Aspects of Data Preparation in the Joint German & Austrian Simmental Evaluation

Reiner Emmerling
Bavarian State Research Centre for Agriculture, Institute of Animal Breeding, Grub, Germany

Introduction

Since 2002 genetic evaluations for all traits in Simmental and Brown Swiss are performed across country borders of Germany and Austria. The joint population is called DEA in the breed specific publications and documentation of Interbull Centre. Currently DEA participates in international evaluations for production traits and udder health in both breeds. Additionally, Brown Swiss data for conformation and longevity traits is included in the Interbull routine evaluation.

Up to now the fusion of the both subpopulations has been restricted to the genetic evaluations. The collection and management of raw data is done in separate computing centres of milk recording organisations. Data for both breeds are available from four data centres, three in Germany (Grub, Stuttgart and Verden) and one in Austria (Vienna). The publishing of breeding values and the responsibility for the breeding programs is still situated separately in the political divided regions (country borders) and breeding organisations in Germany and Austria.

The work load of genetic evaluations is divided between the evaluation centres in Grub (Bavaria), Stuttgart (Baden-Württemberg) and Vienna (Austria). The Bavarian State Research Centre in Grub is responsible for the milk production traits (Emmerling *et al.,* 2002), milkability (Sprengel *et al.,* 2001), somatic cell count and conformation, whereas EBZI in Stuttgart is doing the beef evaluation (Schild *et al.,* 2003) and the ZAR in Vienna is responsible for functional traits, comprising longevity (Fürst *et al.,* 2002a), fertility (Fürst *et al.,* 2002b), calving ease and stillbirth (Fürst *et al.,* 2003).

The separate databases in the participating regions increase the data management work when data has to be merged and checked for national and international genetic evaluations. As an example, the preparation and verification of national data, before it is passed to Interbull, is described in this paper for the milk production traits.

General aspects

The data preparation and checking cannot be restricted to the final results, the estimated national breeding values. The time consuming and computationally demanding evaluation work on the one hand and the strict time schedules of national and international genetic evaluations on the other makes it indispensable to discover data problems as early as possible. This should enable the evaluation centre to localize the sources of data problems in time and, if necessary, should enable the data centres to deliver new or updated raw data.

The DEA system consists of three main preparation and verification steps: establishing the joint pedigree file, the creation of joint data files for genetic evaluation and post processing of estimated breeding values.

Joint pedigree file

One main issue of international evaluations is linking of animals between different (national) subpopulations. In order to establish these links a unique identification system has turned out to be very advantageous. Fortunately, the paticpating countries in the DEA system had worked with (or at least had stored) the 15-digit life time ear tag for each animal already in their databases, before the DEA collaboration started in the year 2000. This 15-digit identification comprises the 12 digit animal identity code and 3 digits for the numerical country code of origin (ICAR, 2005).

Up to the year 2002 the herd book based German identification system was used for bulls in the international SIM and BSW evaluation. In 2002 the identifications were reconverted to the lifetime ear tag codes. These unique identifications are also used in the whole data preparation process. This has considerably decreased the amount of data preparation work and leads to a significantly lower risk of missing links between the subpopulations in Bavaria, Baden-Württemberg and Austria. The responsibility for the correctness of lifetime ear tag animal identifications is in the responsibility of the data centres, which have all the relevant information about the registered animals.

However, the habit of converting foreign identifications to national ones in the eighties and nineties has led to some pseudo identifications in the data bases. Participants of the joint evaluation tried to identify and recode these cases to the lifetime ear tag. Additionally, milk recording centres have recently started to develop new data storage solutions jointly across country borders. This leads to an intensive exchange of information between these institutions and therefore single cases of recoded identifications come up from time to time. As a consequence the pedigree and data files are build up from scratch in each evaluation, four times a year. The current joint pedigree file of DEA Simmental contains all male and female records for the different evaluations and comprises currently 27.7 million records.

