

# An integrated environment for genetic evaluation: databases, variance component estimation and genetic evaluation

Eildert Groeneveld  
Institute for Animal Science  
Federal Agricultural Research Center  
Mariensee, Germany

Multivariate genetic evaluation considers jointly many different traits from different sources on current and past animals. As a routine operation, investment in a central data repository or an integrated database creates a flexible, safe and efficient environment for multi trait BLUP. With a central database all data editing can be recorded in scripts thereby documenting the data editing process. An open source approach is presented using the APIIS framework. Generic procedures and programs are presented, which allows the creation of integrated databases, supports the pre-BLUP and post-BLUP data processing steps, which are all centered around the integrated database. An automated evaluation of BLUE and BLUP is presented.

## 1 Introduction

All modern breeding programs use three constituent components data collection, analysis of data, genetic evaluation of data and finally selection based on results from the genetic evaluation.

## 2 Creating an integrated database

Genetic evaluation of animals has always been a heavily data based technique. As such, animal breeders are used to working with computers on a large scale, to implement statistical procedures in program code and push the limits of what can be done with

computers ever further. Particularly, with the introduction in BLUP (best linear unbiased prediction), historical data obtained a whole new value, as they were to be included in every genetic evaluation. With the advent of animal models, the computational demands increased even further. When multivariate genetic evaluations became a feasible proposition, the demand and necessity to manage more and diverse data became even more obvious. While large scale computing centers tended to be sufficiently staffed to deal with this situation, smaller organizational units found and still find it rather difficult to meet the demand for properly organized data repositories. As a result, these are often insufficient, containing data with limited accuracy, thereby making proper genetic evaluation difficult and cumbersome. To address this situation the initiative for an open source framework [7] for data management for individual animal records was launched and is proving its first utility [4, 3].

So far most of the data used in genetic evaluations have been generated in traditional performance recording schemes as are produced from milk recording programs in cattle or litter recording schemes in pigs and other species. Some of the data used also originate in laboratories which are an integral part of the recording scheme infrastructure like milk recording labs, which produce fat and protein contents of samples and values like somatic cell counts.

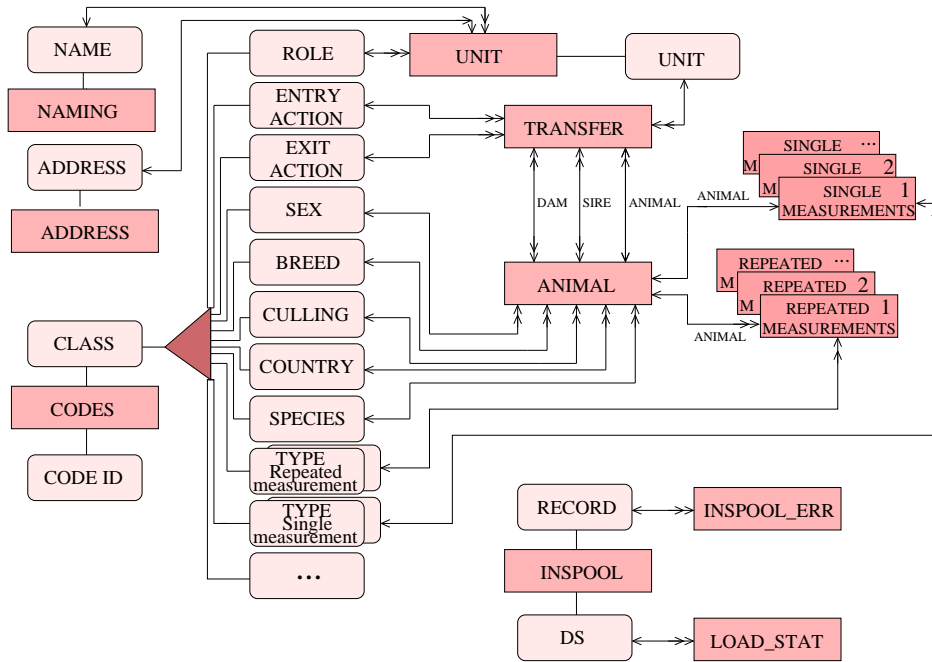
With the advent of genotyping information generated in molecular genetic labs, a whole new class of data is generated and needs to be managed. Here, we have a number of options: firstly, molecular genetic data may get handled only as results coming from the labs where we are not concerned with the data management in the labs themselves. Thus, for animals investigated the labs will supply only a number of genotypes for each animal investigated. Alternatively, we include the data management in the labs themselves into our consideration. Operationally, the two setups can be viewed as one: the data management support could be omitted where not appropriate, while it would be included, where labs become a constituent component of the complete operation. It is this setup that Groeneveld et al. focused on in their MoLabIS project [4].

