genetic distances are usually estimated from the genotypic information of a set of neutral loci. Weitzman [16, 17] developed an algorithm for the estimation of the genetic diversity within a set of elements based on pairwise distances under the assumption that all elements are distinct and are obtained from a single founder population by fission. The application of this algorithm to a set of breeds makes it possible to rank the breeds for their priority for conservation according to their contribution to the Weitzman diversity. This diversity measure has a number of nice mathematical and biological properties [15, 16] and was used recently in several breed genetic diversity studies [2, 11, 13]. However, as pointed out by Eding and Meuwissen [3] and Caballero and Toro [1], the use of the Weitzman diversity measure on a within-species breed level might be inappropriate because it ignores migration between breeds, which is unrealistic, and also ignores within breed diversity.

Instead of genetic distances, Eding and Meuwissen [3] and Caballero and Toro [1] used average kinships between and within breeds for the description of genetic diversity. Kinship is defined as the probability that two gametes randomly drawn from a population are identical by descent. Following this, the average kinship between two breeds is an estimate of the fraction of alleles that these breeds have in common. In order to prioritise breeds for conservation, Eding *et al.* [4] defined a core set that is built by relative contributions of the breeds under consideration in order to minimise the mean kinship in this core set. The core set maximises the variance of a hypothetical quantitative trait that can be found in a hypothetical population obtained from interbreeding the conserved breeds [4]. The breeds are ranked for their priority for conservation according to their relative contribution to the core set. Eding and Meuwissen [5] described a method that estimates average kinships between breeds using similarities of genetic marker alleles. By way of simulations, the authors demonstrated that the accuracy of the estimation of the breed's

contribution to the core set from estimated kinships is only moderate, indicating that there is scope for improvement. However, the core set method of Eding *et al.* [4] does not explicitly consider the variance that can be found between breeds. An analogue of the core set method for prioritising breeds for conservation has been presented by Caballero and Toro [1]. Using a similar approach, Piyasatian and Kinghorn [12] defined genetic diversity as the amount of allelic variation that can be found within and between subdivided breeds. A breed is preferred for conservation if it contributes significantly to the total allelic variation. The conceptual differences between the core set diversity method and the approach of Piyasatian and Kinghorn [12] on the one hand and the Weitzman diversity on the other hand, is that the former methods consider both the between and within breed variation and they account for possible migration between breeds. Fabuel *et al.* [7] showed that the two approaches can produce different results.

The aim of this paper was to put forward a conservation criterion that values the differences between breeds more than the core set method of Eding *et al.* [4] does. For this purpose an algorithm was introduced that estimates the relative contribution of breeds to a core set in order to maximise the total additive genetic variance of a hypothetical quantitative trait. The relative contributions were used to rank the breeds for conservation priority. The method was compared to the core set method of Eding *et al.* [4] using simulated and real data. Additionally, bootstrap based methods were introduced, substantially improving the accuracy of contribution estimates.

2. MATERIALS AND METHODS

2.1. Core set methods

Assume a population with n animals and each animal i has a breeding value u_i for a hypothetical quantitative trait. The additive genetic variance within the population is:

$$\text{var}(u_w) = \frac{1}{n} \sum_{i=1}^{n} \left[\text{var}(u_i - \overline{u}) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\text{var}(u_i) - \text{var}(\overline{u}) \right],$$

where $\text{var}(u_i)$ is the variance of the breeding value u_i of animal i and $\text{var}(\overline{u})$ is the variance of the population mean of the breeding values. The averaging

is because individual animals may have different $\text{var}(u_i)$. Ignoring the total additive genetic variance of the trait, σ_u^2, since it multiplies the results by a constant, yields:

$$\text{var}(u_w) = \frac{1}{n} \sum_{i=1}^{n} \left[A_i - \overline{A} \right], \tag{1}$$

where $\overline{A}$ denotes the average of the elements of the numerator relationship matrix, $\mathbf{A}$, with dimension $n \times n$. A_i is the diagonal element i of $\mathbf{A}$, this element corresponds to one plus the inbreeding coefficient of animal i. The elements of the numerator relationship matrix are twice the elements of the kinship matrix [8]. In the following, this outline is transferred from a single breed level to a multiple breed level.

Assume a set S of N breeds with a known average kinship matrix, $\mathbf{M}$, of dimension $N \times N$ as described by Eding and Meuwissen [3,5]. The off-diagonal elements of $\mathbf{M}$ are the average kinships between breeds and correspond to the inbreeding coefficient of putative offspring from the corresponding between breed mating. The diagonal elements of $\mathbf{M}$ are the average within breed kinships and correspond to the inbreeding coefficients of putative offspring from within breed mating. Following (1), the additive genetic variance of a hypothetical quantitative trait within set S can then be described by

$$\text{var}(u_S) = \frac{1}{N} \sum_{i=1}^{N} \left[(1 + M_i) - 2\overline{M} \right], \tag{2}$$

where M_i is the within breed kinships of breed i obtained from the diagonal elements of $\mathbf{M}$ and $\overline{M}$ denotes the mean of all the elements of $\mathbf{M}$. A one is added to M_i, because the A_i elements in (1) correspond to one plus the inbreeding coefficient. In order to maximise the total genetic variance, a core set S_{mvt} is formed by relative contributions of the N breeds of S. This core set is termed maximum variance total (MVT) core set method in the following. The contributions of the breeds to the MVT core set are stored in a vector, $\mathbf{c_{mvt}}$, of dimension N. The total genetic variance conserved by the MVT core set, $\text{Var}_{\text{total}}(S_{\text{mvt}})$, is then calculated by:

$$\text{Var}_{\text{total}}(S_{\text{mvt}}) = 1 + \mathbf{c'_{mvt}}\mathbf{F} - 2\mathbf{c'_{mvt}}\mathbf{Mc_{mvt}}, \tag{3}$$

with $\mathbf{F}$ being a vector of dimension N that contains the within breed kinship, i.e. $\mathbf{F} = \text{diag}(\mathbf{M})$. In order to find the solutions for $\mathbf{c_{mvt}}$ under the restriction $c_{mvt}(i) \geq 0$ and $\sum_{i=1}^{N} c_{mvt}(i) = 1$, a Lagrangian equation is set up (the one is omitted since it does not affect the optimum solution):

