An Evaluation of Milk Recording, Somatic Cell Counts and Reproductive Performance in a Large Cohort of Dairy Herds in England and Wales

Aurélien Madouasse, DVM.

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy

December 2009

Abstract

Three years of milk recording data from 2,128 dairy herds from England and Wales were used to describe herd, lactation, and milk production characteristics ; to describe somatic cell count patterns and their link to bulk milk somatic cell count and to investigate the relationship between milk quantity and composition at the start of lactation and the calving to conception interval.

All the data collected by the National Milk records between the 1^{st} of January 2004 and the 31^{st} of December 2006 were available. A selection was operated to retain samples taken on two consecutive milkings from herds which recorded on average monthly for the complete three years. From the 19,893,093 recordings in 5,714 herds initially available, 8,211,988 recordings from 2,128 herds were selected for further analysis. This represented data for approximately 16 % of the dairy herds in activity in England and Wales in December 2006.

In Chapter 3 these data aggregated at the herd, lactation and month levels were described. Calvings and milk production followed a seasonal pattern. The number of calvings was the highest in September and the lowest in May with 80 % more calvings in September than in May. The peak was more pronounced for heifers and flattened out with successive lactations. Seventy three percent of cows calved again and the median interval between consecutive calvings was 391 days. There were variations in the quantity and composition of milk produced per month of the year. Milk production was highest in May (26.5 kg) and lowest in October (24.1 kg). Butterfat was stable, close to 4 % from October to March and reached a minimum at 3.7 % in June and July. Protein stayed between 3.2 and 3.3 % all the year. Geometric mean somatic cell count was between 177,000 and 180,000 between October and March and reached 205,000 cells/mL in July and August.

In Chapter 4 individual cow milk yield and constituents were described. Distributions, lactation curves and cumulative quantities were investigated. Between 5 and 305 days in milk, the mean milk yield, percentage of butterfat, percentage of protein, fat to protein ratio and somatic cell count (geometric mean) were 26.4 kg, 3.96 %, 3.29 %, 1.21 and 90,000 cells/mL. Lactation curves for milk yield were lower and flatter in parity one than in later parities. For parity one cows, the peak occurred around 50 days in milk and production at the peak was 27.9 kg. For later parities, the peak occurred between 38 and 41 days in milk and was between 33 and 38 kg. Between 50 and 305 days in milk, milk production decreased by 3.2 kg/100 days in parity one cows and by between 5.7 and 7.4 kg/100 days for later parities. Lactation curves for percentage of butterfat, percentage of protein and somatic cell count followed an inverse shape to the one

of milk yield. The fat to protein ratio increased from week 2 to 4 in lactation and decreased between week 4 and week 12 of lactation. Between 5 and 305 days in milk, first lactation cows produced 7,358 kg of milk, 284 kg of fat and 235 kg of protein ; cows in later lactation produced 8,483 kg of milk, 327 kg of fat and 272 kg of protein. High yielding herds were characterised by higher peaks, lower persistency, and slightly lower concentration in butterfat and protein.

In Chapter 5 patterns of somatic cell counts aggregated at the herd-year level were described and their contribution to an estimated bulk milk somatic cell count quantified using linear mixed models. Patterns were defined using a threshold of 200,000 cells/mL to categorise cows as having a low or a high somatic cell count and combined over two consecutive milk recordings. Cows of parity one and greater than one and cows in their first month of lactation and later in lactation were described separately. Predictions made by the model were tested against observed data and the model used to predict a different dataset. The selected model predicted the data accurately. Cows staying above 200,000 cells/mL for 2 consecutive milk recordings were the main contributors to bulk milk somatic cell count.

In Chapter 6, the model developed in Chapter 5 was tested at the test-day level and the probability of transition between below and above 200,000 cells/mL, dry period and culling were modelled on 3 datasets. Seven test-days were randomly sampled from a random sample of 100 herds. The first six test-days were used for parameter estimation. The seventh test-days as well as one test-day per herd in 100 randomly selected herds were used for validation. Overall, the model using somatic cell count categories predicted bulk milk somatic cell count well and most coefficients were close to the ones estimated in Chapter 5. The state transition model described the transitions well. The probability of moving or staying above 200,000 cells/mL increased with stage of lactation. The probability of being above 200,000 cells/mL was higher for cows above 200,000 cells/mL on the previous test-day and for cows of parity greater than one. However, states predicted for each individual cow were not useful in predicting bulk milk somatic cell count.

In Chapter 7, the milk quantity and composition on the first two test-days of lactation were used to model and predict the calving to conception interval. Multilevel discrete-time survival models were used. There was an association between milk quantity and composition at the start of lactation and the probability of conception before 145 days in milk. This probability increased with lower milk production on the second test-day, higher percentage of protein on the second test-day and higher percentage of lactose on the first test-day. Positive associations were of a limited magnitude but also significant with the percentage of protein on the first test-day, the percentage of butterfat on the first test-day and somatic cell count on both test-days. Characteristics of milk production on the second test-day of lactation were of more importance, probably because they were related to the production at the peak. While there

was a good agreement between observed and predicted probabilities of conception at the cow-lactation level, predicted probabilities of conception aggregated at the herd level were not useful in ranking individual herds.

Acknowledgements

These three years spent in the UK have been a wonderful scientific as well as human experience. This will, I hope, be extremely valuable in the career I intend to pursue as a researcher in Veterinary Epidemiology.

First and foremost, I would like to thank Prof. Martin Green for guiding me through this project, for open-mindedness and for helping me to communicate my research. The help of Prof. William Browne was also extremely important especially, but not exclusively, in setting the more complex statistical models. Thanks to D^r Jonathan Huxley for advice on clinical aspects of the project and to D^r Andrew Bradley for useful inputs.

I am grateful to the University of Nottingham for sponsoring me. This research was jointly funded by the School of Veterinary Medicine and Science and the School of Mathematical Science.

Thanks to John Wall, Henry Richardson, Ben Bartlett and Bridget Rymer from the National Milk Records for technical help on milk recording.

Thanks to Peter Watkins and Lars Nygaard from Foss Electric for technical explanations on the functioning of the Milko-Scan 605 and the Fossomatic 360.

A substantial part of my work was carried out using the statistical package R in which I had no knowledge when starting this project. The useful and gratuitous help of the members of the French-speaking forum *Groupe des utilisateurs du logiciel R* is acknowledged.

Thanks to my family, friends from Brittany, friends from the National Veterinary School of Nantes, fellow postgraduate students and staff at the Nottingham School of Veterinary Medicine and Science, people I have learn to know in Britain and elsewhere. I can't name you all, but you know who you are!

To Aude and Quentin, for constant support and for following me around the world.

"All right," said Deep Thought. "The Answer to the Great Question ..."
"Of Life, the Universe and Everything ..."
"Is ... "
"Forty-two," said Deep Thought, with infinite majesty and calm. ...
"Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years work?"
"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

Prediction is very difficult, especially about the future.

Niels Bohr

Contents

List of Figures

List of Tables

List of Abbreviations

BMSCC	Bulk Milk Somatic Cell Count
g	Gram
hg	Hectogram
IMI	Intramammary Infection
kg	Kilogram
ln	Natural logarithm
MCMC	Markov Chain Monte Carlo
mL	Milliliter
MVN	Multivariate Normal Distribution
NMR	National Milk Records
SCC	Somatic Cell Count

Chapter 1

General Introduction

1.1 Background to milk production in dairy cows

Cow's milk is produced by four anatomically independent mammary glands. At the cellular level, milk is synthesised by lactocytes organised in globular structures called acini. Once secreted, milk is collected in acini, ducts and a cistern until suckling or milking occurs. Milk is the exclusive source of water, carbohydrates, fat, protein and minerals for the young calf during its first weeks of life. Hence, the amount of each constituent has to be balanced in order for the calf to survive. For farmers, variations in milk quantity and composition are of interest on many grounds. Payment is made on the basis of milk volume and fat content which, in Europe, are constrained on an annual basis by quotas. A supplement is paid for protein content and a penalty when somatic cell count is above a threshold (typically 200,000 cells/mL) so that farmers have a direct financial incentive in monitoring these constituents. Cattle breeding programs using cow information data have been developed throughout the world in order to improve profitability of dairy herds by enhancing individual cow milk production. Thus, there is a strong motivation for the dairy industry as a whole to obtain reliable data on milk production on a large scale. Milk recording consists of the regular, usually monthly, recording of data from all the lactating cows milked in a dairy herd. Results of these tests are used by farmers, farm advisors and breeding companies. This thesis will focus on variations in milk quantity, fat, protein and somatic cell count and their link to bulk milk somatic cell count, and reproduction.

1.2 Milk recording

Milk recording is used worldwide to assess individual cow performance. It was developed at the start of the 20^{th} century. The creation of the *American Dairy Association* in 1906 and the early history of milk recording in the United States are described by Fraser (1933). Information

on the history of milk recording in the UK can be found on the National Milk Records website (NMR, 2009). Before 1943, milk recording was carried out by different milk recording societies which were then amalgamated into in the National Milk Records by the Milk Marketing Board. By 1953, the NMR recorded 25 % of cows and 16.2 % of herds in England and Wales. The measure of butterfat and then protein were undertaken. The measure of somatic cell count was introduced in 1990 and within 2 years, 65 % of NMR samples were tested. Today all samples are tested for SCC. The NMR is now a private company, and recently, other competing milk recording companies have appeared in the UK. The information produced by milk recording is of importance to dairy farmers in the monitoring of cow performance and hence in decision making. Another major use of these data is the evaluation of bulls' breeding values. Finally, these data are used in epidemiological studies. The relatively standard way in which milk recording is carried out worldwide makes the data comparable between countries to some extent (Weigel *et al.* , 2001).

In this thesis, a large sample of milk recording data collected by the National Milk Records in England and Wales between January 2004 and December 2006 will be used. The data and data selection used to generate the dataset used throughout this thesis are described in Chapter 2.

1.3 Milk Composition

Milk quantity, butterfat content, protein content, lactose content and somatic cell count are routinely measured as a part of milk recording. Substantial variations in milk composition exist between cow breeds. For example, Channel Island breeds (Jersey, Guernsey) milk contains more fat and protein than Holstein and Friesian which are the object of this thesis. To give an overview, Cerbulis & Farrell (1975) determined the milk composition for various breeds of cows in the United States. Holstein cows' milk contained 3.07 $\pm$ 0.43 % of (true) protein, 3.73 $\pm$ 0.32 % of butterfat and 4.93 $\pm$ 0.61 % of lactose. As an example, the same figures for the Jersey breed were 4.07 $\pm$ 0.49, 5.42 $\pm$ 0.53 and 4.99 $\pm$ 0.34.

1.3.1 Lactose

Lactose is a disaccharide made of one molecule of galactose and one molecule of glucose which condense in the presence of the enzymes galactosyltransferase and α-lactalbumine (Linzell & Peaker, 1971). It is the principal milk constituent affecting milk osmotic pressure. Other constituents contributing to this pressure are K^+, Na^+ Cl^- and other sugars (Shennan & Peaker, 2000). The osmotic pressure causes an influx of water in the lumen of the alveoli determining milk volume.

1.3.2 Protein

Milk nitrogen can be divided in 3 fractions: casein, whey protein and non protein nitrogen which represent approximately 77.9, 17.2 and 4.9 % of total milk nitrogen (Cerbulis & Farrell, 1975). While some of the whey proteins such as immunoglobulins and serum albumin originate from the bloodstream, most of the milk proteins are synthesised by the mammary gland. The amino acids used for these syntheses are either derived from the bloodstream or synthesised by the gland (Akers, 2002).

1.3.3 Milk Fat

Milk lipids are organised as fat globules surrounded by a membrane called the milk fat globule membrane (**MFGM**). The MFGM prevents globule coalescence thus stabilising the emulsion. At the centre of the globule, the lipid core is mainly composed of triacylglycerols (**TAG**) synthesized in the rough endoplasmic reticulum. These TAG assemble in microdroplets which are coated with protein and polar lipids and expulsed from the cell by a budding process (Bauman *et al.* , 2006). The lipid composition of milk has been extensively reviewed by Jensen (2002). TAG represent 97.5 % of the total lipid weight (Kurtz, 1974). Fatty acids entering the composition of TAG have two different origins, namely *de novo* synthesis by the mammary gland and uptake from the bloodstream. Fatty acids present in the blood originate either from the diet or from lipolysis. They vary both in terms of length and saturation. Bovine milk contains at least 400 different fatty acids, 12 of which being present as more than 1 % of the total triglyceride (**TG**) weight. Hence milk contains several thousands of different TG, most of them only as trace (Jensen, 2002). Lactocytes synthesise fatty acids of length 4 to 14 while fatty acids of length 18 are all blood supplied. C_{16} can be both synthesised or uptaken. Ruminants utilise acetate and to a lesser extent β-hydroxybutyrate as a source of carbon for the synthesis of fatty acids as opposed to glucose in other species (Hawke & Taylor, 2002).

1.3.4 Somatic Cells

The main cell populations in milk are immune cells. Lee *et al.* (1980) reported the main cell population to be macrophages while there was no secretory epithelial cells and very few ductal epithelial cells. Sarikaya *et al.* (2006) identified lymphocytes as the most numerous cells in milk with somatic cell count of less than 12,000 cells/mL and macrophages in milk with somatic cell count between 12,000 and 100,000 cells/mL. Inflammation is initiated by the release in chemo-attractants by macrophages and epithelial cells which results in the recruitment polymorphonuclear neutrophil leukocytes (Paape *et al.* , 2002). Inflammation of the mammary gland is mostly of bacterial origin. The main factor of variation of somatic cell count is infection

(Schepers *et al.* , 1997), and somatic cell count is used as a marker of mastitis.

1.3.5 Determination of milk composition by the National Milk Records

The instruments used by the National Milk Records between 2004 and 2006 were the Milko-Scan 605 for the determination of protein, butterfat and lactose contents, and, Fossomatic 360 for somatic cell count (Foss Electric, Hillerød, Denmark). The Milko-Scan 605 uses different sets of filters to detect infrared wavelengths distinctive of the constituent measured. The analysers need to be calibrated against reference methods for each constituent (Barbano & Clark, 1989; Kaylegian *et al.* , 2006). For protein, the application was calibrated for the measurement of total nitrogen and the reference technique for the calibration was the Kjeldahl method. The conversion from nitrogen to protein content was operated using a factor of 6.38. The Milko-Scan 605 measured butterfat directly. In this case, the reference against which the equipment was calibrated was the Rose-Gottlieb method. Lactose content was measured after de-proteinisation and removal of the butterfat from the sample. The calibration standard for lactose was BS 1741 Section 7.1. With the Fossomatic 360 (McKenna, 1994; Miller *et al.* , 1986), cell number per unit of volume is measured by flow cytometry. Cells are stained with ethidium bromide, a DNA specific dye, and exposed to light of a specific wavelength as they are lined up in a flow. In response they emit light pulses of a different wavelength which are counted one at a time.

1.4 Physiological stage and milk production

1.4.1 Lactation curves

Lactation curves describing changes in milk production with stage of lactation have been extensively studied. The purposes of such studies have been, for example, to adjust observed performance for stage of lactation and season (Wilmink, 1987), to split the phenotypic variations between additive genetic, permanent environmental, herd test-day, herd and residual sources (Silvestre *et al.* , 2009) or to model the effects of disease on performance (Appuhamy *et al.* , 2009). Milk yield is known to increase from calving to 6-8 weeks postpartum and to decrease thereafter. The maximum milk production is referred to as the peak. Milk constituent concentrations follow an inverse shape, decreasing from calving until the peak in yield and increasing thereafter. There is no recent description however of lactation curves in the UK despite the fact that cows have changed greatly in terms of genetics.

Variations in individual cow milk production in England and Wales with stage of lactation are described in Chapter 4.

1.4.2 Transition between the dry period and lactation

The onset of lactation follows calving so that there is a dramatic change in cow metabolism between the dry period and early lactation (Bauman & Currie, 1980). The energy balance as well as the quantity and composition of the milk produced during the first 180 days of lactation in a Dutch dairy herd were measured by de Vries & Veerkamp (2000). On average, cows were in Negative Energy Balance (**NEB**) during the first 41.5 days of lactation and 82.5 % of cows experienced a period of NEB. The most significant change in milk composition associated with NEB was an increase in the fat percentage during the first month of lactation. The change in fat percentage between the first and second month of lactation had the highest correlation with the total energy deficit during early lactation. Grieve *et al.* (1986) found a negative correlation between energy balance (**EB**) and milk percentage of fat, a positive correlation between EB and percentage of protein and a negative correlation between EB and the fat to protein ratio (**FPR**). In their study, the fat to protein ratio was the better indicator of EB. Based on measurement of β-hydroxy-butyrate (**BHBA**) for 1,333 cows in 93 Canadian herds, Duffield *et al.* (1997) identified 14.1 % of cows as being in subclinical or clinical ketosis in early lactation. In another recent study carried out in Ontario higher serum BHBA was also associated with higher milk fat percentage and lower milk protein percentage (Duffield *et al.* , 2009).

Therefore, most cows are in NEB at the start of lactation and this has repercussion on milk composition and can lead to ketosis. Among the consequences of NEB is its negative impact on reproduction. In de Vries & Veerkamp (2000), NEB was associated with a delay in the resumption of the postpartum luteal activity. Lower reproductive performance has been associated with the genetic selection for increased milk production (Lucy, 2001; Butler & Smith, 1989). Extensive work has been carried out on this subject, but the precise relationship between energy status and reproduction is still only partly understood.

In Chapter 7, we use milk recording data to predict the interval between calving and conception.

1.5 Somatic Cell Count and Mastitis

1.5.1 Background on Mastitis

Mastitis is an inflammation of a mammary gland, mostly of bacterial origin. Depending on the degree of adaptation of the pathogen to the mammary gland, mastitis causing bacteria have classically been divided between contagious and environmental. An example of a purely contagious bacteria is *Streptococcus agalactiae* which can only survive in the mammary gland and hence is transmitted from cow to cow (Keefe, 1997). At the other end of the spectrum is *Escherichia coli* which persists in the environment and infects the cow occasionally, though

it is not generally adapted to survive in the mammary gland. This distinction has become less clear recently because studies have shown that some *Escherichia coli* persist in the udder subclinically (Dopfer *et al.* , 1999; Bradley & Green, 2001). Another distinction is made between major pathogens which elicit a strong inflammatory response and minor pathogens which elicit a milder response. In all cases, mastitis has adverse effects such as decrease in milk production, financial penalties from treatments, milk withdrawal and death or culling (Seegers *et al.* , 2003).

A survey conducted in 97 herds from England and Wales on mastitis incidence and aetiology (Bradley *et al.* , 2007) found a mean incidence of clinical mastitis of 47 cases per 100 cows per year when estimated from farm records and of 71 cases per 100 cows per year when estimated from the intervals between samples collected for bacterial analysis. This highlights the facts that despite decades of effort to control the disease, it remains an important problem in British dairy herds, and that dairy farmers might underestimate the real incidence in their herds. In this same study, the main aetiological agents for mastitis were *Streptococcus uberis* (23.5 %) and *Escherichia coli* (19.8 %), but 26.5 % of samples produced no growth. The main agent isolated from sample with a somatic cell count greater than 200,000 cells/mL were *coagulase-negative staphylococci* (15 %), *Streptococcus uberis* (14 %) and *Corynebacterium species* (10 %). *Staphylococcus aureus* and *coagulase-positive staphylococci* represented 10 % of these samples and 39 % did not yield any growth.

1.5.2 Somatic Cell Count

For decades now, somatic cell count has been used as a marker for mastitis. In the UK, it was introduced by the National Milk Records in June 1990 and within two and a half years over 65 per cent of all NMR samples were tested for cell count. Today, all samples are tested for SCC (NMR, 2009). In this respect, milk recording allows farmers to routinely check cow SCC levels to identify individuals likely to carry an infection. Following bacterial challenge, immune cells, mainly neutrophils, are recruited from the blood stream. Because infection can trigger the recruitment of millions of cells/mL, the overall distribution of SCC is right skewed. Thus, a log transformation is usually applied to study SCC (Ali & Shook, 1980) and the geometric mean is preferred to the arithmetic mean SCC.

Djabri *et al.* (2002) carried out a meta-analysis of the increase in SCC associated with the main bacteria causing mastitis. The geometric mean SCC in negative quarters, quarters harboring IMI with *Staphylococcus aureus*, *Streptococcus agalactiae*, *Streptococcus dysgalactiae*, *Streptococcus uberis*, coliforms, staphylococci other than *S. aureus* and *Corynebacterium bovis* were 68 000, 357 000, 857 000, 547 000, 1 024 000, 1 151 000, 138 000 and 105 000 cells/mL respectively.

Different bacteria also cause different patterns of SCC during lactation. Looking at the link

between clinical mastitis of particular etiologies and SCC, de Haas *et al.* (2004) observed that in the case of IMI caused by *Esherichia coli*, SCC was close to the levels in uninfected cows and returned quickly to normality after the clinical episode. This pattern was similar to that observed in cases of mastitis with culture negative samples. In cases of clinical mastitis caused by *Staphylococcus aureus*, SCC was elevated before and remained elevated after the case. The effects observed in cases of mastitis caused by *Streptococcus dysgalactiae*, *Streptococcus uberis*, and other Streptococci were similar and characterised by a continuous increase until the case and high SCC levels following it. Using the same dataset as well as data collected in the UK, Green *et al.* (2004) found that *Staphylococcus aureus* and *Streptococcus uberis* clinical cases were associated with increased standard deviation of the log SCC during lactation. *Escherichia coli* was characterised by higher coefficients of variation. Increased lactation geometric mean SCC were associated with increased risks of *Staphylococcus aureus* mastitis and a reduced risk of *Escherichia coli* mastitis.

1.5.3 Diagnosis of Mastitis: problems and application at the herd level

The diagnosis of mastitis is complicated by the lack of sensitivity of bacteriology. In de Haas *et al.* (2004), no bacterial growth occurred in 22 % of cows with clinical mastitis. The associated lactations had characteristics close to the ones with a clinical mastitis caused by *Escherichia coli*. In Bradley *et al.* (2007), twenty six percent of samples of milk from clinical mastitis did not result in bacterial growth. It can be stated that there is no gold standard for the detection of mastitis. With bacteriology used as the 'best' available test, it is hard to determine the true SCC distribution in uninfected quarters. Therefore, it is common in epidemiological studies to include SCC as well as bacteriology in the definition of subclinical cases (Dohoo & Leslie, 1991).

In practice, fixed values are used to identify cows likely to bear an infection. A threshold of 200,000 cells/mL is commonly used in the UK (Bradley & Green, 2005) as in other countries (Pantoja *et al.* , 2009) but alternative thresholds are also in use. For example, the Dutch milk recording program uses a threshold of 150,000 cells/mL for primiparous and a threshold of 250,000 cells/mL for multiparous cows (de Haas *et al.* , 2008). The resolution of mastitis problems in a herd does not require the certain identification of all cases of mastitis but the identification of the population of individuals at greatest risk of IMI. To achieve this goal, herd performance is evaluated and compared to benchmarks. No recent reference values have been provided in the UK regarding SCC values and herd performance.

In Chapter 3, patterns of SCC in England and Wales between 2004 and 2006 are explored. In Chapter 4 lactation curves for individual cow SCC are presented. In Chapter 5, two thresholds are tested for the prediction of herd bulk milk somatic cell count and the contribution of cows in the categories defined by these threshold quantifies. Finally, in Chapter 6 the probabilities

of transition of individual cows across this threshold are predicted and these predictions used to predict test-day BMSCC.

1.6 Statistical Techniques used throughout this Thesis

In this thesis, the associations between milk quantity and composition and bulk milk somatic cell counts or the interval from calving to conception are studied with statistical models (see Chapters 5, 6 and 7). Linear models and generalized linear models are used in both frequentist and Bayesian frameworks. This section is an introduction to the statistical techniques used starting with likelihood functions which are central to frequentist as well as Bayesian approaches, followed by an outline of the differences between these two approaches. Because it is increasingly encountered in veterinary literature, emphasis is put on Bayesian model building and checking. Finally, linear mixed models and generalized linear mixed models are introduced.

1.6.1 Likelihood Functions

As the name indicates, a likelihood function measures the likelihood of observing a certain outcome under a given distribution. They are pivotal in both frequentist and Bayesian analysis. In the next two paragraphs, this concept is illustrated with examples of a discrete distribution and of a continuous distribution.

Example 1: discrete outcome

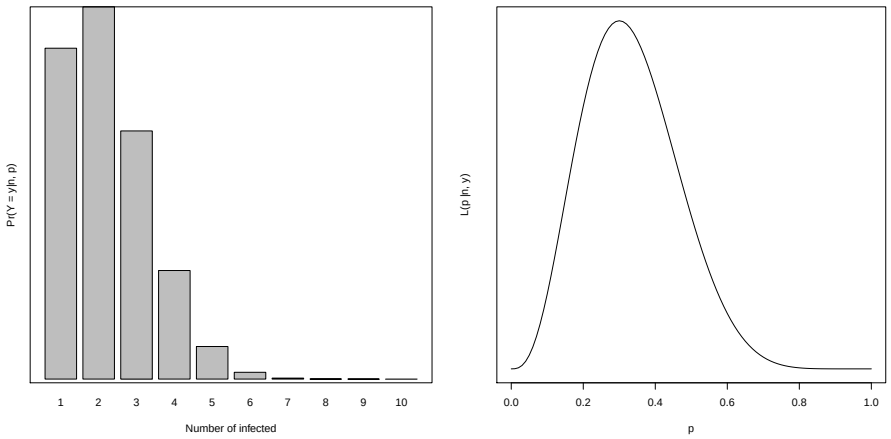
The binomial distribution models a sequence of n Bernoulli trials. In these trials there are two possible outcomes which can be labelled *Success* and *Failure*. In Epidemiology, a classical application of this distribution is the sampling of a group of individuals to determine the prevalence of a disease in a population. For a given prevalence in the general population, different samples will give various numbers of positives. These random variations around the *real* prevalence are described by the binomial probability mass function. Given the prevalence p of the disease in the population, if there are n animals sampled (*trials*), the probability of observing y individuals with the disease (*Success*) is:

$$Pr(Y = y|n, p) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \tag{1.1}$$

The problem of statistical inference is the inverse one. In this case, y is known, and the aim is to get an estimate of p, that is find the value of p for which $Pr(p|n, y)$ is maximum (Myung, 2003). By removing the parts of Equation 1.1, which do not depend on p the likelihood function

Figure 1.1: Binomial Distribution $B(10, 0.2)$

(a) Probability mass function. *(b) Likelihood function.*



can be written as:

$$L(y|p, n) \propto p^y(1-p)^{n-y} \quad (0 \leq p \leq 1) \tag{1.2}$$

We take the example of a disease which has a prevalence of 20 % in a population. Ten individuals are sampled. The individuals with the disease are numbered 1 and the individuals who are disease-free 0. The vector of observed values is $z = 0001100100$. The likelihood function is a function of p:

$$L(y = 3|p, n = 10) = p^3(1-p)^7 \quad (0 \leq p \leq 1) \tag{1.3}$$

Figure 1.1a shows the probability mass function for $Pr(Y = y|n = 10, p = 0.2)$. Making p vary in Equation 1.3 gives the curve in Figure 1.1b. In this case, the likelihood function peaks at 0.3 because 3 individuals out of ten are positive but the distribution of possible values is large because of the small sample size.

Example 2: continuous outcome

The same idea can be applied to continuous data. The probability density function for the normal distribution can be written as:

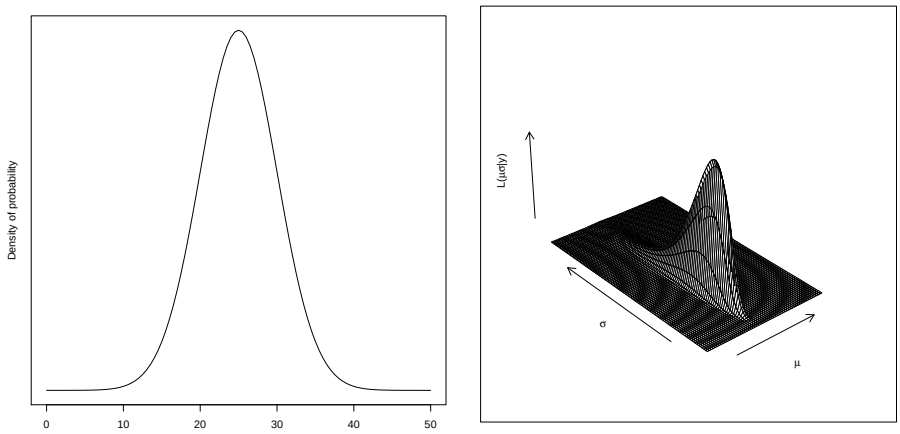
$$Pr(y|\mu, \sigma) = f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \tag{1.4}$$

For a vector y_i where i is the i^{th} observed value, the likelihood function will be the product:

$$L(y_i|\mu, \sigma) = \prod_i f(y_i|\mu, \sigma) \tag{1.5}$$

Figure 1.2: Normal Distribution $N(25, 25)$

(a) Probability density function *(b) Likelihood function*



As an example, we generate a vector from a normal distribution $y \sim N(25, 25)$ which could be a sample of dairy cow milk yields. This vector is

$y = 22.2 \quad 23.8 \quad 32.8 \quad 25.4 \quad 25.6 \quad 33.6 \quad 27.3 \quad 18.7 \quad 21.6 \quad 22.8$. The probability density function for this distribution is plotted in Figure 1.2a. In this case, there are two parameters to model, so that the likelihood is a surface. The plot of this surface is shown in Figure 1.2b. The maximum likelihood estimate is the highest value of this surface. In this example the maximum likelihood was reached for $\mu = 26$ and $\sigma = 4.5$.

Maximum likelihood and model fit

While particular values of the functions are unimportant, variations are of interest for the estimation of model parameters and assessment of model fit. Transformations which preserve the variations but simplify the function are used. Taking the logarithm changes the product into a sum. Hence the deviance is defined as $-2log(likelihood)$. Minimizing the deviance is equivalent to maximizing the likelihood. Penalties can be added to the deviance to account for the number of parameters in the model. For example, the Akaike Information Criterion is defined as:

$$AIC = -2log(L(\theta|y_i)) + 2K \tag{1.6}$$

where θ is the vector of parameters of interest, y_i the observed values, and K the number of parameters in the model.

1.6.2 Frequentist or Bayesian?

Two different approaches are used in statistical inference, namely frequentist and Bayesian.

The Frequentist approach

In the frequentist paradigm, probabilities are considered as frequencies e.g. the number of *Successes* out of a sequence of trials when the number of trials tends to infinity. As sample size increases the parameters' estimates approach the *true value*. Parameters estimations are mostly carried out by maximum likelihood estimation, but also least squares related techniques. Various algorithms can be used for this purpose.