The adaptable platform independent information system (APIIS) was developed as a framework of procedures and software to support the rapid creation of integrated databases built on individual animal records [3]. With its generic design it allows to handle all animal identification systems be they based on unique lifetime numbers or recurring animal identifications that may be repeated over time. With this setup it is possible to handle all species with different identifications like cattle using the EU unique identification and the recurring numbering systems with renumbering at some time during the life span of pigs and sheep.

The basic principle of APIIS allows the creation information derived from individual animal records, be they daily gain, milk yield or type classifications. The generic structure of this setup is given in figure 1.

Depending on the data collected, the resulting database structure may be more or less complicated. However, for many situations a standard setup will be sufficient. This is based on generic structure of any APIIS implementation. For each individual animal database a large portion of the database can be considered fixed. These are, for instance, the blocks that center around address handling. Also the animal numbering scheme as

Figure 1: Generic Structure of APIIS



represented by the tables TRANSFER and ANIMAL can be considered fixed. Nearly all recording schemes of individual animal records have a block that deals with SERVICE - be it AI or natural service - and the reproduction area.

Many of the other performance data recorded can be grouped into “single animal records” and “multiple animal records”. The former is a one to one relation with the animals: there is only one measurement on an animal. An example might be the genotype for a certain micro satellite. “Multiple animal records” are usually recorded through the life of the animal. Therefore, all records are discriminated from one another through the animal identification (which must always and under all circumstances be unique) and the date at which the measurement was taken. Typical examples are weights and laying records. If this setup is used then the procedures developed in APIIS allows rapid design and implementation of any testing scheme. Support for this process is provided in the RapidAPIIS framework that is currently under development to allow non IT specialized persons to setup up customized databases. As an outcome a standalone database will be created that allows data entry through GUI (graphical user interface) forms under the control of a complete set of business rules.

As APIIS is developed using solely Open Source software, the creation of a database under its umbrella does not incur any licensing costs. While this may not be an issue for large computing centers, it does present a problem for many organizations which do not have the financial power to afford commercial products like database systems.

### 3 Integrating genetic evaluation

As will have become clear from the above, all data used for breeding an population management should be contained in that one and only database for general use. This implies that no other data files exist anywhere on the computer that are being used unless they are derived from the central database.

Clearly, for many purposes not all data (which includes historical data on animals that have long ceased to exist) are needed for some evaluation or data processing. There may be an investigation on one line only, or only certain years may be of interest. The process of creating these subsets always implies a selection or editing procedure. Often this is done manually by – for instance – starting from a spreadsheet, then manually some rows are chosen and others deleted. The resultant spreadsheet is then saved under a new name. The more often this procedure is repeated the more difficult it becomes the keep track of the editing or selection / modification actions. Also, it gets increasingly difficult to keep track of which file means what and even what the original may be.

With an integrated database a simpler mode of operation is possible. Firstly, it is clear that only the integrated database is the “original”. This implies that - by definition - all other files that may exist have to be copies or derived form the original. If now, additionally, all extractions and editing for further investigations that require separate files as input are done through scripts or little programs, then the editing logic is fully contained in this extraction script. As a result, the extracted files can be deleted at any time and recreated by simply rerunning the extraction script. While this does require knowledge about how to write these scripts, the resulting operation structure is simple, consistent and straight forward. One other result from this setup is that a procedure can be repeated quickly, as no human intervention is required to execute the individual steps. Thus, whole series of steps can be executed automatically. In routine operations this script based automation of perhaps long and involved processes, like the extraction of data, the computation of BLUP and their further processing, this is a large advantage for effective data processing. It contrasts to manual point and click operations. This is only fast for the first time, but if to be repeated it becomes very inefficient compared to script based operations.

#### 3.1 Model definition and variance component estimation

A first step for Animal model genetic evaluation is the development of the statistical model followed by the estimation of the required covariance matrices. While this is an involved process, it does not have to be done regularly. Software packages are available to actually perform the covariance component estimation like VCE [6, 5] or ASREML [1] and others. As the procedures involved are identical to those required for genetic evaluation, everything there applies to the data preparation steps required for covariance component estimation.