$$L\left(\text{Var}_{\text{total}}(S_{\text{mvt}})\right) = \mathbf{c'_{mvt}}\mathbf{F} - 2\mathbf{c'_{mvt}}\mathbf{Mc_{mvt}} - \lambda(\mathbf{c'_{mvt}}\mathbf{1_N} - 1), \tag{4}$$

with $\mathbf{1_N}$ a vector of dimension N containing ones and λ as the Lagrangian multiplier. In order to maximise (4), the first derivative with respect to $\mathbf{c_{mvt}}$ is set to zero and solved for $\mathbf{c_{mvt}}$:

$$\frac{\partial L}{\partial \mathbf{c_{mvt}}} = \mathbf{F} - 4\mathbf{Mc_{mvt}} - \lambda\mathbf{1_N} = 0,$$

$$\mathbf{Mc_{mvt}} = \frac{1}{4}\left(\mathbf{F} - \lambda\mathbf{1_N}\right),$$

$$\mathbf{c_{mvt}} = \frac{1}{4}\left(\mathbf{M^{-1}F} - \lambda\mathbf{M^{-1}1_N}\right). \tag{5}$$

Because $\mathbf{c}'_{mvt}\mathbf{1_N} = 1$ it can be written:

$$\mathbf{c}'_{mvt}\mathbf{1_N} = \frac{1}{4}\left(\mathbf{1}'_N\mathbf{M^{-1}F} - \lambda\mathbf{1}'_N\mathbf{M^{-1}1_N}\right) = 1.$$

Solving for λ yields:

$$\lambda = \frac{\mathbf{1}'_N\mathbf{M^{-1}F} - 4}{\mathbf{1}'_N\mathbf{M^{-1}1_N}}. \tag{6}$$

The solution for $\mathbf{c_{mvt}}$ is obtained by substituting (6) into (5):

$$\mathbf{c_{mvt}} = \frac{1}{4}\left(\mathbf{M^{-1}F} - \frac{\mathbf{1}'_N\mathbf{M^{-1}F} - 4}{\mathbf{1}'_N\mathbf{M^{-1}1_N}} \cdot \mathbf{M^{-1}1_N}\right). \tag{7}$$

This core set method is compared to the core set method of Eding *et al.* [4], which maximises the variance in a hypothetical population that could be bred from the conserved breeds, *i.e.* the offspring variance. The Eding *et al.* [4] core set method will be termed the maximum variance offspring core set method (MVO). The MVO method minimises the average kinship within the core set, $\overline{f}_{\min}(S_{mvo})$. In brief, the relative breed contribution vector, $\mathbf{c_{mvo}}$, is calculated under the restriction $c_{mvo}(i) \geq 0$ and $\sum_{i=1}^{N} c_{mvo}(i) = 1$ as follows:

$$\overline{f}_{\min}(S_{mvo}) = \mathbf{c}'_{mvo}\mathbf{Mc_{mvo}}. \tag{8}$$

The solution for $\mathbf{c_{mvo}}$ is:

$$\mathbf{c_{mvo}} = \frac{\mathbf{M^{-1}1_N}}{\mathbf{1}'_N\mathbf{M^{-1}1_N}}. \tag{9}$$

The offspring genetic variance maximised by this core set method is calculated by:

$$\text{Var}_{\text{offspring}}(S_{\text{mvo}}) = 1 - \overline{f}_{\text{min}}(S_{\text{mvo}})$$
$$= 1 - \mathbf{c}'_{\mathbf{mvo}}\mathbf{M}\mathbf{c}_{\mathbf{mvo}}. \tag{10}$$

See Eding *et al.* [4] for the full derivation of this core set method.

Under the assumption that all breeds are descendants from a single and non-inbred base breed, both genetic variances, the total variance and the offspring variance, can be seen as the variance relative to the variance of the base breed. Both core sets use the same average marker kinship matrix $\mathbf{M}$, whose estimation is described below.

2.2. Estimation of the average kinship matrix M from genetic marker data

Ideally, the average kinship matrix $\mathbf{M}$ is calculated using pedigree information, but often no appropriate pedigree information is available. In these cases $\mathbf{M}$ can be estimated using molecular marker information [1, 3, 5]. Here a method of Eding and Meuwissen [5] was used and is described in the following. A similarity index between all pairs of individuals genotyped for a marker k ($S_{xy,k}$) was calculated as $S_{xy,k} = \frac{1}{4}[I_{11} + I_{12} + I_{21} + I_{22}]$, where I_{ij} is an indicator variable which is 1 when allele i in the individual x and allele j in the individual y are identical, and otherwise it is 0. The average similarities between breed i and j for locus k ($\widetilde{S}_{ij,k}$) were estimated by:

$$\hat{\widetilde{S}}_{ij,k} = \frac{1}{n_{ij}} \sum_{l=1}^{n_{ij}} S_{ij,k}(l), \tag{11}$$

where n_{ij} is the number of combinations of individuals between breed i and j. The expectation of $\widetilde{S}_{ij,k}$ is $E(\widetilde{S}_{ij,k}) = f_{ij} + (1 - f_{ij})s_k$, where f_{ij} is the average kinship between breed i and j and s_k is the probability of the alleles being Alike In State (AIS). By subtracting both sides from one and taking the natural logarithm of this, it is possible to set up a log-linear model for the estimation of kinships from marker data [5]:

$$\ln(1 - \widetilde{S}_{ij,k}) = \ln(1 - f_{ij}) + \ln(1 - s_k) + e_{ij,k}$$
$$\Leftrightarrow$$
$$y_{ij,k} = a_{ij} + b_k + e_{ij,k}, \tag{12}$$

where $e_{ij,k}$ is the error term. In matrix notation the model becomes

$$\mathbf{y} = \mathbf{X_a}\mathbf{a} + \mathbf{X_b}\mathbf{b} + \mathbf{e}, \tag{13}$$

where $\mathbf{y}$ is a vector with the observations and $\mathbf{X_a}$ and $\mathbf{X_b}$ are incidence matrices relating the observations to the effects a ($a_{ij} = \ln(1-f_{ij})$) and b ($b_k = \ln(1-s_k)$) stored in the vectors $\mathbf{a}$ and $\mathbf{b}$, respectively, and $\mathbf{e}$ is the vector of errors. The observations show the following expected variance [5]:

$$\mathrm{Var}\left[\ln(1 - \widetilde{S}_{ij,k})\right] \approx \frac{\hat{f}_{ij} + (1 - \hat{f}_{ij})\hat{s}_k}{4n_{ij}(1 - \hat{f}_{ij})(1 - \hat{s}_k)}. \tag{14}$$