The merging of pedigree records of different origins includes a stepwise selection strategy, if one record is delivered from more than one source. This strategy consists of a decision tree, where the decision is dependent on the origin of the animal, the origin of the record and the completeness of the record.

All genetic evaluations in the DEA system are based on a single pedigree file, which is prepared in one place. Before the joint pedigree file is distributed to the three evaluation units, the file must be checked very thoroughly. It is compared to the previous joint pedigree file. The comparison comprises statistics of missing and new sires by birth year, changes in pedigree, birth year and name of the bulls. Each record is tagged with its data centre of origin and from which data centres copies of the record were delivered. This simplifies the identification of sources when problems with single records come up. If irregularities are identified, national data centres are contacted.

Joint data files

The second step is the joining of data files delivered by the four contributing data centres and preparation of input data for the evaluation. The first task in this step is the controlled merging of data with following checks of the consistency and plausibility of the data. This is done for the separate traits by the responsible evaluation centres.

Controlled merging describes the part where data is stepwise combined under application of precisely defined rules for the case where a record is delivered from more than one source. The plausibility checks comprise the control of minimum and maximum requirements for a bundle of different variables in the data files. The monitoring of *excluded records* jointly with the information from which origin each single record came from is an important step. The comparison of statistics from previous and current evaluations gives an overview over changes in the supplied data. If unexpected changes are discovered in these statistics, thorough analyses of delivered data are initiated.

Of course, besides the monitoring of excluded records the *prepared data* has to be analysed very thoroughly. For these data simple statistics are calculated for key variables over various subgroups of data. As an example, the frequency and mean of yield observations over time and subgroups of defined fixed effects are calculated. Automatically generated graphical diagrams and tables give a brief overview over prepared data. Graphical overviews are suitable for frequencies over year-month of production or calving within regions and lactations. Furthermore, in milk production traits plotting lactation curves for subgroups of cows are recommended.

Beside these monitoring tasks individual statistics about incoming information for each animal are calculated. These statistics are partly used for publication, like number of test days, lactations and herds. Additionally, they are very valuable for the verification of observed changes in breeding values.

Breeding value estimation

The key application of the DEA evaluation system is the BLUP solver MiX99 (Lidauer et al. 1999). The biological yield traits milk, fat and protein are calculated univariately in multiple lactation random regression models (Emmerling et al. 2002) with accounting of heterogeneous variances (Lidauer *et al.,* 2002). The mixed model system for Simmental consists currently of 144 million equations and the runtime for one trait is around 68 hours on a 6-way IBM-6F1 computer with 600-MHz processors. Some runtime performance indicators are collected during the evaluation in order to compare the runtime behaviour over consecutive evaluations. These variables comprise the requirements of technical resources and the solving behaviour. Erroneous data, incorrect classification of fixed effects or genetic groups would lead to significant changes in these runtime indicators.

Besides the breeding values, some by-products of the evaluation are produced and stored in a local data base for later verification of breeding values. These by-products comprise statistics of individual daughter deviations (Lidauer *et al.,* 2005), residuals and correction factors for heterogeneous variances within different time and region classified subgroups of data.

Post processing of data

The verification of results from the evaluation is starting right after first evaluation run is done. Also for this stage, it is important to identify problems as early as possible. The time schedule for evaluation covers a time buffer for replication of single evaluation runs. The length of this time buffer is a compromise between safety issues and the aim of including the most recent yield data in the genetic evaluation.

The verification of national results consists in the first step of the general verification of national bull and cow breeding values. In the second step a closer analyses of individual animals is done.

Step one comprises the comparison of estimated breeding values from the current and the previous evaluation. Correlations, means and variation of results and changes grouped by subgroups of data (e.g. birth year, origin, gender) are calculated. These outputs help to identify groups of animals or individual animals with unexpected changes. A useful tool for analysing bull ebv's is the 'verify' program from Interbull, that covers verification procedures described by Klei *et al.* (2002). The output delivers an overview over changes between two consecutive evaluations. The outputs of these first step analyses directly hint to questionable bull groups and individual bulls with irregularities.