## 3.2 Computation of BLUPs for selection

Once all data on a breeding program are stored in a normalized database, extraction of any subset is usually straight forward. In animal breeding we need normally a data and a pedigree file. These two files are not independent, as the pedigree file to be used in genetic evaluation should be derived from that actual data used in the analysis. Thus, for a data record only those pedigree records should be used, that are genealogically connected to it. This can be done by a recursive procedure starting from the animal in the data record collection all its ancestors through backward recursion.

Furthermore, some columns of the data files may need to be modified depending on the statistical model chosen for the analysis. For instance, if seasons are used instead of months, these need to be created during the extraction process. Creation of genetic groups is also often required depending perhaps on birth years of animals.

All these editings have one common aspect: the changes that are made from the original data in the database to the file that is used for genetic evaluation is rule based and these rules can be conveniently be put into the extraction program.

As part of APIIS there are two programs that are grouped around the genetic evaluation. The first one supports the extraction process from the database to create input for the BLUP software (here PEST [2]), while the second loads all solutions for fixed and random effects back into the database and performs a post BLUP data analysis. The latter is set up in a way that consecutive sets of solutions are stored in the database. In this way the changes in BLUP (and of course also BLUE) can be followed up using the database access mechanisms.

The ever growing amount of data that goes into BLUP runs results in proofs being repeatedly recomputed. With increasing information, the BLUPs will tend to get more accurate. However, this also means that they will change. As a result of this – in particular if the BLUPs go down with time – one will want to investigate how they and the BLUEs have developed over time. With all solutions from each run to be loaded into the database, this procedure has all the information at hand to answer questions.

### 3.2.1 pre BLUP processing

As mentioned above, a generalized program is available which supports the created of BLUP input files. Its main features are:

- a simple SQL command per trait and effect
- only internal database numbers are being used for animal ID and also for codes. Thus, by virtue of APIIS these are always unique even if external codes and animal IDs get reused
- consistent pedigrees are generated on the basis of the data set extracted
- allows the definition of genetic groups
- as an option recoded data and pedigree files can be created as are required by some packages like VCE for REML variance component estimation

### 3.2.2 Storing BLUEs and BLUPs in the database

In BLUP genetic evaluation only a small fraction of the solutions are indeed required for selection. These may indeed be only a hand full of animals from tens of thousands that were included in the BLUP run. Thus, in postprocessing other information like test dates are required to print the meaningful lists for selection. This is best done within the database, requiring solutions from the mixed models to be loaded back into the database. Also, if for consistency reasons, internal database codes are used throughout the BLUP computations, for use in the field the external codes have to be available. For this process, again, the database is required. Thus, after BLUEs and BLUPs have been computed from the Mixed Model Equations these will be loaded into the database in a uniform format. This is then the basis for further analyses in the post processing step.

### 3.2.3 post BLUP processing

Both, the extraction script and the model definition for the computations of the BLUPs contain all information required to perform a widely automated post processing. Its objective is the generation of statistics which allows a quick assessment of the actual data structure used in the genetic evaluation. One of the problems in using mixed model methodology lies in the fact, that the actual solutions of interest, mostly the BLUPs say little about the data structure used for its computation. If, for instance, many fixed effect class have only very few levels like 1 or 2, BLUPs will be computed even though they may not be useful, because the model used is not appropriate relative to the underlying data structure. Thus, if the actual datastructure is not observed and checked, meaningless BLUPs may be computed but this fact may go undetected.

Here, the post processing can help to generate a number of statistics that may give a quick overview about the underlying data structure and thus help to pinpoint weak spots. The beauty of the procedure lies in the fact that no explicit specification about the traits and effects used has to be made.

The following gives some parameters estimated automatically from the information stored thus far:

- biggest differences in BLUPs together with structural information like number of half sibs, animals own performance, number of offspring and its average performance
- statistics on fixed effects like minimum and maximum number of records per level
- covariables plotted as histograms
- information specified per breeder
- the genetic trend plotted for animals with a birthdate
- trends in fixed effects (BLUE) in comparison to the population

The statistics are generated as high quality pdf documents, some samples of which are given in the appendix. The first examples gives an excerpt from a set of statistics describing the data structure used in the genetic evaluation. Here, for all the effects their number of levels are reported. As can be seen, for each effect the average number of levels together with their minimum and maximum count are given, followed by a list of effect codes that have less than three observations. The next table gives information on how much BLUPs change from one genetic evaluation to the next. This may be the change from one week to the following. The differences in BLUP together with structural information on the sources of information, i.e. number of full and half sibs along with ancestor records are given. This will give some insight into what can be reasonably expected in terms of change of BLUPs.