Taking these variances into account a weighted log-linear model was formulated using the following equations:

$$\begin{bmatrix} \mathbf{X'_a W^{-1} X_a} & \mathbf{X'_a W^{-1} X_b} \\ \mathbf{X'_b W^{-1} X_a} & \mathbf{X'_b W^{-1} X_b} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X'_a W^{-1} y} \\ \mathbf{X'_b W^{-1} y} \end{bmatrix}, \tag{15}$$

where $\mathbf{W}$ is a diagonal matrix that contains the corresponding expected variances of the observations. Because the weights are obtained from the estimates, the final solutions were estimated iteratively (100 iterations). The design matrices showed dependencies and a generalised inverse of the coefficient matrix was used, with the consequence that the solutions were not unique. Therefore, the solutions to $\mathbf{a}$ were restricted by setting the highest a_{ij} (*i.e.* the smallest kinship) to zero. This was aimed at defining the breed that existed just before the first fission event as the base breed. After solving this system, the kinships were calculated by back transformation, *i.e.*:

$$\mathbf{f} = 1 - e^{\mathbf{a}}, \tag{16}$$

where $\mathbf{f}$ is a vector (dimension $0.5N(N + 1)$) containing the kinships. These kinships were finally transferred into the average kinship matrix $\hat{\mathbf{M}}$. See Eding and Meuwissen [5] for a more detailed description.

2.3. Contribution vector estimation from $\hat{\mathbf{M}}$

The breed contribution vectors to the MVT and MVO core sets (*i.e.* $\hat{\mathbf{c}}_{\mathbf{mvt}}$ and $\hat{\mathbf{c}}_{\mathbf{mvo}}$), respectively, were calculated from $\hat{\mathbf{M}}$ and $\hat{\mathbf{F}}$ ($\hat{\mathbf{F}} = \mathrm{diag}(\hat{\mathbf{M}})$) as outlined in (7) and (9). If breeds showed negative contributions, the most negative contribution was set to zero and the contribution vector was recalculated without the corresponding breed. This was repeated until no further negative contribution estimates were observed. We will refer to this method of contribution vector estimation as the Eding *et al.* method (EEA), because it is analogous to the idea of Eding *et al.* [4].

2.4. Contribution vector estimation from $\hat{\mathbf{M}}$ by a bootstrap approach

As stated in the introduction, the accuracy of $\hat{\mathbf{c}}_{\mathbf{mvo}}$ is only moderate when estimated by EEA [5]. In order to obtain a higher accuracy of the contribution vectors for both core set methods ($\hat{\mathbf{c}}_{\mathbf{mvt}}$ and $\hat{\mathbf{c}}_{\mathbf{mvo}}$, respectively), a non-parametric bootstrap method was tested as described in the following. A bootstrap sample was generated by performing two sampling steps. In one step, N individuals were sampled with replacement out of the pool of N original individuals within breed. In the second step, L marker loci were sampled with replacement out of the pool of L original marker loci across breeds. Hence, a bootstrap sample consisted of sampled individuals and sampled marker loci. Throughout the study 100 bootstrap samples were generated ($B = 100$). For each bootstrap sample, average similarities were calculated using (11) and the log-linear model was set up using (12) and (13) to obtain the kinship of the bootstrap sample. However, it was not possible to run the weighted log-linear model with the expected variances obtained from (14), because for each bootstrap sample around 100 iterations would have to be performed – an undertaking that would be computationally too demanding. The similarities from the bootstrap samples were therefore analysed using two different methods.

The first method used model (13) and the equations of (15) but calculated the solutions without weights (*i.e.* omitting matrix $\mathbf{W}$ from the equations in (15)). The solutions were restricted such that the highest a was zero (a restriction is needed in the fixed effects equations (15)) and the kinships were estimated from the solution vector using (16) and stored in the kinship matrix. The breed contribution vectors of the bootstrap sample to both core set methods were obtained using (7) and (9). As for the EEA method, negative contributions were subsequently set to zero, starting with the most negative one, and the contribution vectors were recalculated without the corresponding breed. The final solutions of the vectors $\hat{\mathbf{c}}_{\mathbf{mvt}}$ and $\hat{\mathbf{c}}_{\mathbf{mvo}}$ were obtained from the mean of the solutions from the B bootstrap samples. This method of contribution vector estimation is termed the bootstrap method (BM).

The second method again used model (13) and the equations of (15) but used the reciprocal value of the empirical variance of the observations. This variance was estimated one time by bootstrapping as shown in the appendix, and then was used for the analysis of all bootstrap samples. The final solutions of the vectors $\hat{\mathbf{c}}_{\mathbf{mvt}}$ and $\hat{\mathbf{c}}_{\mathbf{mvo}}$ were obtained as described for the BM method. This method of contribution vector estimation is termed the weighted bootstrap method (WBM).

2.5. Application to simulated data

To test and to compare the methods outlined above, two series of Monte Carlo simulations were carried out. The number of replicates was always 10 (test simulation showed that this number is sufficient to obtain reliable results). In the first series, a phylogenetic situation was simulated over 50 generations. A base breed was simulated that consisted of 50 individuals, the size of which was kept constant throughout the simulation. For each individual, a number of L unlinked genetic marker loci was assumed ($L = 10$ and 20, respectively) and the alleles were randomly assigned to the individuals. The number of alleles per locus in the base breed was randomly chosen from the interval 25/50. From a number of test simulations it was found that this interval would produce realistic numbers of alleles still segregating after 50 generations, although in exceptional cases fixation occurred. Each next generation was generated by randomly assigning sires and dams from the current generation as parents of the individuals of the next generation. New breeds were randomly generated by base breed fission between generations 10 and 49 by sampling sires and dams randomly as parents of the individuals of the new breed. The effective size of the new breed was randomly chosen from the interval 24/76 and was kept constant throughout the simulation. For each breed the male/female ratio was one and an even number of breed sizes was therefore chosen. In total, 15 breeds were simulated for each replicate, one base breed and 14 breeds that were formed by fission from the base breed.

