G3, 2022, 12(2), jkab450
https://doi.org/10.1093/g3journal/jkab450
Advance Access Publication Date: 7 January 2022
Investigation

# Influence of model selection and data structure on the estimation of genetic parameters in honeybee populations

Manuel Du ,* Richard Bernstein , Andreas Hoppe , and Kaspar Bienefeld

Breeding and Behavior, Institute for Bee Research Hohen Neuendorf, 16540 Hohen Neuendorf, Germany

*Corresponding author: Institute for Bee Research Hohen Neuendorf, 16540 Hohen Neuendorf, Germany. Email: manuel.du@hu-berlin.de

## Abstract

Estimating genetic parameters of quantitative traits is a prerequisite for animal breeding. In honeybees, the genetic variance separates into queen and worker effects. However, under data paucity, parameter estimations that account for this peculiarity often yield implausible results. Consequently, simplified models that attribute all genetic contributions to either the queen (queen model) or the workers (worker model) are often used to estimate variance components in honeybees. However, the causes for estimations with the complete model (colony model) to fail and the consequences of simplified models for variance estimates are little understood. We newly developed the necessary theory to compare parameter estimates that were achieved by the colony model with those of the queen and worker models. Furthermore, we performed computer simulations to quantify the influence of model choice, estimation algorithm, true genetic parameters, rates of controlled mating, apiary sizes, and phenotype data completeness on the success of genetic parameter estimations. We found that successful estimations with the colony model were only possible if at least some of the queens mated controlled on mating stations. In that case, estimates were largely unbiased if more than 20% of the colonies had phenotype records. The simplified queen and worker models proved more stable and yielded plausible parameter estimates for almost all settings. Results obtained from these models were unbiased when mating was uncontrolled, but with controlled mating, the simplified models consistently overestimated heritabilities. This study elucidates the requirements for variance component estimation in honeybees and provides the theoretical groundwork for simplified honeybee models.

Keywords: REML estimates; genetic parameters; honeybees; maternal and direct effects; computer simulation; animal model; mating control

## Introduction

Breeding efforts in honeybees, as in other agricultural species, rely on the fact that selection traits are partly determined by genetic features and that superior animals can thus pass on their qualities to their offspring. In order to determine the extent to which this is the case, it is necessary to measure the magnitude of genetic variance as a contributor to the phenotypic variance in a population. Thus, estimates of genetic and residual variance components are important to judge the prospect of success in selective breeding schemes (Lush 1937).

In addition to their importance for evaluating the possibilities for breeding, genetic and residual variance estimates also belong to the input data of best linear unbiased prediction (BLUP) breeding value estimation (Henderson 1975) and are thus an integral part of modern techniques of genetic evaluation, including that of honeybees (Bienefeld et al. 2007; Brascamp and Bijma 2014; Hoppe et al. 2020).

Estimates of variance components in livestock populations typically rely on the animal model (AM), assuming that the phenotype $y_{ij}$ of an individual $j$ in an environment $i$ is determined by a fixed environmental effect $b_i$, a random genetic effect $u_j$ of the individual, and a random residual effect $e_{ij}$:

$$y_{ij} = b_i + u_j + e_{ij}. \quad \text{(AM)}$$

Based on this model and the relationships between the involved individuals, the variances of $u_j$ (additive genetic variance, $\sigma^2_A$) and $e_{ij}$ (residual variance, $\sigma^2_E$) are typically estimated as restricted maximum likelihood (REML) values, assuming that $u_j$ and $e_{ij}$ are normally distributed (Patterson and Thompson 1971). There exist several algorithms to derive REML estimates, the two most widely used probably being Expectation Maximization REML (EMREML) (Dempster et al. 1977; Mäntysaari and van Vleck 1989) and Averaged Information REML (AIREML) (Madsen et al. 1994; Johnson and Thompson 1995; Gilmour et al. 1995). The EMREML algorithm is generally deemed reliable to return exact REML estimates; however, it often takes many iterations to converge (Meng and van Dyk 1998; Thompson et al. 2005; Misztal 2008). In contrast, AIREML terminates much faster but also requires a higher degree of regularity of the likelihood function. It may thus fail to yield accurate REML estimates, particularly if the data set is small or irregular (Misztal 2008; Masuda 2019).