These cases are thoroughly investigated together with the output statistics that already had been produced in the data preparation of the DEA system. Almost all questionable cases can be explained with the information already collected in the preparation programs.

Besides the 'verify' program some closer verifications of individual national results are routinely performed. Single bulls with deviations of more than one fourth of a genetic standard deviation are selected for a closer check. In first analyses it is analysed how the input information (pedigree, yield data), the reliability and persistency breeding value have changed for the single animals. If these analyses do not come up with coherent explanations, the individual animals have to be analysed more in detail.

For this reason an internal data base with results and by-products from several past breeding value estimations is used. Internal reports summarize the information and compare key figures over consecutive evaluations. These key figures comprise phenotypic data and individual daughter deviations summarized for different lactation stages of a bulls daughters or the cows own performance, respectively. This information helps to understand, which data is causing the changes in estimated breeding values.

Conclusion

The strategy of data preparation and checking in the DEA system is to identify problems as soon as possible in the process of preparing data for the evaluation. Of course, the quality of data is dependent on the internal checks within the data processing centres. For these we require high standards and correctness, which leads to a low probability of data problems in the breeding value estimation. Nevertheless, in some cases data problems appear. A thorough checking of the supplied data is mandatory to identify single identification errors and erroneous data.

Within a joint evaluation, where data is delivered from more than one data source, a uniquely defined identification system of animals turned out to be very useful. It helps to avoid problems that arise with renumbering of identifications in the evaluation unit and lowers the risk of missing links between the national subpopulations.

In order to explain changes in breeding values it turned out to be efficient, to calculate various statistics for individual animals or progeny groups already in the data preparation procedures. These statistics can save a significant amount of time in the explanation and verification of individual changes in estimated breeding values.

References

Emmerling, R., Lidauer, M. & Mäntysaari, E.A. 2002. Multiple lactation random regression test-day model for Simmental and Brown Swiss in Germany and Austria. *Interbull Bulletin 29,* 111-117.

Fürst, C. & Egger-Danner, C. 2002a. Joint genetic evaluation for functional longevity in Austria and Germany. *7th World Congress on Genetics Applied to Livestock Production.* CD-ROM communication, n°01-16.

Fürst, C. & Egger-Danner, C. 2002b. Joint genetic evaluation for fertility in Austria and Germany. *Interbull Bulletin 29,* 73-76.

Fürst, C. & Egger-Danner, C. 2003. Multivariate genetic evaluation for calving ease and stillbirth in Austria and Germany. *Interbull Bulletin 31,* 47-51.

ICAR 2005. International agreement of recording practices. http://www.icar.org/docs/Rules%20and%20regulations/Guidelines/Guidelines_2005_final_low_resolution.pdf.

Klei, B., Mark, T., Fikse, F. & Lawlor, T. 2002. A method for verifying genetic evaluation results. *Interbull Bulletin 29,* 178-182.

Lidauer, M., Emmerling, R. & Mäntysaari, E.A. 2002. Accounting for heterogeneous variance in a test-day model for joint genetic evaluation of Austrian and German Simmental cattle. *7th World Congress on Genetics Applied to Livestock Production.* CD-ROM communication, n°20-09.

Lidauer, M., Mäntysaari, E.A., Pedersen, J. & Strandèn, I. 2005. Model validation using individual daughter deviations –statistical power. *Interbull Bulletin 33,* 195-199.

Lidauer, M., Strandèn, I., Mäntysaari, E.A., Pösö, J. & Kettunen, A. 1999. Solving large test-day models by iteration on data and preconditioned conjugate gradient. *J. Dairy Sci. 82,* 2788-2796.

Schild, H.J., Niebel, E. & Götz, K.-U. 2003. Across country genetic evaluation of beef traits in Middle European dual purpose breeds. *Interbull Bulletin 31,* 158-162.

Sprengel, D., Dodenhoff, J., Götz, K-U., Duda, J. & Dempfle, L. 2001. International genetic evaluation for milkability. *Interbull Bulletin 27,* 35-40.