During the second series of simulations the phylogenetic structure was fixed. A base breed of size 50 was bred for 50 generations as described above ($L = 20$). From this, four new breeds were generated at generation 10 by fission as outlined above. In order to obtain different within breed kinships after 50 generations, the effective size of the breed was chosen to be 10, 20, 30 and 40, respectively. The male/female ratio was one.

In generation 50 the genotypic data of the breeds were used for the estimation of the breed contribution vectors for both core sets by use of the EEA, BM and WBM methods as outlined above. In addition, the full pedigree information was recorded during the simulation and this was used to estimate the true average kinships between the breeds using path analysis [8]. The calculated true kinships were corrected in order to define the breed that existed just before the first fission as the base breed [5]:

$$f'_{ij} = \frac{f_{ij} - f_{\min}}{1 - f_{\min}}. \tag{17}$$

This transformation was done in order to make these kinships comparable to the marker estimated kinships obtained from (15) and (16). From the true average kinship matrix, the true breed contributions were obtained using (7) and (9). The correlation between the estimated and the true contribution vectors and the mean square error of the estimated contributions was calculated, serving as an empirical measure of the ability of EEA, BM and WBM to estimate correct breed core set contributions.

The total variance and the offspring variance within the two core sets were estimated using (3) and (10) and using the correct kinships. Additionally, for the second series of simulations the individual contributions of the five breeds to the core set, averaged over all replicates, were recorded. This revealed the differences between the MVT and the MVO core set methods regarding the breed contributions as a function of their within breed kinships and regarding the variance conserved, respectively.

2.6. Application to real data

To compare the two core set methods using a real case, the field data of Eding and Meuwissen [5] were reanalysed. The data set consisted of 10 Dutch breeds that were genotyped for 11 microsatellite markers and are described in detail by Eding and Meuwissen [5]. Unfortunately only allele frequencies were available. Therefore, the similarity between population i and j at locus k with n alleles was calculated as described by Eding and Meuwissen [3]:

$$\hat{\tilde{S}}_{ij,k} = \sum_{l=1}^{n} p_{ik,l} p_{jk,l},$$

where $p_{ik,l}$ is the allele frequency of allele l at locus k in population i. The breed contribution vectors for both core set methods were calculated by EEA and BM. The bootstrap samples could only be generated by sampling the loci (instead of sampling the loci and the animals), because no information regarding the animals was available. For this reason it was also not possible to apply the WBM. Note that the estimation of the empirical variance of the observations required the bootstrap sampling of the individuals (Appendix). To show the relationship between the breeds, genetic distance between breed i and j was computed as $\hat{d}_{ij} = \hat{f}_{ii} + \hat{f}_{jj} - 2\hat{f}_{ij}$ [3]. These distances can be interpreted as twice the Nei minimum distance corrected for the allele frequency in the base breed. They were represented graphically in a dendrogram using the Neighbour-Joining algorithm of the Phylip software [10]. For comparison purposes, the Weitzman diversity was estimated using the genetic distances

obtained and the software of Cañón and García as applied in [2]. The breeds were ranked for conservation priority using the estimated contributions to the total Weitzman diversity measure.

3. RESULTS

3.1. Results from the simulations

As expected, when comparing the variance conserved by the MVT and by the MVO core set methods, respectively, the MVT method conserved more total variance and the MVO method more offspring variance. However, the differences between the two core set methods are only minimal (Tabs. I and II). This is regardless of the phylogenetic situation simulated.

The accuracy of the WBM contribution vector is the highest in both core set methods. These contribution estimates showed the highest average correlations with the true contributions and on average the lowest mean square error (MSE of the WBM estimates roughly 0.01 lower compared to the BM and 0.02 lower compared to the EEA estimates, Tabs. I and II). Most total variance in the MVT core set and most offspring variance in the MVO core set were conserved, respectively, when the contribution vectors were estimated by WBM. Generally, as expected, the accuracy of the contribution vector estimates was a function of the number of loci included in the analysis. The WBM produced estimates with a significantly higher accuracy in phylogenetic situations with 10 loci than the EEA with 20 loci (Tab. I). The accuracies of the contribution vector estimates are higher in the MVT core set compared to the corresponding estimates of the MVO core set (Tabs. I and II).

All breeds contributed to the MVO core set (results from the true contributions, Tab. I), but when the contributions were estimated with EEA, many breeds erroneously showed a zero contribution estimate. This was also reported by Eding *et al.* [4]. This problem was less marked when the BM method was used and did not occur when the WBM method was applied. In contrast, a substantial fraction of breeds did not contribute to the MVT core set (results from the true contributions, Tab. I), but nearly all breeds showed a contribution estimate greater than zero, when using BM and WBM. However, these estimates were in general only very small.

In Table II, the within breed kinships of the five simulated breeds (results from the second series of simulation, fixed phylogenetic situation and fixed and unequal population size) are given as an average over all replicates. Note that all between breed kinships were very similar and low (in general < 0.01),

Table I. Correlation (Corr) between the estimated and the true core set contributions, mean square error (MSE) of contribution estimates and actual conserved total genetic variance (Var_{total}) and offspring genetic variance ($Var_{offspring}$) as functions of the core set, of the number of simulated loci (L) and of the method of contribution estimation.