The Bayesian approach

In the Bayesian paradigm, probabilities are treated as beliefs. A prior belief is combined with observed data to get a new estimate of the parameter, here called θ (which can be a vector e.g. μ and σ in the above example), on which inference is made. This is expressed by Bayes' rule:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{1.7}$$

where $p(\theta|y)$ is called the posterior distribution. It is the probability distribution for θ after having observed the data. $p(\theta)$ quantifies the ideas we have about the parameter θ before having observed any data and is called the prior. $p(y|\theta)$ is the likelihood function (see Section 1.6.1), that is the probability of having observed the data given a particular value of θ. $p(y)$ is a normalising constant. It is the integral $\int p(\theta)p(y|\theta)d\theta$ for continuous data or the sum $\Sigma_\theta p(\theta)p(y|\theta)$ for all values of θ for discrete data (Gelman *et al.* , 2003). Hence, in the Bayesian paradigm, new data are used to update a prior belief. This belief can be based on personal opinion, experts' opinion or literature review. Strong priors will require more data to be moved. It is also possible to give priors that are as uninformative as possible to let the data drive the parameters' estimates. In this case, parameters estimates are very close to maximum likelihood estimates.

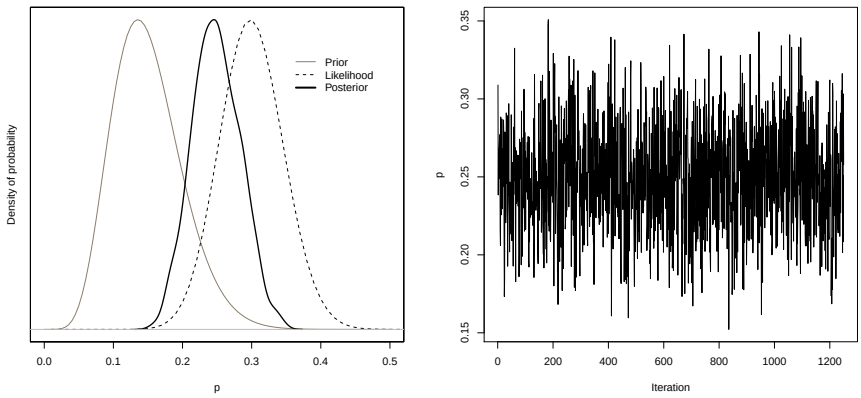
An example

To illustrate these ideas we take a simple example. A frequentist and a Bayesian vet go onto a farm. They estimate mastitis prevalence in the herd using the number of cows with SCC above 200,000 cells/mL they observe. On the day of the visit, 30 cows out of 100 are above the threshold. The frequentist vet concludes that the prevalence in the herd is 30 % and making the

Figure 1.3: Example of Bayesian inference for a binomial proportion.

(a) Prior, posterior and likelihood densities.

(b) Samples from the posterior distribution for the proportion p.



hypothesis that this is the true prevalence, sampling 100 similar herds would give proportions between 21 % and 39 %, 95 % of the time. The Bayesian vet has come to the herd many times, and from previous experience, he thinks that the prevalence of SCC $>$ 200,000 in the herd is centred around 15 % with most of the possible values between 5 and 25 %. Figure 1.3a shows the prior and posterior distributions as well as the likelihood function. Figure 1.3b presents samples from the posterior distribution of p. The Bayesian vet has updated his knowledge about high SCC prevalence on this farm. The posterior distribution is higher than the prior because there were enough data to move it upwards, possibly because the prevalence has increased recently. However, because the prior distribution was narrow, the posterior estimate is not as high as the frequentist one. On a following visit, this posterior distribution can be used as a prior.

Bayesian parameters' estimation

In Bayesian inference, parameter estimation is complicated by the fact that, for most models, the normalising constant $p(y)$ cannot be estimated by numerical integration. Parameters have to be estimated by simulation, using Markov Chain Monte Carlo (**MCMC**). Thus the Bayesian approach did not have many practical application before the 1990s because it was too computer intensive. The principle of MCMC is to sample θ iteratively from the posterior distribution. Initial values are provided for θ and at each iteration, these values are updated. The new values for θ only depend on the previous iteration (Markov property), and, after a variable number of iterations converge to the target (posterior) distribution. Samples from this distribution can then be summarised by calculating means or quantiles. There is no guarantee at any time that a chain has converged to the target distribution. One common solution to this problem is to

run many chains in parallel and to check that they have converged to the same distribution.

Bayesian model checking

To assess model fit, the Deviance information criterion was developed by Spiegelhalter *et al.* (2002) as an equivalent to AIC for Bayesian models. The problem in such models is the estimation of the number of parameters in the model. It is replaced by a variable called pD which is obtained by calculating the posterior mean of the deviance minus the deviance of the posterior means.

Another way to check the results given by the model is prediction. Posterior prediction is used to predict the data used in parameters' estimation ; in cross-validation a different dataset is used for prediction. Discrepancy variables are then tested against the observed data as a way to evaluate model fit (Green *et al.* , 2009; Gelman *et al.* , 2000).

Which approach to use?

Both frequentist and Bayesian approaches have advantages and disadvantages. The frequentist approach is more efficient and quicker computationally. The Bayesian approach requires more computer power and time to run the simulations but it allows the inclusion of prior information and the generation of predictions incorporating the uncertainty in all parameters of the models. When analysing correlated data (Section 1.6.3), Gelman & Hill (2007) recommend to carry out an initial frequentist analysis and then to fit a full Bayesian model. This was done in Chapter 5. Bayesian software such as WinBUGS allow to fit very complex models, which would be hard to fit in other statistical packages, with a relative ease. For this reason, model exploration in Chapter 6 was carried out directly in WinBUGS. In Chapter 7, a large dataset was used. The time required to run the corresponding models in WinBUGS would have rendered their use impractical and a frequentist approach was preferred.

1.6.3 Linear Mixed Models

Linear models are the basis for all the models used in this thesis. They describe linear associations between an outcome and covariates. Such models can be written as:

$$Y_i = \alpha + X_i^T \beta + \epsilon_i \tag{1.8}$$

where Y_i is the i^{th} observed datum, α the model intercept, X the matrix of predictors, β the associated vector of coefficients and ϵ the vector of errors. The model estimates of α and β are

$\hat{\alpha}$ and $\hat{\beta}$. The estimates for the errors are called residuals. They are calculated as follows:

$$\begin{aligned} \hat{\epsilon_i} &= Y_i - (\hat{\alpha} + X_i^T \hat{\beta}) \\ \hat{\epsilon_i} &= Y_i - \hat{Y_i} \end{aligned} \tag{1.9}$$

The residuals are assumed to be normally distributed, uncorrelated and their variance constant for all levels of the covariates. In longitudinal data such as collected by the National Milk Records, and more generally for data originating from farms, data are correlated within the units under study. For example, milk recording data can be assumed to be correlated within cows and cow data correlated within herds. This can be accounted for by using Linear Mixed Models, also called Multilevel models. Instead of adding a fixed coefficient for each upper level unit (one coefficient per herd for example), it is assumed that these effects are distributed as a normal distribution with mean 0 and standard deviation σ. Each unit of the level is attributed a random coefficient from this distribution which is a compromise between the population mean effect and the individual unit effect weighted by the number of observations in the unit (Rasbash *et al.* (2009) - p 39). In these models, the observed variations can be partitioned between the different levels. In the simplest of these models, only the intercept is modelled as random. This can be written as:

$$\begin{aligned} Y_{ij} &= \alpha + X_i^T \beta + u_j + \epsilon_{ij} \\ u_j &\sim N(0, \sigma_u) \\ e_{ij} &\sim N(0, \sigma_e) \end{aligned} \tag{1.10}$$

where i and j indicate the lower and upper levels, u_j is the vector of upper level residuals and ϵ_{ij} the bottom level residuals. These models can be made more complex by modelling different slopes for each level:

$$\begin{aligned} Y_{ij} &= \alpha + X_i^T (\beta + \upsilon_j) + u_j + \epsilon_{ij} \\ \upsilon_j, u_j &\sim MVN(0, \Sigma) \\ e_{ij} &\sim N(0, \sigma_e) \end{aligned} \tag{1.11}$$

where υ_j and u_j are the random residuals associated with the intercept and the slope(s) respectively, which are drawn from a multivariate normal distribution.

1.6.4 Generalized Linear Mixed Models

In a lot of instances, the assumptions or prerequisite of linear regression do not hold. In this thesis, binomial and multinomial outcomes, to which linear regression cannot be applied, will be considered. *Generalized Linear models* were introduced by Nelder & Wedderburn (1972) to model such data. Typically the outcome is related to the linear predictor through a link function $g(.)$ which is modelled as a function of linear predictors as described above. As for linear mixed model, random coefficients can be modelled.

Logistic regression

When the outcome follows a binomial distribution, it is common to model the log of the odds of the proportion. Such models are called logistic models, of which specifications are:

$$\begin{aligned} Y_i &\sim Binomial(p_i, n) \\ ln(\tfrac{p_i}{1-p_i}) &= \alpha + X_i^T \beta \end{aligned} \tag{1.12}$$

Where Y_i is the outcome which can take values 0 and 1, p and n the parameters of the associated binomial distribution, α the intercept of the regression, X_i the matrix of covariates and β the vector of associated coefficients. The interpretation of the coefficients of such regression is not straightforward. Exponentiating the whole expression yields:

$$\begin{aligned} \tfrac{p_i}{1-p_i} &= e^{\alpha + X_i^T \beta} \\ \tfrac{p_i}{1-p_i} &= e^{\alpha} \prod_1^i e^{X_i \beta} \end{aligned} \tag{1.13}$$

So that the exponential of a coefficient gives the odds ratio of p associated with each unit of increase in the covariate considered. These odds ratio must be multiplied and not added to estimate the effect of a set of covariates. A convenient way to explore a model is to set all the covariates to their mean and to plot the prediction for a covariate of interest for a range of plausible values.

Multinomial logit regression

When the outcome can fall in one of n categories, the sampling variations can be described by a multinomial distribution. The multinomial distribution can be written:

$$Y_i \sim Multinomial(p_{ik}, n) \tag{1.14}$$

where p_k is the probability for Y_i of being in category k and n the number of observations in Y. The multinomial distribution is the generalisation of the binomial distribution to k categories. Thus, the principle of the multinomial logit regression is to use one of the categories as the reference and to model the odds of being in any of the other $k-1$ categories against being in the reference category. This can be written as:

$$ln(\tfrac{p_{ik}}{p_{iK}}) = \alpha_k + X_i^T \beta_k \tag{1.15}$$

where p_{ik} is the probability of being in category k, p_{iK} is the probability of being in the reference category, α_k the intercept for the k^{th} category, X_i the matrix of covariates and β_k the vector

of associated coefficients for the k^{th} category. As in the binomial logit models coefficients are odds ratio of being in category k compared to category K.

Chapter 2

Data Selection

2.1 Introduction

In the UK, the main provider of milk recording is a private company called the National Milk Records (NMR). All data collected between the 1^{st} January 2004 and the 31^{st} December 2006 were purchased from the NMR for this project. These included herds which started or stopped milk recording at some point during these three years, data imported from the previous herd when cows were bought and other peculiarities such as factored data. The aim of the data selection process was to obtain a subset of homogeneous data from regularly recorded herds. The present chapter details the data exploration carried out and the set of criteria applied to select the dataset used throughout this thesis.

2.2 Data Collection

2.2.1 Traditional milk recording

The procedure milk recorders must follow is described in a booklet distributed to milk recorders (NMR, 2008). Milk recording is usually carried out on all the lactating cows of a herd on a monthly basis. Exceptions are cows under antibiotic treatment, cows which are suckling, cows purchased within 4 days of the visit, cows that have calved or aborted within 4 days of the visit and cows yielding less than 3 kg of milk in the 24 hour recording period. A first milk sample is collected in the evening and a second one in the following morning. Preliminary data collection is operated before the first sampling to collect or update cow historical information such as date of calving, date of service and other historical events. A list of the cows present in the herd is available from previous recordings and cows are assigned a recording line number in order of ascending herd identification number on this list. Sample pots are marked up with the cow numbers and sorted by recording number within boxes prior to the milking. Any samples

from new cows and freshly calved heifers are placed after the last numerically ordered animal. Because the time between morning and evening milking can be different to the time between evening and morning milking a varying volume of the milk collected is sampled using sampling dippers. For the usual two sampling scheme, 5 sample dippers are available. These are coded as 12, 13, 14, 15 and 16 representing increasing volumes. If the interval is of 12 hours, 13 is used on both milkings. When this interval is greater than 12 hours, 12 is used on the evening milking and the sampling dipper used during the following morning corresponds to the entire part of the number of hours elapsed between the milkings e.g. 13 for everything between 13:00 and 13:59. As a cow is milked, a volume of milk proportional to the volume produced is stored in a jar attached to the milking unit. The milk collected is a pool of the milk given by the four quarters. Once the cow has finished milking, her number is identified, the weight of milk is read on the jar, in kg to the nearest 0.2, and written on the sample bottle top with a chinagraph pencil or a permanent marker pen. It can be noticed here that even though milk yield is recorded in kg, what is actually measured is a volume. The milk in the jar is agitated for ten seconds and approximately one third of the volume released. A sample of milk is taken with the sample dipper and poured into the sample pot. The remainder of the milk in the jar is then released.

2.2.2 Alternative milk recording

Some herds do not or cannot comply with the usual two consecutive milking monthly recordings. NMR proposes alternative ways of recording:

- One sampling time: in some cases, farmers can ask to have milk samples taken at one instead of the usual two consecutive milkings. This can be done on a regular basis or to retest some cows of the herd between regular tests. In this case, the data have to be factored, that is to say daily milk volume and constituents are extrapolated from one sampling point. More variation can be expected from one sampling point compared to the average of two consecutive ones.
- 3, 6 or 8 weekly sampling intervals: farmers can choose to have their cows recorded at different time intervals.
- *Do It Yourself* sampling: Farmers can supply milk yield, milk sample and data themselves to NMR. This is also possible for 3, 6 or 8 week time intervals. Data from approved automatic recording systems can be collected via the farm computer or from a box in the parlour.

2.3 Initial Dataset

The data were sent to the School of Veterinary Medicine and Science on a DVD containing a Microsoft Access 2003 database. Minimal information was provided on an accompanying A4 sheet. The database contained all the recordings performed by the NMR between the 1^{st} of January 2004 and the 31^{st} of December 2006. The data were in two tables: one for the recordings data (`NMR_recordings`) and one for the cows data (`NMR_animals`). The names, types of data and descriptions for each field of each table are in Tables 2.1 and 2.2. The recordings table had 19,893,093 lines from 1,247,427 cows in 5,714 herds. The minimum, maximum and number of missing values associated with each of these tables are in Tables 2.3 and 2.4. Throughout this thesis, weight of milk was converted to kg by dividing the provided values by 10. Protein, butterfat and lactose contents were converted to percentages by dividing the provided values by 100.

2.4 Data selection

After an initial phase of data exploration, it was realised that the number of cows per herd was decreasing over time when the opposite trend was expected. Some cows had not been included when the dataset was built. The data and data selection described here are from a dataset subsequently provided by the NMR containing all the recordings performed between the 1^{st} January 2004 and the 31^{st} December 2006. The aim of the data selection process was to select a subset of homogeneous data. We were interested in monthly regular recordings from herds milk recording for the complete three years. The data selection process is described in the order the selection criteria that were applied.

2.4.1 Missing data

Some cows were not recorded because of the reasons mentioned above (dry, antibiotic treatment, ...). In such cases, it was expected for the field authentic_recordings to be set to N. Out of the 19,893,093 lines of data initially available, 40,861 were flagged as unauthentic records. However, in the remainder of the dataset, cows could be recorded as present (authentic_recording = Y) even though all fields either contained the value 0 or were empty. The number of lines which either were empty or contained the value 0 for weight of milk, somatic cell count and percentage of butterfat are indicated in Table 2.5. In more than 99.9 % of the cases, when weight of milk was null or empty, somatic cell count and percentage of butterfat were also null or empty. Lines where weight of milk was null or empty were recoded as missing data.

Table 2.1: Descriptions of the fields in the table NMR_recordings

Field	Data Type	Description
herd_identity	Number	Herd identity
animal_identity	Number	Animal identity
calving_date	Date/Time	Calving date
recording_date	Date/Time	Recording date
authentic_recording	Text	Cow present on the recording date
weight_of_milk	Number	Milk quantity in hectograms
bfat_percent	Number	Butterfat content (g/hg)
protein_percent	Number	Protein content (g/hg)
lactose_percent	Number	Lactose content (g/hg)
cell_count	Number	Somatic cell count (1,000 cells/mL)
bfat_factored_flag	Text	Data factored for bfat_percent

Table 2.2: Description of the fields in the table NMR_animals

Field	Data Type	Description
herd_identity	Number	Herd identity
county	Text	County in which the herd is located
animal_identity	Number	Cow identity
animal_breed	Number	Cow breed
date_of_birth	Date/Time	Cow date of birth
sire_breed	Number	Sire breed
sire_identity	Number	Sire identity
latest_calving_date	Date/Time	Latest calving date in the dataset
latest_recording_date	Date/Time	Latest recording in the dataset
current_lactation	Number	Parity at the last recording date

Table 2.3: Minimum, maximum and missing values for the table NMR_recordings

Field	Min	Max	Missing
herd_identity	28	24,675	0
animal_identity	1	6,727	0
calving_date	21/12/1994	28/12/2006	15
recording_date	01/01/2004	21/12/2006	0
authentic_recording[a]	N	Y	0
weight_of_milk	0	998	0
bfat_percent	0	1,470	0
protein_percent	0	897	0
lactose_percent	0	899	0
cell_count	0	9,999	838,011
bfat_factored_flag[a]	N	Y	0

[a]Y: Yes ; N: NO

Table 2.4: Minimum, maximum and missing values for the table NMR_animals

Field	Min	Max	Missing
herd_identity	28	24,675	0
county	-	-	0
animal_identity	1	6,727	0
animal_breed	1	93	0
date_of_birth	09/11/1966	2/12/2005	58,659
sire_breed	1	95	0
sire_identity	-	-	0
latest_calving_date	27/06/199	19/03/2007	0
latest_recording_date	01/01/2004	31/12/2006	0
current_lactation	1	56	0

Table 2.5: Repartition of the data between missing or null (0) and positive values (> 0) when authentic_record = Y

Weight of milk	n	Cell count	n	Butterfat	n
0	757,031	0	757,016	0	756,986
				> 0	30
		> 0	15	0	5
				> 0	10
> 0	19,095,201	0	133,337	0	570
				> 0	132,767
		> 0	18,961,864	0	3
				> 0	18,961,861

2.4.2 Selection of herds recording for most of the study

Herds starting or stopping milk recording might be different from other herds and may provide insufficient information for robust analysis. The aim of this part was to select herds recording for the majority of the study. For each herd, the first and last recording date in the dataset were identified. The numbers of herds having their first and last recording at each month of the study are presented in Table 2.6. The first step consisted in the removal from the dataset of 1,779 herds for which either the first test-day occurred after February 2004 or the last test-day occurred before November 2006.

Table 2.6: Number of herds starting and ending milk recording at each month of the study

Month	First recording date	Last recording date
Jan 04	5157	29
Feb 04	71	26
Mar 04	21	50
Apr 04	12	85
May 04	20	47
Jun 04	15	41
Jul 04	11	37
Aug 04	12	64
Sep 04	16	47
Oct 04	14	44
Nov 04	14	37
Dec 04	12	20
Jan 05	8	28
Feb 05	10	38
Mar 05	17	33
Apr 05	9	49
May 05	11	23
Jun 05	20	49
Jul 05	20	40
Aug 05	9	48
Sep 05	20	47
Oct 05	20	41
Nov 05	17	35
Dec 05	13	23
Jan 06	10	24

Continued on next page

Table 2.6: Number of herds starting and ending milk recording at each month of the study

Month	First recording date	Last recording date
Feb 06	7	31
Mar 06	19	36
Apr 06	10	38
May 06	15	52
Jun 06	13	34
Jul 06	16	42
Aug 06	11	58
Sep 06	16	55
Oct 06	11	61
Nov 06	13	696
Dec 06	10	3592

2.4.3 Identification of monthly regular test-days and removal of factored data

The remaining herds had between 18 and 364 test-days and the total number of test-days (a day on which at least one cow in the herd was recorded) was 208,242. A peculiarity of the database was that when a herd bought cows, all the previous recordings of these cows present in the database were retrieved in the new herd. This meant that a particular recording could have been carried out in a different herd than the one recorded in the database i.e. the herd the cow was bought from. These recordings were not flagged and had to be identified individually. The distribution of the number of cows recorded for a herd per test-day is shown in Figure 2.1. For illustration purposes, the distribution was truncated at 300 cows with 2,433 (1.17 %) tests with more than 300 cows. This distribution is clearly bimodal with 27,161 test-days with only one cow recorded, a decrease in the number of cows recorded until around twenty cows and a subsequent increase. Test-days with less than 20 cows recorded were mainly data imported from a previous herd when cows had been bought and were therefore removed from the data. This represented 67,058 tests. The interval between consecutive test-days were re-calculated and herds with intervals smaller than 20 days were checked individually. Most of the time, the difference between the regular and imported test-days were obvious since the number of cows bought was smaller than the number of cows regularly recorded. In these cases, all the tests with the number of cows recorded being smaller than a threshold identified visually were discarded. In other instances, a single threshold would have discarded regular tests. For these herds, tables with cow identity and recording dates were generated. The date a group of cows entered the

herd was identified and all prior recordings removed for these cows. This resulted in the deletion of 3,390 test-days. The intervals between consecutive tests and number of cows recorded for two herds are presented as an example in Figure 2.2. The left part of the figure represents a herd for which a single threshold value allowed the identification of imported recording dates. The right part of the figure shows a herd for which there was an overlap between the number of cows regularly recorded and the number of cows imported. A table of the recording dates and the identities of the cows recorded was generated. As they were incorporated into the herd, cows of different origins were recorded together and the corresponding recordings selected.

The traditional milk recording scheme involves two consecutive milkings i.e. evening and the following morning. Farmers can choose to have milk recording on only one milking instead of two. In this case, the data need to be corrected to account for this. The resulting data were flagged as *Factored* with the variables *bfat_factored_flag*. Test-days with at least one factored recording were removed from the data. At this stage, 10,231,300 recordings in 93,644 test-days from 3,001 herds were remaining in the dataset. The number of test-days for each of the 3 years were calculated for each herd. If a herd had less than 10 tests for either 2004, 2005 or 2006, it was removed from the dataset. After this step, 8,762,480 recordings in 81,603 test-days from 2,302 herds were remaining. The highest number of test-days in the remaining herds was of 13 per year.

2.4.4 Breeds

Important differences in milk quantity and composition exist between cow breeds. The NMR records cows' breeds. The breeds' names used in the data and the number of recordings per breed are presented in Table 2.7. Cows of the Holstein breeds represented more than 90 % of all cows and all recordings. Farms which milked mainly cows of breeds other than Holstein or Friesian that is with less than 80 % of recordings from the breeds Holstein and/or Friesian were discarded because it was intended to investigate data from Holstein-Friesian herds. The final dataset contained 8,211,988 recordings from 2,128 herds.

Table 2.7: Number of recordings for each breed as defined in the NMR data. Cows of the breeds Holstein and/or Friesian are sorted by decreasing number of occurences in the dataset. Cows of other breeds which have more than 1,000 cows follow sorted by decreasing number of recordings.

Code	Breed	Numb rec	% rec
1	HOLSTEIN	8023801	91.57
12	BRITISH HOLSTEIN	101741	1.16
Continued on next page			

Table 2.7: Number of recordings for each breed as defined in the NMR data. Cows of the breeds Holstein and/or Friesian are sorted by decreasing number of occurences in the dataset. Cows of other breeds which have more than 1,000 cows follow sorted by decreasing number of recordings.

Code	Breed	Numb rec	% rec
20	BRITISH FRIESIAN	71929	0.82
63	DUTCH HOLSTEIN FRIESIAN	31604	0.36
71	FRENCH HOLSTEIN FRIESIAN	4274	0.05
65	AMERICAN HOLSTEIN	4235	0.05
64	CANADIAN HOLSTEIN	4200	0.05
60	GERMAN HOLSTEIN-FRIESIAN	3566	0.04
61	DANISH HOLSTEIN-FRIESIAN	894	0.01
15	RED & WHITE FRIESIAN	515	0.01
72	ITALIAN HOLSTEIN-FRIESIAN	458	0.01
62	NEW ZEALAND HOLSTEIN-FR.	445	0.01
47	AUSTRALIAN HOLSTEIN-FR.	22	0.00
54	SPANISH HOLSTEIN FRIESIAN	14	0.00
52	SCANDINAVIAN HOLSTEIN	13	0.00
4	JERSEY	214638	2.45
3	AYRSHIRE	99585	1.14
5	GUERNSEY	74849	0.85
2	DAIRY SHORTHORN	46933	0.54
31	BROWN SWISS	27780	0.32
24	MEUSE-RHINE-ISSEL	17751	0.20
28	MONTBELIARDE	8515	0.10
66	EUROPEAN JERSEY	7521	0.09
59	SCANDINAVIAN RED - IMPORT	2201	0.03
23	SIMMENTAL/FLECKVIEH	1685	0.02
29	NOT KNOWN (CATTLE)	1463	0.02
42	WATER BUFFALO	1447	0.02
37	BELGIAN BLUE	1116	0.01
33	LIMOUSIN	1012	0.01
	Other	8273	0.09

Figure 2.1: Number of cows recorded per test-day

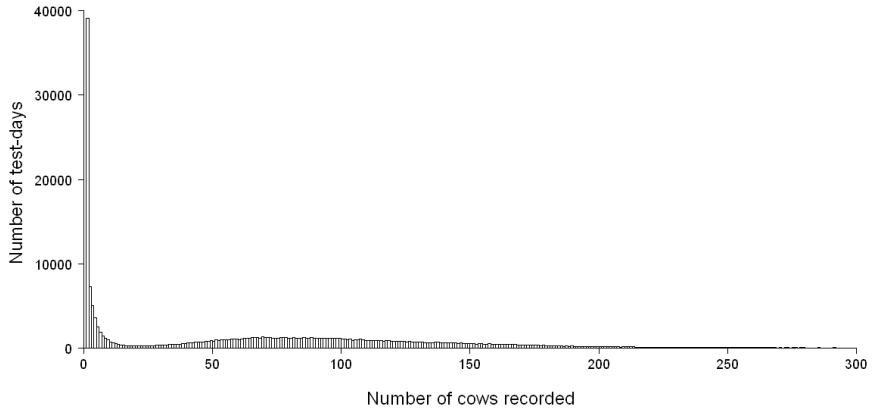
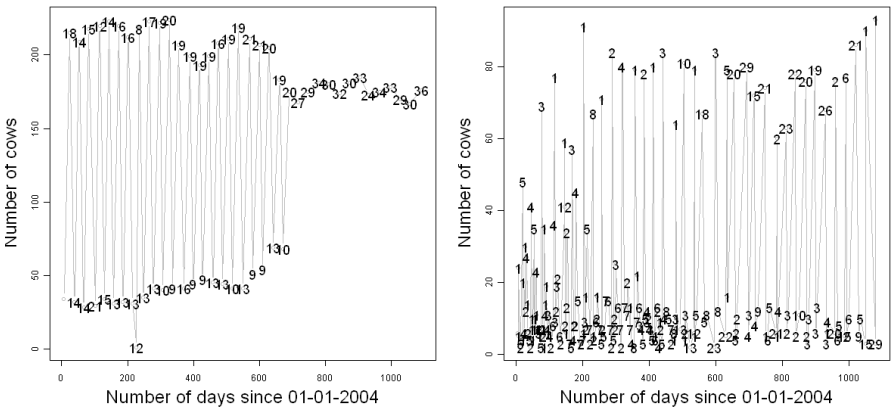


Figure 2.2: Interval between consecutive test-days and number of cows recorded per test for two herds which bought cows. The numbers on the plots indicate the number of days since the previous recording date in the dataset. The herd on the left bought 68 cows on day 691 of the study. The recordings for these cows carried out in their previous herd were imported and indistinguishable from the recordings carried out in the current herd. They were easily identified because the number of cows bought was much lower than the number of cows usually recorded. For the herd on the right, the distinction between recordings operated in another herd and recordings operated in the current herd was harder because the number of cows bought was close to the number of cows regularly recorded. In such cases a table was created with cow identities and recording dates to identify the dates new cows joined the herd.



2.5 Final dataset

The final dataset contained 8,211,988 recordings in 992,625 lactations from 483,747 cows in 2,128 herds. Some missing values remained in the dataset. However, because some of this work will focus on test-days or on some particular aspects of the data, all the data selected until this stage were kept and excluded appropriately in later stages. Milk yield was the minimum required information indicating that a cow was present on the recording date considered, and was greater than 0 for all the recordings. The NMR's Guide to Recording/Sampling specifies that cows yielding less than 3 kg should not be collected, but 278 lines had milk yield of less than 3 kg. These were considered as valid. There were 27 lines with both the percentage of butterfat and the percentage of protein equal to 0. In 26 of these, cell count was 0 or missing. There was no line with missing values for either percentage of butterfat or percentage of protein. The highest values recorded for milk yield, percentage of butterfat and percentage of protein were 99.8 kg, 9.62 %, and 8.74 %. There were 20,103 lines were cell count was missing. There were 29,688 lines with cell count equal to 0 and only 865 lines with cell count equal to a thousand. In some instances missing data for cell count were recorded as 0. Values of 0 were recoded as missing data.

2.6 Discussion

The objective of this thesis is to describe the variations in milk recording data and to relate them to udder health and reproduction. From data available on the DairyCo (2009a) website, which publishes statistics on farming in the UK, the number of dairy herds in England and Wales went from 16,189 in January 2004 to 13,270 in December 2006. It is hard to know how much the present sample is representative of these herds. The total number of herds present in the initial NMR dataset was 5,714. This means that for these years, NMR had data for between 35 and 45 % of dairy herds in England and Wales. The present dataset contains data for 16 % of the dairy herds in activity in December 2006. Hence, a large quantity of longitudinal data was available for this work. These data were collected by trained technicians as a part of routine controls of cow performance for which farmers paid and hence were willing to get reliable data. All the samples were analysed in the same location with the same equipment. However, not all the data present in the initial dataset could be used and some information was missing. For example, no flag indicating whether a recording was a regular recording, an imported recording or a *Do It Yourself* recording was present. Milk recording data are used for epidemiological studies as well as for the estimation of breeding values. These recordings could represent a source of bias in these studies. Based on the distribution of the number of cows per test-day, the removal of all tests with less than twenty cows was undertaken to remove the

majority of recordings originating from the previous herd the cows were in. By doing this it is possible that regular tests from small herds were also discarded. In this case, either the whole herd was removed or only a few tests were removed since herds with less than 10 test-days for any of the three years were subsequently removed. In the remaining herds, the identification of data from other herds was based on unusual intervals between consecutive recording dates. For herds with at least one abnormally short interval between consecutive recording dates, the number of cows recorded and the interval between recordings were plotted for the complete period under study. This was assessed visually. In most cases, the distinction was easy since the number of cows imported was clearly lower than the usual number of cows recorded. In other instances the situation was less obvious but checking recording dates at the cow level allowed the identification of cows which had been for the longer period in the herd. However, it is possible that some regular test-days have been discarded or some imported test-days kept in the dataset. Even if it was the case, this would represent a negligible proportion of recordings in the dataset.