The following tables give a graphical representation of covariables and some plots derived from the solutions to the mixed model solutions. The latter may be given as the difference between a subsection of the population, i.e. a herd relative to the average of the population. For BLUE, this can be used as a management support to the herd: the difference to population average indicates what management changes may be able to achieve. On the other hand, the plots of BLUP over time give an indication of the genetic trend in the population.

## **4 Conclusion**

Much of the data processing has to be done on a regular schedule be it for data entry or also for genetic evaluation. In an integrated system these activities can be automated and be based on scripts. This contrast to a mode of operation which does the steps required manually, possibly through a seemingly easy point and click graphical user interface. While this may give a quick start it is efficient in regular and repeated operations.

## **Acknowledgement**

Ralf Fischer has written much of the pre and postprocessing software and supplied the examples in the appendix which is gratefully acknowledged.

# Appendix

breeding value estimation from 23rd September 2005

## Analysis of breeding value estimation from 23rd September 2005

### 1 Effects

#### 1.1 Breeding value estimation tbw1

##### 1.1.1 Effect: hwr

hwr	count	min	avg	max
	62	1	40.03	431
classes	1	≤ 5	≤ 10	≤ 20
number of effects in this class	11	20	6	7

The following effects have less than 3 members:

'168 '230 '240 '354' '445 '362 '199 '189 '167 '215' '218' '149' '123' '295' '29' '202' '133' '256' '227' '198'

##### 1.1.2 Effect: sjm

sjm	count	min	avg	max
	35	14	45.60	74
classes	1	≤ 5	≤ 10	≤ 20
number of effects in this class	0	0	0	32

The following effects have less than 3 members:

##### 1.1.3 Effect: bjg

bjg	count	min	avg	max
	449	1	14.68	174
classes	1	≤ 5	≤ 10	≤ 20
number of effects in this class	46	162	73	59

The following effects have less than 3 members:

382-1998-1' 396-1995-3' '461-2003-3' '410-2002-4' '432-2003-1' '461-1995-2' 461-2002-2'  
 396-1998-2' 430-2003-2' '432-2002-3' '460-1999-4' '441-1995-2' 396-1997-3' '402-2002-1'  
 406-1996-2' 400-1997-1' '304-2003-2' '468-1995-2' '430-2002-2' 385-2002-1' 468-1998-1'  
 460-1997-2' 394-2002-3' '448-1999-2' '443-2003-4' '432-2000-2' 434-2002-1' 468-1995-1'  
 448-2002-1' 435-2003-1' '432-2003-2' '432-2003-1' '435-1997-2' '434-1999-2' 468-1997-3'  
 406-1998-1' '461-1999-2' '432-2003-1' '441-1996-2' '432-2002-1' 382-1997-4' '432-2003-3'  
 430-1995-1' '454-2003-1' '435-1995-1' '432-1999-3' '394-2002-4' '402-2003-1' '432-1998-2'  
 402-1999-1' 461-1999-1' '434-1995-2' '402-2000-3' '394-1997-2' '461-2000-2' 441-1996-4'  
 380-2001-1' 468-1996-3' '440-1998-1' '410-1996-3' '431-1995-2' '461-1997-3' 408-1998-1'  
 448-1996-4' 452-2001-1' '461-1997-2' '432-2003-1' '402-1998-3' '402-1995-2' 460-1995-1'  
 434-2000-1' '435-1995-1' '435-1996-1' 394-1996-1' '435-2000-3' '432-2000-1' 402-1995-1'  
 448-2000-2' 430-2001-1' '402-1997-1' '432-2002-1' '435-1999-2' '442-1998-2' 425-2000-1'  
 441-1995-4' 410-1995-2' '300-1995-1' '435-2001-1' '448-2000-4' '402-2002-2' 432-1996-1'  
 441-1998-1' '431-1995-1' '440-1997-2' '402-1995-1' '430-2002-1' '448-1998-3' 410-2000-1'  
 461-2003-1' 461-1997-1' '402-2001-1' 398-1996-3' '402-1995-3' '435-2001-4' '400-2000-3'

