Proc. Assoc. Advmt. Anim. Breed. Genet. 20:519-522

HANDLING A SUBSET OF A LARGE DAIRY INDUSTRY DATASET FOR QUANTITATIVE GENETIC ANALYSES OF EXTENDED LACTATION TRAITS IN AUSTRALIAN DAIRY CATTLE

M. Abdelsayed[1], H.W. Raadsma[1], and P.C. Thomson[1]

[1] ReproGen Animal Bioscience Group, Faculty of Veterinary Science,
The University of Sydney, 425 Werombi Road, Camden NSW 2570, Australia

SUMMARY

Handling large-scale industry data is challenging both in the preliminary screening of data for validation, and also in the subsequent statistical analyses to obtain estimates of fixed effects, genetic parameters and estimated breeding values (EBVs). This paper summarises how a small subset from large industry-derived dairy data records have been explored and edited to be ready for genetic analysis prior to the analysis of a much larger data set, with a focus on lactation persistence and extended lactation traits in Australian dairy cattle. To cope with the large data volume (>158 million test-day records) test-day data were randomly divided into data-subsets, each of which could be effectively managed through a semi-automated quality control procedure for removal of extraneous outlier records. The methods as applied to one such data subset are reported here. The goals of this research were an investigation of extended lactation, but it was apparent that there were many instances of the calving date not being recorded, and hence initially being flagged as an extended lactation. Methods were derived for detecting potential "double lactations", based on comparing single- and double-lactation curve models. These techniques are illustrated using test-day records from a sample of ~1 million cows recorded between 1985 and 2010 obtained from the Australian Dairy Herd Improvement Scheme (ADHIS).

INTRODUCTION

In many areas of scientific study, and in society more generally, there has been a huge growth in the amount of data collected, the so-called "big data" phenomenon (Howe *et al*. 2008), particularly evident in the life sciences. In livestock industries, automatic recording of data has resulted in the availability of extraordinarily large industry data sets. They offer the promise of providing measures of new complex traits, as well as extremely accurate estimates of genetic parameters and breeding value predictions. However to attain these goals, some substantial difficulties need to be overcome. Firstly, automated methods of data cleaning need to be developed. Secondly, computationally efficient methods of data analysis need to be developed, as routinely used statistical methods may not "scale up" to deal with the massively increased data volume (Jacobs 2009). Also, before conducting any comprehensive analysis it is essential to explore the data to understand the overall trends, but more importantly the type of errors that can be encountered. This paper describes some approaches developed using a subset from large dairy industry data records for the analysis of test-day records from Australian dairy cattle milk yield data. A subset was used initially to trial different methods, for ease of handling and to obtain set steps which then can be applied automatically on the larger data file.

MATERIALS AND METHODS

For this study, dairy industry data were obtained from the Australian Dairy Herd Improvement Scheme (ADHIS) consisting of ~158 million test day records ranging from 1985 up to 2010. To handle the large volume of data, the dataset was randomly split into eight subsets, each subset consisting of ~1 million cows with ~20 million test day records, giving eight separate subsets. Splitting the data in this way ensured that all test-day records from the one cow were kept in the

same subset, so that all lactations from the one cow were in the same data file. Hence, the editing and analysis were then conducted on each of the eight data subsets separately.

For this paper and preliminary analysis, a randomly selected sample of 10,000 cows' data was selected; these were further filtered to only include Holstein Friesians (6,018 cows, 29,882 lactations). Analysis of this small data set would allow decisions of modelling techniques and automated data screening to be developed, and then applied to each of the eight large data subsets.

Graphical and numerical summaries from the sample data set revealed the necessity of log-transformation of some of these traits (e.g. SCC and lactose). Extreme outliers were filtered out (on transformed data where necessary) using a criterion linked to the number of records, specifically if it is more than k standard deviations away from the trait mean, where $k = |\Phi^{-1}(1/n)|$, n is the number of data records, and $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal distribution. For example for n = 30,287 test-day records as in the current example, k = 3.99, so four standard deviations from the mean would be an appropriate cut-off.

The ADHIS database records the calving date, and from this, days in milk (DIM) can be calculated for a particular recorded test-day. As a primary reason for undertaking this study was to explore variation in lactation curve shape, particularly those related to lactation persistence and extended lactation (Abdelsayed *et al*. 2013), cow-lactation data sets were removed from the analysis when fewer than three test-days records were available. Also, any test-day record beyond 750 DIM was excluded, as was any test-day record at birth (0 DIM).