In the UK, more than 95 % cows were of the breeds Holstein, Friesian or Holstein/Friesian between 2004 and 2006 (The Center for Dairy Information, 2009). Major differences in milk production exist between breeds. The number of recordings originating from breeds other than Holstein and/or Friesian in the dataset was not sufficient to infer anything about them and herds milking predominantly these other breeds were discarded. Some other breeds remain in the dataset but represent less than 20 % of a herd. This was done to be able to use data on a herd basis and reflects the normal UK situation for normal Holstein-Friesian herds.

The fact that weight of milk is recorded in kg when what is actually measured is a volume can be misleading. A litre of milk is usually estimated to weigh 1.033 kg. It is possible that some variation exists around this mean depending on the quantity of milk solids a particular sample contains. The difference is probably minor, but since a volume is measured, it would make sense to record it as a volume and let other parties operate the conversion. We are not aware of the reason for the recording of a volume as a weight. Milk will be referred as weight of milk in kg throughout this thesis and the data left unchanged. However, this fact must be kept in mind.

2.7 Conclusion

A large sample of milk recording data representing approximately 16 % of the herds in activity in England and Wales at the end of 2006 was selected on the basis of monthly intervals between recordings, completeness of the data, more than 80 % of Holstein-Friesian cows recorded on average and a number of cows recorded greater than 20.

Chapter 3

Distributional Characteristics of Herds, Lactations and Milk Production

3.1 Introduction

In some European countries, milk production is seasonal when the demand for dairy products remains stable throughout the year (Hennessy & Roosen, 2003) which results in additional costs for the storage or transport of milk (McErlean, 1999). There is no recent description of such seasonality in England and Wales in the scientific literature. It is important for the dairy industry to be aware of the observed patterns of variations. These patterns are the result of individual cow characteristics, mostly stage of lactation and parity (Silvestre *et al.* , 2009), as well as other factors specific to the season such as feed composition. In the UK, changes in the number and structure of dairy herds are also ongoing. For example, there was a decrease of 13.1 % in the number of dairy farms between June 2004 and June 2006 while in the same period, the number of cows decreased by 4.3 % (DairyCo, 2009a).

Milk recording contains nationwide information on individual cow milk quantity and composition which can easily be used to derive the monthly variations in milk supply. Reliable data on dates of calving and parity are recorded so that calving patterns can be described with the same data. There is no recent description of the variations in milk production and calving patterns in England and Wales. This chapter describes the general characteristics and changes in herd, lactation and milk production data in a large sample of dairy herds from England and Wales using milk recording data collected between 2004 and 2006. The aim is to provide current patterns and information for the UK dairy industry.

3.2 Materials

The selection and general description of the data analysed in this chapter are detailed in Chapter 2. Specific steps of the analysis required the aggregation of the data at the herd, lactation, test-day or month level.

3.3 Methods

3.3.1 Herd characteristics

At the herd level, the number of cows recorded, the quantity of milk produced per test-day and their variations per month of the study were determined. In order to quantify the variations during the three years, arithmetic means per month were calculated for each variable and for the 36 months available (Jan 2004 to Dec 2006) as well as for the 12 calendar months (Jan to Dec). For each of the 36 months of the study, the difference between the mean calculated for a particular month (n = 36) and the mean for a specific month of the year (n = 12), was evaluated. In order to identify trends, these differences were smoothed using locally weighted regression as implemented in the R loess function. With this type of regression, the dependent variable is smoothed as a function of the independent variable in a way similar to a moving average (Cleveland & Devlin, 1988). The degree of smoothing is controlled by a parameter α and a value for α of 0.75 was used here.

3.3.2 Lactation characteristics

The data available at the lactation level were the date of calving and parity. Lactations for which calving occurred before the 1^{st} January 2004 were excluded for the analysis. The proportion of cows calving of each parity was calculated. For each heifer calving between the 1^{st} January 2004 and the 31^{st} December 2006, the age at first calving was calculated by subtracting the date of birth from the date of calving. The numbers of cows calving per month of the study were plotted. For cows of parity one to four and greater than 4, the number of cows calving each month as a proportion of the cows calving over the whole study were plotted. Finally, intervals between consecutive calvings were investigated. The Kaplan-Meier survivor function for the probability of calving given the previous date of calving was computed using the R survfit function (Therneau, 2009). All the cows that had not calved by the 30^{th} November 2006 were censored.

3.3.3 Average milk quantity per cow and bulk milk composition per calendar month

The mean quantity produced per cow per test-day, the bulk milk concentration for the percentage of butterfat, the percentage of protein and somatic cell count were calculated as follows: the concentrations of each constituent were multiplied by the quantity of milk produced for each cow recording, and, these quantities were summed up and divided by the total quantity of milk produced on a given test-day. The mean and standard deviations of these test-day bulk milk estimates were calculated. The arithmetic mean of these estimates across all farms for each milk variable were calculated for each month between January and December. The distribution of the mean test-day bulk milk somatic cell count was right skewed. Taking the natural logarithm of BMSCC resulted in a normal distribution and the geometric mean was therefore deemed more meaningful than the arithmetic mean.

3.4 Results

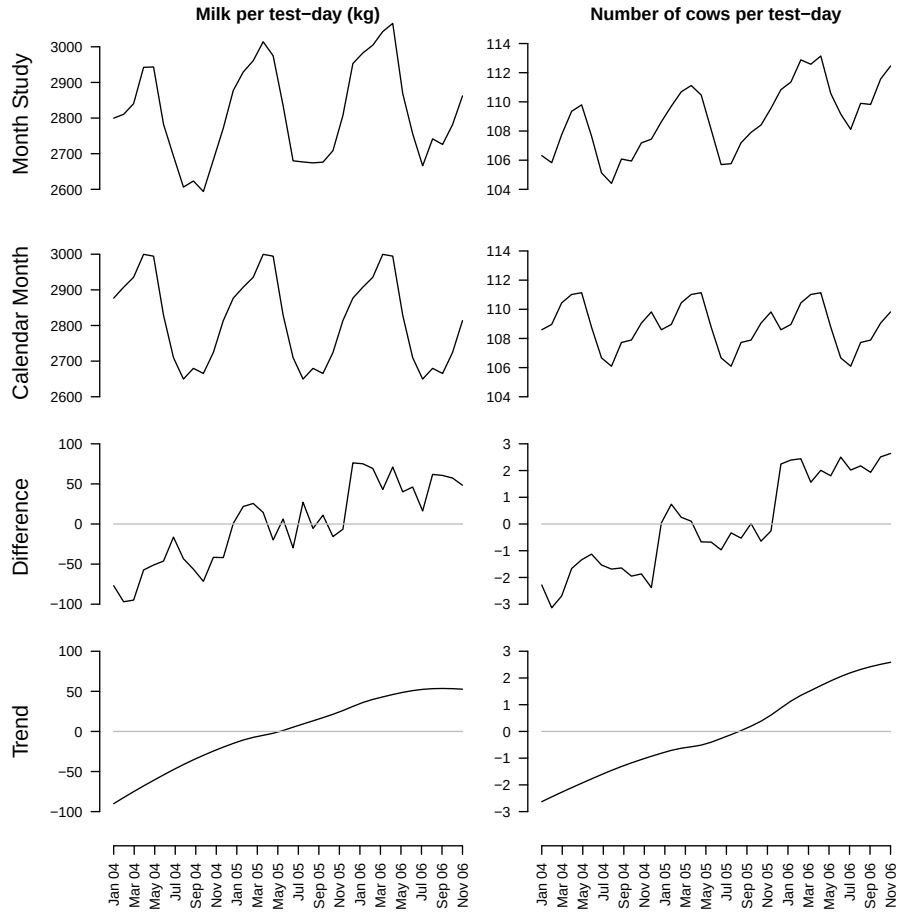
3.4.1 Herd characteristics

Data were available for 2,128 herds. The changes in mean quantity of milk produced and the mean number of cows recorded per test-day between January 2004 and December 2006 are presented in Figure 3.1. On average, the month with the highest milk production was April and the months with the lowest production were August and October. Milk production was 13.2 % higher in April than in August. The highest numbers of cows per test-day were recorded in April and May and the lowest number in August. For both milk production and the number of cows recorded, there was an increase from 2004 to 2007: this increase was of 143 kg of milk and and 5.2 cows per test-day between January 2004 and December 2006.

3.4.2 Lactation characteristics

Data were available for 769,086 lactations of which 226,102 were first lactations. Date of birth was missing for 4,379 (1.9 %) of these first lactations. The first quartile, median, mean and third quartile for the age at first calving were 787, 879, 906 and 997 days. The distribution of the parities of the cows calving is presented in Figure 3.2. Parity 1, 2, 3 and 4 represented 22.8 %, 21.7 %, 17.2 % and 13.4 % of all calvings respectively. The number of calvings per month and parities of the cows calving are presented in Figure 3.3. The number of calvings in a month varied during the year. The month with the lowest and highest numbers of calvings were May and September respectively, with 80 % more calvings in September than in May. The peak became less pronounced with successive parities. There were 23,209 calvings in December

Figure 3.1: Changes in the mean milk quantity of milk produced and mean number of cows recorded per test-day between January 2004 and 2007. *Month study* represents the mean per month of the study (January 2004 to December 2006). *Calendar month* represents the mean for January to December, regardless of the year. *Difference* is the difference between the means for *Month study* and *Calendar Month*. *Trend* is the *Difference* smoothed using a loess regression in order to identify trends in the variation between 2004 and 2006.



2004, 23,840 in December 2005 and 5,597 in December 2006. Only a proportion of the calvings that occurred in December 2006 had been recorded when the database was built. For the following calculations, the number of calvings in December 2006 was replaced by the mean of December 2004 and December 2005. The total number of calvings for 2004, 2005 and 2006 were 263,613, 261,989 and 261,427 respectively which represents a decrease of 0.6 % between 2004 and 2005 and a further decrease of 0.2 % between 2005 and 2006. In the same time period the number of calvings originating from parity one cows increased by 3.8 % between 2004 and 2005 and by 7.4 % between 2005 and 2006. The equivalent figures were -0.68 % and +2.58 % for parity 2, -3.38 % and -3.28 % for parity 3, -3.88 % and -6.08 % for parity 4 and -0.88 % and -4.18 % for parity greater than 4. The cumulative Kaplan-Meier survivor curve for the interval between consecutive calvings is presented on Figure 3.4. Seventy three percent of cows calved again before December 2006. Of the cows which recalved, 0.27 % calved less than 260 days after the previous calving indicating either that some abortions had been recorded as calvings or misrecording of calving dates. Percentages of cows calving before 350, 400, 450, 500, 550 and 600 days after the previous calving were 11, 37, 53, 62, 66 and 69 respectively. Of all the cows which calved more than 260 days after their previous calving, the mean interval between consecutive calvings was 411 days and the median was 391 days.

3.4.3 Average milk quantity per cow and bulk milk composition per calendar month

The arithmetic mean (standard deviation) for milk per cow, bulk milk percentage of butterfat and bulk milk percentage of protein for all recordings from 2004 to 2006 were 25.2 (4.6) kg, 3.92 (0.32) %, 3.25 (0.14) %. The geometric mean (standard deviation) for bulk milk somatic cell count was 187,528 (1,561) cells/mL. Variations in mean milk quantity and composition by calendar month are presented in Figure 3.5. Mean milk yield was constant from January to March at 25.7 kg, increased to 26.4 kg in April and May, decreased from May to reach 24.7 kg in October and increased from October to January. Mean butterfat was constant at 4 % from October to March and decreased from March to June to a minimum of 3.7 %. Variations in the percentage of protein were small compared to variations observed in butterfat. The mean protein content was constant between February and July at 3.2 %, reached a maximum of 3.3 % in November and then decreased until February. The arithmetic mean bulk milk somatic cell count was between 190,000 and 195,000 cells/mL from October to March, increased to 220,000 cells/mL in July and August and decreased from August to October. Between year variations were negligible for butterfat and protein, and limited for the quantity of milk produced per cow (Figure 3.6). There was an increase of approximately 9,000 cell/mL in the estimated BMSCC between January 2004 and March 2005 and a further increase of 4,000 cells/mL between

Figure 3.2: Distribution of the parities of cows calving during the study.

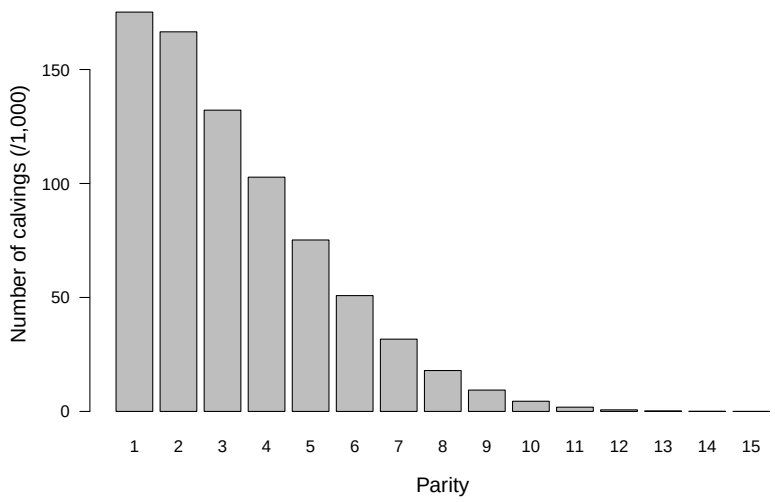
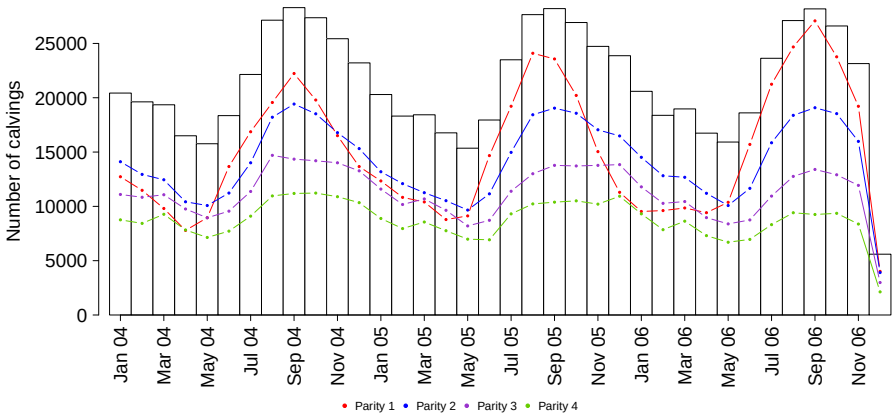


Figure 3.3: Distribution of the number of calvings per month of the study (bars) and proportion of the calvings originating from cows of parity 1 to 4 and greater than 4 (lines).



November 2005 and September 2006 (Figure 3.6).

3.5 Discussion

Our sample contained data for 16 % of the 13,270 dairy herds of England and Wales that remained in business in 2006. The notable features of these data related to the annual patterns of milk yield and calvings.

There was an estimated increase of 143 kg of bulk milk and 5.2 cows recorded per test-day between January 2004 and December 2006. Based on the number of herds and the number of cows present in England and Wales in June 2004 and June 2006, the average number of cows per herd went from 103.8 to 109.4 (DairyCo, 2009a). This increase roughly matches what was observed in the present study. Only cows milk recorded were considered here and the average herd size was higher than the figures derived from DairyCo data. Hence, the mean herd size in this study was higher than the general population and herds that carry out milk recording, may be different in structure and output to non-recorded herds.

The global quantity and composition of milk per calendar month reflect the contribution of cows at different stages of lactation and parities under similar environmental conditions. Given the important role of lactation stage on milk quantity and composition, the month to month variations in the number of calvings are likely to play a role in the monthly output. This has consequences on the overall milk supply which is at its highest in May and is at its lowest between September and December. For example, the daily milk supply was around 20 % higher in May than in November in 2004 (DairyCo, 2009b). Milk price follows the opposite trend with higher prices when the quantity delivered was at its lowest.

There are only two recent studies on calving patterns in Britain (Robinson & Christley, 2006; Mitchell *et al.* , 2005). However, these studies did not discriminate between beef and dairy cattle, and there are nearly as many beef as dairy cows in the United Kingdom (DairyCo, 2009a). Hence there were no recent data on calving patterns in UK dairy cows. In the present study, there were large month to month variations in the number of calvings with a peak of calvings in September and a minimum in May. This resulted in changes in the structure of the milking cow population during the year. This unequal repartition of calvings was probably the result of farmers' objectives and their abilities to manage reproduction. The month at first calving determines calving patterns in later lactations and there was a peak in heifers' calvings in September with a minimum in April and May. A widespread objective is of one calving per cow per year, but the median interval between consecutive calvings was 391 days. The consequence of this is progressive shift and smoothing in the calving peak with successive lactations. It would be possible to change the overall calving pattern by having more spring

Figure 3.4: Cumulative Kaplan-Meier survivor curve for the interval between consecutive calvings.

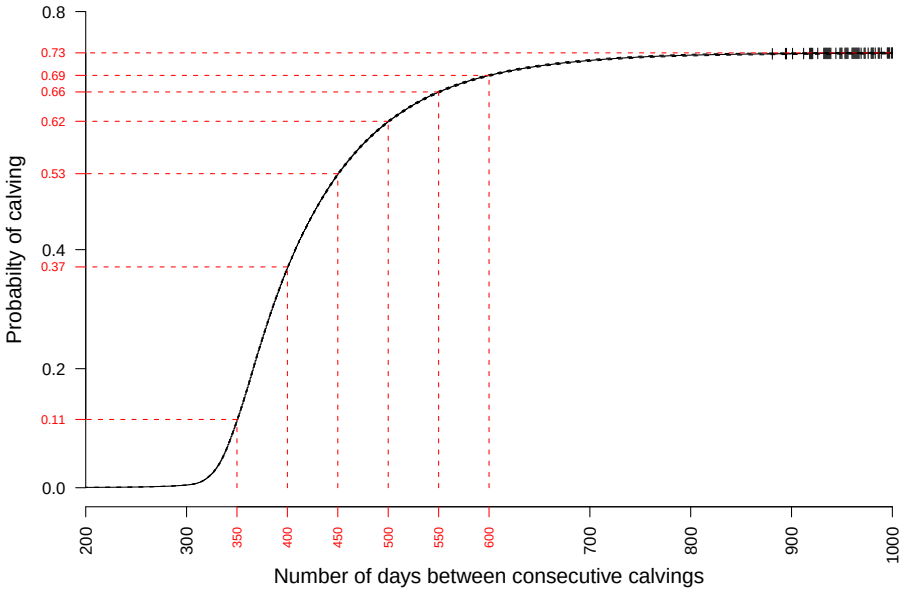


Figure 3.5: Arithmetic mean of test-day bulk milk quantity and composition and geometric mean bulk milk somatic cell count per calendar month between January 2004 and December 2006.

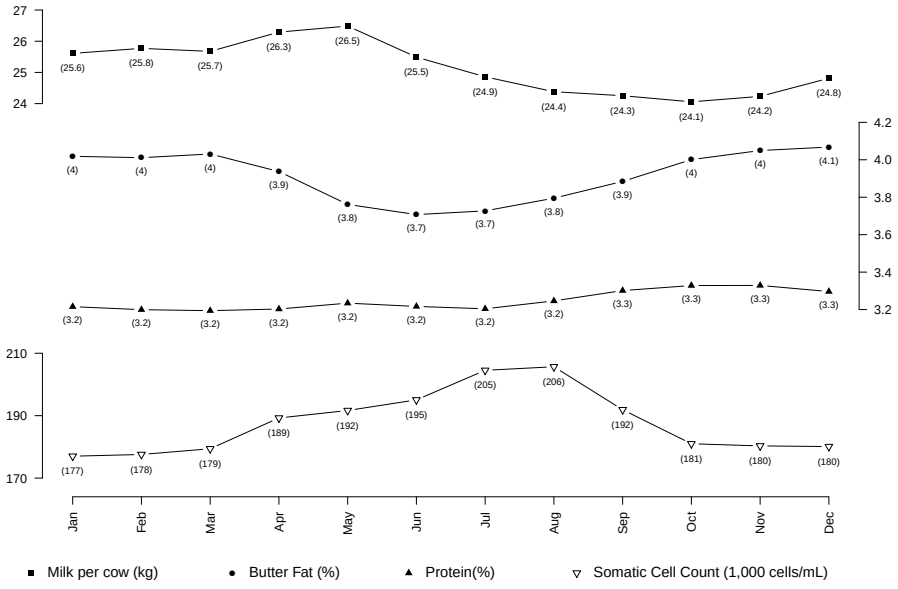
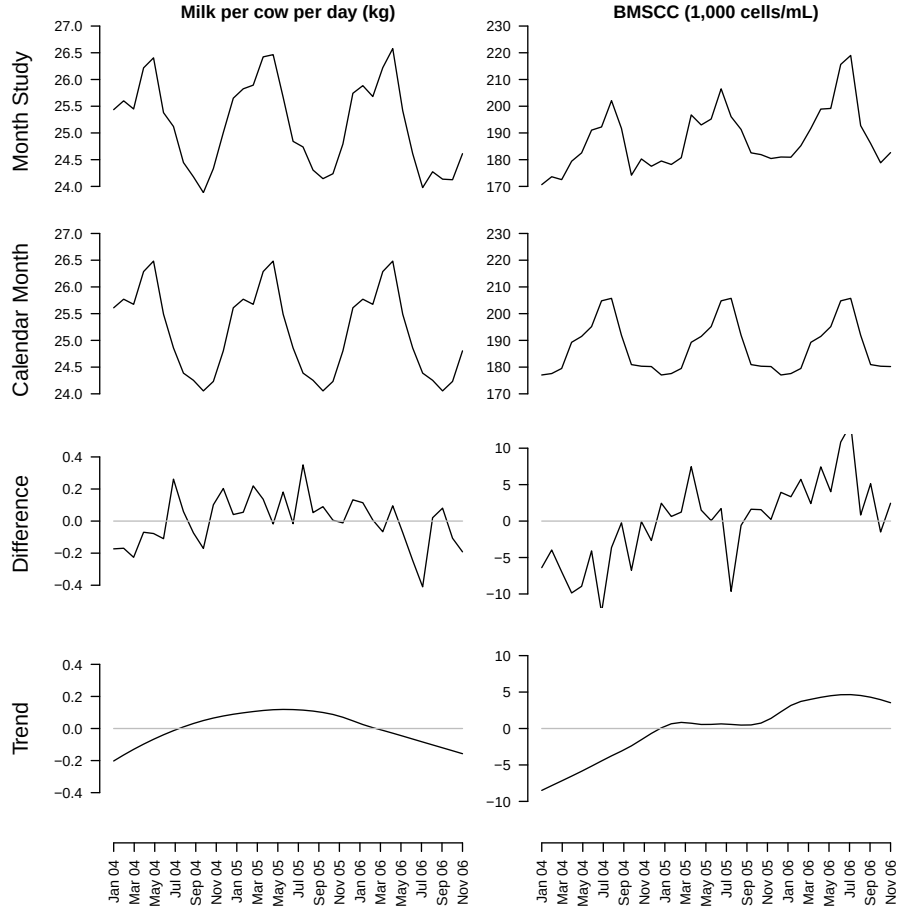


Figure 3.6: Evolution in the arithmetic mean milk yield per cow per day and the geometric mean estimated bulk milk somatic cell count between January 2004 and December 2006. *Month study* represents the mean per month of the study (January 2004 to December 2006). *Calendar month* represents the mean for January to December, regardless of the year. *Difference* is the difference between the means for *Month study* and *Calendar Month*. *Trend* is the *Difference* smoothed using a loess regression in order to identify trends in the variation between 2004 and 2006.



calvings. In the current context, this could involve delaying the first insemination from winter to summer for some of the heifers, although this may cause some management difficulties.

To our knowledge, this is the first study to use a nationwide milk recording dataset to investigate milk production characteristics. Given the worldwide availability of such data, it would be possible to replicate this analysis in other countries or in the UK at regular intervals. This would give interesting indications on spatial and time variations in milk production characteristics. Moreover, from reproduction and milk production parameters, it would be possible to simulate different scenarios to assess the effects of changing the calving patterns on the national or dairy plant level milk supply. This would be of interest in order to adapt the supply to the demand.

3.6 Conclusion

Calvings and milk production follow a seasonal pattern in England and Wales. There was 80 % more calvings in September than in May. Overall milk production increased from October to April and decreased from April to August. Herd size increased by 5.2 cows recorded and herd milk yield by 143 kg of milk between 2004 and 2006.

Chapter 4

Distributional Characteristics of Individual Cow Milk Yield and Constituents

4.1 Introduction

Lactation curves have been the object of extensive research, mainly to improve the determination of breeding values used in genetic selection (Silvestre *et al.* , 2006). The general shape and variance components of these curves for milk yield, fat and protein contents and somatic cell count are well characterised. Basically, after an initial rise from calving until the peak of lactation which happens around 50 days after calving, milk yield decreases towards the end of lactation. Fat and protein contents and somatic cell counts curves have an inverse shape decreasing between calving and lactation peak and increasing thereafter (Silvestre *et al.* , 2009; Caccamo *et al.* , 2008). A biological model for lactation curves fitted on UK data was proposed by Albarrán-Portillo & Pollott (2008), but this model analysed only variations in milk yield and its main goal was to derive variance components. Descriptions of current individual cow values for milk production in England and Wales are not available in the recent literature. These may vary between countries or between low production and high production level herds within England and Wales because of differences in climate, management, production systems and breeding programs.

Moreover, there has been interest for some time in the use of the fat to protein ratio as a measure of negative energy balance at the start of lactation. Grieve *et al.* (1986) found it was a better predictor of the energy status of the cow than either the percentage of fat or the percentage of protein on its own. Podpecan *et al.* (2008), using a sample of 51 high yielding dairy cows, identified a cut-off value for the fat to protein ratio of 1.34 applied to milk samples collected

between 75 and 90 days in milk to be optimal to predict whether a cow would have conceived by 120 days in milk and a cut-off value of 1.44 for the prediction of a conception before 140 days in milk. However, the lactation curve for this parameter has not been described.

Knowledge of current individual cow production characteristics is of interest to the dairy industry in order to put individual performance in context and to understand the main reasons for variation in milk production.

The purpose of this chapter is to describe the distributions in individual cow milk quantity and composition using a large cohort of UK dairy herds.

4.2 Materials

The selection and general description of the data analysed in this chapter are detailed in Chapter 2.

4.3 Methods

4.3.1 Milk quantity and milk constituents distributions

The distributions of cow milk yield, percentage of butterfat, percentage of protein and somatic cell count are described for the period 5 to 305 days in lactation for cows of different parities and for herds of different production levels. NMR does not collect data from cows prior to 5 days in milk.

4.3.2 Lactation curves

Lactation curves were plotted for different parities, months of calving and herd levels of production between 5 and 400 days after calving. The purpose of these graphs was to illustrate the mean of the parameter of interest each day after calving. An arithmetic mean was used for all parameters, except somatic cell count, because the underlying distributions were approximately Gaussian. The distribution of somatic cell count is reported to be right skewed (Ali & Shook, 1980). Thus, while most readings are below 200,000 cells/mL, values as high as 10 million cells/mL are recorded. The geometric mean was used for somatic cell count in order to give less weight to these extreme values. For the fat to protein ratio, the distribution of the percentage of recordings above 1.4 per week from calving was determined because this value has been recommended to monitor cows' energy status at the start of lactation in the UK (Cook *et al.* , 2006). For parity, five categories were initially considered i.e. parity one to four and

parity greater than four. Since the properties of parities greater than one were very similar, these parities were considered as one group when appropriate. Weight of butterfat and weight of protein were calculated by multiplying milk yield by the percentages of butterfat and protein respectively. For the purpose of identifying milk production and days in milk at the peak, lactation curves for milk yield were smoothed using local regression. This was done in R using the loess function (R Development Core Team, 2009; Cleveland & Devlin, 1988). The degree of smoothing can be controlled by changing the number of neighbouring points used to estimate each smoothed value. In R, this is controlled by the α parameter. Several values were tested for α, and the smoothed curves plotted against the observed data. A value of 0.1 for α was used. Persistency was defined as the slope of the milk yield lactation curve between 50 and 305 days in milk. It was estimated by fitting a straight line between the mean values for milk yield and days in milk using least square linear regression as implemented in the R lm function (R Development Core Team, 2009). The fit of the linear models were assessed using the R squared value. In order to determine the extent of the variations in milk quantity and composition with the month of the year, mean parameters were calculated per day in milk, per month of calving and for parity one and greater than one when relevant. These mean parameters were smoothed with local regression as described above and displayed graphically. This allowed a visual comparison of the impact of calendar month on cows at the same stage of production. Herd-year milk production was categorized according to the level of production as follows. A mean milk yield per recording between 5 and 305 days in milk was calculated for every herd in each of the years 2004, 2005 and 2006. The values of the 10^{th}, 25^{th} , 50^{th}, 75^{th} and 90^{th} percentiles of the distribution of this mean were determined and each herd-year categorized as belonging to the interval [0-10], [10-25], [25-50], [50-75], [75-90], [90-100]. Lactation curves were compared for each one of these six subsets.

4.3.3 Cumulative quantities

Cumulative quantities of milk, butterfat and protein were calculated over lactation from 5 days in milk onwards. For butterfat and protein, a weight in kg was calculated for each recording by multiplying the weight of milk by the percentage of each constituent divided by 100. These quantities were then averaged over the category under investigation and cumulated over the required stage of lactation.

4.4 Results

4.4.1 Distributions

The distributions of daily milk yield per cow, percentage of butterfat, percentage of protein, fat to protein ratio and somatic cell count between 5 and 305 days after calving are presented in Figure 4.1. The mean milk yield between 5 and 305 days was 26.4 kg; ninety percent of the readings were between 12.4 kg and 42.4 kg. The mean percentage of butterfat was 3.96 % and ninety percent of the readings were between 2.83 % and 5.18 %. For the percentage of protein, the mean was 3.29 % and ninety percent of the readings were 2.78 and 3.91 %. The mean fat to protein ratio was 1.21 with ninety percent of the readings between 0.89 and 1.54 and 14.2 % of the ratios were above 1.4. The somatic cell count distribution was highly right skewed. The arithmetic mean was 218,000 cells/mL, the geometric mean was 90,000 cells/mL, the median was 82,000 cells/mL. Thirty three percent of readings were below 50,000 cells/mL, 56 % were below 100,000 cells/mL, 76 % were below 200,000 cells/mL and 3.9 % were above one million.