Core set	L	Method	MSE	Corr	n_{zero}[1]	Var_{total}[2]	$Var_{offspring}$[2]
MVT	10	TRUE	0	1	5.0	1.197	0.926
		EEA	0.057	0.618	7.4	1.176	0.917
		BM	0.047	0.676	0.7	1.179	0.919
		WBM	0.035	0.803	0.1	1.185	0.924
	20	TRUE	0	1	3.9	1.188	0.925
		EEA	0.049	0.641	7.2	1.172	0.918
		BM	0.038	0.683	0.7	1.177	0.921
		WBM	0.032	0.767	0.2	1.179	0.922
MVO	10	TRUE	0	1	0	1.185	0.931
		EEA	0.066	0.589	6.4	1.151	0.916
		BM	0.052	0.651	0.6	1.160	0.919
		WBM	0.038	0.754	0.1	1.169	0.923
	20	TRUE	0	1	0	1.178	0.930
		EEA	0.057	0.653	5.6	1.152	0.918
		BM	0.045	0.712	0.1	1.160	0.921
		WBM	0.036	0.835	0	1.169	0.924

Results from the first series of simulation, averaged over all replicates. [1]n_{zero} number of breeds that showed zero contribution to the core set; [2]Var_{total} and $Var_{offspring}$, actual conserved total genetic variance and offspring genetic variance estimated by the use of the contribution vectors and the true kinship matrix.

because of the correction in (17). The kinships between breeds become constant after fission [3], and in this series of simulations all breeds were formed by fission at the same time. The breeds with low kinships contributed more to the MVO core set. This was expected, because this core set method is aimed at keeping the average kinship of the set as small as possible [4]. In contrast, the true contributions of the breeds with a higher kinship are slightly larger in the MVT core set. This could not be observed from the estimated contributions, probably due to sampling errors.

In the two initial series of simulations, we sampled both the animals and the loci during the bootstrap approach. In order to investigate which sampling is responsible for the improved performance of the bootstrap approach, an additional simulation was carried out. The method is the same as that used in the

Table II. Mean square error (MSE) of contribution estimates, actual conserved total genetic variance (Var_{total}) and offspring genetic variance ($Var_{offspring}$) as functions of the core set and of the method for contribution estimation, and contribution estimates as a function of the within breed kinship (f_i).

Core set	Method	MSE	Var_{total}[1]	$Var_{offspring}$[1]	Core set contributions of breeds of size n				
					Breed 1	Breed 2	Breed 3	Breed 4	Breed 5
					$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
					$f_i = 0.86$	$f_i = 0.64$	$f_i = 0.48$	$f_i = 0.40$	$f_i = 0.33$
MVT	TRUE	0	1.323	0.881	0.221	0.211	0.201	0.190	0.176
	EEA	0.089	1.277	0.873	0.175	0.200	0.173	0.207	0.245
	BM	0.052	1.310	0.883	0.187	0.219	0.185	0.194	0.216
	WBM	0.047	1.312	0.888	0.178	0.205	0.193	0.207	0.217
MVO	TRUE	0	1.288	0.899	0.112	0.152	0.197	0.243	0.296
	EEA	0.090	1.206	0.878	0.070	0.088	0.187	0.277	0.378
	BM	0.084	1.233	0.884	0.094	0.125	0.151	0.251	0.379
	WBM	0.066	1.247	0.889	0.088	0.126	0.182	0.268	0.326

Results from the second series of simulation, averaged over all replicates. [1]Var_{total} and $Var_{offspring}$, actual conserved total genetic variance and offspring genetic variance estimated by the use of the contribution vectors and the true kinship matrix.

Table III. Correlation (Corr) between the estimated and the true core set contributions and mean square error (MSE) of contribution estimates as a function of the number of simulated loci (L), of the bootstrap method of contribution estimation and of the subject of sampling during bootstrapping (either only animals or only loci).

Core set	L	Method	Sampled	MSE	Corr	n_{zero}[1]
MVT	10	BM	animals	0.058	0.594	4.9
			loci	0.049	0.672	1.2
		WBM	animals	0.061	0.654	4.9
			loci	0.036	0.783	0.2
	20	BM	animals	0.047	0.571	4.0
			loci	0.039	0.691	1.0
		WBM	animals	0.065	0.544	5.2
			loci	0.035	0.769	0.1
MVO	10	BM	animals	0.066	0.592	3.0
			loci	0.054	0.643	0.6
		WBM	animals	0.066	0.675	5.0
			loci	0.040	0.749	0.2
	20	BM	animals	0.056	0.598	2.7
			loci	0.046	0.678	0.7
		WBM	animals	0.069	0.638	4.7
			loci	0.037	0.801	0.0

Results averaged over all replicates, the number of simulated breeds was 15. [1]n_{zero}, number of breeds that showed zero contribution to the core set.

first two series of simulations with the exception that either animals only or loci only were bootstrapped. The results are presented in Table III and show that the improvement stems mainly from the sampling of the loci. It seems that sampling of the animals only did not generate enough bootstrap variance to obtain significantly improved contribution estimates.

3.2. Results from field data

From the estimated kinship matrix from the field data (Tab. IV) it can be seen that the Heck breed shows the highest level of within breed kinship followed by Galloway, Dutch Black Belted and Improved Red Pied. This is also visualised in the dendrogram of the genetic distances obtained from the kinships (Fig. 1), where these breeds show the longest branch length. The highest mean kinships were also observed for these three breeds. The Holstein Friesian
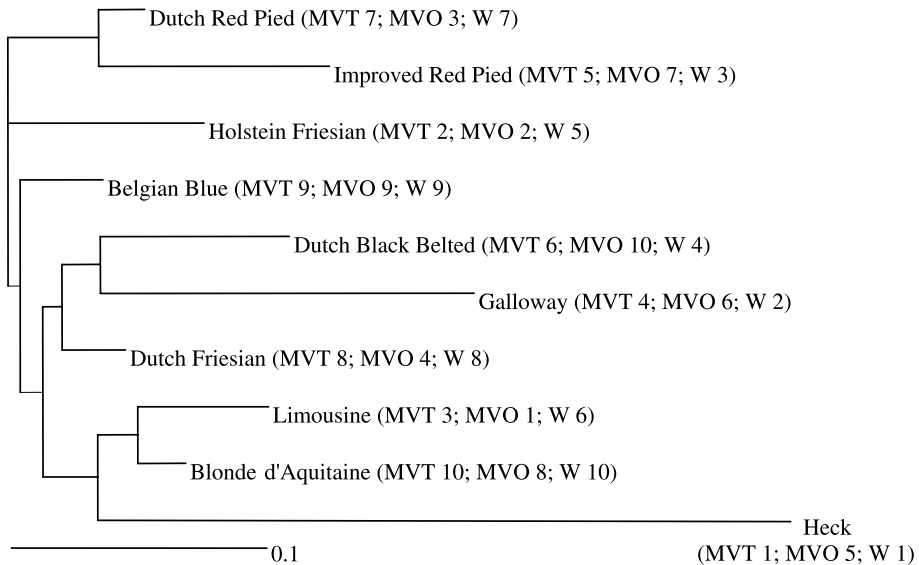
Figure 1. Neighbour-Joining dendrogram of the 10 cattle breeds obtained from genetic distances derived from the between breed kinships. In parentheses the ranking number for prioritisation for conservation obtained by the total variance core set diversity (MVT), by the offspring variance core set diversity (MVO) and by the distance based Weitzman diversity (W). The results of the BM bootstrap approach (Tab. V) were used for the two core set rankings.