An exploratory plot of the average milk yield over days in milk revealed a number of apparently extended lactations were in fact new lactations for which the calving date had not been recorded. This was important to ascertain, as this could substantially affect genetic estimates of lactation persistence and extended lactation traits. To identify this, a hypothesis testing method was used fitting one- and two-lactation Wood (1967) models to the test-day data. The Wood model as implemented here has the form $W(t; k, b, c) = \exp(k + b\log_e t - ct)$, where t is DIM, $W(\cdot)$ is the model-based mean yield, and k, b, and c are parameters that describe the lactation curve. Hence an observed milk yield for cow-lactation i on DIM j can be modelled as $y_{ij} = W(t_{ij}; k_i, b_i, c_i) + \varepsilon_{ij}$, where ε_{ij} represents a random error associated with the test-day record.

Single vs double lactation screening was conducted as follows, for each cow-lactation:

1. Fit a single-lactation Wood model to cow-lactation data set i, modelled as a single lactation: $y_{ij} = W(t_{ij}; k_i, b_i, c_i) + \varepsilon_{ij}$, and save the Residual SS (=RSS_0, reduced model).
2. Fit a two-component Wood model for lactation data set i, modelled as a double lactation, assuming the second lactation commenced 365 days after the first:

$$y_{ij} = \begin{cases} W(t_{ij}; k_i, b_i, c_i) + \varepsilon_{ij} & 0 < t_{ij} < 365 \\ W(t_{ij} - 365; k_i, b_i, c_i) + \varepsilon_{ij} & t_{ij} \geq 365 \end{cases}$$

   and save the Residual SS (=RSS_1, full model). This model assumes that the second curve has the same shape as the first, just separated by 12 months.
3. Calculate an F statistic and P-value based on comparing RSS_0 and RSS_1.
4. If $P < 0.1$, that could be sufficient evidence for a second lactation.

However, the specific threshold P-value needs to be evaluated to achieve a balance of false positives / false negatives. To assist this process, a false discovery rate approach can be used, and the q-value method of Storey and Tibshirani (2003) has been adopted here.

In the present study, the fitted Wood model was then used to summarise various characteristics of the lactation curve shape, such as persistence and extended lactation (Abdelsayed *et al*. 2013), and these are derived from estimates of k, b and c from each cow-lactation. Consequently cow-lactations with extreme or infeasible estimates of these were excluded, based either on the outlier method mentioned above, or in the case of c, excluding any cow-lactation with negative estimates (which would imply ever-increasing yield over a lactation).

RESULTS AND DISCUSSION

Figure 1 shows the average milk yield for all test-day records observed on each calculated DIM. The form of the curve is as expected for a lactation curve up to approximately day 270. Beyond that however, the smaller second peak is evidence of a number of cows not having their calving dates recorded, resulting in a false 'extended lactation'. This is also supported by the increase in variability of these means, not entirely explained by sampling fluctuations of fewer lactations recorded at the particular DIM. The wide variation in mean yield beyond 600 days however is a reflection of the fewer records extending that far.

For each cow-lactation that extended beyond 365 days, single- and double-lactation models were fitted to the test-day data, and the *P*-value calculated as a means of assessing if the fit of a double-lactation model was a better fit than a single lactation model. Lactations with $P < 0.1$ were considered for possibility of being a double lactation: two sample plots are shown in Figure 2, one being almost certainly a double lactation ($P \approx 0$), the other probably better considered a single lactation (with $P = 0.084$). Sample lactation curves were scrutinised, and $P = 0.06$ threshold was adopted: this corresponds to a false discovery rate of just in excess of 5% ($q = 0.054$); a distribution of *P*-values of 1,150 sample cow-lactations is shown in Figure 3.

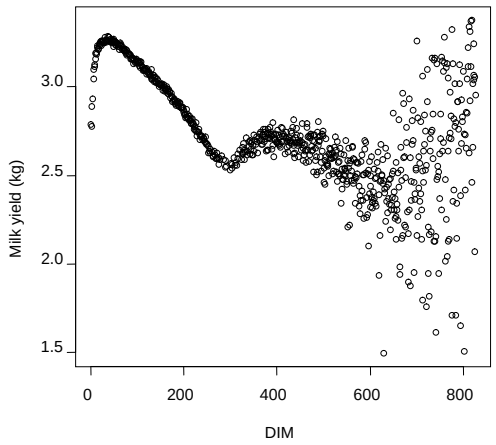

Figure 1. Mean milk yield (kg) vs days in milk (DIM). Evidence for calving date not being recorded is indicated by the second smaller peak around Day 400.