4.4.2 Stage of lactation and Parity

Stage of lactation had a major impact on all variables. The effects of parity were pronounced for milk yield and somatic cell count while the variations exhibited for the percentage of butterfat and the percentage of protein were limited. For milk yield and somatic cell count, the curves were lower and flatter for parity one compared to other parities (Figures 4.2 and 4.3). For parity 1, 2, 3, 4 and greater than 4, the peak in mean milk yield occurred at days 51, 38, 39, 41 and 41, and was of 27.9, 33.8, 36.9, 37.6 and 36.0 kg, and, between 50 and 305 days in milk, milk production decreased by 3.2 kg, 5.7 kg, 7.0 kg, 7.4 kg and 7.2 kg per 100 days. For all linear models used to estimate these slopes, the R squared was greater than 0.99. The lowest concentrations were observed at around 50 days in milk for butterfat when the mean value was approximately 3.7 % and between 35 and 40 days in milk for protein when the mean value was close to 3 %. The fat to protein ratio peaked during the 4^{th} week after calving and decreased until the end of lactation. Percentiles of its distribution per week after calving for all parities are provided in Table 4.1. Butterfat and protein yields were highest at the start of lactation and did not decrease until the peak in milk yield. Cumulated milk production, weight of butterfat, and weight of protein were calculated separately for parity one and parities greater than one, and are given in Table 4.2. Between day 5 and 305 after calving, parity one cows produced on average 7,358 kg of milk, 284 kg of butterfat and 235 kg of protein. For the same period, parities greater than one produced 8,483 kg of milk, 327 kg of butterfat and 272 kg of protein.

Figure 4.1: Distribution of milk yield (kg), butterfat (%), protein (%), fat to protein ratio and somatic cell count (1,000 cells/mL) between 5 and 305 days after calving.

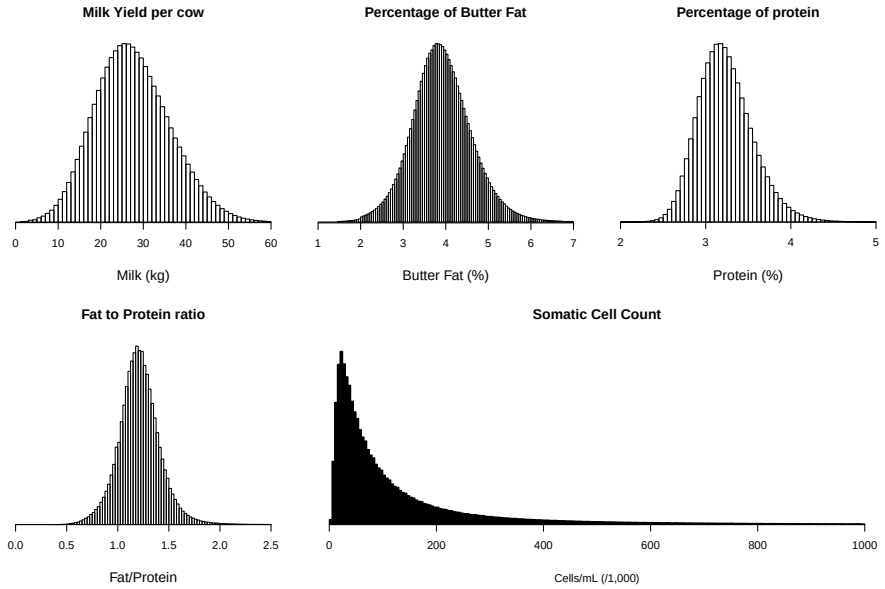


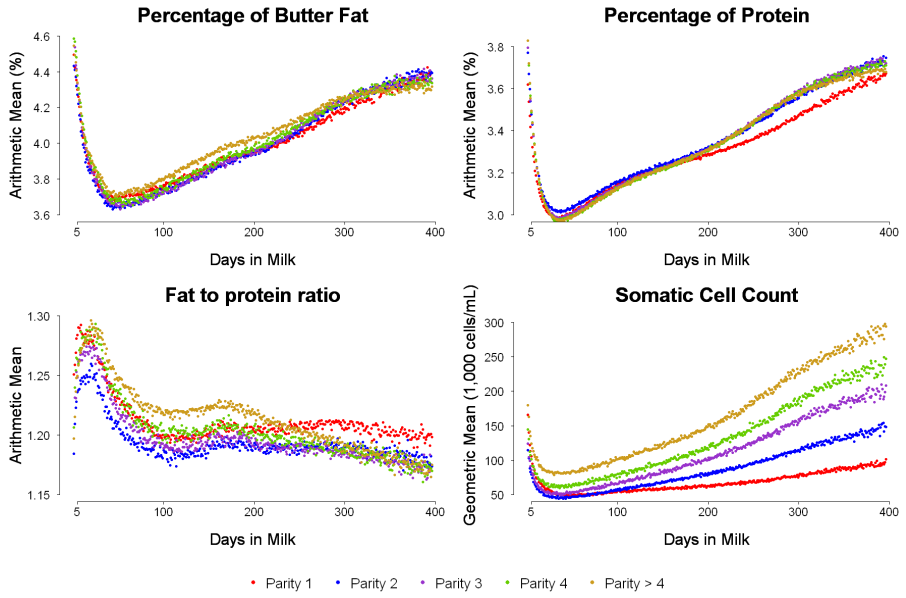
Table 4.1: Quantiles and percentage > 1.4 for the fat to protein ratio per week from calving

Weeks from	Quantile					
calving	10%	25%	50%	75%	90%	> 1.4 (%)
2	0.98	1.11	1.24	1.40	1.57	24.61
3	0.99	1.11	1.25	1.40	1.58	25.63
4	0.99	1.12	1.26	1.41	1.58	26.43
5	0.98	1.11	1.25	1.40	1.56	24.97
6	0.97	1.10	1.24	1.38	1.54	22.75
7	0.96	1.09	1.23	1.37	1.51	20.45
8	0.96	1.09	1.22	1.36	1.50	19.21
9	0.96	1.09	1.22	1.35	1.49	18.02
10	0.95	1.08	1.21	1.34	1.48	17.14
11	0.95	1.08	1.21	1.34	1.47	16.34
12	0.95	1.08	1.20	1.33	1.46	15.51

Table 4.2: Cumulated production from day 5 after calving (kg).

	Parity 1			Parity > 1		
Days in Milk	Milk	Fat	Protein	Milk	Fat	Protein
15	261.2	11.1	8.7	344.7	14.7	11.9
30	660.2	26.5	20.8	866.1	34.9	27.9
45	1075.5	41.9	33.2	1403.2	54.9	43.9
60	1492.9	57.3	45.7	1935.9	74.3	59.8
75	1905.0	72.4	58.2	2454.0	93.2	75.5
90	2311.4	87.4	70.8	2956.8	111.7	90.9
105	2709.3	102.2	83.2	3442.4	129.6	105.9
120	3101.5	116.9	95.5	3912.8	147.2	120.7
135	3486.6	131.4	107.8	4367.4	164.3	135.1
150	3865.9	145.8	119.9	4807.9	181.0	149.1
165	4238.5	160.1	131.9	5233.9	197.3	162.8
180	4605.1	174.2	143.7	5645.8	213.3	176.1
195	4964.7	188.2	155.4	6043.5	228.8	189.1
210	5318.3	202.0	167.0	6427.1	243.8	201.7
225	5664.3	215.6	178.4	6795.4	258.3	213.9
240	6002.3	228.9	189.6	7147.6	272.4	225.7
255	6331.4	242.0	200.5	7482.8	285.9	237.0
270	6650.2	254.8	211.3	7800.6	298.8	247.9
285	6958.9	267.3	221.7	8102.1	311.2	258.4
300	7259.1	279.6	232.0	8389.7	323.1	268.5
305	7357.6	283.7	235.4	8483.2	327.1	271.7

Figure 4.2: Effects of days in milk and parity on milk constituents concentration.



4.4.3 Stage of lactation and Month of calving

The impact of calving month was limited for milk yield and somatic cell count. It was more pronounced for the percentage of butterfat and the percentage of protein and was substantive for the fat to protein ratio. The effects of the month of calving on lactation curves for the mean percentage of butterfat, the mean percentage of protein and the mean fat to protein ratio are presented in Figure 4.4. Generally, cows tended to give more milk with a lower content of fat and protein from March to June. The opposite effect was observed from September to December.

4.4.4 Effect of herd production level

The characteristics of the six groups of herds categorized by milk per cow per year are presented in Table 4.3. As herd milk production went up, the mean number of cows per test-day was higher, from around 60 cows in group one to 110 cows in groups six. The means for butterfat percentage, protein percentage, and somatic cell count were lower as milk per cow increased. The difference between the mean for group one and six reached 0.07 % for butterfat and 0.08 % for protein. The estimated quantities produced between 5 and 305 days in milk were 155 kg of fat and 129 kg for protein higher in group six compared to group one. The mean somatic cell count was 28,600 cells/mL lower in group six than in group one. The lactation curves showed differences between groups. For milk yield, the ascending phase was longer from group one to six. In group one, after reaching the maximum production at 20 days in milk, milk yield levelled until day 50 while in group six milk yield increased until day 50 and started to decrease from day 60 (Figure 4.5). For parity one cows, the production decreased by 2.6 kg, 3.0 kg, 3.3 kg, 3.3 kg, 3.3 kg and 3.3 kg per 100 days in milk between day 50 and 305 after calving in group one to six respectively. For parity greater than one the corresponding figures were 4.8 kg, 5.7 kg, 6.4 kg, 7.0 kg, 7.6 kg, 8.1 kg. All the R^2 were greater than 0.96. The lactation curves were similar in shape for butterfat and protein (Figure 4.5). The lactation curves were equally spaced between yield groups for milk yield and somatic cell count, the difference was attenuated from the low yield groups to the high yield groups.

4.5 Discussion

Average milk production characteristics between 2004 and 2006 in this sample of cows from England and Wales was of 26.4 kg of milk per cow containing on average 3.96 % of butterfat and 3.29 % of protein.

Although the shape of the lactation curves for these constituents have been described previously,

Figure 4.3: Effects of days in milk and parity on milk constituents concentration.

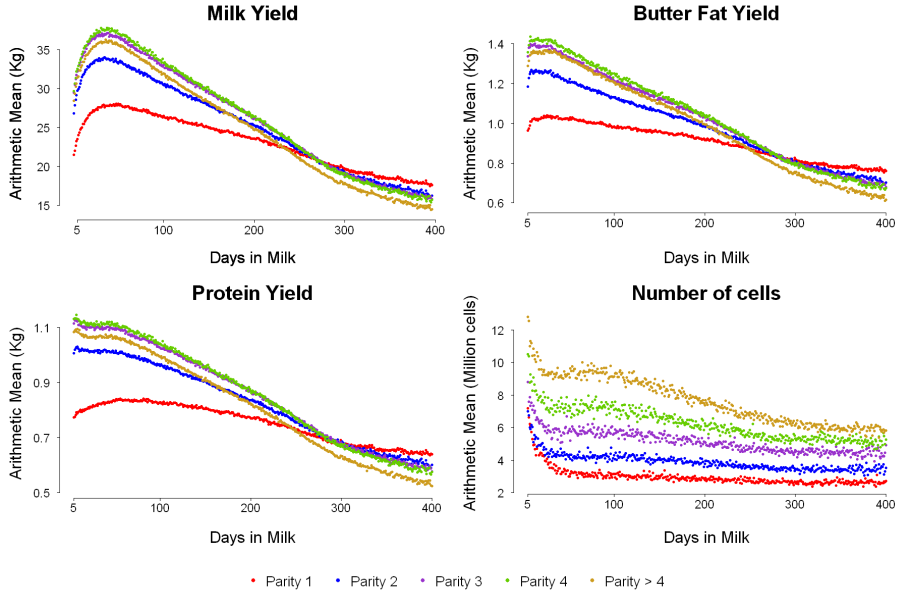


Figure 4.4: Variations in mean butterfat (%), protein (%) and fat to protein ratio with stage of lactation and calendar month.

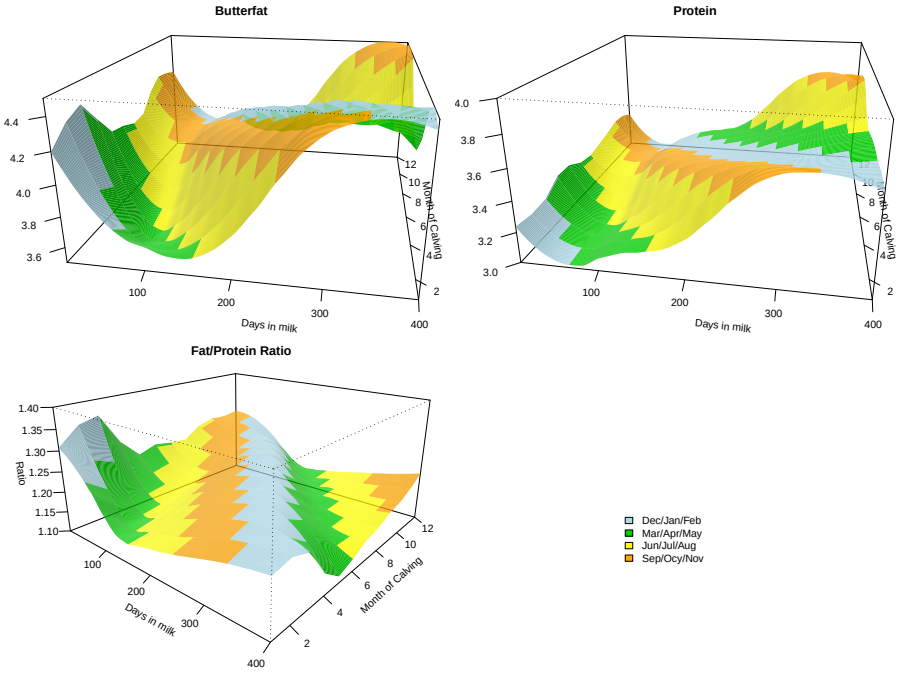


Table 4.3: Characteristics of the 6 herd production groups between 5 and 305 days in milk.

Group	Number	Cows/test		Milk/cow		Butterfat		Protein		Fat to protein		Cell count		Cumulated Production		
	of Herds	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Milk	Fat	Protein
1	638	61	18.7	19.7	6.4	3.97	0.72	3.27	0.36	1.22	0.20	104.2	3.3	5871	229	189
2	957	71	20.1	23.0	6.9	3.96	0.70	3.26	0.34	1.22	0.20	92.2	3.3	6843	267	221
3	1596	87	21.2	25.7	7.5	3.93	0.71	3.26	0.33	1.21	0.20	86.4	3.4	7650	296	246
4	1596	101	22.8	28.5	8.0	3.91	0.70	3.24	0.32	1.21	0.20	82.4	3.4	8468	327	272
5	957	108	23.5	30.9	8.5	3.89	0.71	3.22	0.31	1.21	0.21	78.6	3.4	9202	353	293
6	639	110	24.7	33.8	9.1	3.87	0.73	3.19	0.30	1.22	0.22	75.6	3.4	10081	384	318

values observed in the UK were not available. Milk quantity and composition were predominantly affected by stage of lactation. There were significant differences in the shape of the curves for milk yield and somatic cell count between first lactation cows and older cows. For milk yield, the peak occurred at around 50 days in milk and was 28 kg of milk while for later parities, the peak occurred at around 40 days in milk and was between 33 and 38 kg. After the peak, milk yield decreased by 3.2 kg of milk per 100 days in parity ones while this decrease was between 5.7 and 7.4 per 100 days in older cows. As a result, after 300 days in milk, primiparous cows were giving more milk than multiparous.

The lactation curve for milk yield for parity one cows was characterized by a lower peak and a higher persistency. Milk production depends on the number of milk producing cells as well as the activity of each one of these cells, which vary with lactation stage (Capuco *et al.* , 2003). Capuco *et al.* (2001) investigated mammary epithelial cell dynamic in multiparous cows between 14 days and 240 days in milk and reported that the increase in milk production they observed between 14 days and 90 days in milk was due to an increase in the activity per cell while the decline between 90 and 240 days in milk was caused by cell loss through apoptosis. Miller *et al.* (2006) hypothesized that the higher persistency in primiparous cows was the result of a higher rate of cell proliferation in the mammary gland mediated by mitogenic or survival factors. This and the fact that the curves for butterfat and protein concentration were similar across parities would suggest that the mammary gland develops during the first lactation. Curves for later parities were similar, and the quantities produced reached a maximum at parity four. A mathematical model based on biological assumptions was fitted on UK milk yield recording data by Albarrán-Portillo & Pollott (2008). They reported a peak yield of 28.7 kg at 34.6 days for parity one cows and a peak yield of 37.1 at 33.7 days in later parities. The values for the production at the peak were essentially similar to the present study, but the peak happened much earlier for parity one cows. This could be because the peak was longer and less pronounced for parity one cows so that milk production does not vary to a great extent between 40 and 50 days in milk.

The butterfat and protein concentrations were marginally affected by parity but appeared to be influenced to a greater extent by the month of the year. Butterfat was the most variable showing the lowest concentrations between March and June when, in the UK, most cows are at pasture and highest concentrations between October and February when most cows are fed indoor. The percentage of protein was lowest between May and July and the highest between September and December. The effect was less pronounced than for butterfat.

The fat to protein ratio is increasingly used as a marker of negative energy balance. A threshold of 1.4 during the first month of lactation is commonly used by veterinary practitioners as a marker of negative energy balance (Cook *et al.* , 2006) and the percentage of readings above

it at different stages of lactation were examined in this article. The effects of this ratio on reproductive performance have been recently studied by Podpecan *et al.* (2008). They used the value of the fat to protein ratio between 75 and 90 days in milk for the prediction of the calving to conception interval using fixed thresholds. The choice of this particular time frame is interesting since the ratio varies less during this interval than at the beginning of lactation. The lactation curves presented for this parameter suggest that it is greatly affected by stage of lactation especially around the start of lactation. This could reflect the changes in energy balance as lactation proceeds. Hence when looking at this parameter, lactation stage should be taken into account and more research is needed in this area (see Chapter 7). We suggest that using different thresholds according to stage of lactation may be worth consideration. The calculation of an average for all the cows at similar stage of lactation within a herd has been recommended by Cook *et al.* (2006) who state that when the percentage of cows with a ratio greater than 1.4 on the first test-day after calving exceeds 40, a ketosis problem is likely.

Comparing the low to the high yielding herd groups, curves for milk yield were characterized by a higher and longer ascending phase followed by a smaller persistency and there was a limited decrease in the percentages of fat and protein. Cows in high yielding herds are under a higher energy demand, especially during the first three months of lactation. This is compatible with the fat to protein ratio reflecting energy balance since it was higher at the beginning of lactation in high producing groups, then lower between 100 and 250 days in milk, after which the curves merged. This would indicate that higher yields are achieved despite higher energy deficits during the first three months of lactation. The effect of calendar month on the fat to protein ratio was large. Cows calving between December and February had higher curves and thus, possibly a greater energy deficit at the start of lactation. Since the fat content of milk is affected by diet, this could also be a consequence of the type of diet given during this period. Further studies are required to evaluate the joint effects of lactation stage, parity, energy balance and diet on the fat to protein ratio.

The mean milk production at 400 days in milk was still over 20 kg in high yielding herds and between 10 and 15 kg in low yielding herds even though the persistency was slightly higher in the low yielding group. Reproduction is an issue in modern dairy herds and the optimum interval between calvings is still debated (Arbel *et al.* , 2001). The lactation is interrupted in late pregnancy because the mammary gland needs a period of rest to maximise milk production in the subsequent lactation. Depending on herd production level, different objectives for the interval between calvings could be adopted, based on individual farm financial returns, if these can be accurately calculated. Moreover, energy deficit has been linked to longer calving to conception interval (Jorritsma *et al.* , 2000). Since cows in high yielding herds are under a greater energy deficit at the start of lactation, early services may be less likely to result in a pregnancy. More

studies are needed in the UK to investigate optimal interval lengths between calvings according to the herd mean milk production and their consequences on calving pattern. This could be done by simulating different management strategies for reproduction incorporating different lactation curves between and within herds based on herd production level and individual cow characteristics such as production at the peak.

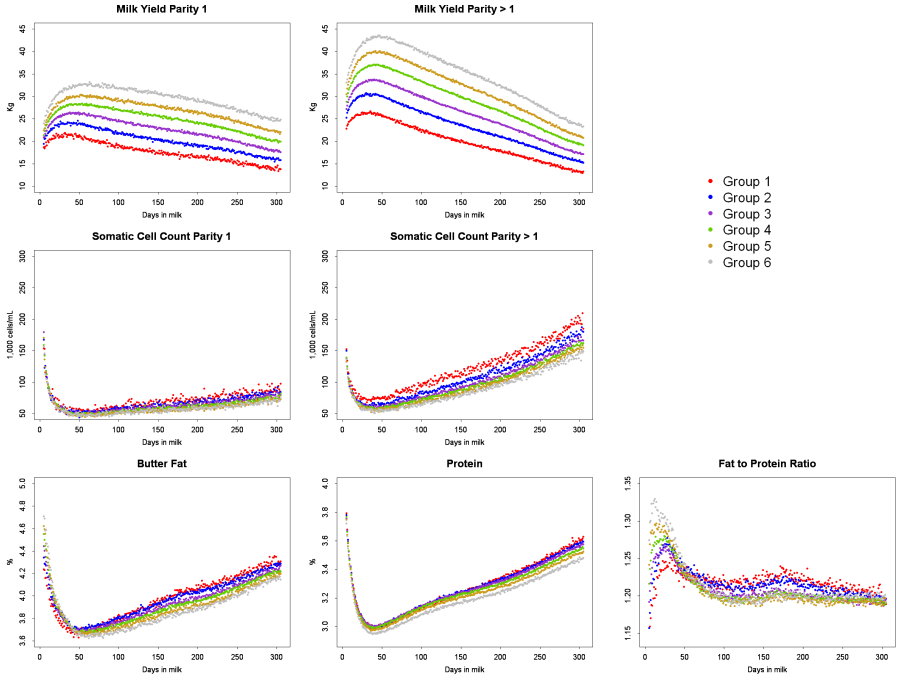
The lactation curve for somatic cell count was flatter and lower for parity one cows than for other parities. This is in agreement with Schepers *et al.* (1997). With successive parities, the curves had the same shape but on average tended to increase by a fixed amount. While other milk constituents are secreted by lactocytes, somatic cells are recruited from the bloodstream. This recruitment can be massive in cases of infection and infection is the main cause of somatic cell count increase (Schepers *et al.* , 1997). There is a big overlap in SCC between infected and uninfected cows and various thresholds are used to diagnose subclinical infections. No bacteriological data were available for this study so that the relative part played by infection and physiological processes cannot be determined. The similarities in the shapes of the curves for somatic cell counts and other milk constituents seem to indicate that there is a physiological increase in SCC over lactation. Green *et al.* (2006a) modelled the impact of milk yield on SCC and showed that there was an inverse linear relationship between SCC and milk yield suggesting that a dilution effect might mitigate SCC as milk yield goes up. It could explain partly the increase in SCC observed towards the end of lactation as milk yield decreases as well as the observed decrease in geometric mean somatic cell counts from low yielding to high yielding herds. Similarly, infection has been associated with decreased milk production (Hortet & Seegers, 1998) so that an increase in infection prevalence, and hence SCC, during lactation could result in lower milk yields. The reasons for the increase in SCC with parity are less clear. It could be due to a regular increase in the prevalence of intramammary infection with parity, more cells as a result of previous infections or undetermined physiological factors. The relative roles played by infection and dilution in milk yield and SCC variations according to stage of lactation and parity remain to be clearly quantified.

4.6 Conclusion

Milk quantity and composition are greatly affected by stage of lactation, calving month and cow parity and important patterns within England and Wales' dairy herds were highlighted. Whilst energy demands of dairy cows are high at the start of lactation the use of the fat to protein ratio to evaluate the energy balance requires clarification particularly with respect to the impact of lactation stage on the normal range of fat to protein ratio expected. Although large variations in herd production levels existed, the shapes of the different lactation curves

were similar in herds with different milk yields.

Figure 4.5: Lactation curves for milk yield, somatic cell count, percentage of butterfat, percentage of protein and fat to protein ratio in herds categorised according to their annual yield per cow. The geometric mean was used for somatic cell count and the arithmetic mean for the other parameters.



Chapter 5

Prediction of Bulk Milk Somatic Cell Counts from Cow Somatic Cell Count Categories I: Herd-Year Level

5.1 Introduction

Individual cow somatic cell counts (**SCC**) are widely used to identify cows likely to have an intramammary infection (Bradley & Green, 2005; Dohoo & Leslie, 1991). At the herd level, bulk milk somatic cell count (**BMSCC**) is used to estimate herd mastitis prevalence (Emanuelson & Funke, 1991). There is also ongoing motivation for farmers to control BMSCC because a penalty is applied on milk price when BMSCC gets above certain values, typically 200,000 cells/mL in the UK, and, in the European Union, milk is not saleable if the geometric mean BMSCC exceeds 400,000 cells/mL for more than two or three consecutive months, depending on the frequency of measurement (Council of the European Communities, 1992). Although the main factor increasing SCC is infection, it is also affected by other factors among which stage of lactation and parity (Schepers *et al.* , 1997; Brolund, 1985). SCC decreases between calving and the lactation peak and then increases towards the end of lactation, and, the lactation curve is lower and flatter in first lactation cows compared to older cows (Chapter 4). However, the use of a single threshold of 200,000 cells/mL regardless of parity and stage of lactation is commonly applied to categorise cows as uninfected or infected in the UK and used to diagnose herd problems (Bradley & Green, 2005). Dohoo & Leslie (1991) found this threshold to be the best to detect new infections and Dohoo & Morris (1993) used it to study infection prevalence and dynamics in Prince Edward Island dairy herds. The same threshold was used by Cook *et al.* (2002) to investigate infection status during the dry period. Pantoja *et al.* (2009) concluded that cows with SCC $\geq$ 200,000 cells/mL at both drying-off and first recording after calving

were 20.4 and 5.6 times more likely to be infected by a major pathogen or a minor pathogen respectively than being uninfected.

When having to devise and justify mastitis plans to farmers, 3 steps are necessary: (i) Define indices to look at ; (ii) Determine the distribution for these indices ; (iii) Estimate the gain resulting from improving herd performance in a given area. Applied to mastitis and SCC, a threshold of 200,000 cells/mL has the advantage of simplicity and movements across this threshold can reveal infection dynamics (Dohoo & Leslie, 1991). But, given the significantly lower SCC values observed in first lactation cows, it may be argued that a lower threshold may reflect better the prevalence of infections in primiparous cows. Milk recording databases contain a large amount of data which can be used to describe SCC at a population level. BMSCC is a readily available outcome, in which farmers have a strong interest, to quantify the impact of corrective action.

The purposes of this chapter were (i) using two thresholds of 100,000 cells/mL and 200,000 cells/mL on two consecutive monthly milk recordings, to determine which transitions were best able to discriminate between herds based on their impact on BMSCC (ii) to describe the variability in the percentage of herds undergoing these transitions in a large sample of dairy herds from England and Wales.

5.2 Materials

SCC was coded in thousand cells per mL and could take any integer value between 1 and 9,999 i.e. between 1,000 and 10 million cells/mL. In the NMR database, missing values are coded as 0 or NULL. There were 240,791 lines where SCC was equal to 0 or NULL which were not used in the analysis. It was not possible to know whether these values were missing at random. Test-days with more than 5 % of null or missing data and herd-years with less than ten test-days were excluded from further analysis. For this analysis, 7,770,956 recordings from 2,128 herds were available.

5.3 Methods

5.3.1 Categories of SCC levels and transitions

Each SCC record was categorized with respect to lactation number as follows: Recordings in parity one cows were coded as *Heifers* and recordings in later parities were coded as *Cows*. Two categories of recordings reflecting stage of lactation were created: *Calving* for SCC readings in the first 29 days after calving and *Lactating* for later recordings. These were combined to

create four categories: *Heifers Calving* **(Hc)**, *Heifers Lactating* **(Hl)**, *Cows Calving* **(Cc)** and *Cows Lactating* **(Cl)** (Table 5.1). Three categories measuring the level of SCC were used. SCC were labelled as Low (**L**), Medium (**M**) and High (**H**) when below 100,000 cells/mL, between 100,000 and 200,000 cells/mL and above 200,000 cells/mL respectively as used by Green *et al.* (2006b). When at least two recordings were available for a cow, the change in the level of SCC between consecutive recordings was monitored. Hence, 9 transitions between the three levels of SCC were possible (Table 5.2) for *Cows Calving*, *Heifers Lactating* and *Cows Lactating*. *Heifers Calving* only had one recording and were classified as either Low, Medium or High. Herd SCC patterns vary between consecutive test-days (Dohoo & Morris, 1993) and between calendar month (Green *et al.* , 2006b). These sources of variation were not of interest for this analysis and the number of cows in each category were summed over a year for each herd. Thus for each herd-year, the numbers of recordings in each of the 30 categories were available.

5.3.2 Estimation of Bulk Milk Somatic Cell Counts

Herd-year BMSCC were estimated as follows. The number of cells contributed by a cow was estimated by multiplying the cell count by the weight of milk on a given recording date. These numbers were added for all the cows of a herd-year and divided by the sum of the weights of milk produced by these cows. BMSCCs are presented in thousand cells/mL.

5.3.3 Statistical Analysis

Statistical models were constructed with the aim of identifying the impact of the percentages of recordings in various combinations of the categories defined in Section 5.3.1 on the estimated BMSCC. That is, an evaluation was made of which SCC categories most influenced herd-year BMSCC. Seven competing models were tested. In all cases, the model specifications were:

$$\begin{aligned} BMSCC_{ij} &= \alpha + \Sigma X_{ij}\beta^T + u_j + e_{ij} \\ u_j &\sim N(0, \sigma_j^2) \\ e_{ij} &\sim N(0, \sigma_{ij}^2) \end{aligned} \tag{5.1}$$

where the subscripts i and j denote the i^{th} year in the j^{th} herd, α the regression intercept, X_{ij} the covariates relating to a herd-year, β^T the vector of coefficients for covariates X_{ij}, u_j the herd residuals and e_{ij} the herd-year residuals.

In models 1 to 4 the percentages of herd-year recordings in the categories defined in Section 5.3.1 were considered. Depending on the model, the Medium category was grouped either with the Low or the High category for a given group of cows. For example, when a single threshold of 200,000 cells/mL was considered, the numbers of cows in the Low and Medium SCC categories

were added to define a new Low category. In models 5 to 7, new variables were created. These variables were aimed at measuring the proportion of a cow category moving from a Low to a High (**Rais**) or a High to a Low (**Dimin**) level of SCC as a percentage of the number of cows eligible for this movement. In terms of the thresholds used, Model 5, Model 6 and Model 7 corresponded to Model 2, Model 3 and Model 4 respectively. The 7 models are explained below and presented in Table 5.3.