breeding value estimation from 23rd September 2005

## Analysis of breeding value estimation from 23rd September 2005

### 1 Maximum change in breeding values

#### 1.1 Breeding value estimation tbw1

##### 1.1.1 Trait: rmlf

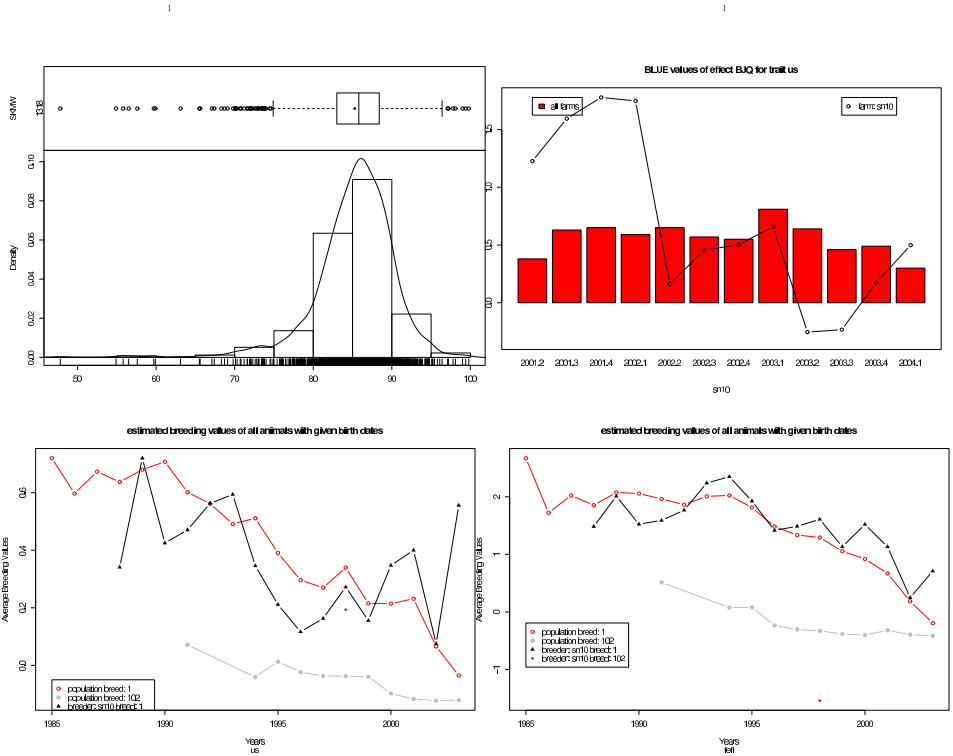
animal	BLUP		diff	EL	ancestors	pat. half sib		mat. half sib	
	new	old				new	old	new	old
19350154	14.22	2.75	11.25	-	0 0 0	0 0	0 0	0 0	
19355739	13.86	2.53	11.23	-	0 0 0	0 0	1 1 2.95	1 1 62.03	
19350076	14.44	3.69	10.84	-	0 0 0	0 0	0 0	0 0	
1935560817	13.82	3.25	10.57	-	0 0 0	0 0	0 0	0 0	
19350512	13.85	3.69	10.26	-	0 0 0	11 60.88	10 59.25	0 0	
1929007510	10.22	.65	9.57	-	0 0 0	0 0	1 1 64.49	0 0	
193114306	13.27	3.59	9.28	-	0 0 0	0 0	0 0	0 0	
1929022433	10.21	1.33	8.49	-	0 0 0	14 69.23	8 59.29	3 63.28	
1930001	12.25	3.09	9.18	-	0 0 0	11 60.88	10 59.25	1 53.25	
193572922	12.29	3.24	9.07	-	0 0 0	11 60.88	10 59.25	0 0	
19350564	-7.67	-6.1	-1.25	-	0 0 0	1 35.29	1 35.29	0 0	
1935003	-3.68	2.93	-6.58	-	0 0 0	9 59.05	8 59.08	2 63.28	
19350471	-3.02	2.34	-5.35	-	0 0 0	0 0	0 0	0 0	
193500910	-1.81	3.03	-4.22	-	0 0 0	0 0	0 0	0 0	
19350066	-1.24	3.09	-4.33	-	0 0 0	11 60.88	10 59.25	1 53.25	
193572923	-0.51	3.24	-3.73	-	0 0 0	11 60.88	10 59.25	0 0	
19350453	-0.3	2.43	-2.26	-	0 0 0	0 0	0 0	0 0	
19355338	-0.3	2.44	-2.57	-	0 0 0	0 0	0 0	1 62.03	
19350519	.48	3.52	-3.48	-	0 0 0	11 60.88	10 59.25	0 0	
19353071	.25	3.52	-2.97	-	0 0 0	0 0	0 0	0 0	