The application of the AM to the honeybee bears the problem that, unlike in other livestock species, a phenotype record is usually not attributed to an individual bee but to a colony, i.e. a

Received: November 16, 2021. Accepted: December 22, 2021
© The Author(s) 2022. Published by Oxford University Press on behalf of Genetics Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

collective of bees with different genetic properties. In particular, the distinction between queen and worker group has proven important, because the same genetic set-up has different influences on a trait, depending on the caste it is expressed in (Bienefeld and Pirchner 1990). The model which to date represents the honeybee biology most accurately was introduced by Brascamp and Bijma (2014), refining several earlier approaches (Chevalet and Cornuet 1982; Bienefeld et al. 1989, 2007). We call this model the colony model (CM). It assumes that for a colony, consisting of a queen $q$ and a worker group $w$ in an environment $i$, the phenotype $y_{iqw}$ is determined by

$$y_{iqw} = b_i + u^{Q}_{q} + \overline{u}^{W}_{w} + e_{iqw}. \qquad \text{(CM)}$$

The concept is similar to that of maternal and direct effects in other livestock species (Willham 1963): The genetic effect $u$ splits up into a (maternal) queen component $u^{Q}$ and a (direct) worker component $u^{W}$. The genetic effect of the worker group, $\overline{u}^{W}_{w}$, is equipped with a bar to indicate that it is formed as the average genetic effect of the individual worker bees in the colony. When the CM is used for parameter estimation, REML estimates are calculated for the queen effect genetic variance, $\sigma^2_{A^Q}$, the worker effect genetic variance, $\sigma^2_{A^W}$, the residual variance, $\sigma^2_E$, as well as the genetic covariance between queen and worker effects, $\sigma_{A^{QW}}$. While $\sigma^2_{A^Q}$ and $\sigma^2_E$ are precisely the variances of $u^{Q}_{q}$ and $e_{iqw}$, respectively, the variance of $\overline{u}^{W}_{w}$, being the variance of an averaged value, is reduced to $a_{ww}\sigma^2_{A^W}$, where the reduction factor $a_{ww}$ is the average relationship between two workers in a colony (Brascamp and Bijma 2014, 2019). If a queen mates freely with many mutually unrelated drones, most of her workers are maternal half-sibs and thus $a_{ww}$ is little larger than 0.25. In breeding schemes with controlled mating on mating stations, the value for $a_{ww}$ has been estimated between 0.317 (Brascamp et al. 2018) and 0.427 (Bienefeld et al. 1989).

The CM has yielded unbiased REML estimates of variance components with simulated data (Brascamp et al. 2014; Brascamp and Bijma 2019) and has been applied successfully in the estimation of genetic parameters in real honeybee populations (Bienefeld and Pirchner 1990; Brascamp et al. 2016; Hoppe et al. 2020). However, it turned out that this model has high demands on data quality. In case of imperfect data, algorithms for parameter estimation in honeybees have repeatedly failed to converge or yielded unreasonable results. This was observed for the AIREML algorithm (Andonov et al. 2019; Guichard et al. 2020), as well as older and supposedly more reliable algorithms, including EMREML (Willam and Eßl 1993; Zakour et al. 2012).

As a remedy, several recent studies used simplified honeybee models for parameter estimations, attributing the entire genetic component of the phenotype either to the queen (queen model, QM) or to the worker group (worker model, WM), while ignoring the other caste (Andonov et al. 2019; Facchini et al. 2019; Guichard et al. 2020, 2021):

$$y_{iq} = b_i + u^{(QM)}_{q} + e^{(QM)}_{iq}, \qquad \text{(QM)}$$

or

$$y_{iw} = b_i + \overline{u}^{(WM)}_{w} + e^{(WM)}_{iw}. \qquad \text{(WM)}$$

These simplified models lead to increased numeric stability of the estimation procedures, also when the data quality is poor (Willam and Eßl 1993; Guichard et al. 2020, 2021). They are easily justified for traits that are clearly attributed to only one caste (de Graaf et al. 2020; Facchini et al. 2021). However, most commercially important honeybee traits, such as honey yield or gentleness, are commonly influenced by the queen and her workers. For these traits it is to date unclear, how parameter estimates that were obtained with the simplified models QM and WM compare to estimates using the complete model CM. Sometimes, $u^{(QM)}_{q}$ from the QM is directly associated with $u^{Q}_{q}$ from the CM, although there exists a general sentiment that $u^{(QM)}_{q}$ likely also captures some worker effect information due to the close relationship between a queen and her workers (Guichard et al. 2020, 2021). The same holds for $u^{(WM)}_{w}$ and $u^{W}_{w}$. In simulation studies on mammals, it has been shown that neglecting existing maternal effects in parameter estimations can lead to severely biased variance estimates for the direct effects (Robinson 1996; Clément et al. 2001).

Data paucity, which has been made responsible for the failure to estimate genetic variance components under the CM in several honeybee populations, appears in different forms.