- **Model 1** contained all 30 categories (3 for Hc, 9 each for Cc, Hl and Cl). It was the full model against which the other simpler models were compared.
- **Model 2 and 5** considered a single threshold of 200,000 cells/mL for all 4 cow categories, to define Low and High categories.
- **Model 3 and 6** considered a threshold of 100,000 for *Heifers lactating* and a threshold of 200,000 cells/mL for all other cow categories.
- **Model 4 and 7** considered a single threshold of 200,000 cells/mL and the cow categories *Heifers lactating* and *Cows lactating* were grouped together.

Models were initially fitted by Restricted Maximum Likelihood and explored in R (R Development Core Team, 2009) using the lmer function from the lme4 package (Bates & Maechler, 2009). The same models were then estimated in a Bayesian framework using Markov Chain Monte Carlo as implemented in WinBUGS (Lunn *et al.* , 2000). An example of WinBUGS model is presented in Appendix A. One thousand five hundred of the 2,128 herds were randomly selected for parameter estimation. In Bayesian inference, prior information is incorporated in the models. In this case, no prior knowledge was assumed and vague priors of mean 0 and variance 1,000 were put on the models' parameters. For all models, 3 chains were run for 30,000 iterations and the first 20,000 discarded. One iteration in 100 was stored for further analysis. Thus, for each variable in the model, 300 values from the posterior distribution were available.

5.3.4 Models checking

Different procedures were applied for model checking, as follows:

Deviance Information Criterion

The Deviance Information Criterion (**DIC**) was used to assess model fit (Spiegelhalter *et al.* , 2002). Lower values of DIC indicate models that fit the data better. A difference of 10 is usually considered to discriminate well between models.

Table 5.1: Definition of cow categories.

Cow Category	Days in Milk	Parity	Abbreviation
Heifers Calving	< 30	1	Hc
Heifers Lactating	≥ 30	1	Hl
Cows Calving	< 30	> 1	Cc
Cows Lactating	≥ 30	> 1	Cl

Table 5.2: Definition of the three SCC levels Low (L), Medium (M) and High (H) and transitions between these levels between consecutive recordings. These are combined with the cow categories defined in Table 5.1. Heifers Calving can only be Low, Medium or High since only one recording is available for these animals.

		SCC 2		
		$< 100{,}000$	$100{,}000 \geq \geq 200{,}000$	$> 200{,}000$
SCC 1	$< 100{,}000$	LL	LM	LH
	$100{,}000 \geq \geq 200{,}000$	ML	MM	MH
	$> 200{,}000$	HL	HM	HH

Table 5.3: Description of the 7 models used to describe the association between herd-year SCC patterns and BMSCC.

Model	Cow Catgeories	SCC Levels	Transitions
1	Hc, Cc, Hl, Cl	L: < 100 ; M: 100-200; H: > 200	LL, LM, LH, ML, MM, MH HL, HM, HH
2	Hc, Cc, Hl, Cl	L: ≤ 200 ; H: > 200	LL, LH, HL, HH
3	Hl	L: < 100 ; H: ≥ 100	
	Hc, Cc, Cl	L: ≤ 200 ; H: > 200	
4	Hc, Cc, l:(Hl + Cl)	L: ≤ 200 ; H: > 200	
5	Hc, Cc, Hl, Cl	L: ≤ 200 ; H: > 200	Rais $= \frac{LH}{LL+LH}$ Dimin $= \frac{HL}{HL+HH}$
6	Hl	L: < 100 ; H: ≥ 100	
	Hc, Cc, Cl	L: ≤ 200 ; H: > 200	
7	Hc, Cc, l:(Hl + Cl)	L: ≤ 200 ; H: > 200	

Summary of the posterior distribution

Median and 95 % credible intervals were calculated from the posterior distribution. The 95 % credibility interval of the posterior distribution was the interval between the 2.5^{th} and the 97.5^{th} percentiles for the posterior distribution.

Predictions

Posterior predictions and cross-validation predictions were generated as follows. At each iteration of the MCMC algorithm, each BMSCC used for parameter estimation was predicted back (Posterior prediction - Training data) as well as the BMSCC in the remaining 628 herds (Cross validation - Validation data). In the case of the validation data, at each iteration, a herd effect was generated from the herd random effects distribution (Normal(0, σ_u)) and combined with the parameter values to predict BMSCC. For each predicted BMSCC, residuals were computed by subtracting the median of the prediction from the observed value. The 95 % credibility interval of the prediction was the interval between the 2.5^{th} and the 97.5^{th} percentiles for the predicted values.

Sensitivity and specificity of the predictions

Finally, the sensitivity and specificity of the models' predictions were assessed as follows. Calculated and predicted BMSCC were categorised as below or above 200,000 cells/mL. True positives were BMSCC $>$ 200,000 cells/mL for which the median value predicted by the model was $>$ 200,000 cells/mL. True negatives were BMSCC $\leq$ 200,000 cells/mL for which the median predicted value was $\leq$ 200,000 cells/mL. Sensitivity was the percentage of BMSCC $>$ 200,000 which were true positives. The specificity was the percentage of BMSCC $\leq$ 200,000 which were true negatives.

5.4 Results

5.4.1 Somatic Cell Count distributions

Somatic cell count distributions are shown in Figure 5.1. The number of SCC readings per class increased steeply between 1 and 22,000 and decreased gradually afterwards. The distribution tail reached 10 million cells per mL which is the maximum value recordable. The highly right skewed distribution explains the difference between the arithmetic mean (223,200) and the geometric mean (93,900). Fifty five percent of SCC were below 100,000 cells/mL ; 75 % below 200,000 ; 11.9 % above 400,000 and 5.2 % above 800,000. The test-day BMSCC distribution

was narrower with a maximum of 1,621,000 cells/ml. The difference between the arithmetic mean (205,500) and the geometric mean (187,600) was also reduced and 8.1 % of test-days had a BMSCC below 100,000 ; 54.3% below 200,000 and 3.7 % above 400,000 cells/mL. The herd-year SCC was more limited in range and approximated to a normal distribution. The arithmetic mean was 205,900 ; the geometric mean 194,500 and 3.4 % of herd-years were below 100,000 ; 51.3% below 200,000 and 1.4% greater than 400,000 cells/mL.

5.4.2 Somatic Cell Count patterns

The individual cow and herd-year characteristics associated with each cow SCC category are presented in Tables 5.4 and 5.5. *Cows lactating*, *Heifers lactating*, *Cows calving* and *Heifers calving* represented 71.8 %, 21.3 %, 5.1 % and 1.8 % of recordings. During lactation, the patterns were different between primiparous and multiparous cows. Cows remaining Low for two consecutive recordings were the most prevalent categories, they represented 59.9 % of heifers and 41.4 % of cows. The percentage of cows in the L and M categories becoming H was 7.3 % for parity ones and 13.8 % for later parities. The percentage of cows in the H becoming L or M was 48.2 % for parity ones and 27.4 % for later parities. The percentage of cows in the H category went from 41.9 to 22.1 between the last recording before the dry period and the first recording after. That is 71.4 % of the cows which were High before the dry period were Low or Medium after this period. Sixty five percent of heifers calved in the L ($<$ 100,000) category and 18.9 % in the H ($>$ 200,000) one. At the herd-year level, there were large variations between herds. The intervals between the 10^{th} and 90^{th} percentiles were of 42.2 % for the percentage of heifers calving L, 36.2 % for the Heifers lactating staying L, 35.8 % for the Cows lactating staying L and 32.1 % for the Cows calving staying L.

5.4.3 Models

Model fit was generally good. Between 97 and 98 % of predicted values were in the 95 % credibility interval for the training dataset and between 94.5 and 95.5 % for the validation dataset (Table 5.6). The distribution and standard deviation of the residuals for models 2 to 4 compared to model 1 are presented in Figure 5.2. As expected, the residuals' standard deviation was higher for the validation dataset than for the training dataset. When modelling the association between the percentage of herd-year recordings in each cow SCC category and BMSCC, the model with the lowest DIC was the full model (Model 1) and was followed by the model with a single threshold of 200,000 for cows and heifers. Having a different threshold for *Cows lactating* and *Heifers lactating* resulted in poorer model fit. The coefficients and credibility intervals for model 1 and 2 are presented in Table 5.7 and 5.8. The coefficients represent the

Figure 5.1: Somatic Cell Count distribution at the Cow, test-day and herd-year levels.

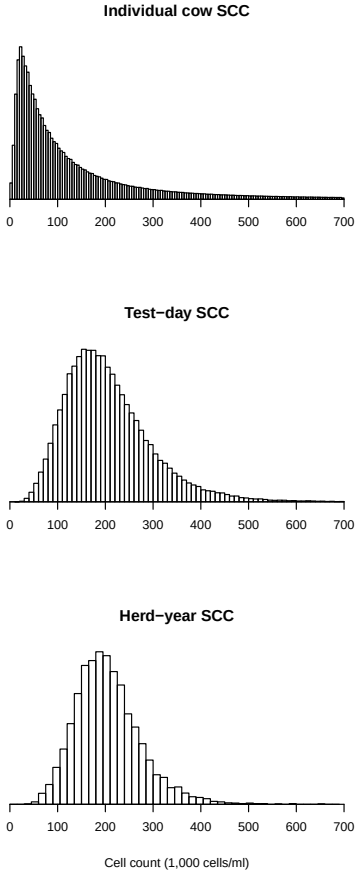


Table 5.4: Individual cow characteristics associated with each cow SCC category as defined in Table 5.1 and 5.2.

		SCC [a]		Milk [b]	
Category [c]	n [d]	Mean	sd	Mean	sd
Hc.L	85755	39.4	1.8	25.9	5.6
Hc.M	21486	136.2	1.2	24.4	5.9
Hc.H	24881	553.0	2.3	23.2	6.3
Hl.LL	950329	36.6	1.8	24.1	6.5
Hl.LM	127859	130.7	1.2	21.7	7.0
Hl.LH	48636	408.5	2.1	21.7	7.5
Hl.ML	107841	54.7	1.7	23.2	6.8
Hl.MM	104331	138.8	1.2	20.8	6.9
Hl.MH	53037	339.4	1.8	19.9	7.6
Hl.HL	49356	47.8	1.7	24.4	6.8
Hl.HM	45679	145.5	1.2	21.6	7.1
Hl.HH	102061	479.4	2.0	20.6	7.5
Cc.LL	88658	30.2	2.0	34.2	7.5
Cc.LM	14385	136.7	1.2	32.4	8.0
Cc.LH	17930	560.5	2.3	31.7	8.2
Cc.ML	64928	34.0	1.9	34.7	7.6
Cc.MM	14794	137.6	1.2	32.6	8.2
Cc.MH	20857	567.5	2.3	31.7	8.3
Cc.HL	88219	36.3	1.9	34.6	7.8
Cc.HM	26054	139.1	1.2	32.5	8.1
Cc.HH	45687	619.4	2.4	31.2	8.4
Cl.LL	2205414	37.7	1.9	29.4	8.8
Cl.LM	459327	132.1	1.2	24.9	8.8
Cl.LH	200056	419.1	2.1	25.4	10.0
Cl.ML	289054	59.0	1.7	27.2	9.1
Cl.MM	479389	142.4	1.2	22.7	8.3
Cl.MH	349287	346.4	1.8	21.1	9.4
Cl.HL	141705	51.2	1.8	29.3	9.4
Cl.HM	235646	149.7	1.2	24.2	9.2
Cl.HH	997930	530.6	2.1	21.4	9.7

[a] Geometric mean and standard deviation. In 1,000 cells/mL.
[b] Arithmetic mean and standard deviation.
[c] See Tables 5.1 and 5.2.
[d] Number of observations

Table 5.5: Herd year characteristics associated with each cow SCC category as defined in Table 5.1 and 5.2. Figures are presented as percentages of cow categories.

		Percentiles				
Category [a]	Mean	10	25	50	75	90
Hc.L	64.5	43.5	54.5	66.7	75.8	85.7
Hc.M	16.1	0.0	8.8	15.0	22.2	29.6
Hc.H	18.5	0.0	9.7	16.7	25.0	33.3
Hl.LL	59.9	41.1	51.1	60.9	70.1	77.3
Hl.LM	8.0	4.4	6.0	7.8	9.8	11.7
Hl.LH	3.1	1.0	1.8	2.8	4.0	5.4
Hl.ML	6.7	3.6	5.0	6.6	8.3	10.0
Hl.MM	6.6	2.1	3.7	6.0	8.7	11.6
Hl.MH	3.4	0.9	1.9	3.0	4.5	6.1
Hl.HL	3.1	1.1	1.9	2.9	4.0	5.3
Hl.HM	2.9	0.7	1.5	2.6	3.8	5.3
Hl.HH	6.4	1.1	2.8	5.3	8.8	12.6
Cc.LL	23.4	8.3	14.3	21.9	31.0	40.5
Cc.LM	3.7	0.0	1.4	3.2	5.6	7.8
Cc.LH	4.6	0.0	2.0	4.0	6.7	9.6
Cc.ML	17.0	9.1	12.6	16.7	21.1	25.5
Cc.MM	3.9	0.0	1.7	3.4	5.6	8.0
Cc.MH	5.4	0.0	2.8	5.0	7.6	10.5
Cc.HL	22.9	10.9	15.9	22.1	29.2	36.1
Cc.HM	6.9	1.4	3.4	6.2	9.5	12.9
Cc.HH	12.0	3.3	6.5	10.9	16.2	22.2
Cl.LL	41.4	23.8	31.9	40.9	50.9	59.6
Cl.LM	8.5	6.7	7.6	8.5	9.5	10.4
Cl.LH	3.7	2.3	2.9	3.5	4.3	5.2
Cl.ML	5.4	3.8	4.5	5.3	6.1	7.0
Cl.MM	9.0	5.4	6.9	8.8	10.8	12.8
Cl.MH	6.5	3.7	5.0	6.5	7.9	9.2
Cl.HL	2.6	1.4	1.9	2.5	3.2	3.9
Cl.HM	4.4	2.5	3.3	4.2	5.3	6.4
Cl.HH	18.6	8.6	12.5	17.6	23.3	29.5

[a] See Tables 5.1 and 5.2.

increase in BMSCC per unit of increase in the percentage of a particular cow category at the herd year level. Hence, it represents the contribution of each cow category to BMSCC.

Model 1 (Table 5.7) considered the three SCC levels L, M and H for all the cow categories. Both the highest and lowest coefficients were associated with cows calving. After *Cows lactating* LL, the cow category with the greatest number of cows was *Cows lactating* HH with a mean of 13.4 % of herd-year recordings. For each percent of increase in the number of cows in this category, the model predicted an increase of 7,890 cells/mL in BMSCC making these cows the largest contributors to BMSCC. Each percent of increase in the *Cows lactating* HL resulted in a predicted increase in BMSCC of 5,630 cells/mL when the mean SCC in this category was 51,200 cells/mL (Table 5.4). This was because herds with a high proportion of cows H also had a high proportion of cows moving from H to L. This effect was also observed in Model 2 (Table 5.8).

In Model 5 to 7 the impact of the variations in risks of going from a Low level of SCC to a High level (Rais) or from a High level to a Low level (Dimin) were modelled. In this case, considering a single threshold of 200,000 cells/mL and grouping *Heifers lactating* and *Cows lactating* resulted in the lowest DIC (Model 7 - Table 5.6). The results of the corresponding model are presented in Table 5.9. In these models, the intercept was higher than in the herd-year percentage models (Models 1 to 4) for which the intercept were the *Lactating cows* staying L. Each percent of increase in the percentage of *Rais* had a greater impact on BMSCC than an identical decrease in *Dimin*.

5.5 Discussion

The dynamics of SCC across a threshold of 200,000 cells/mL between consecutive recordings predicted BMSCC accurately when aggregated at the herd-year level. Using an alternative threshold of 100,000 cells/mL for primiparous cows during lactation did not improve the prediction. Considering cows moving from a low to a high SCC level as a percentage of cows initially low or the cows moving from a high to a low level as a percentage of the cows initially high, resulted in poorer model fit and ability to predict than considering the percentage of the herd in a SCC category. These figures can be used to compare herds as a proxy for the proportion of cows becoming infected or clearing infection. Figures summarising the variability in the indices which are of practical interest are grouped in Table 5.10. This table presents the herd-year means, percentiles 10, 25, 50, 75 and 90 for the percentages of cows below 200,000 cells/mL and for the various transition across this threshold tested for the categories *Heifers calving*, *Heifers lactating*, *Cows calving* and *Cows lactating*.

The characteristics of the best performing herds can be considered as realistic targets and are

Table 5.6: Model checking for models 1 to 7. For each model, the Deviance Information Criterion (DIC) was computed. The percentage of observed BMSCC in the 95 % credibility interval provided by the models were calculated, and the abilities of the models to classifiy BMSCC regarding a thresholds of 200,000 cells/mL were assessed by calculating the sensitivities (Se) and specificities (Sp).

Models		Training Dataset			Validation Dataset		
	DIC	% Predicted in 95 % CI	Se	Sp	% Predicted in 95 % CI	Se	Sp
1	41022	97.4	93.0	90.9	95.1	89.8	86.5
2	41171	97.7	92.5	91.5	94.7	86.5	86.6
3	41317	97.5	92.7	90.7	95.0	87.1	86.3
4	41220	97.5	93.1	91.3	95.2	87.1	86.4
5	42629	97.7	92.2	89.9	95.4	84.6	83.5
6	42724	97.6	92.0	89.7	95.5	84.8	83.5
7	42507	97.8	92.7	90.1	95.3	85.6	83.8

Figure 5.2: Density of the residuals computed as median predicted value - observed value for models 2 to 7 compared to model 1. Solid lines are predictions from the training data, dotted lines are predictions from the validation data. Lines representing model 1 are in grey. Lines for models to which it is compared are in black.

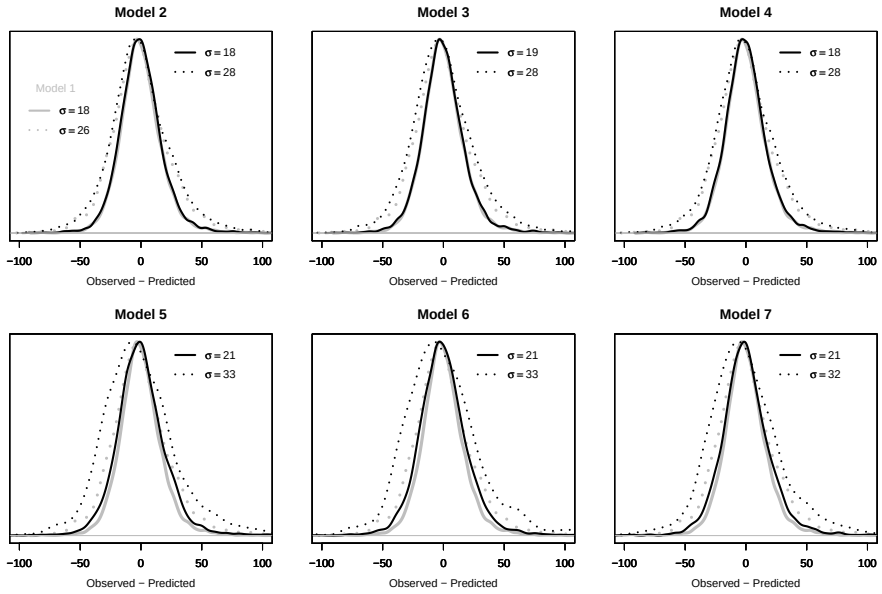


Table 5.7: Model 1: Association between the percentage of recordings in each level of SCC and BMSCC - Full model.

		95 % Credibility Interval	
Variable [a] [b]	Coefficient	2.5 %	97.5 %
Intercept	17.98	7.74	26.55
Hc.L	-1.67	-2.91	-0.16
Hc.M	2.63	-0.75	5.56
Hc.H	9.61	6.60	12.78
Cc.LL	-0.12	-1.71	1.88
Cc.LM	1.64	-1.81	5.51
Cc.LH	13.57	10.26	16.72
Cc.ML	-1.60	-3.59	0.31
Cc.MM	-1.33	-4.94	2.09
Cc.MH	11.61	8.39	14.63
Cc.HL	-7.14	-8.68	-5.58
Cc.HM	0.06	-2.79	3.15
Cc.HH	19.43	17.55	21.73
Hl.LL	0.13	-0.09	0.39
Hl.LM	0.60	-1.31	2.31
Hl.LH	7.03	4.40	9.65
Hl.ML	1.34	-0.77	3.32
Hl.MM	-0.19	-1.39	1.12
Hl.MH	2.52	-0.11	5.17
Hl.HL	4.91	2.45	7.80
Hl.HM	4.30	1.25	6.81
Hl.HH	6.22	5.24	7.08
Cl.LM	2.58	1.54	3.75
Cl.LH	6.26	5.18	7.48
Cl.ML	1.41	0.16	2.67
Cl.MM	-0.75	-1.30	-0.25
Cl.MH	-1.06	-2.28	0.28
Cl.HL	5.63	4.29	6.99
Cl.HM	2.58	1.22	3.97
Cl.HH	7.89	7.64	8.07
σ_u	17.00	16.00	18.06
σ_e	20.68	20.18	21.16

[a] See Table 5.1 for definition of cow categories.
[b] See Table 5.2 for definition of inflammation categories.

Table 5.8: Model 2: Association between the percentage of recordings in each level of SCC and BMSCC.

		95 % Credibility Interval	
Variable [a] [b]	Coefficient	2.5 %	97.5 %
Intercept	39.23	32.24	46.53
Hc.L	-0.98	-2.20	0.20
Hc.H	8.99	5.91	12.27
Cc.LL	-0.24	-1.29	0.68
Cc.LH	16.16	13.49	18.61
Cc.HL	-6.72	-7.85	-5.40
Cc.HH	19.20	17.10	21.37
Hl.LL	-0.08	-0.27	0.11
Hl.LH	5.54	3.44	7.78
Hl.HL	6.19	4.03	8.38
Hl.HH	5.12	4.13	6.02
Cl.LH	2.03	1.28	2.93
Cl.HL	4.91	4.00	5.79
Cl.HH	7.09	6.89	7.29
σ_u	18.79	17.83	19.89
σ_e	20.88	20.37	21.44

[a] See Table 5.1 for definition of cow categories.
[b] L:$\leq$ 200 and H:>200 (Table 5.3).

Table 5.9: Model 7: Association between the dynamics of SCC and BMSCC.

		95 % Credibility Interval	
Variable [a] [b]	Coefficient	2.5 %	97.5 %
Intercept	161.50	153.79	167.45
Hc.Rais	0.29	0.23	0.35
Cc.Rais	0.82	0.73	0.91
Cc.Dimin	-0.54	-0.60	-0.48
l.Rais	9.31	9.07	9.58
l.Dimin	-1.68	-1.79	-1.57
σ_u	24.12	22.87	25.32
σ_e	24.11	23.49	24.69

[a] See Table 5.1 for definition of cow categories. l: Hl + Cl.
[b] $Rais = \frac{LH}{LL+LH}$; $Dimin = \frac{HL}{HL+HH}$ with L:$\leq$ 200 and H:>200 (Table 5.3).

Table 5.10: Herd-year means and percentiles 10, 25, 50, 75 and 90 for the variables describing movements across a threshold of 200,000 cells/mL.

		Percentiles				
Variables[a b c]	Mean	10	25	50	75	90
Hc.L	81.3	66.7	75.0	82.8	90.0	100.0
Cc.L	77.9	65.0	72.2	78.8	84.8	89.6
Hl.L	87.2	78.1	83.4	88.3	92.3	95.2
Cl.L	71.3	58.1	65.0	71.9	78.6	83.9
Cc.LL	48.0	29.2	37.9	47.8	58.1	67.2
Cc.LH	10.1	3.4	6.1	9.4	13.3	17.6
Cc.HL	29.8	15.2	21.7	29.1	37.5	45.5
Cc.HH	12.1	3.3	6.5	10.9	16.2	22.2
Hl.LL	81.2	69.2	76.2	82.5	87.8	91.8
Hl.LH	6.4	2.8	4.2	6.0	8.1	10.5
Hl.HL	6.0	2.7	4.0	5.6	7.5	9.6
Hl.HH	6.4	1.1	2.8	5.4	8.8	12.6
Cl.LL	64.3	49.2	57.0	64.9	72.5	78.7
Cl.LH	10.2	6.8	8.4	10.2	11.9	13.4
Cl.HL	7.0	4.5	5.6	6.8	8.2	9.5
Cl.HH	18.6	8.6	12.5	17.6	23.3	29.5
Cc.Rais	17.9	6.4	11.1	16.7	23.3	31.2
Cc.Dimin	72.0	54.3	63.3	72.7	81.5	88.9
Hl.Rais	7.6	3.0	4.6	6.8	9.6	12.9
Hl.Dimin	53.6	31.2	40.0	51.0	65.2	80.0
Cl.Rais	14.3	8.0	10.5	13.7	17.2	21.0
Cl.Dimin	29.4	19.4	23.3	28.2	34.1	40.8

[a] See Table 5.1 for definition of cow categories.
[b] L: $\leq$ 200,000 cells/mL ; H: $>$ 200,000.
[c] The definition of *Rais* and *Dimin* are similar to the ones used in Model 5 (Table 5.3).

therefore helpful for monitoring schemes in health and production management. The percentage of cows with a SCC greater than 200,000 cells/mL across a herd-year was 25 %, but large variations were present between herds. In the 10 % best herds, the percentage of readings above 200,000 cells/mL during lactation were 4.8 % for parity ones and 16.1 % for later parities. The same figures were 21.9 % and 41.9 % for herds at the 90^{th} percentile. In a similar study in Prince Edward Island, Dohoo & Morris (1993) observed a mean percentage of recordings greater than 200,000 cells/mL of 29 %. The distribution of SCC in that study had a longer tail, with 10 % of the herd-years having more than 49 % of the recordings above 200,000 cells/mL, but similar values to this current study up to the 25^{th} percentile were observed. Another study conducted between 1992 and 1995 on 300 herds in The Netherlands (Olde Riekerink *et al.* , 2007), indicated that 23.8 % of individual cow recordings were above 200,000 cells/mL. The differences between these studies may relate partly to different economic incentives or legal constraints. Since 1992, in the European Union, milk is not saleable if the geometric mean BMSCC exceeds 400,000 cells/mL for more than two or three consecutive months, depending on the frequency of measurement (Council of the European Communities, 1992). In the UK a financial penalty is generally deducted from the milk price when the monthly BMSCC is above a certain value. A common BMSCC threshold for this is 200,000 cells/mL, calculated as a rolling three monthly geometric mean. The distribution of test-day BMSCC identified in this research indicated that 45.7 % of test-days were above 200,000 cells/mL suggesting that either penalties would be paid or that milk would be discarded from the tank in a substantial proportion of UK dairy herds.

As it was not possible to get BMSCC from the dairy companies, these were estimated from individual cow recordings. These estimates might be different from the *true* BMSCC. It was not possible to check the extent of this disagreement in the present study. The relation between BMSCC as measured and as estimated here was investigated in 246 herds in The Netherlands. Lievaart *et al.* (2007) concluded that estimated BMSCC was a better estimator of the prevalence of cows with a SCC greater than 250,000 cells/mL. Lievaart *et al.* (2009) found measured BMSCC to be on average 49,000 cells/mL lower than BMSCC estimated from individual cow recordings. Higher differences were associated with differences in management. A difference between measured and estimated BMSCC greater than 20 % was 2.4 times more likely in herds feeding high SCC milk or milk with antibiotic residues to calves. Because BMSCC estimated from individual cow recordings takes into account milk which is produced but not sold, BMSCC as estimated in this study would reflect more accurately herd infection levels than measured BMSCC.

Restricting the *calving* period to the first 29 days of lactation allowed us to evaluate the variability in the management of the heifers' rearing and dry period. A similar approach was used

by Cook *et al.* (2002) in the USA. As in the present study, they found an elevated percentage of the cows which had high SCC on the last milk recording prior to drying off with a low SCC on the first recording calving. SCC decreases very quickly at the start of lactation and Dohoo (1993) recommended to avoid the use of SCC collected before 9 days in milk in primiparous and 11 days in milk in multiparous cows for the diagnosis of infection. The National Milk Records do not normally collect milk prior to 5 days in milk. Therefore, data collected as early as 5 days in milk were available in this dataset and used in the present study. The threshold values used in this study were not aimed at measuring the true infection status of a particular cow but at getting estimates of the distribution in a large population.

In this chapter, models were first explored in a frequentist framework and then developed in a Bayesian framework using MCMC. This is the approach recommended by Gelman & Hill (2007). Bayesian models are especially suited when prior knowledge about the parameters modelled are available. No priors were incorporated in this analysis, however, posterior distributions for the models' parameters will be used to derive priors for similar models at the test-day level in chapter 6. A further motivation for a Bayesian analysis was the ease with which predictions accounting for the uncertainty in the models' parameters can be generated. These were used to check both the models' assumptions and the ability of the models to predict BMSCC in a different dataset.

5.6 Conclusion

The variability in herd dynamics of SCC across a threshold of 200,000 cells/mL between consecutive recordings can be used to analyse and predict BMSCC. During lactation, the percentage of cows above 200,000 cells/mL was lower in first lactation than in later ones. Different target values should be considered for primipara and multipara during lactation. A set of targets for cows in different SCC categories can be obtained for UK dairy herds, derived from percentile values shown in Table 5.10.

Chapter 6

Prediction of Bulk Milk Somatic Cell Counts from Cow Somatic Cell Count Categories II: Transition Between Categories and Prediction at the Test-Day Level

6.1 Introduction

A threshold of 200,000 cells/mL is commonly used to categorise cows as having a Low or High level of SCC, in the UK (Bradley & Green, 2005) as well as in other countries (Pantoja *et al.* , 2009). In Chapter 5, movements of individual cows between Low ($\leq$ 200,000 cells/mL) and High ($>$ 200,000 cells/mL) levels between consecutive recordings, aggregated at the herd-year level, were described at the start of and during lactation for first and later lactation cows. Herd-year percentages in each category were shown to predict BMSCC accurately. However, the ability of these percentages to predict individual test-day BMSCC was not assessed. At the test-day level, more variability can be expected because some categories can have only a few cows, and the contribution of these categories will be harder to predict. This is why Dohoo & Morris (1993) recommended the use of one year of data when looking at somatic cell count patterns in herds of 40 to 60 cows in Canada.

High and Low SCC levels can be considered as states between which the probability of transition can be modelled. Lactation curves for somatic cell count have been described in Chapter 5. For our data, SCC increased from 50 days in milk until the end of lactation and with successive lactations. Thus cows are more likely to move to a high level of SCC as lactation progresses.

Furthermore, herds can vary in their abilities to prevent cows from moving to High SCC levels by preventing IMI or to facilitate a return to Low SCC levels by successfully treating mastitic cows. Another way to mitigate the impact of mastitis on BMSCC could be to dry or to cull cows, and it is possible that infected cows will be dried or culled earlier (Grohn *et al.* , 1998). Low and High SCC levels, dry period and culling can be seen as competing risks because they are mutually exclusive. Steele *et al.* (2004) provided a statistical framework to describe transitions between such events.