and Limousine breeds contributed considerably to the MVT core set, despite their low within breed kinships and thus their short branch length. This was due to their comparatively low mean kinship, and consequently these two breeds were also the major contributors to the MVO core set (Tabs. IV and V). The third breed of the top three contributors was different in the two core sets. The Heck breed showed the largest contribution to the MVT core set but only a small contribution to the MVO core set. The high MVT contribution of the Heck breed is due to its high within breed kinship and its comparatively low mean kinship. The opposite was true for the Dutch Red Pied (low contribution to MVT and high contribution to MVO core set). As a result of the estimation of the Weitzman diversity, the conservation priorities of the breeds according to their contribution to the Weitzman diversity are given in Figure 1. As expected, the priority ranking was in agreement with the branch lengths of the corresponding breed.

Four breeds showed zero contributions to the MVT core set using EEA and only one using BM (Tab. V). However, the EEA zero contribution breeds received only a very small (between 0.002 and 0.03) value using BM. A similar

Table IV. Estimated kinships within and between the 10 cattle breeds using a weighted log-linear model.

Breed	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Belgian Blue	**0.104**									
(2) Dutch Red Pied	0.055	**0.088**								
(3) Dutch Black Belted	0.076	0.050	**0.179**							
(4) Limousine	0.027	0.026	0.036	**0.089**						
(5) Holstein Friesian	0.049	0.027	0.048	0.001	**0.102**					
(6) Galloway	0.038	0.030	0.088	0.045	0.041	**0.213**				
(7) Dutch Friesian	0.062	0.040	0.086	0.027	0.044	0.061	**0.099**			
(8) Improved Red Pied	0.049	0.076	0.063	0.033	0.041	0.058	0.052	**0.172**		
(9) Blonde d'Aquitaine	0.043	0.038	0.038	0.052	0.025	0.036	0.042	0.046	**0.085**	
(10) Heck	0.059	0.032	0.055	0.022	0.000	0.029	0.056	0.006	0.059	**0.310**
Mean (within and between)	**0.056**	**0.046**	**0.072**	**0.036**	**0.038**	**0.064**	**0.057**	**0.060**	**0.046**	**0.063**

Table V. MVT and MVO core set contribution vectors for the 10 cattle breeds estimated either by EEA or by BM.

	MVT core set		MVO core set	
	EEA	BM	EEA	BM
Belgian Blue	0	0.002	0	0.032
Dutch Red Pied	0	0.028	0.207	0.115
Dutch Black Belted	0.011	0.049	0	0.021
Limousine	0.186	0.169	0.359	0.304
Holstein Friesian	0.242	0.209	0.321	0.229
Galloway	0.135	0.139	0	0.051
Dutch Friesian	0	0.004	0.037	0.086
Improved Red Pied	0.169	0.133	0	0.049
Blonde d'Aquitaine	0	0	0	0.039
Heck	0.257	0.267	0.075	0.072

pattern can be observed in the results of the MVO core set (Tab. V). EEA estimated five, but BM no zero contributions. However, in this core set the EEA zero contribution breeds received considerable contributions when using BM (between 0.02 and 0.05). When the breeds were ranked according to their contributions estimated either by EEA or by BM, there was a general agreement in the order within the core set. In both core sets only the fourth and fifth breeds were changed in their order due to small differences in the contribution estimates estimated by EEA and BM, respectively.

4. DISCUSSION

This study introduces the MVT core set method for prioritising breeds for conservation in order to optimally maintain total additive genetic variance of a hypothetical quantitative trait. The method uses kinships between and within breeds and for their estimation some of the ideas of Eding and Meuwissen [3] and Eding *et al.* [5] were implemented. The following discussion focuses on the differences and similarities of the MVT core set and the MVO core set proposed by Eding *et al.* [4] and Caballero and Toro [1] and on the applied methods for the breed core set contribution estimation.

4.1. MVT *versus* MVO core set

The differences between the two core set methods are demonstrated by the following simple hypothetical example. Consider two unrelated breeds A

and B, and a breed AB are available for conservation, where AB is obtained by crossbreeding of A and B. The MVO method will not give extra value to A and B when AB is already conserved, whilst the MVT method would give extra value to A and B. The latter is because, although A and B contain no alleles that are not present in AB, they may contain genotype and allele combinations that are not present in AB, because in AB the alleles of A and B are mixed up. Thus, it may be easier to find particular genotype and allele combinations when A and B are conserved instead of AB, *i.e.* the genetic variance is more accessible in a set containing A and B compared to a set containing only AB.

The differences between the two core sets become obvious when focussing on the individual breed contributions. From the optimisation term (formula (3)) it follows that the MVT core set prioritises breeds with a high within breed kinship and a low average between breed kinship. In contrast, the MVO core set method favours breeds that show a low average kinship both within and between breeds. This was shown by the results of the second series of simulations (Tab. II). Considering the breed contributions to the two core sets from all simulated configurations and all replicates, the Spearman rank correlation was around 0.7 (averaged over all replicates), indicating that the two core sets prioritise different breeds only to some extent. From formulae (3) and (10), it follows that this extent is also a function of the differences of the within breed kinships of the breeds.