not entry represent the count of observations | average

##### 1.1.2 Trait: RZ

animal	BLUP		diff	EL	ancestors	pat. half sib		mat. half sib	
	new	old				new	old	new	old
1935023378	16.30	10.62	24.68	-	0 0 0	10 718.20	12 618.08	1 156.20	
29430072	-12.26	-6.58	-22.82	-	0 0 0	11 574.23	95 573.08	0 0	
2923003	-18.20	-8.20	-18.98	-	0 0 0	0 0	0 0	0 0	
19350423	20.12	10.63	18.49	-	1 611.00	2 1619.00	1 1057.00	1 701.00	
2935030	-2.54	-9.28	-6.58	-	0 0 0	4 569.00	2 470.00	8 586.38	
19350682	24.93	8.17	16.74	-	0 0 0	59 1630.02	42 633.33	8 822.20	
19290075	18.22	11.23	16.29	-	0 0 0	2 701.00	0 0	34 685.20	
193572918	27.65	21.68	16.21	-	0 0 0	115 617.71	59 612.29	4 638.75	
19290081	22.84	6.64	16.20	-	0 0 0	14 540.29	11 508.91	7 547.27	
193500881	16.13	10.29	15.82	-	0 0 0	40 1600.02	30 1611.30	8 638.20	
19291615	-16.10	-2.85	-19.25	-	0 0 0	37 1688.02	29 1615.26	7 577.29	
1935350	19.96	7.15	12.81	-	0 0 0	0 0	0 0	0 0	
29291828	-17.84	13.06	-11.80	-	0 0 0	18 594.29	12 620.75	6 592.33	
29291010	-19.22	-6.51	-21.76	-	0 0 0	15 568.20	11 509.30	29 616.71	
19290510	11.28	7.62	-11.36	-	0 0 0	0 0	0 0	0 0	
2929285	16.28	10.56	-11.29	-	0 0 0	0 0	0 0	0 0	
19290810	18.15	7.63	-11.28	-	0 0 0	0 0	0 0	0 0	
19290649	10.76	7.17	-10.77	-	0 0 0	0 0	0 0	0 0	
193501035	-21.44	-2.28	-20.77	-	0 0 0	53 1686.08	34 1611.23	29 609.00	
1935037510	-21.27	7.13	-28.48	-	0 0 0	53 1686.08	34 1611.23	3 566.67	

not entry represent the count of observations | average



## References

- [1] Cullis-B.R. Wellham S.J. Thompson R. Gilmour, A.R. *ASReml Reference Manual 2nd edition, Release 1.0 NSW Agriculture Biometrical Bulletin3, NSW Agriculture, Locked Bag, Orange, NSW 2800, Australia*, 2002. ISBN 0-7347 1078 X, ISSN 1038-1201.
- [2] E. Groeneveld, M. Kovač, and T. Wang. PEST, a general purpose BLUP package for multivariate prediction and estimation. In *4th World Congress on genetics applied to livestock production, Edinburgh*, number XIII, pages 488–491, 1990.
- [3] Eildert Groeneveld. An adaptable platform independent information system in animal agriculture: Framework and generic database structure. *Livestock Production Science*, 87:1–12, 2004.
- [4] Eildert Groeneveld, Ralf Fischer, and Špela Malovrh. "MoLabIS" a labs backbone for storing, managing and evaluating molecular genetics data. In *Bioinformatics Open Source Conference, Brisbane/Australien*, 27.-28. Juni 2003.
- [5] Milena Kovač, Eildert Groeneveld, and Luis Alberto García-Cortés. VCE-5, a package for the estimation of dispersion parameters. In *Proceedings of the 7th World Congress on Genetics applied to Livestock Production (WCGALP), Montpellier, France*, volume 33, pages 741–742, 19.-23. August 2002. ISBN 2-7380-1052-0.
- [6] A. Neumaier and E. Groeneveld. Restricted Maximum Likelihood Estimation of Covariances in Sparse Linear Models . *Genet. Sel. Evol.*, 1(30):3–26, 1998.
- [7] Eric S. Raymond. *The Cathedral & the Bazaar*. O'Reilly UK, 2001. ISBN 0596001088.