The aims of this chapter were to (i) assess the ability of the model developed in Chapter 5 to predict test-day BMSCC (ii) model the probability of transition between 2 SCC levels, dry period and culling between consecutive test-days using multilevel multistate competing risks models and thereby to predict test-day BMSCC from the predicted percentage of a herd in each SCC transition category.

6.2 Materials

A sample of 200 herds was randomly selected from the 628 herds which were not used for parameter estimation in Chapter 5. In 100 of these herds, 7 consecutive recording dates were randomly selected between June 2004 and June 2006. The first 6 months and the last 6 months of the data were not sampled, so that cows moving in/out of the herd or from/to a dry period could be identified by their previous or next recording respectively. For each of the 100 herds selected at this stage, the first six recording dates were separated from the last recording date. These two datasets were labeled *dataset 1* and *dataset 2* respectively. In the remaining 100 herds, one recording date was selected between June 2004 and June 2006 at random in each herd. This dataset was called *dataset 3*.

6.3 Methods

6.3.1 Prediction of test-day BMSCC

Definition of cow-SCC categories

For each recording in the 3 datasets, cows were grouped using the same procedure as in Chapter 5. Briefly, SCC levels were categorised as Low (**L**) when $\leq$ 200,000 cells/mL or High (**H**) when $>$ 200,000 cells/mL. The transition between SCC categories between 2 consecutive test-days was used to construct 4 SCC transition categories (**LL**, **LH**, **HL**, **HH**). Cows of parity 1 were categorised as *Heifers* and cows of parity $>$ 1 as *Cows*. Recordings occurring earlier than 30 days in milk were labeled as *calving* and recordings happening after 29 days in milk as

Table 6.1: Definition of the cow categories used to model BMSCC. Parity and days in milk on test-day n are used to define the cow category and SCC values on test-day $n-1$ and n used to define the SCC transition category.

Parity_n	Days in milk_n	SCC_{n-1}	SCC_n	Category [a]
1	< 30	-	$< 200{,}000$	Hc.L
1	< 30	-	$\geq 200{,}000$	Hc.H
1	≥ 30	$< 200{,}000$	$< 200{,}000$	Hl.LL
1	≥ 30	$< 200{,}000$	$\geq 200{,}000$	Hl.LH
1	≥ 30	$\geq 200{,}000$	$< 200{,}000$	Hl.HL
1	≥ 30	$\geq 200{,}000$	$\geq 200{,}000$	Hl.HH
≥ 1	< 30	$< 200{,}000$	$< 200{,}000$	Cc.LL
≥ 1	< 30	$< 200{,}000$	$\geq 200{,}000$	Cc.LH
≥ 1	< 30	$\geq 200{,}000$	$< 200{,}000$	Cc.HL
≥ 1	< 30	$\geq 200{,}000$	$\geq 200{,}000$	Cc.HH
≥ 1	≥ 30	$< 200{,}000$	$< 200{,}000$	Cl.LL
≥ 1	≥ 30	$< 200{,}000$	$\geq 200{,}000$	Cl.LH
≥ 1	≥ 30	$\geq 200{,}000$	$< 200{,}000$	Cl.HL
≥ 1	≥ 30	$\geq 200{,}000$	$\geq 200{,}000$	Cl.HH

[a] The first letter is for the cow category (H: Parity 1 ; C: Parity > 1), the second letter is for days in milk (c: < 30 ; l: ≥ 30) and the last 2 letters are for SCC transition between test-day $n-1$ and n (L: Low ; H: High ; LL: Low on test-day $n-1$ to Low on test-day n...)

lactating. These three different types of categorisation based on SCC, parity and stage of lactation were combined to define 14 categories to which individual cow recordings were assigned (Table 6.1).

Estimation of Bulk Milk Somatic Cell Counts

Test-day BMSCC were estimated as follows: The number of cells contributed by a cow was estimated by multiplying the cell count by the weight of milk on a given recording date. These numbers were added for all the cows recorded on the same recording date and divided by the sum of the total weight of milk produced by these cows. BMSCCs are presented in thousand cells/mL.

Statistical Analysis

Test-day BMSCC was modelled as a function of the percentage of the cows recorded on a test-day in the 14 categories defined in Table 6.1 using linear mixed models. The model specifications were:

$$\begin{aligned} BMSCC_{ij} &= \alpha + \Sigma X_{ij}\beta^T + u_j + e_{ij} \\ u_j &\sim N(0, \sigma_j^2) \\ e_{ij} &\sim N(0, \sigma_{ij}^2) \end{aligned} \tag{6.1}$$

where the subscripts i and j denote the i^{th} test-day in the j^{th} herd, α the regression intercept, X_{ij} the covariates relating to a herd-year, β^T the vector of coefficients for covariates X_{ij}, u_j the herd residuals and e_{ij} the herd test-day residuals.

Parameter estimation was carried out using Markov chain Monte Carlo in WinBUGS (Lunn *et al.* , 2000). *Dataset 1* was used for parameter estimation. Priors were put on the parameters as follows. α and β were given normal priors for which the mean and standard deviation were the mean and twice the standard deviation of the sample from the posterior distributions from Model 2 in Chapter 5. Uniform priors on 0 - 100 were used for σ_j and σ_{ij} (Table 6.4). For all models, 3 chains were run for 30,000 iterations and the first 20,000 discarded. One iteration in 100 was stored for further analysis. Thus, for each variable in the model, 300 values from the posterior distribution were available.

Predictions

At each iteration of the MCMC algorithm BMSCCs were predicted for the three datasets. For iteration n, the predicted BMSCC was:

$$BMSCC_{ij}^n = \alpha^n + \Sigma X_{ij}\beta^{nT} + u_j^n \tag{6.2}$$

where α^n, β^n and u_j^n were the values of α, β and u_j at iteration n. For *Dataset 3*, u_j^n was generated from the normal distribution $N(0, \sigma_j^{n^2})$.

Model checking and validation

Medians and 95 % credibility intervals were calculated for the posterior distribution of each parameter as well as for the predicted BMSCC. Prior and posterior distributions were compared.

Medians and 95 % credibility intervals were calculated for the distribution of predicted BMSCC. Observed BMSCC were plotted against the median of the predicted ones. For each predicted BMSCC, residuals were computed by subtracting the median of the prediction from the observed value.

The sensitivity and specificity of the models' predictions were assessed as follows: Calculated and predicted BMSCC were categorised as below or above 200,000 cells/mL. True positives were BMSCC $>$ 200,000 cells/mL and for which the median value predicted by the model was

Table 6.2: Definition of the 5 states used in the state transition model.

State	Cow Recording	Lactation Recording	SCC
Low	First $\leq \leq$ Last	First $\leq \leq$ Last	≤ 200
High	First $\leq \leq$ Last	First $\leq \leq$ Last	> 200
dry	First $\leq \leq$ Last	$>$ Last	-
first [a]	First - 1	First - 1	-
last [a]	Last + 1	Last + 1	-

[a] *first* only occurs as a previous state and *last* as a current state. The same codes were used for these two states in the models.

$>$ 200,000 cells/mL. True negatives were BMSCC $\leq$ 200,000 cells/mL and for which the median predicted value was $\leq$ 200,000 cells/mL. Sensitivity was the percentage of BMSCC $>$ 200,000 which were true positives. The specificity was the percentage of BMSCC $\leq$ 200,000 which were true negatives.

6.3.2 State Transitions

States definition

Two somatic cell count states were defined and individual cow SCC categorised accordingly. Cow SCC readings $\leq$ 200,000 cells/mL were categorised as Low (**L**) and SCC readings $>$ 200,000 cells/mL as High (**H**). Three more states regarding lactation status were defined. At the cow level, the recording date preceding the first recording was labeled as first and the one following the last as last. For a cow, the last recording in a lactation and the first recording date in the following lactation were identified and the recordings happening in the herd between these two dates were labeled as dry period for this cow. The definition of the five different states used is shown in Table 6.2.

Statistical Analysis

The transitions between these states between consecutive recordings were modelled with multilevel multistate competing risks models (Steele *et al.* , 2004). In these models, each state is modelled as a function of the state a cow was in on the previous recording date as well as of covariates of interest. The number of cows in a given state followed a multinomial distribution and the logarithm of the probability of being in any of these states divided by the probability of being in the reference state was modelled. This model is analogous to a logistic regression model for a binary outcome and is referred to as a multinomial logit model (Agresti, 2002). Thirteen transitions were possible since a cow could not move from *first* to *dry*, *first* to *last* or *dry* to *last*. The model's specification was as follows:

$$
\begin{aligned}
State_{ijk} &\sim Multinomial(\pi_{ijk}) \\
ln(\tfrac{\pi_{ijk}}{\pi_{1jk}}) &= \sum_{i'=1}^{4} I[State^{i'}_{i(j-1)k}](\alpha_i^{i'} + \sum X_{ijk}\beta_i^{i'} + u_{ik}^{i'}) \\
u_{ik}^{i'} &\sim MVN(0, \Sigma_u)
\end{aligned}
\tag{6.3}
$$

where $State_{ijk}$ was the i^{th} State (1: Low ; 2: High ; 3: dry ; 4: last) a cow could be in on the j^{th} recording date in herd k. The log-odds of the probabilities π_{ijk} of being in one of these states was modelled as a function of the State i' (1: Low ; 2: High ; 3: dry ; 4: first) the cow was in on the $(j-1)^{th}$ recording date that is her State on the previous recording date in the herd. $I[State^{i'}_{i(j-1)k}]$ was an indicator variable taking the value 1 when cow was in State i' on the previous recording in the herd, 0 otherwise. $\alpha_i^{i'}$ was the vector of regression intercepts, X_{ijk} the matrix of predictors, $\beta_i^{i'}$ the associated coefficients, $u_{ik}^{i'}$ the herd effect for the probability of transition from $State_{i'(j-1)}$ to $State_{ij}$ in herd k and Σ_u the variance-covariance matrix for the herd random effects.

Model Building

A first model including only previous states and no other covariates was built. This model was then extended to include parity and days in milk. Parity was coded as 1 or greater than 1 because the lactation curve is different between first lactation and older cows (Chapter 4).

Gestation length in Holstein cows is of approximately 280 days (Norman *et al.* , 2009) and the dry period is typically of 2 months. Thus, cows usually do not move to dry period at the start of lactation. It was assumed that after a certain stage of lactation, the chance of moving to dry period or culling was increasing with days in milk. Cutpoints were used to model different slopes for different parts of lactation. Days in milk and squared days in milk were tested in the models and different coefficients were used before and after the cutpoints.

Parameter Estimation

Parameter estimation was carried out using MCMC in WinBUGS. Three chains were run in parallel for each model tested. In the models including parity and stage of lactation, each iteration was taking between 25 and 30 seconds. Depending on whether the chains had converged or not, the first 5,000 or 10,000 iterations were discarded. The models were then run for 10,000 more iterations. The final model was run for 20,000 iterations and the first 10,000 iterations discarded.

Model Checking, Predictions and Simulations

Models' DIC were recorded and used as indicators of individual model fit. To see whether the models predicted the data well, individual cow states were predicted in *Dataset 1*. Parameter values saved from each WinBUGS iteration were used. Every 1 in 100 iterations from the models, that is 100 iterations per chain, were imported into R (R Development Core Team, 2009). Each iteration was used to predict individual cow state given the previous state the cow was in. Thus, each line of *Dataset 1* was predicted 300 times. Median, 2.5^{th} percentile and 97.5^{th} predicted values were plotted against observed mean for all the transitions per day in milk for cows of parity one and greater than one.

Based on these investigations, initial models were altered and a final model was selected. Because the parameters of such models were hard to interpret, predicted values were generated in order to compare the probabilities of transitions per day in milk, for parity one and greater than one cows. Three hundred iterations were imported into R as described above. Median values were calculated for all the fixed parameters of the model because only average effects were of interest. A function calculating the probability of being in any of the 4 possible states given the previous state, the number of days in milk and the parity was written and used to predict probabilities of transition.

BMSCC Prediction from Predicted States

For the final model, individual cow states were predicted (as described above) in *Dataset 1*, *Dataset 2* and *Dataset 3*. For each line of data, the previous state and three hundred predicted current states were available. For each prediction in each test-day, the percentage of the herd in each combination of states on 2 consecutive recordings was calculated. These test-day percentages were used to predict BMSCC using the same model specifications as in Section 6.3.1. However, the model had to be slightly modified because the previous SCC level of cows moving out of a dry period was not available. On the first recording after the dry period, cows could only be Low or High with this new model. All the other categories and model specifications were the same as in Section 6.3.1. For this model, using WinBUGS, 3 chains were run for 30,000 iterations and the first 20,000 discarded. Three hundred iterations were imported into R as described above. From these 300 iterations, BMSCC was predicted for the 300 predicted percentages in each category per test-day. Median, 2.5^{th} and 97.5^{th} percentiles were calculated for each predicted BMSCC and plotted against the observed values.

Table 6.3: Distribution of the number of cows recorded per test-day in the 3 datasets.

	Min	1^{st} Quartile	Median	Mean	3^{rd} Quartile	Max
Dataset 1	31	80	110	117.3	139	285
Dataset 2	35	81.75	111	119	145.2	281
Dataset 3	35	86.25	119	138.7	159.2	649

6.4 Results

6.4.1 Data description

Dataset 1 was the training dataset containing 100 herds with 6 test-days each. *Dataset 2* contained a seventh test-day from the same herds. *Dataset 3* contained data for one test-day per herd for 100 different herds. *Dataset 1*, *2* and *3* had 70,382, 11,895 and 14,669 lines respectively. The distribution of the number of cows recorded in the 3 datasets is presented in Table 6.3. The median number of cows recorded were 110 and 111 in *datasets 1* and *2* respectively. This number was higher in *dataset 3* with a median of 119.

6.4.2 BMSCC prediction

Summaries of the prior and posterior distributions are presented in Table 6.4. In most cases, posterior means were close to prior ones. Differences were important for *Cows calving LL* (prior mean: -0.25 ; posterior mean: -2.35), *Cows lactating LH* (prior: 2.05 ; posterior: 6.73), *Cows lactating HL* (prior: 4.86 ; posterior: 0.93). Mean residuals were close to 0 for the training dataset (*dataset 1*), but the model tended to predict values slightly higher than observed for both validation datasets (*datasets* 2 and *3*) (Figure 6.1 and Table 6.5). The standard deviations of the predictions were between 48.6 for dataset 3 and 54.6 for dataset 2. The percentage of observed BMSCC in the 95 % credibility intervals were of 97 %, 96 % and 98 % for datasets 1 to 3 respectively. The percentages of BMSCC correctly predicted above 200,000 cells/mL (sensitivity) were 82.4 %, 77.8 % and 79.3 % and the percentages of BMSCC correctly predicted below 200,000 cells/mL (specificity) were 86.9 %, 84.8 % and 95.2 % for datasets 1 to 3.

6.4.3 State transition

Baseline State Transition Model

Of the cows with a SCC reading, 75.6 % of *Cows lactating*, 78 % of *Cows calving* and 82 % of *first* (mostly *Heifers calving*) were Low. The probabilities of moving to any of the four possible states given the initial state as observed and as predicted by the baseline model are presented

Table 6.4: Prior and posterior distribution. All variables were given normal prior except σ_u and σ_e which were given uniform priors on 1-100.

			Posterior				
	Prior					Credibility Interval	
Parameter	Mean	sd	Mean	sd	Median	2.5 %	97.5 %
Intercept	39.20	3.70	39.20	0.53	39.21	38.18	40.25
Hc.L	−1.00	0.64	−0.96	0.89	−0.96	−2.71	0.72
Hc.H	8.94	1.53	9.25	0.79	9.26	7.63	10.85
Cc.LL	−0.25	0.52	−2.35	0.80	−2.33	−3.93	−0.74
Cc.LH	16.10	1.27	15.90	0.87	15.91	14.28	17.63
Cc.HL	−6.71	0.63	−5.62	0.89	−5.64	−7.35	−3.93
Cc.HH	19.23	1.14	17.30	0.84	17.30	15.64	18.93
Hl.LL	−0.08	0.10	0.16	0.31	0.14	−0.42	0.75
Hl.LH	5.56	1.07	6.48	0.84	6.51	4.78	8.14
Hl.HL	6.16	1.16	4.48	0.79	4.47	3.02	6.01
Hl.HH	5.11	0.48	5.98	1.15	5.96	3.90	8.29
Cl.LH	2.05	0.42	6.73	0.55	6.74	5.68	7.78
Cl.HL	4.86	0.47	0.93	0.63	0.90	−0.31	2.16
Cl.HH	7.09	0.11	6.41	0.37	6.40	5.68	7.15
σ_u	Unifom(0, 100)		30.44	3.49	30.40	23.70	37.51
σ_e	Unifom(0, 100)		57.38	1.91	57.39	53.76	61.22

Figure 6.1: Observed versus predicted values by the model

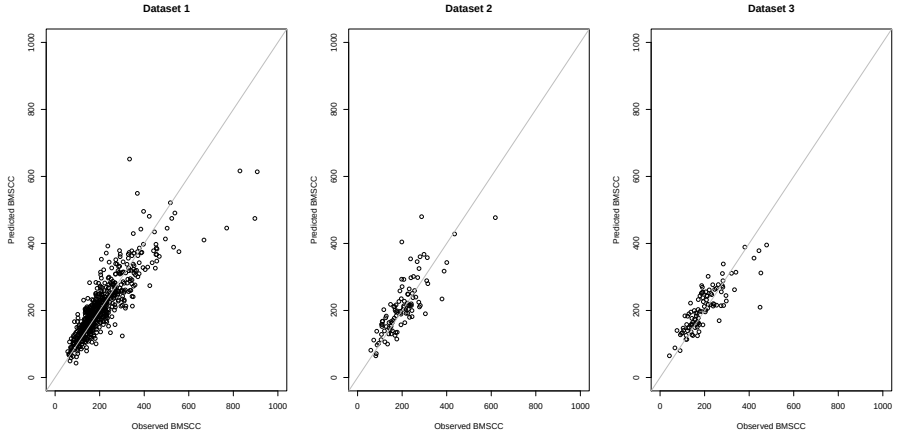


Table 6.5: Results of the prediction of test-day BMSCC. Residuals were calculated as Observed - Median predicted BMSCC. The percentage of observed BMSCC between the 2.5 and 97.5 percentile of the predictions were calculated. The sensitivity and specificity of the prediction to classify BMSCC as greater or smaller than 200,000 cells/mL were also calculated.

	Dataset 1	Dataset 2	Dataset 3
Mean residuals	−0.17	−9.63	-6.64
Standard deviation residuals	54.10	54.62	48.64
% predicted 95 % CI	97.00	96.00	98.00
Sensitivity	86.92	84.78	95.24
Specificity	82.35	77.78	79.31

in Table 6.6. Cows ≤ 200,000 cells/mL had a 82.3 % chance to be ≤ 200,000 cells/mL on the following test-day and a 10.7 % of being > 200,000 cells/mL, while cows > 200,000 cells/mL had a 57.2 % chance of staying > 200,000 cells/mL and a 25.6 % chance of moving ≤ 200,000 cells/mL.

Table 6.6: Probability of transition from *Previous* to *Current* state as a proportion of the number of cows in the same *Previous* state in *Dataset 1*: numbers, Observed probabilities, Median and Credibility intervals predicted by the model are shown.

State			Probability of transition			
					Credibility Interval	
Previous	Current	n	Observed	Median	2.5 %	97.5 %
Low	Low	37,259	0.822	0.822	0.819	0.825
Low	High	4,870	0.107	0.107	0.105	0.110
Low	dry	2,487	0.055	0.055	0.053	0.057
Low	culled	720	0.016	0.016	0.015	0.017
High	Low	3,770	0.258	0.257	0.251	0.264
High	High	8,349	0.570	0.570	0.563	0.579
High	dry	1,718	0.117	0.117	0.113	0.123
High	culled	798	0.055	0.054	0.051	0.058
dry	Low	2,647	0.283	0.283	0.274	0.292
dry	High	745	0.080	0.079	0.075	0.085
dry	dry	5,967	0.638	0.638	0.627	0.646
first	Low	863	0.820	0.821	0.797	0.842
first	High	189	0.180	0.179	0.158	0.203

Final State Transition Model

Based on changes in observed probabilities of transition and probabilities predicted by the models, a final model was selected. For each transition, a different set of intercepts was used for cows of parity one and for cows of parity greater than one. Herd random effects were also

introduced to allow for herd variability. The probability of moving to a dry period increased from 250 days in milk onwards. Hence a cutpoint was placed at this lactation stage and different slopes modelled before and after 250 days in milk. Once in the *dry* state cows could only stay in this state as days in milk were reset to 0 with the start of a new lactation. For cows moving out of a dry period, a cutpoint was set at 100 days in milk to allow for a probability of 1 for cows to stay *dry* after the dry period had been initiated. Linear and quadratic terms were introduced for days in milk so that non linear effects with stage of lactation could be accounted for. Figures 6.2 and 6.3 show the probabilities of transition predicted by the final model versus the observed probabilities of transition to any state given the previous state a cow was in for parity 1 and greater than 1 respectively. Observed probabilities were scattered because there could be only a few observations per calculated proportion. However, overall, the model was describing the data well.

Probabilities of moving from *Low* and *High* to *Low*, *High*, *dry* and *Culled* were calculated from the median of the parameters estimated by the final model between 30 and 500 days in milk for cows of parity one and greater than one. These probabilities are shown in Figure 6.4. The risk of moving to *dry* started to increase a few days before 250 days in milk and reached a maximum at around 450 days in milk. Changes in probabilities of transition were close to linearity between 30 and 200 days in milk and between 300 and 400 days in milk. Probabilities of transition from *Low* and *High* at 30 and 200 days in milk and at 300 and 400 days in milk are presented in Tables 6.7 and 6.8 respectively. Generally, cows were more likely to move to *High* and less likely to move to *Low* as lactation was progressing. The probability of moving to *High* was always higher for cows of parity greater than 1 compared to primiparous cows. The probability of being *Low* was higher for cows already *Low* and the probability of being *High* higher for cows already *High*. The probability of moving to a dry period was approximately 4 % higher at 300 days for cows which were *High* than for cows that were *Low*, but the increase in the risk of moving to *dry* increased more rapidly thereafter for the *Low* groups.

BMSCC Predictions from the Final State Transition Model

Predicted states were used to generate BMSCC predictions (Figure 6.5). BMSCC was not predicted very accurately. For *Dataset 1*, *Dataset 2* and *Dataset 3*, only 57 %, 61 % and 64 % of observed BMSCC were in the 95 % credibility intervals respectively.

6.5 Discussion

This study validated the herd-year model developed in Chapter 5 to the test-day level. Thus even on a single recording date, the contribution of groups of cows categorised according to

Figure 6.2: Observed and predicted probabilities of transition per day in milk for first lactation cows. Grey dots represent the proportion of cows that moved from *Previous State* to *Current State*. Plain (dashed) lines are the median (2.5^{th} and 97.5^{th} percentile).

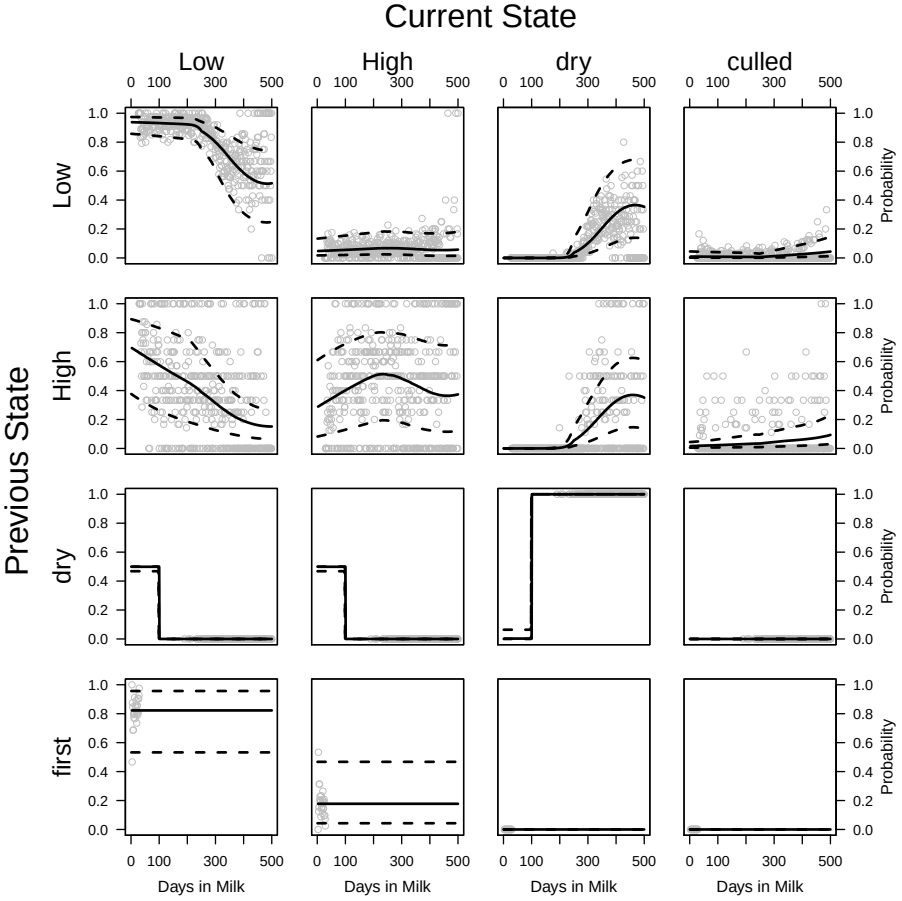


Table 6.7: Probability of moving to *Current* state given the *Previous* state a cow was in: Predicted probabilities at 30 and 300 days in milk and variations.

State			Day in Milk		
Previous	Current	Parity	30	200	Variation (%)
Low	Low	1	0.94	0.93	-1.5
Low	Low	>1	0.92	0.86	-6.3
Low	High	1	0.05	0.06	30.4
Low	High	>1	0.07	0.13	83.1
Low	Culled	1	0.01	0.01	-19.9
Low	Culled	>1	0.01	0.01	-23.8
High	Low	1	0.67	0.47	-29.9
High	Low	>1	0.49	0.28	-41.8
High	High	1	0.32	0.50	57.2
High	High	>1	0.48	0.67	38.3
High	Culled	1	0.02	0.03	70.1
High	Culled	>1	0.03	0.04	41.1

Figure 6.3: Observed and predicted probabilities of transition for cows of parity greater than 1. Grey dots represent the proportion of cows that moved from *Previous State* to *Current State*. Plain (dashed) lines are the median (2.5^{th} and 97.5^{th} percentile).

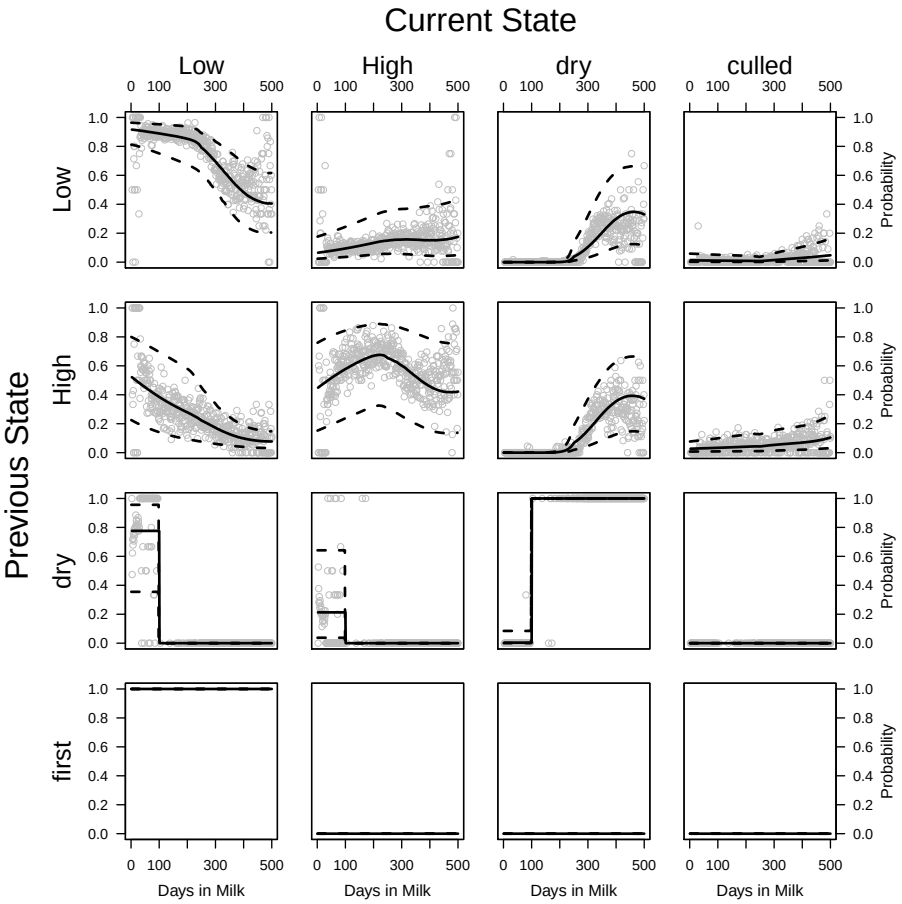


Figure 6.4: Predicted probabilities of transition from below and above 200,000 cells/mL to below and above 200,000 cells/mL, dry period and culling between calving and 500 days in milk for parity 1 (solid lines) and greater than 1 (dashed lines) cows.