Assuming that the dendrogram in Figure 1 corresponds to a phylogenetic tree, it then appears that the MVO core set method prioritises breeds that are as close as possible to the base breed whereas the MVT core set approach prefers breeds that have drifted further away from the base breed. In terms of allele frequencies, the MVO core set avoids extreme frequencies and therefore, maximises the possible directions of selection. The MVT core set prioritises breeds with more extreme frequencies and thus prioritises breeds that already show different combinations of genotypes. Hence, the MVT core set method aims to conserve breeds that show large differences in the respective population mean of a hypothetical quantitative trait. This makes the MVT core set method attractive, because the efficiency of upgrading a breed by introducing genetics from another breed is a function of the difference in the respective population means. The MVT core set thus enables a faster reaction on putative changed conditions compared to the MVO core set.

The conceptual differences between the two core set diversity methods on the one hand and the Weitzman diversity on the other hand is that both core set methods do account for the within breed variance and do account for migration between breeds. The ignorance of these two aspects was seen as the main point

of criticism of the application of Weitzman diversity on a within-species breed level [1, 3, 4], but in the conservation of farm animal genetic resources we are acting on this level. The numerical results of the field data analysis demonstrate the differences between the two core set diversity methods and the Weitzman diversity (Fig. 1). Whereas the latter focuses exclusively on the between breed variance, and hence prioritises breeds for conservation that are most distantly from the other breeds, the MVT core set method also accounts for the within breed variance. Indeed this is the reason why the MVT core set gives substantial weights to the Holstein Friesian and the Limousine breeds, but not the Weitzman diversity (Fig. 1). Nevertheless, Eding *et al.* [4] showed that the nice mathematical and biological properties of the Weitzman diversity [15, 16] also hold in the MVO core set, if the 'monotonicity in distance property' (*i.e.* diversity should increase as any of the distances increase) is replaced by a kinship argument: the offspring variance in a set of breeds increases when the kinship between or within breeds decreases. Obviously, if the inbreeding in any of the breeds increases (*i.e.* distances increase), the total variance will be increased in the MVT core set method. Accordingly, the MVT core set method fulfils all of the original Weitzman criteria.

Simianer *et al.* [15] developed formulae for the estimation of marginal diversities of breeds using the Weitzman unit of diversity together with breed extinction probabilities. The marginal diversity of breed *i* is defined as the change of the conserved diversity when the extinction probability of breed *i* is changed by one unit [15]. With the use of marginal diversities it is possible to compare different conservation strategies. The authors state that it is possible to also use other diversity units as long as these fulfil the Weitzman criteria [15]. Therefore, the estimation of the marginal diversities is also possible in the core set methods given that estimates for the extinction probabilities are available. In this case, marginal diversities could replace the relative breed contributions as criteria for conservation priority. The advantage of this prioritisation is that the extinction dynamics of the breeds would be considered. Furthermore, Simianer *et al.* [15] suggested an approach of how diversity can be combined with other conservation criteria like *e.g.* special trait characteristics in order to generate a more comprehensive conservation criterion. Their outlines also seem to be valid for the diversity units considered by the two core set methods. Further work is needed to demonstrate the expansion of the presented core set methods to the ideas proposed by Simianer *et al.* [15].

The diversity approach of Piyasatian and Kinghorn [12] uses allelic variation as a criterion for conservation. Allelic variation does not account for the extra variance due to inbreeding (*i.e.* due to a combination of alleles), and

therefore this method is closer to the MVO core set method rather than to the MVT core set method. However, Piyasatian and Kinghorn [12] recognise that a higher between breed variance is valuable for conservation and therefore gives a five times higher weight to the between breed variance compared to the within breed variance.

Throughout this study we focussed on the selection of breeds for a conservation plan rather than on the management of conserved breeds in order to maintain genetic variance. Generally, this management is a question of the selection of individuals as parents for the next generation and a question of the mating scheme of these individuals. It seems advisable to mate the breeds selected for the conservation plan only within breed in order to maintain the combinations of alleles and genotypes within the breeds, and hence, to conserve the between breed variance.

4.2. Accuracy of the contribution vector estimation

All three methods for the contribution vector estimation used marker based kinships obtained by a log-linear model as proposed by Eding and Meuwissen [5]. The advantage of this model is the simultaneous use of all available information to correct the average similarities for the AIS probabilities, and hence to obtain the kinships. It was shown that this model produces kinship estimates that show a high accuracy with a small upward bias [5]. Nevertheless, in the same study, the authors found that the accuracy of the contribution vector estimates is only moderate, which is due to the sampling errors in the kinship matrix, despite the high accuracy of the kinship estimates. In order to obtain contribution vectors with a higher accuracy, two bootstrap methods were tested and the results of the simulations (Tabs. I and II) clearly show the superiority of the bootstrap based methods (BM and WBM, respectively) over EEA in both core sets. Not only were the correlations higher and the mean square errors lower but the number of breeds with an estimated zero contribution was also closer to the real value. Remember that the true zero contribution values in the MVT core set received in general only a very small estimate when using the bootstrap based methods; from a practical point of view these values are negligible. See for example the contributions of the Belgian Blue and of the Dutch Friesian breed to the MVT core set (Tab. V). The large fraction of false zero contribution estimates was one of the main problems reported by Eding *et al.* [4] in contribution vector estimation. It now seems that the use of the bootstrap based methods provides a solution to this problem. The reason is that for each contribution estimate a distribution was generated by

bootstrapping, and because of the recursive nature of the optimisation algorithms (*i.e.* the most negative estimate was set to zero and the contribution vector estimation was repeated without the corresponding breed), the lower bound of the bootstrap distribution was zero. Thus, if at least one contribution estimate of the distribution showed a positive value, the final contribution estimate was small but positive. The significantly better performance of WBM over BM emphasised the advantage of weighting the observations in the log-linear model accordingly. Eding and Meuwissen [5] found that the benefit from the use of the weighted log-linear model is due to the fact that less informative loci receive a lower weight. The present study shows that instead of using the expected variance (formula (14)), it is also possible to use the empirical variance estimated by bootstrapping as shown in the appendix.