- The mating behavior of the honeybee differs from that of other agricultural species. A young queen mates only once, soon after hatching, with several drones from other hives and afterwards uses the collected semen to fertilize eggs for the rest of her life. For the breeder, this mating behavior is typically not observable, leading to uncertain paternity and thus incomplete pedigrees. Several studies for other livestock species have shown that incomplete pedigree information can yield biased estimates of variance components (Cantet et al. 2000; Clément et al. 2001; Id-Lahoucine and Casellas 2017). The situation for the honeybee can be ameliorated by the use of artificial insemination or isolated mating stations, which provide a certain degree of paternal pedigree information (Bienefeld et al. 1989; Brascamp and Bijma 2014; Uzunov et al. 2017). But although it has been shown that these strategies substantially enhance genetic response (Plate et al. 2019b; Du et al. 2021a), many honeybee populations are still bred without or with incomplete mating control (Andonov et al. 2019; De la Mora et al. 2020; Maucourt et al. 2020).
- Apiary sizes play a crucial role in separating fixed and random effects, because colonies from the same apiary are usually attributed the same fixed effect. In comparison to other agricultural species, contemporary groups in honeybees tend to be small (Andonov et al. 2019; Bieńkowska et al. 2020), which is likely to harm the accuracy of estimated genetic parameters (Swalve 1995; Strabel and Swaczkowski 1999; Vasconcelos et al. 2008).
- Performance tests for some honeybee traits, such as the pin-test for hygienic behavior, are laborious and therefore not recorded by all breeders, leading to incomplete phenotype data (Hoppe et al. 2020). Studies in other agricultural species have shown that missing phenotype data hampers the estimation of genetic parameters and that the disentanglement of maternal and direct genetic variances is particularly compromised (Gerstmayr 1992; Maniatis and Pollott 2003; Heydarpour et al. 2008).

While it is known that these factors have a negative influence on genetic parameter estimations, it is unclear to what extent this is the case in honeybee populations.

In this study, we derive a theoretical framework that allows the comparison of genetic parameter estimates under the CM with those under the simplified queen and worker models (QM

Downloaded from https://academic.oup.com/g3journal/article/12/2/jkab450/6500294 by guest on 24 September 2023

Strabel BT, Szwaczkowski T. The use of test day models with small size of contemporary groups. J Anim Breed Genet. 1999;116(5): 379–386.

Swalve HH. The effect of test day models on the estimation of genetic parameters and breeding values for dairy yield traits. J Dairy Sci. 1995;78(4):929–938.

Thompson R, Brotherstone S, White IMS. Estimation of quantitative genetic parameters. Philos Trans R Soc Lond B Biol Sci. 2005; 360(1459):1469–1477.

Tsuruta S, Misztal I. THRGIBBSF90 for estimation of variance components with threshold and linear models. In: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production. Brazil: Belo Horizonte; 2006.

Uzunov A, Brascamp EW, Büchler R. The basic concept of honey bee breeding programs. Bee World. 2017;94(3):84–87.

van der Werf JHJ, Meuwissen THE, de Jong G. Effects of correction for heterogeneity of variance on bias and accuracy of breeding value estimation for Dutch dairy cattle. J Dairy Sci. 1994;77(10): 3174–3184.

van Tassel CP, Van Vleck LD, Gregory KE. Bayesian analysis of twinning and ovulation rates using a multiple-trait threshold model and Gibbs sampling. J Anim Sci. 1998;76(8):2048–2061.

Vasconcelos J, Santos F, Bagnato A, Carvalheira J. Effects of clustering herds with small-sized contemporary groups in dairy cattle genetic evaluations. J Dairy Sci. 2008;91(1):377–384.

Visscher PM, Hill WG. Heterogeneity of variance and dairy cattle breeding. Anim Sci. 1992;55(3):321–329.

Willam A, Eßi A. Schätzung von Populationsparametern für verschiedene Merkmale bei der Honigbiene (*Apis mellifera carnica*). Apidologie. 1993;24(4):355–364.

Willham RL. The covariance between relatives for characters composed of components contributed by related individuals. Biometrics. 1963;19(1):18–27.

Yabe S, Iwata H, Jannink JL. Impact of mislabeling on genomic selection in cassava breeding. Crop Sci. 2018;58:1470–1480.

Zakour MK, Ehrhardt K, Bienefeld K. First estimate of genetic parameters for the Syrian honey bee *Apis mellifera syriaca*. Apidologie. 2012;43(5):600–607.