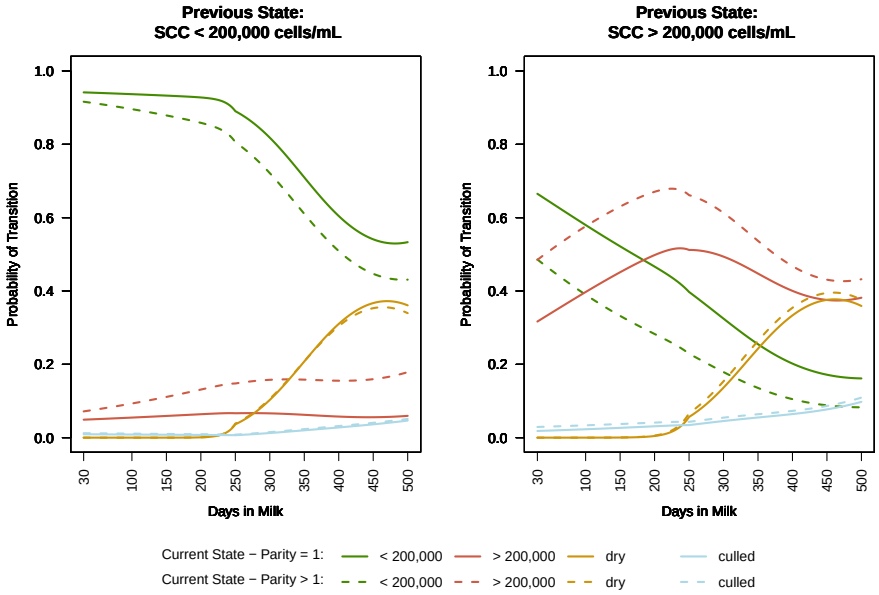
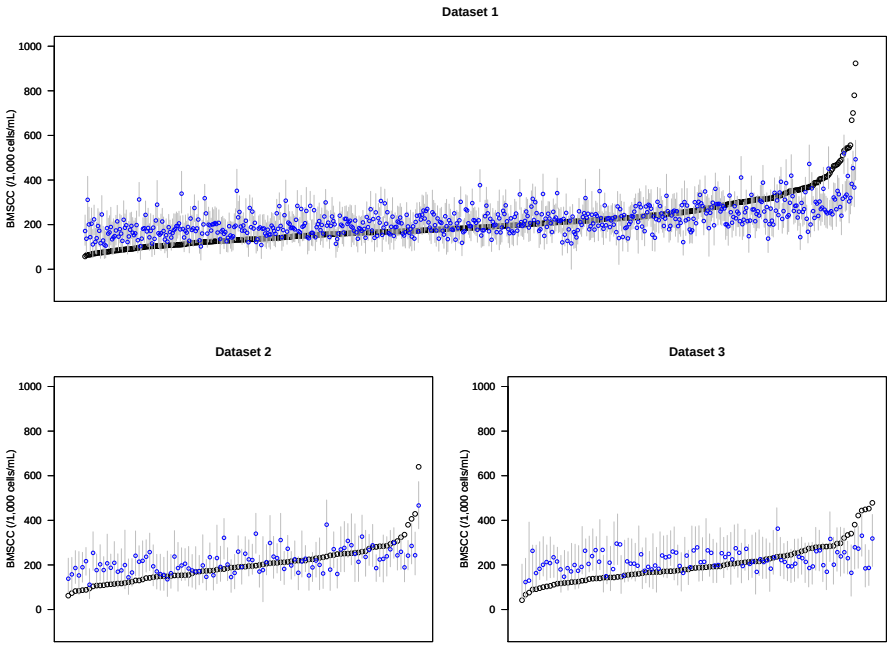


Table 6.8: Probability of moving to *Current* state given the *Previous* state a cow was in: Predicted probabilities at 300 and 400 days in milk and variations.

State			Day in Milk		
Previous	Current	Parity	300	400	Variation (%)
Low	Low	1	0.82	0.60	-26.1
Low	Low	>1	0.72	0.51	-29.4
Low	High	1	0.07	0.06	-13.2
Low	High	>1	0.16	0.16	-1.6
Low	Dry	1	0.10	0.31	201.1
Low	Dry	>1	0.11	0.30	187.5
Low	Culled	1	0.01	0.03	120.4
Low	Culled	>1	0.02	0.03	110.4
High	Low	1	0.32	0.20	-37.8
High	Low	>1	0.18	0.10	-41.3
High	High	1	0.49	0.40	-18.9
High	High	>1	0.61	0.47	-23.7
High	Dry	1	0.14	0.33	144.3
High	Dry	>1	0.15	0.35	130.4
High	Culled	1	0.05	0.06	42.0
High	Culled	>1	0.05	0.07	33.9

Figure 6.5: Observed versus predicted BMSCC. Predictions were made from *States* predicted by the final model combined with observed *Previous States* and aggregated at the test-day level. *Dataset 1* was the training dataset with data from 6 test-days in 100 herds, *Dataset 2* was a validation dataset with the test-day following the sixth test-days in the previous 100 herds and *Dataset 3* contained one test-day per herd in 100 herds. Black dots, blue dots and grey bars represent the observed BMSCC, the median predicted BMSCC and the 95 % credibility interval for the prediction.



SCC, stage of lactation and parity to BMSCC can be calculated. This can be done for didactic purpose, in order to show farmers what is the impact of mastitis in their herd as well as to set targets and to estimate the result of their implementation.

A state transition model constructed to explore the probabilities of transition between the Low and High levels of SCC used in this model as well as dry period and culling was shown to describe these transitions well. However, the use of the predicted individual cows SCC levels was not useful to predict BMSCC, even when including herd random effects in both the state transition and BMSCC prediction steps (*Dataset 1*). The state transition model was designed to account for stage of lactation and parity in the prediction of SCC level. Lactation curves for SCC indicate that there is an increase in SCC from the lactation peak towards the end of lactation, and that SCC levels increase with successive lactations, especially between the first and second lactation (Schepers *et al.* (1997), Chapter 4). This increase could be the result of either a physiological process, an increase in IMI prevalence or both. It is unclear what are the roles played by these two phenomena. It is possible that while the model was describing physiological variations in these parameters, it was not capturing the probability of acquiring an IMI. Infection is the main factor increasing SCC (Schepers *et al.* , 1997), and, while physiological variation is moderate, an IMI could have a much bigger impact on BMSCC. Thus a model accounting for this physiological variation alone would not be able to predict BMSCC. Looking at figures aggregated at the herd level over one or more test-days or studying changes in BMSCC between consecutive measurements (Lukas *et al.* , 2005) would then be more useful to predict BMSCC.

Predictions from the percentages of a herd in each cow infection category had sensitivities and specificities greater than 75 % even with one test-day. The median number of cows recorded in this study was between 110 and 120 and in 75 % of test-days, more than 80 cows were recorded. So on average, each cow represented less than one percent of a test-day and individual variations within a category were averaged. Means of the prior and posterior distributions were usually close. One notable feature was the inversion in the relative importance of cows going from below to above and above to below 200,000 cells/mL in lactating cows for both first and later lactations: at the herd-year level, the coefficients associated with going from above to below 200,000 cells/mL were positive and higher than the coefficients associated with going from below to above 200,000 cells/mL. At the test-day level, the former were negative and smaller. This was related to the time period considered when looking at either herd-year or test-day estimates. When looking at prolonged periods of time, the numbers of cows going from below to above and from above to below will be correlated, because for the prevalence of high SCC to remain stable, 'cures' have to compensate for 'new infections'. This is not the case at the test-day level where prevalence can increase or decrease between consecutive recordings and 'cures' will, by

definition, be associated with lower SCC and hence a lower contribution.

6.6 Conclusion

The herd dynamics of SCC around a threshold of 200,000 cells/mL can be used to predict test-day BMSCC. The use of individual cow information on stage of lactation and parity was not useful in predicting individual cow movement across this threshold and subsequently predicting test-day BMSCC.

Chapter 7

Use of Individual Cow Milk Recording Data at the Start of Lactation to Predict the Calving to Conception Interval

7.1 Introduction

The transition between dry period and the onset of lactation results in a sudden and massive increase in the energy demand for dairy cows (Bauman & Currie, 1980). At this stage, the appetite of the cow is limited, the amount of energy exported in milk cannot be covered by food and most cows will experience a period of negative energy balance (**NEB**). In an experimental herd in the Netherlands, de Vries & Veerkamp (2000) found an average period of NEB of 41.5 days. This NEB has been linked to poor reproduction (Patton *et al.* , 2007; Buckley *et al.* , 2003). Energy exported and severity of NEB increase with milk yield. There is also a negative correlation between milk yield and the resumption of luteal activity (Veerkamp *et al.* , 2000), oestrous expression (Cutullic *et al.* , 2009) and days open (Haile-Mariam *et al.* , 2003; Abdallah & McDaniel, 2000).

Many studies have measured the association between energy balance and milk composition. Four of these studies were carried out on 2 experimental farms at Lelystad, the Netherlands (Heuer *et al.* , 2000; de Vries & Veerkamp, 2000; Heuer *et al.* , 1999; Grieve *et al.* , 1986) and one in 93 dairy farms in Canada (Duffield *et al.* , 1997). It was constantly found that in cases of NEB there is an increase in the milk percentage of butterfat, a decrease in the percentage of milk protein and an increase in the fat to protein ratio. Short-chain and medium-chain fatty acids are synthesized *de novo* by the mammary gland, long-chain fatty acids are taken up from

the bloodstream (Bauman *et al.* , 2006). Cows in negative energy balance mobilise fat reserves and the milk percentage of C16:0 and C18:0 fatty acids increases (Stoop *et al.* , 2009). For de Vries & Veerkamp (2000), the difference in milk fat percentage between the first 2 months of lactation was the best indicator of energy balance. In their study, the percentage of protein was only changed by a small amount with varying energy balance so that neither the percentage of protein or the fat to protein ratio were useful in the prediction of the energy status. For Grieve *et al.* (1986) the fat to protein ratio was the best indicator in predicting the energy status. Heuer *et al.* (1999) identified cows with a fat to protein ratio greater than 1.5 to be at greater risk of ketosis while Duffield *et al.* (1997) found none of these indicators to be useful in the screening for ketosis although an association was present.

Few studies have looked at the association between milk composition at the start of lactation and the probability of conception. Kristula *et al.* (1995) identified a negative association between the first insemination pregnancy rates and both the percentage and weight of butterfat on the first milk recording of lactation in 1,640 cows from 22 United States dairy herds. Using data from 51 high yielding cows in Slovenia, Podpecan *et al.* (2008) found a threshold of 1.44 for the fat to protein ratio able to identify 91.7 % of cows with a calving to conception interval greater than 140 days.

Milk quantity and composition vary with lactation stage as well as, to a lesser extent, with parity, season and other factors (Chapter 4). Lactation curves have been extensively studied for milk yield, percentage of butterfat and percentage of protein (Silvestre *et al.* , 2009). But the current variations for the fat to protein ratio according to stage of lactation have not been described. Because all cows will go through the same variation with stage of lactation, this variation must be taken into account when comparing monthly milk samples. Moreover, milk quantity and composition, vary with the month of the year (Chapter 4) and it is not known whether the observed variation is due to different energy status or whether it is the result of specific seasonal factors which should be accounted for. Heuer *et al.* (2000) included days in milk and parity as confounders in their models, however, this assumes that the variables included vary together with stage of lactation and parity.

Therefore, milk quantity and composition at the start of lactation can reflect individual cow energy balance. However, their usefulness in the prediction of the calving to conception interval needs clarification. The purpose of the present chapter is to use a large dataset of milk recording data to investigate the calving to conception interval as a function of milk quantity and composition at the start of lactation.

7.2 Materials

The data used in this Chapter are described in Chapter 2. Lactations for which calving occurred in 2004 and 2005 and that had at least two milk recordings between 5 and 65 days in milk were used. There were data for 525,602 lactations in 366,939 cows from 2,128 herds. Calvings occurring in 2004 in a random sample of 500 of these herds were used to model the calving to conception interval. The remaining 1,627 herds were used for model checking. The number of herds and time frame in the dataset used for parameter estimation were a compromise between data size and the ability to model herd variability. This dataset contained 40,514 lactations from 39,585 cows.

7.3 Methods

7.3.1 Calving to Conception Interval

For each lactation, the date of calving was available. When two consecutive calving dates were available for the same cow, the interval between calvings was calculated. The date of conception was estimated by subtracting 280 days from the date of calving (Norman *et al.* , 2009). The calving to conception interval was the difference between the date of conception and the date of the previous calving. These intervals were categorised as follows. Conceptions occurring before 20 days in milk were deemed unusual and not included in the analysis. Because the impact of negative energy balance on conception was expected to be more important at the start of lactation, and, in order to ease computation, conceptions occurring after 144 days in milk were censored. The first interval was 20-60 days after calving. Each of the subsequent intervals was of 21 days. Five intervals were considered for the analysis: [20-60] ; [61-81] ; [82-102] ; [103-123] ; [124-144].

7.3.2 Statistical Analysis

The association between milk quantity and composition at the first 2 recordings after calving and the probabilities of conception in each interval were modelled using discrete time survival models (Yang & Goldstein, 2003). The probability of a conception at each interval, from interval 1 (20 to 60 days in milk) to interval 5 (124 to 144 days in milk) were modelled and cows were

censored at the time of conception. The model specification was:

$$\begin{aligned} Y_{tjk} &= Bernoulli(p_{tjk}) \\ ln(\tfrac{p_{tjk}}{1-p_{tjk}}) &= \sum_{t=1}^{5} I_{tjk}(\gamma_t + \delta_{tk}) + X_{jk}\beta \\ \delta_{tk} &\sim MVN(0, \Sigma_\delta) \end{aligned} \tag{7.1}$$

with $Y_{tjk} = 1$ when a conception occurred in interval t for lactation j in herd k and 0 otherwise ; p_{tjk} the associated probability of conception in interval t; I an indicator variable taking the value 1 in interval t, 0 otherwise ; γ_t the log-odds of a conception in interval t ; δ_{tk} the herd effect for the log-odds of a conception in interval t ; X_{jk} the matrix of predictors and β the vector of associated coefficients and Σ_δ the covariance matrix for the herd random effects. Since cows were censored in the intervals following conception, in cows that did not recalve, the outcome was coded as 0 for the 5 intervals. Because the date of conception was determined by subtracting 280 days from the date of calving, it was impossible to know whether a cow was pregnant from 280 days before she left the herd. Cows were censored from 280 days before their last recording in the dataset because this was making their pregnancy status unknown. Models were estimated in MLwiN using Iterative Generalised Least Squares (Rasbash *et al.* , 2009).

7.3.3 Covariates

Variables related to milk quantity and composition in the first two recordings of lactations, for lactations with at least 2 recordings between 5 and 65 days in milk, were considered as covariates. The covariates considered were milk yield, percentage and weight of butterfat, percentage and weight of protein, percentage and weight of lactose and somatic cell count on both the first and second test-days of lactation. Since variations in these variables occurred with lactation stage, parity and season, variables standardised for either stage of lactation or stage of lactation and day of the year were tested in the models. Interactions between the covariates and the natural logarithm of the interval number were tested to allow variations with time. Variables were kept in the model if the associated coefficient was at least twice the standard error and if its inclusion resulted in a decrease in the deviance.

7.3.4 Standardisation for stage of lactation and time of the year

All the recordings occurring between 5 and 70 days in milk were extracted from the original dataset. There were 1,582,488 recordings from 798,763 lactations from 441,320 cows in 2,128 herds available. For each variable, the mean and standard deviation per day in milk (5 to 70) and per week of the year (1 to 52) were calculated. These values were smoothed using linear

models. The effects of stage of lactation were accounted for using polynomials. The models' specifications for the stage of lactation mean and standard deviation were:

$$Y_i = \alpha + \sum_{p=1}^{P} Dim_i^p \beta_p \tag{7.2}$$

where Y_i was either the mean or the standard deviation of the variable modelled for day in milk i, α the model intercept, Dim the days in milk, P the highest power retained in the model and β_p the coefficient associated with Dim^p. Powers 1 to 6 were tested for each variable.

The effects of the time of the year were modelled using sine and cosine functions. Each week was converted back to a number of days using the formula:

$$Day = Week \times 7 - 3.5 \tag{7.3}$$

The models' specification for the stage of lactation and time of the year standardisation was:

$$Y_{ij} = \alpha + \sum_{p=1}^{P} Dim_i^p \beta_p + \sum_{t=1}^{T} (sin(2\pi t \frac{Day_j}{365})\delta_t + cos(2\pi t \frac{Day_j}{365})\gamma_t) \tag{7.4}$$

where Y_{ij} were the mean or standard deviation of the variable considered on day in milk i and for day of the year j. The transformation applied to the days of the year produced a yearly periodicity. Interactions between days in milk and sine and cosine were tested. Variables were retained when the p-value was smaller than 0.05 and when the residuals displayed graphically showed no systematic patterns. The predictions were also plotted to detect overfitting.

These models were used to standardise the observed values for either stage of lactation or stage of lactation and season as follows:

$$Standardised_{ij} = \frac{(Observed_{ij} - \bar{X}_{ij})}{\bar{\sigma}_{ij}} \tag{7.5}$$

where $Observed_{ij}$ was the observed value, $\bar{X}_{ij}$ the estimated mean and $\bar{\sigma}_{ij}$ the estimated standard deviation for day in milk i (1 to 60) at day of the year j (1 to 365). Thus, the corrected values were centred around 0 and scaled to have a standard deviation of 1.

7.3.5 Model checking and validation

To check final models, predictions were used. The predictions were performed using fixed as well as random effects estimated by the model. For each covariate, bins were created based on quantiles, each of 20 bins containing 5 % of the data. For each bin in each interval, the difference between the observed and predicted proportions of conceptions and the corresponding

confidence intervals were plotted. Based on model assumptions, approximately 5 % should have fallen outside the confidence interval.

Models were also checked using a cross validation procedure. For the 500 herd dataset used for parameter estimation, both herd random effects and fixed effects were used to generate predicted probabilities of conception. For the data from the 1,628 herds not used for parameter estimation, only the fixed effects were used. For each cow not pregnant at the start of an interval, the probability that she would conceive during this interval was calculated. The predicted probability P that a cow would have conceived by the end of interval 5 was calculated as:

$$P = p_1 + \sum_{i=2}^{5} \prod_{j=1}^{i-1} (1 - p_j) p_i \tag{7.6}$$

where p_i and p_j were the probabilities that a cow conceived during interval i and j. The probabilities that a cow would conceive at any of the 5 intervals given that she had not conceived on the previous ones were summed up for the 5 intervals. The percentiles 1 to 99 were calculated for the predicted probabilities of conception at each interval as well as for the P. These percentiles were used to create 100 categories of increasing predicted probability of conception per interval and by then end of interval 5. The observed proportions of cows that had conceived at each interval as well as by the end of interval 5 were calculated for each category. For any of the 5 intervals, cows that had already conceived at the beginning of the interval considered were not included in the calculation of this observed proportion because they were not at risk. Predicted proportions were plotted against the mean observed proportions of conception in each category. Finally, the mean predicted and observed probabilities of pregnancy by the end of interval 5 were calculated at the herd level.

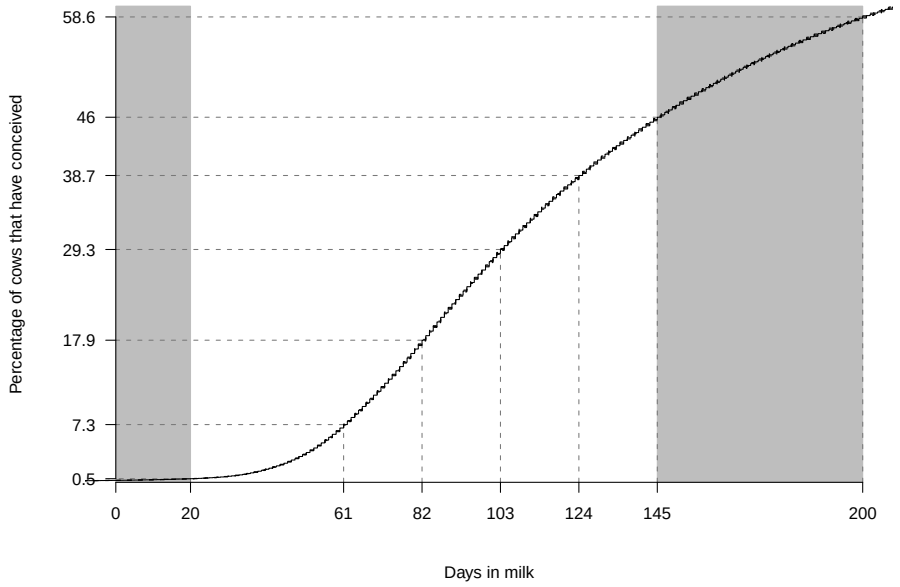
7.4 Results

7.4.1 Calving to Conception Interval

Overall, 73 % of the cows recalved in the full dataset (see Chapter 3). The cumulative Kaplan-Meier survivor curve for the calving to conception interval for the lactations for which calving occurred in 2004 and 2005 is presented in Figure 7.1. The percentage of cows which had conceived by day 20, 61, 82, 103, 124 which were the start of interval 1 to 5 and the end of interval 5 were 0.5, 7.3, 17.9, 29.3, 38.7 and 46 respectively.

The distribution of the percentage of cows that conceived during each interval, of those eligible, in the 500 herds used for parameter estimation is presented in Figure 7.2. For cows eligible to become pregnant in each interval, the herd median (interquartile range) percentages of concep-

Figure 7.1: Cumulative Kaplan-Meier survivor curve for the interval calving to conception between calving and 200 days in milk for 2,128 herds between 2004 and 2005. The values on the x-axis are the days at which each interval starts and the values on the y-axis the corresponding percentage of cows that have conceived up to this stage. Greyed areas were not modelled.



tion were 6.4 (2.6-10.9), 12.9 (8.4-17.5), 16.7 (11.8-21.5), 15.9 (12.2-21.4) and 15.3 (10.3-20.0) for intervals 1 to 5.

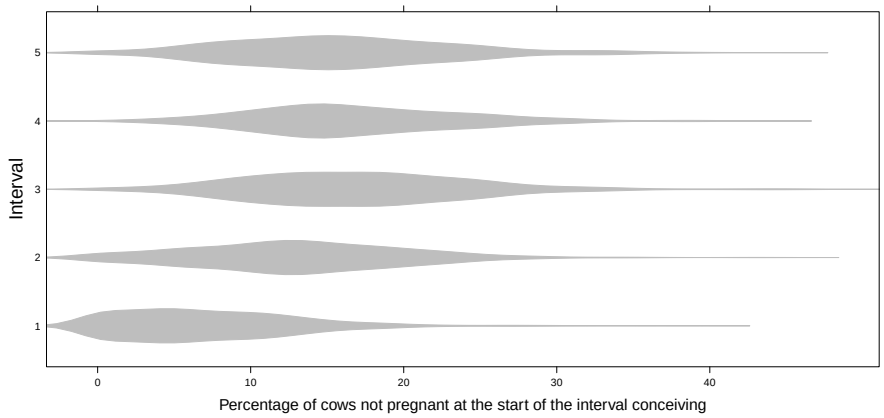
7.4.2 Corrections for stage of lactation and season

Descriptions of changes in milk production with stage of lactation between 5 and 60 days in milk and season are presented in Figures 7.3, 7.4 and 7.5. Parameter estimates of the different models for stage of lactation and stage of lactation and season are presented in Appendix C. Different curves were used for parity one and greater parities for milk yield and somatic cell count (Chapter 4). A greater mean was associated with a higher standard deviation, except for lactose. Variations in the mean with season were important for the fat to protein ratio and for the percentage of butterfat.

7.4.3 Models

A model with variables corrected for stage of lactation or stage of lactation and season was constructed. The results of this model are presented in Table 7.1. The baseline probability of conception was accounted for by using a different coefficient for each of the five intervals. Because all the variables included were centred, the exponentials of these coefficients represent the odds-ratio of a conception during the interval considered. The odds-ratio of conception

Figure 7.2: Violin plot (Hintze & Nelson, 1998) of the distribution of herd percentages of conception on each interval. The width of each violin represents the density of the percentages of conception on the x-axis.



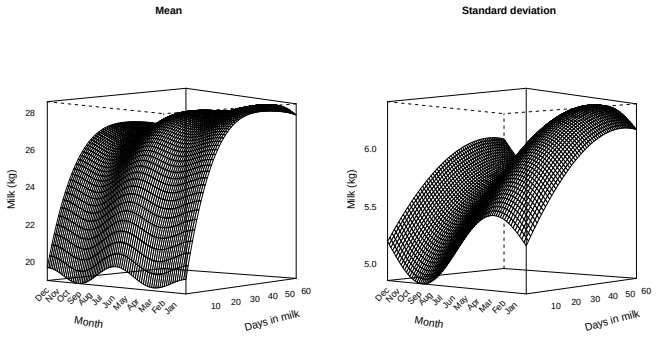
increased from interval one to interval three and decreased thereafter. The herd variances for the probability of conception decreased from interval one to interval four and was similar for interval four and five. The odds-ratio associated with the variables retained ranged between 0.928 (Milk second test-day) and 1.164 (percentage of protein second test-day) indicating relatively mild effects, even though in the case of weight of milk, a quadratic term reinforced this effect. Interactions with $ln(interval)$ indicated a reduction of some effects with time. The probabilities of conception predicted by the fixed effects of the model for each variable are plotted in Figure 7.6. The largest effects were observed for weight of milk on the second test-day, percentage of protein on the second test-day and lactose on the first test-day.

7.4.4 Model checking and validation

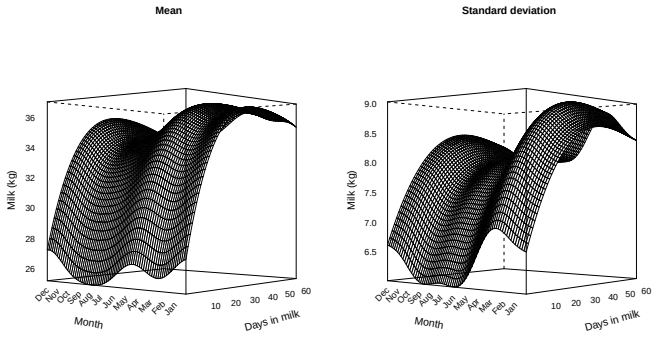
The model predicted the data reasonably well. Figure 7.7 shows the plots of predicted versus observed probabilities of conception using both random and fixed effects to predict the data used for parameter estimation. Figure 7.8 shows the plot of observed versus predicted probabilities using only the fixed effects in the model on the validation dataset. Predictions aggregated at the herd level on the validation dataset are presented in Figure 7.9. The relation between the predicted and observed proportions was linear. Linear regression was carried out on the individual cow and herd proportions presented in Figure 7.7, 7.8 and 7.9 and using the R lm function (R Development Core Team, 2009). The regressions specifications were $observed = \alpha + \beta \times predicted$. Estimates of α, β and the adjusted R squared for these regression models are presented in Table 7.2. A β of 1 indicated that for each increase of 1 unit in the predicted proportions, there was an increase of one unit in the observed proportion. Values greater or

Figure 7.3: Variations in mean and standard deviation for milk yield in parity 1 and greater than 1 and for lactose with stage of lactation between 5 and 60 days in milk and month of the year.

(a) Milk Yield Parity 1



(b) Milk Yield Parity > 1



(c) Lactose

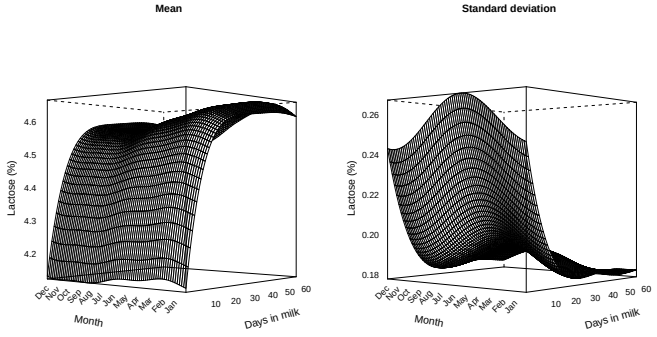
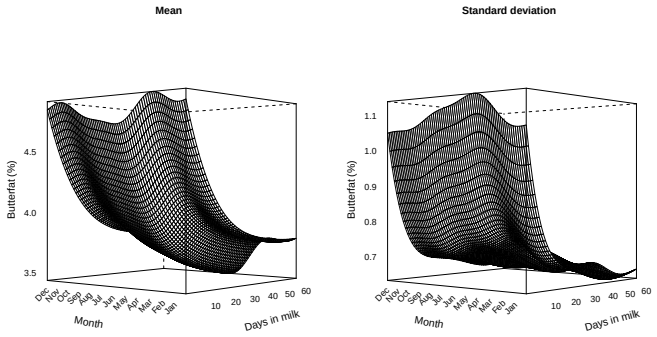
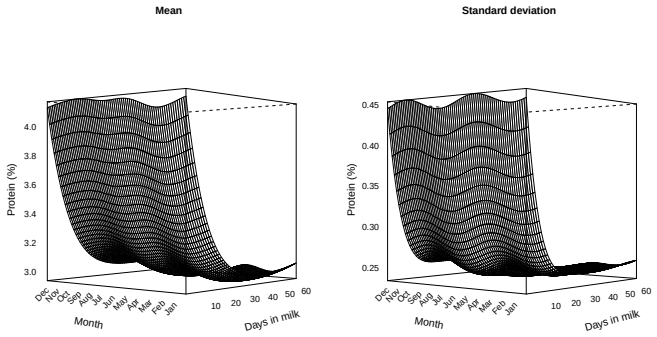


Figure 7.4: Variations in mean and standard deviation for butterfat, protein and fat to protein ratio with stage of lactation between 5 and 60 days in milk and month of the year.

(a) Butterfat



(b) Protein



(c) Fat to Protein ratio

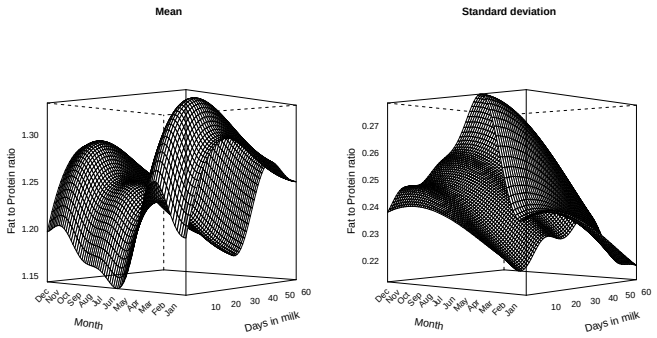
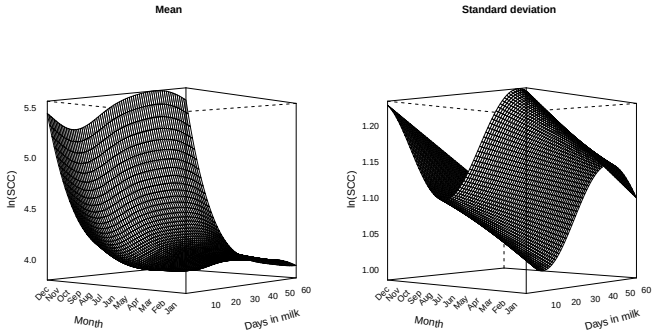


Figure 7.5: Variations in mean and standard deviation for somatic cell count in parity 1 and greater than one with stage of lactation between 5 and 60 days in milk and month of the year.

(a) ln Somatic Cell Count parity 1



(b) ln Somatic Cell Count parity > 1

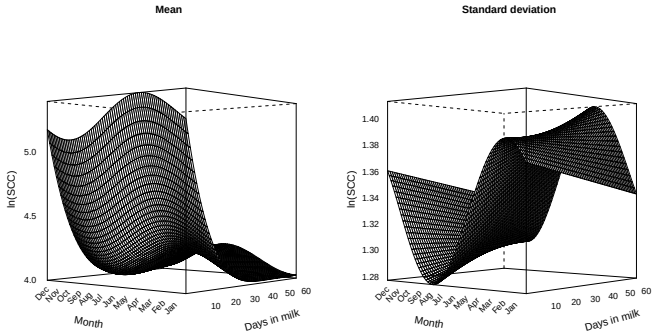


Table 7.1: Results of the model for the influence of milk quantity and composition on the probability of conception between 20 and 145 days in milk.