The higher accuracy of the MVT contribution estimates compared to the corresponding MVO estimates indicates that the MVT core set algorithm tends to be less sensitive to sampling errors in the estimated kinship matrix. The optimisation function of the MVT core set (formula (3)) probably shows a more pronounced extreme than the corresponding function of the MVO core set (formula (8)), hence rendering it easier to find.

5. CONCLUSION

The introduced core set method (MVT) prioritises breeds for conservation by maximising the total genetic variance for a hypothetical quantitative trait in the core set. It was shown that the numeric results of this core set approach and of the MVO core set approach of Eding *et al.* [4] are to some extent similar. The differences were most clearly shown by the results of the field data analysis. The MVT core set approach suggests the conservation of breeds that show comparatively large differences in the respective population mean of a hypothetical quantitative trait. This maximises the speed of achieving selection response for a putative changed breeding objective, which makes the MVT core set method attractive. For the estimation of the core set contribution vectors from molecular marker data, we recommend the use of the weighted bootstrap approach (WBM), because this method produced the most accurate estimates, regardless of the core set method.

REFERENCES

[1] Caballero A., Toro M.A., Analysis of genetic diversity for the management of conserved subdivided populations, Conserv. Genet. 3 (2002) 289–299.

[2] Canon J., Alexandrino P., Bessa I., Carleos C., Carretero Y., Dunner S., Ferran N., Garcia D., Jordana J., Laloë D., Pereira A., Sanchez A., Moazami Gourdazi K., Genetic diversity measures of local European beef cattle breeds for conservation purposes, Genet. Sel. Evol. 33 (2001) 311–332.

[3] Eding J.H., Meuwissen T.H.E., Marker based estimates of between and within population kinships for the conservation of genetic diversity, J. Anim. Breed. Genet. 118 (2001) 141–159.

[4] Eding J.H., Crooijmans R.P.M.A., Groenen M.A.M., Meuwissen T.H.E., Assessing the contribution of breeds to genetic diversity in conservation schemes, Genet. Sel. Evol. 34 (2002) 613–633.

[5] Eding J.H., Meuwissen T.H.E., Linear methods to estimate kinships from genetic marker data for the construction of core sets in genetic conservation schemes, J. Anim. Breed. Genet. 120 (2003) 289–302.

[6] Efron B., Tibshirani R.B., An introduction to the bootstrap, Chapman and Hall, New York, 1993.

[7] Fabuel E., Barragan C., Silio L., Rodriguez M.C., Toro M.A., Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers, Heredity 93 (2004) 104–113.

[8] Falconer D.S., MacKay T.F.C., Introduction to quantitative genetics, Longman, London, 1996.

[9] FAO, World watch list for domestic animal diversity, Third edition, Rome, 2000, http://www.fao.org/dad-is/.

[10] Felsenstein J., PHYLIP (Phylogeny Inference Package) Version 3.5c, University of Washington, 1995, http://evolution.genetics.washington.edu/phylip.html.

[11] Laval G., Iannuccelli N., Legault C., Milan D., Groenen M.A.M., Giuffra E., Andersson L., Nissen P.H., Joergensen C.B., Beeckmann P., Geldermann H., Foulley J.-L., Chevalet C., Ollivier L., Genetic diversity of eleven European pig breeds, Genet. Sel. Evol. 32 (2000) 187–203.

[12] Piyasatian N., Kinghorn B.P., Balancing genetic diversity, genetic merit and population viability in conservation programmes, J. Anim. Breed. Genet. 120 (2003) 137–149.

[13] Reist-Marti S.B., Simianer H., Gibson J., Hanotte O., Rege J.E.O., Weitzman's approach and conservation of breed diversity: an application to African breeds, Conserv. Biol. 17 (2003) 1299–1311.

[14] Ruane J., A critical review of the value of genetic distances in conservation of animal genetic resources, J. Anim. Breed. Genet. 116 (1999) 317–323.

[15] Simianer H., Marti S.B., Gibson J., Hanotte O., Rege J.E.O., An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock breeds, Ecol. Econ. 45 (2003) 377–392.

[16] Thaon d'Arnoldi C., Foulley J.-L., Ollivier L., An overview of the Weitzman approach to diversity, Genet. Sel. Evol. 30 (1998) 149–161.

[17] Weitzman M.L., On diversity, Quart. J. Econ. 107 (1992) 363–405.

APPENDIX

The weighted bootstrap method (WBM) used the reciprocal value of the empirical variance of the observations as weights in the log-linear model. The empirical variance was estimated using the bootstrap approach as described by Efron and Tibshirani [6]. A number of 100 bootstrap samples ($B = 100$) was generated by sampling with replacement n individuals out of the pool of n original individuals within breed. To obtain the observations for each bootstrap sample, the similarities were calculated using (11) of the main text. These similarities were subtracted from one, and the natural logarithm was taken from this difference. Following this, the observations in the current bootstrap sample bi for breed i and j at locus k is $\ln(1 - \hat{\tilde{S}}_{ij,k})_{bi}$, see also formula (12) of the main text. The bootstrap estimate of the variance of $\ln(1 - \widetilde{S}_{ij,k})$ is:

$$\mathrm{Var}_b\left[\ln\left(1 - \hat{\tilde{S}}_{ij,k}\right)\right] = \frac{1}{B-1}\sum_{bi=1}^{B}\left(\ln\left(1 - \hat{\tilde{S}}_{ij,k}\right)_{bi}\right) - \left(\ln\left(1 - \hat{\tilde{S}}_{ij,k}\right)_{b*}\right)^2, \quad (18)$$

where

$$\ln\left(1 - \hat{\tilde{S}}_{ij,k}\right)_{b*} = \frac{1}{B}\sum_{bi=1}^{B}\ln\left(1 - \hat{\tilde{S}}_{ij,k}\right)_{bi}. \quad (19)$$

For each data set, *i.e.* each replicate, these empirical variances were estimated one time and then used for the whole WBM analysis. Thus, for each bootstrap sample analysis the same weights were used.