*Communicating editor: G. de los Campos*

## Appendix

We calculate, which part of a colony's phenotype $y_{iqw}$, formed according to the CM, is projected to the worker group's genetics if the simplified model WM is used for the analysis. Therefore, the random contributions to the phenotype, $r_{iqw} = u_q^Q + \overline{u}_w^W + e_{iqw}$, have to be split up into a part that can be expressed in terms of $\overline{u}_w^Q$ and $\overline{u}_w^W$ thus is attributed to the worker group, and a part that is uncorrelated to these variables and is thus interpreted as a residual. In analogy to Equation (1), the worker group's breeding value for the queen effect, $\overline{u}_w^Q$, is inherited from the queen $q$ and the drones $d$ that mated with $q$:

$$\overline{u}_w^Q = \frac{1}{2}u_q^Q + \overline{u}_d^Q. \tag{7}$$

With $\mathrm{var}(\overline{u}_w^Q) = a_{ww}\sigma^2_{A^Q}$ and $\mathrm{var}(u_q^Q) = \sigma^2_{A^Q}$, Equation (7) yields (because $u_q^Q$ and $\overline{u}_d^Q$ are independent):

$$\mathrm{var}(\overline{u}_d^Q) = \left(a_{ww} - \frac{1}{4}\right)\sigma^2_{A^Q}. \tag{8}$$

As the following calculation shows, $\overline{u}_w^Q$ is therefore uncorrelated to the linear combination $u_w^\perp = \frac{1}{2}u_q^Q - \frac{1}{4a_{ww}-1}\overline{u}_d^Q$:

$$\begin{aligned}\mathrm{cov}\left(\overline{u}_w^Q, u_w^\perp\right) &\overset{\text{Eq. 7}}{=} \mathrm{cov}\left(\frac{1}{2}u_q^Q + \overline{u}_d^Q, \frac{1}{2}u_q^Q - \frac{1}{4a_{ww}-1}\overline{u}_d^Q\right)\\ &\overset{\text{Eq. 8}}{=} \frac{1}{4}\sigma^2_{A^Q} - \frac{1}{4a_{ww}-1}\left(a_{ww} - \frac{1}{4}\right)\sigma^2_{A^Q}\\ &= 0.\end{aligned}$$

Indeed, also $\overline{u}_w^W$ is uncorrelated to $u_w^\perp$:

$$\mathrm{cov}(\overline{u}_w^W, u_w^\perp) = \frac{1}{4}\sigma_{A^{QW}} - \frac{1}{4a_{ww}-1}\left(a_{ww} - \frac{1}{4}\right)\sigma_{A^{QW}} = 0.$$

Thus, by writing $u_q^Q$ as

$$\begin{aligned}u_q^Q &= \frac{1}{2a_{ww}}\cdot\frac{1}{2}u_q^Q + \frac{1}{2a_{ww}}\overline{u}_d^Q + \left(2 - \frac{1}{2a_{ww}}\right)\frac{1}{2}u_q^Q - \frac{1}{2a_{ww}}\overline{u}_d^Q\\ &= \frac{1}{2a_{ww}}\overline{u}_w^Q + \left(2 - \frac{1}{2a_{ww}}\right)u_w^\perp,\end{aligned}$$

we arrive at the desired decomposition

$$\begin{aligned}r_{iqw} &= u_q^Q + \overline{u}_w^W + e_{iqw}\\ &= \frac{1}{2a_{ww}}\overline{u}_w^Q + \overline{u}_w^W + \left(2 - \frac{1}{2a_{ww}}\right)u_w^\perp + e_{ijk}\\ &= \overline{u}_w^{(WM)} + e_{iw}^{(WM)},\end{aligned}$$

with $\overline{u}_w^{(WM)}$ and $e_{iw}^{(WM)}$ as defined in Equations (5a) and (5b).

The above considerations can be formulated more concisely in the language of Hilbert spaces. An inner product of random effects with expectation zero is defined by $\langle v_1, v_2\rangle = \mathrm{cov}(v_1, v_2)$. Then $\overline{u}_w^{(WM)}$ is the orthogonal projection of the combined random effects $r_{iqw} = y_{iqw} - p_i$ onto the subspace spanned by $\overline{u}_w^Q$ and $\overline{u}_w^W$. An orthogonal basis of this subspace is given by the Gram-Schmidt procedure and consists of $v_1 = \overline{u}_w^Q$ and $v_2 = \overline{u}_w^W - \frac{\langle \overline{u}_w^Q, \overline{u}_w^W\rangle}{\langle \overline{u}_w^Q, \overline{u}_w^Q\rangle}\overline{u}_w^Q$. Then, $\overline{u}_w^{(WM)}$ can be calculated as

$$\overline{u}_w^{(WM)} = \frac{\langle v_1, r_{iqw}\rangle}{\langle v_1, v_1\rangle}v_1 + \frac{\langle v_2, r_{iqw}\rangle}{\langle v_2, v_2\rangle}v_2, \tag{9}$$

and

$$e_{iw}^{(WM)} = r_{iqw} - \overline{u}_w^{(WM)}. \tag{10}$$