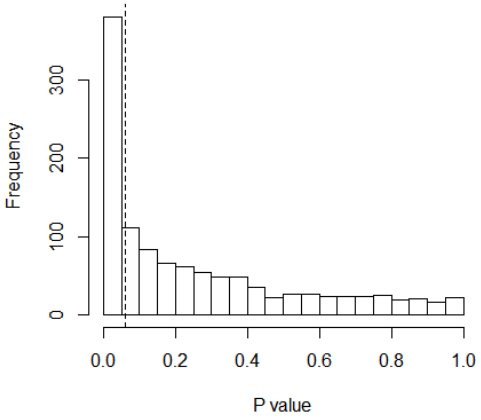

Figure 3. Histogram of *P*-values for testing double- vs single-lactations. The vertical dashed line is drawn at the adopted threshold of $P = 0.06$.

Using this method to classify double lactations, i.e. cows whose second calving was not recorded, there will undoubtedly be some false classifications. It should also be noted that for extended lactations, seasonal influences are apparent, and this has led some authors to model extended lactation with a di-phasic model, even in the absence of "double lactations" (e.g. Vargas *et al*. 2000; Grossman and Koops 2003).

As a result, after such cleaning processes, reliable estimates of the lactation curve parameters (*k*, *b*, *c*) for each cow-lactation have been obtained, and persistency and extended lactation traits can then be derived. These traits can then be used for the standard quantitative genetics analysis (Abdelsayed *et al*. 2013). However, due to computational limitations, when data volumes are large, it may not be possible to fit a large-scale linear mixed model, so it is necessary again to

randomly split the data into further subsets, perhaps using the same data subsets as used in the initial data screening. Because the allocation is random, overall genetic and fixed effect estimates can be obtained by simply averaging across those produced from each data subset, with appropriate weighting.

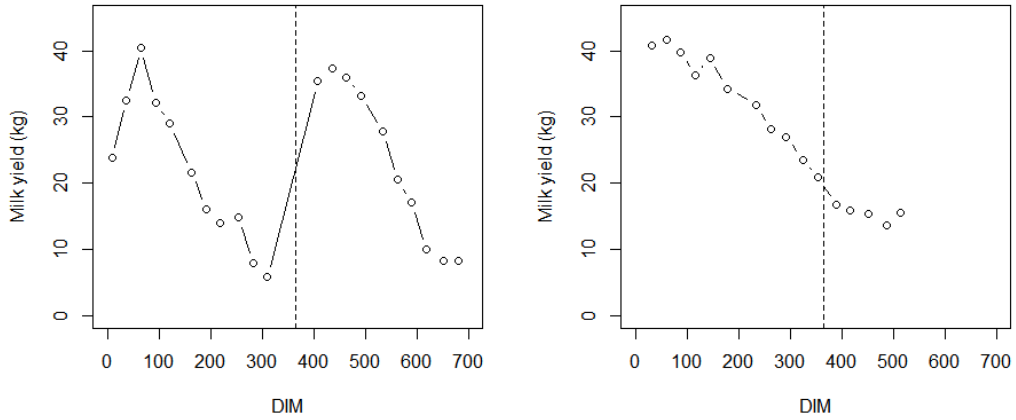

Figure 2. Test-day milk yields of two cow-lactations flagged as possible double lactations. The curve on the LHS, very clearly a double lactation, has $P = 2\times10^{-11}$, whereas the curve on the RHS has $P = 0.084$. The vertical dashed line is drawn at 365 DIM, the potential start of a second lactation.

CONCLUSIONS

Large-scale test-day datasets and other industry data are being routinely collected and in turn new computational and statistical approaches need to be developed to handle these "big data" sets. This paper has described some approaches to this, in the context of test-day records from Australian dairy cattle to assess lactation persistency and extended lactation. In particular, a method for screening for missed second lactations has been outlined, but other data screening and analysis aspects have also been considered. While a methodology has been outlined in this paper to address the problems encountered with extended lactation data, the process is not perfect, and there would be great benefit to investigate how all industry calving data information could be captured, ensuring that lactation length could be accurately evaluated.

REFERENCES

Abdelsayed M., Thomson P.C. and Raadsma H.W. (2013) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **20** (this publ).

Grossman M., and Koops W.J. (2003) *J. Dairy Sci.* **86**: 988.

Howe D., Costanzo M., Fey P., Gojobori T., Hannick L., Hide W., Hill D.P., Kania R., Schaeffer M., St Pierre S., Twigger S., White O., and Rhee S.Y. (2008) *Nature* **455**: 47.

Jacobs A. (2009) *Commun. ACM* **52**: 36.

Storey J.D., and Tibshirani R. (2003) *Proc. Natl. Acad. Sci. USA* **100**: 9440.

Wood P.D.P. (1967) *Nature* **218:** 894.