(a) Fixed effects

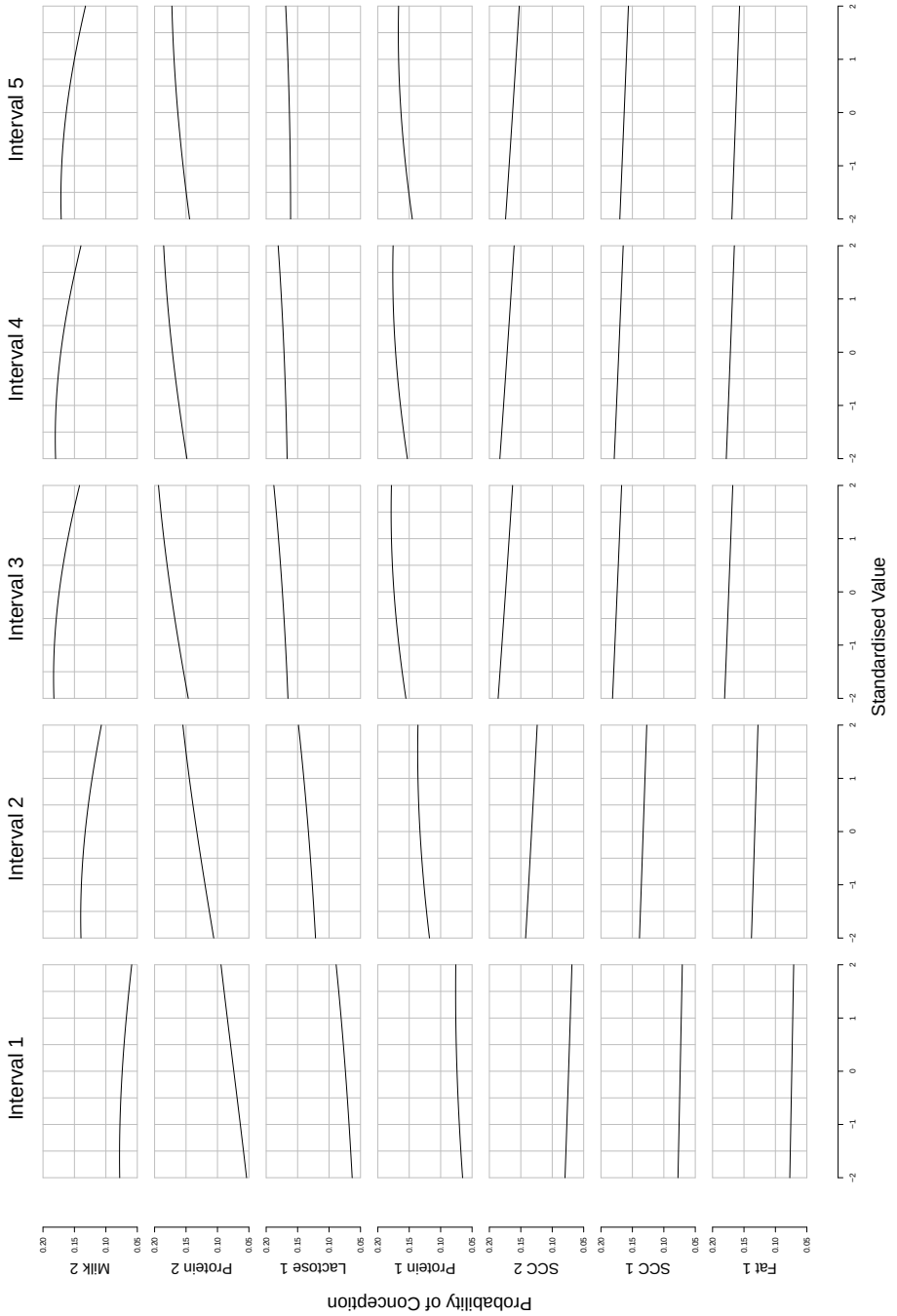
Variable[a]	Test-day	Corrected[b]	β	Standard error	e^{β}
Interval 1 [c]			−2.527	0.039	0.080
Interval 2 [c]			−1.876	0.027	0.153
Interval 3 [c]			−1.556	0.024	0.211
Interval 4 [c]			−1.574	0.022	0.207
Interval 5 [c]			−1.636	0.024	0.195
Weight of milk	2	dim	−0.075	0.010	0.928
Weight of milk^2	2	dim	−0.023	0.006	0.977
Fat	1	dim+seas	−0.023	0.009	0.977
Protein	1	dim	0.041	0.010	1.042
Protein^2	1	dim	−0.014	0.005	0.986
Protein	2	dim	0.152	0.018	1.164
Protein^2	2	dim	−0.010	0.005	0.990
Protein * ln(Interval) [d]	2	dim	−0.061	0.005	0.941
Lactose	1	dim+seas	0.092	0.018	1.096
Lactose^2	1	dim+seas	0.004	0.002	1.004
Lactose * ln(Interval) [d]	1	dim+seas	−0.048	0.016	0.953
Cell	1	dim	−0.025	0.010	0.975
Cell	2	dim	−0.040	0.010	0.961

(b) Variance-covariance matrix of random effects for intervals 1 to 5. Standard errors are in parenthesis.

	1	2	3	4	5
1	0.500 (0.045)				
2	0.184 (0.023)	0.198 (0.020)			
3	0.043 (0.019)	0.113 (0.014)	0.132 (0.015)		
4	0.027 (0.017)	0.079 (0.012)	0.098 (0.011)	0.081 (0.013)	
5	−0.048 (0.019)	0.053 (0.013)	0.095 (0.012)	0.087 (0.011)	0.088 (0.016)

[a] Variable^2: Variable to the square
[b] dim: day in milk ; seas: day of the year
[c] Categorical (1/0)
[d] Interval treated as continuous

Figure 7.6: Probability of conception predicted by the fixed part of the model for each of the variables retained for each interval.



smaller than 1 indicated that the observed proportions were increasing by more and less than 1 for each increase of 1 in the prediction respectively. The R squared values measured the dispersion in the observed values around the fitted lines. Values closer to one were indicative of observed values closer to the fitted lines. The model predicted groups of individual cows correctly: the βs of the linear regression were between 0.7 and 1.2 and the adjusted R squared greater than 0.8 for the first three intervals for both individual cow lactation models. The effects predicted by the model were greater than the observed ones for intervals 4 and 5 with βs of 0.660 and 0.773 and R squared of 0.808 and 0.713 for the full prediction and βs of 0.582 and 0.354 and R squared of 0.758 and 0.575 for the prediction on the validation dataset. Because the model over predicted the probability of conception during the last intervals, the predicted probabilities of individual cow conception by the end of interval 5 were 15.6 and 18.2 % greater than observed. However, the βs were 1.02 for both predictions meaning that the ranking of individual cows was overall accurate. For the validation data, the percentiles 5, 10, 25, 50, 75, 90 and 95 for the predicted probability of conception by the end of interval 5 were 0.44, 0.46, 0.50, 0.53, 0.57, 0.59 and 0.61. The proportion of cows that had conceived by the end of interval 5 for cows predicted between the percentiles 0-5 ; 5-10 ; 10-25 ; 25-50 ; 50-75 ; 75-90 ; 90-95 and 95-100 were 0.24, 0.28, 0.30, 0.34, 0.38, 0.41, 0.43 and 0.45.

The predictions were less accurate at the herd level with adjusted R squared between 0.002 and 0.072.

7.5 Discussion

This study identified an association between milk quantity and composition at the start of lactation and the probability of conception before 145 days in milk. This probability increased with lower milk production on the second test-day, higher percentage of protein on the second test-day and higher percentage of lactose on the first test-day. Positive associations were of a limited magnitude but also significant with the percentage of protein on the first test-day, the percentage of butterfat on the first test-day and somatic cell count on both test-days. Characteristics of milk production on the second test-day of lactation were of more importance, possibly because they were related to the energy balance at peak yield. While there was a good agreement between observed and predicted probabilities of conception at the cow-lactation level, predicted probabilities of conception aggregated at the herd level were not useful in ranking individual herds.

The link between higher milk production and poorer reproduction is well established (Lucy, 2001). Although this study identified associations of the same direction as previous studies (de Vries & Veerkamp, 2000; Heuer *et al.* , 2000; Duffield *et al.* , 1997; Grieve *et al.* , 1986),

Figure 7.7: Observed versus predicted probability of conception at the individual cow level. Predictions were generated from both fixed and random effects on the dataset used for parameter estimation. Each dot is a percentile (1 to 100) of predicted values. Regression lines of observed versus predicted values are plotted. The coefficients of this regression are presented in Table 7.2.

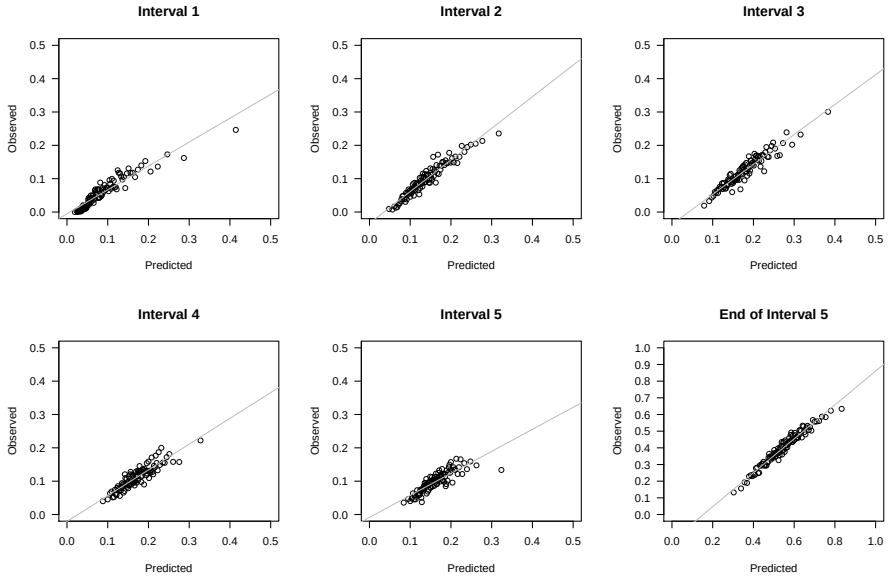


Figure 7.8: Observed versus predicted probability of conception at the individual cow level. Predictions were generated from fixed effects on the validation dataset. Each dot is a percentile (1 to 100) of predicted values. Regression lines of observed versus predicted values are plotted. The coefficients of this regression are presented in Table 7.2.

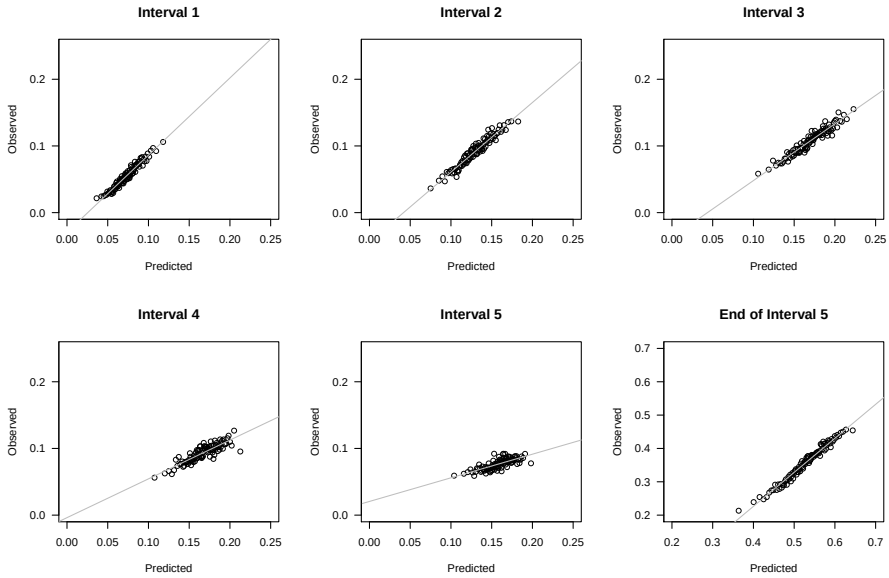
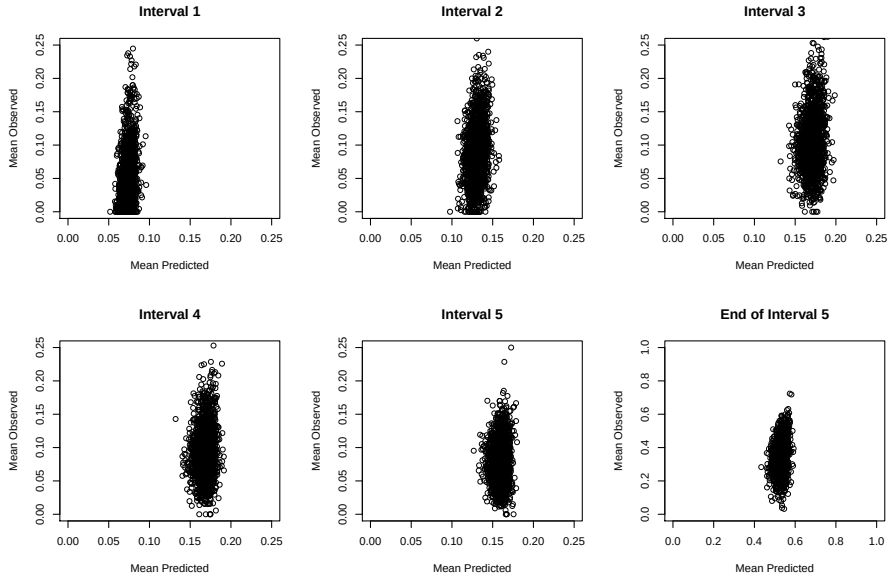


Figure 7.9: Observed versus predicted probability of conception at the herd level. Each dot is a herd of predicted values.



the ranking of the different variables was not the same. The outcome of these previous studies was an estimated energy balance while in the present study an interval calving to conception estimated from recalving dates was used. Kristula *et al.* (1995) found a predominant effect of the percentage and weight of butterfat, but they used data from the first milk recording and looked at the pregnancy rate on the first insemination when in the present study the first two test-days and conceptions before 145 days in milk were used. An association between milk butterfat and probability of conception was present but of a smaller magnitude than the association with percentages of lactose and protein. The effect of energy balance on the percentage of butterfat seemed to be more important at the start of lactation in the study by de Vries & Veerkamp (2000) which agrees with our results. At the start of lactation, as the cow has to adapt her metabolism to the surge in energy demand fat is mobilised. This mobilisation has many implications some of which have been linked to poorer reproduction and health disorders (Duffield *et al.* , 2009; Heuer *et al.* , 1999).

After milk yield, the percentage of protein followed by the percentage of lactose were the most important variables associated with the probability of conception in all the models tested. Both protein and lactose are synthesised in the mammary gland (Linzell & Peaker, 1971) and these processes require energy. Attempts to predict milk protein content from amino acids availability have been of limited success and a link between energetic metabolism and milk protein suggested (Hanigan *et al.* , 2002).

Increasing somatic cell counts on both test-days were associated with lower probabilities of

Table 7.2: Association between predicted and observed probability of conception at interval 1 to 5 and at the end of interval 5 at the individual cow-lactation and herd levels for the years 2004 and 2005. Associations were measured using linear models whose specifications were $Observed = \alpha + \beta \times Predicted$.

(a) Individual cow level - Full prediction [a]

Interval	α	β	Adjusted R^2
1	−0.005	0.716	0.889
2	−0.034	0.949	0.918
3	−0.037	0.897	0.906
4	−0.021	0.773	0.808
5	−0.009	0.660	0.713
End of 5	−0.156	1.017	0.960

(b) Individual cow level - Prediction from fixed effects on validation data [a]

Interval	α	β	Adjusted R^2
1	−0.029	1.157	0.952
2	−0.044	1.044	0.938
3	−0.037	0.851	0.927
4	−0.004	0.582	0.758
5	0.021	0.354	0.575
End of 5	−0.182	1.020	0.974

(c) Herd level - Prediction from fixed effects on validation data

Interval	α	β	Adjusted R^2
1	−0.105	2.160	0.072
2	−0.092	1.396	0.057
3	−0.060	0.977	0.033
4	−0.012	0.633	0.017
5	0.042	0.218	0.002
End of 5	−0.290	1.213	0.068

[a] Data grouped according to percentile of prediction (percentiles 1 to 99 used to define 100 groups).

conception. This is in line with earlier observations (Santos *et al.* , 2004; Schrick *et al.* , 2001). The reaction of individual cows to infection depends on characteristics of the cow as well as characteristics of the pathogen (Burvenich *et al.* , 2003). Cows experimentally infected with *Escherichia coli* showed severer symptoms when they were ketotic than control cows (Kremer *et al.* , 1993). It is not possible to know from this study whether mastitis causes delayed conception or whether there is a common cause for mastitis and delayed conception.

One of the differences between this study and some of the earlier ones was the correction of each variable for days in milk as opposed to having a single coefficient for days in milk in the model. This was justified because the percentage of protein decreases more rapidly than the percentage of butterfat at the start of lactation. As a result, there is an initial increase followed

by a decrease in the fat to protein ratio. Because cows are sampled at random time points, they will all undergo similar variations to a certain extent. A further correction was applied for the season because milk production exhibits variation with the time of the year and the link between this variation and the energy status was unknown. Variables corrected for season were kept when the deviance of the model in which they were included was lower than the deviance of the same model with a correction for stage of lactation alone. This was the case for the percentage of butterfat and the percentage of lactose. Feeding regimen is known to have an effect on milk butterfat (Bauman & Griinari, 2003). In England and Wales, most dairy cows are housed during winter and are fed at pasture when the weather permits it which might account for the observed patterns. The reasons for the variation in lactose with season is less clear.

One limitation of the present study was the estimation of the calving to conception from the re-calving dates. Cows culled early because they were not pregnant were omitted from the present study. Thus, it is possible that the reported associations underestimate the real effects. A further step would be to test the predictions on farm and to see whether they can be used to implement corrective action in cows with a low predicted probability of conception.

7.6 Conclusion

Individual cow information on milk quantity and composition at the start of lactation can be used to predict the individual cow interval calving to conception. Milk yield, percentage of protein on the second test-day ; percentage of lactose, percentage of protein on the first test-day were positively associated with the probability of conception before 145 days in milk. Percentage of butterfat and somatic cell counts on both test-days were negatively associated with the probability of conception before 145 days in milk.

Chapter 8

General Discussion and Conclusion

8.1 Discussion

8.1.1 Structural Aspects of Dairy Farming in England and Wales

Milk production in England and Wales undergoes seasonal variations (Chapter 3). Although not documented in the scientific literature for England and Wales, such variations can result in an inadequacy between milk supply and demand. Because milk production varies with stage of lactation, part of the observed variations can be explained by the uneven distribution of calvings throughout the year. Overall calving patterns depend on calving patterns in heifers as well as of the management of reproduction in later lactations. There is a marked peak in heifers calving in September which determines the overall calving distribution. Because the median interval between consecutive calvings is 391 days, this peak tends to flatten out with successive lactations. It would be interesting to model the impact of altering the calving patterns in heifers on the global milk production, accounting for the performance in reproduction in successive lactations. Parameters for lactation curves and reproduction derived in chapters 3 and 4 could be used for this purpose. After consultation with milk processors, different scenarios could be tested in order to adapt the supply to the demand.

In the UK, changes are ongoing in the number and size of dairy herds. There is a trend for less herds of increasing size. In the sample used in this thesis, this was matched by an increase in the number of 5.2 cows recorded and and 143 kg of milk produced per test-day between 2004 and 2006. This trend has continued since 2006 and the number of dairy herds went from 13,270 in December 2006 to 11,709 in July 2009. It is hard, from milk recording data alone, to anticipate what will be the consequences of such changes in the long run.

8.1.2 Somatic Cell Count and Mastitis

SCC is widely used in the UK and worldwide for both the detection of mastitis and the resolution of herd SCC problems. Various thresholds have been recommended or are used. In Chapter 5, it was shown that a threshold of 200,000 cells/mL can be used to diagnose and set targets for the resolution of herd problems. But, because the lactation curve for SCC is lower and flatter during the first lactation than in later ones (Chapter 4); the same threshold, but different targets should be used for first lactation cows and older cows (Chapter 5). Looking at the movements across this threshold can be used to predict BMSCC and to estimate the expected impact of corrective action. The main contributors to BMSCC are cows staying above 200,000 cells/mL for two consecutive recordings. Even though the movements across this threshold could be modelled and described the variations with stage of lactation and parity accurately, they were not useful in predicting test-day BMSCC (Chapter 6). As could be expected from the lactation curves, the probabilities of moving or staying above 200,000 cells/mL increased with stage of lactation and were markedly higher for cows of parity greater than one. But it is possible that these variations reflect physiological variations and that inflammation results in higher contribution in BMSCC and is intrinsically more unpredictable.

8.1.3 Reproduction

Modern dairy cows are under intense energy demand at the start of lactation which is associated with a higher frequency of health disorders such as ketosis, displaced abomasum, metritis, hypocalcemia or mastitis. Failure to conceive is one of the consequences associated with this negative energy balance. In Chapter 6, an association was observed between milk quantity and composition and the probability of conception. This association was less important as lactation was progressing but was still noticeable at 145 days in milk. Milk yield around the peak was found to be negatively correlated with the probability of conception. It can be hypothesised that energy exports and hence negative energy balance increase with yield. Higher lactose and protein contents, lower fat content were associated with higher probabilities of conception. The negative correlation between fat percentage and energy balance have been observed in previous studies and it seems that when cows mobilise fat in early lactation, milk butterfat increases. Associations with protein have been reported but, to our knowledge, associations with lactose have not. Because the 2 components are synthesised and this synthesis requires energy, it could be that in case of negative energy balance, less energy can be allocated to their synthesis and their concentration in milk is lower. The link between mastitis and reproduction has been observed in previous studies. It is hard to know whether mastitis has detrimental impact on reproduction or whether the probability of showing mastitis signs increases with negative energy balance. All these events have been linked in the past, but association or not even precedence

are proofs of causality.

8.2 Conclusion

For as long as farming has existed, farmers have tried to improve yields by selecting the best performers. With the advent of modern genetic selection, this process has been dramatically improved and has resulted in a massive increase in the quantity produced and a reduction in the genetic variability. In England and Wales, an average cow in her second lactation currently produces around 8,500 kg of milk between calving and 305 days of lactation. This increase in production is marked at the start of lactation, when the cow has to mobilise a vast amount of energy that will be exported in milk. For example, it was estimated in the present research that a multiparous cow exports approximately 900 kg of milk, 35 kg of butterfat and 30 kg of protein during her first month of lactation. Selection on a limited set of parameters has the disadvantage of being prone to unintended consequences because the improvement is operated on what is measured and can, in the same time, degrade unobserved parameters. Decreased resistance to mastitis and poorer reproduction have been some of the consequences of a selection directed mainly at milk yield and constituents. Milk recording is powerful tool to assess and resolve individual herd mastitis problems. The distribution of somatic cell count patterns and their contribution to bulk milk somatic cell, provided in this thesis, can be used to tackle such problems. Regarding reproduction, cows producing more milk at the peak and a milk containing less protein and lactose in early lactation were identified as less likely to conceive before 145 days in milk. In some countries, there have been attempts to mitigate these trends by incorporating new parameters in the selection programs such as somatic cell count, fertility or longevity. However, the metabolic consequences of increasing what the cow exports and their link with endocrine balance and immunity are still poorly understood and a global picture is missing. This also means that genetically 'improved' cows need to be cared by highly trained staff and in standardised conditions. Currently, dairy farming in England and Wales seems to be characterised by increasingly hard economic conditions. It should be ensured that the conditions remain attractive for students to train and get involved in the industry. The country also relies on imported protein to feed these cows. While this is feasible in the context of cheap energy, the prospect oil rarefaction in the medium to long term and its consequences on energy price should be considered.

Appendix A

An example of WinBUGS model

WinBUGS model for model 2 in chapter 5.

Table A.1: WinBUGS functions used in the model.

Function	Description
$\sim$	Stochastic relation
$<-$	Deterministic relation
$vector[i]$	i^{th} value of $vector$
$matrix[i, j]$	Value in the i^{th} row and j^{th} column of $matrix$
$dnorm$(Mean, Precision[a])	Normal distribution
$dunif$(Minimum, Maximum)	Uniform distribution
pow(Number, Power)	Number raised to Power

[a]Precision $= \frac{1}{Variance}$

Table A.2: Data used in the model.

Training dataset	
N	Number of observations
ebmscc	Observed BMSCC
nHcL, nHcH ...	Percentage of herd-years in categories Hc.L, Hc.H ...
Validation dataset	
P	Number of observations
pHcL, pHcH ...	Percentage of herd-years in categories Hc.L, Hc.H ...

Table A.3: Parameters used in the model.

Training dataset	
mubmscc	Mean for BMSCC
tau.bmscc	Precision for BMSCC
beta	Vector of coefficients
pebmscc	Predicted BMSCC
u.herd	Vector of herd intercepts
Validation dataset	
pbmscc	Predicted BMSCC
pmubmscc	Mean for predicted BMSCC

```
model {

## Training dataset
for(i in 1:N){

ebmscc[i]  ~ dnorm(mubmscc[i], tau.bmscc)
pebmscc[i] ~ dnorm(mubmscc[i], tau.bmscc)

mubmscc[i] <- mubmscc1[i] + mubmscc2[i]

mubmscc1[i] <-           beta[1] * nHcL[i] +
                         beta[2] * nHcH[i] +
                         beta[3] * nCcLL[i] +
                         beta[4] * nCcLH[i] +
                         beta[5] * nCcHL[i] +
                         beta[6] * nCcHH[i] +
                         beta[7] * nHlLL[i] +
                         beta[8] * nHlLH[i]

mubmscc2[i] <-           beta[9] * nHlHL[i] +
                         beta[10] * nHlHH[i] +
                         beta[11] * nClLH[i] +
                         beta[12] * nClHL[i] +
                         beta[13] * nClHH[i] +
                         u.herd[hd_id[i]]

}
```

```
# Priors for fixed effects
for(k in 1:13) {beta[k] ~ dnorm(0,0.1)}
tau.bmscc <- pow(sigma.bmscc, -2)
sigma.bmscc ~ dunif(0, 100)

# herd intercept
 for(j in 1:nhd){
 u.herd[j] ~ dnorm(mu.herd, tau.herd)
      }

# priors for random effects
 mu.herd  ~ dnorm(0, .01)

 tau.herd <- pow(sigma.herd, -2)
 sigma.herd ~ dunif(0, 100)


## Validation dataset (Predictions)

for(m in 1:P){
# a herd random effect is generated at each iteration
pu.herd[m] ~ dnorm(mu.herd, tau.herd)

pbmscc[m] ~ dnorm(pmubmscc[m], tau.bmscc)

pmubmscc[m] <- pmubmscc1[m] + pmubmscc2[m]

pmubmscc1[m] <-          beta[1] * pHcL[m] +
                         beta[2] * pHcH[m] +
                         beta[3] * pCcLL[m] +
                         beta[4] * pCcLH[m] +
                         beta[5] * pCcHL[m] +
                         beta[6] * pCcHH[m] +
                         beta[7] * pHlLL[m] +
                         beta[8] * pHlLH[m]
```

```
pmubmscc2[m] <-          beta[9] * pHlHL[m] +
                         beta[10] * pHlHH[m] +
                         beta[11] * pClLH[m] +
                         beta[12] * pClHL[m] +
                         beta[13] * pClHH[m] +
                         pu.herd[m]
}

}
```

Appendix B

Supplement to Chapter 6. Details on the state transition model

B.1 WinBUGS code

```
## State: 1-Low; 2-High; 3-dry; 4-Culled
## pstate: 1-Low; 2-High; 3-dry; 4-first
## Dim: Days in milk / 100
## day100: 1 when days in milk > 100; 0 otherwise
## day250: 1 when days in milk > 250; 0 otherwise
## par2: 1 when parity = 1; 0 otherwise


model
{
    for (i in 1:N) {

        State[i, 1:4] ~ dmulti(pi[i, 1:4], 1)

        for (j in 1:4) {
            pi[i, j] <- p[i, j]/sum(p[i, ])
        }

        p[i, 1] <- 1
```

```
# transition to High
log(p[i, 2]) <- beta[1, i] + beta[2, i] + beta[3, i] + beta[4, i]

## from Low
beta[1, i] <-
  pstate[i, 1] * (
       theta[1] + u1[hd_id[i], 1] + par2[i] * theta[2] +
       (1 - par2[i]) *  (Dim[i] * theta[3] + pow(Dim[i], 2) * theta[4]) +
       par2[i] * (Dim[i] * theta[5] + pow(Dim[i], 2) * theta[6]))

## from High
beta[2, i] <-
  pstate[i, 2] * (
       theta[7] + u1[hd_id[i], 2] + par2[i] * theta[8] +
       (1 - par2[i]) * (Dim[i] * theta[9] + pow(Dim[i], 2) * theta[10]) +
       par2[i] * (Dim[i] * theta[11] + pow(Dim[i], 2) * theta[12]))

## from dry
beta[3, i] <-
  pstate[i, 3] * (
       par2[i] * (1 - day100[i]) * (theta[13] + u1[hd_id[i], 3]) +
       day100[i] * gamma)

## from first
beta[4, i] <-
  pstate[i, 4] * (theta[14] + u1[hd_id[i], 4] + par2[i] * gamma)


# transition to dry
log(p[i, 3]) <- beta[5, i] + beta[6, i] + beta[7, i] + beta[8, i]

## from Low
beta[5, i] <-
  pstate[i, 1] * (
       theta[15] + u1[hd_id[i], 5] + par2[i] * theta[16] +
       (1 - day250[i]) * (2.5 - Dim[i]) * theta[17] +
       day250[i] * (
```

```
          (Dim[i] - 2.5) * theta[18] +
          pow((Dim[i] - 2.5), 2) * theta[19]))

## from High
beta[6, i] <-
   pstate[i, 2] * (
        theta[20] + u1[hd_id[i], 6] + par2[i] * theta[21] +
        (1 - day250[i]) * (2.5 - Dim[i]) * theta[22] +
        day250[i] * (
          (Dim[i] - 2.5) * theta[23] +
          pow((Dim[i] - 2.5), 2) * theta[24]))

## from dry
beta[7, i] <-
   pstate[i, 3] * (
        u1[hd_id[i], 7] + (1 - day100[i]) * theta[25] +
        day100[i] * theta[26])

## from first (impossible)
beta[8, i] <- pstate[i, 4] * gamma


# transition to culling
log(p[i, 4]) <- beta[9, i] + beta[10, i] + beta[11, i] + beta[12, i]


## from Low
beta[9, i] <-
   pstate[i, 1] * (
    theta[27] + u1[hd_id[i], 8] + par2[i] * theta[28] +
    (1 - day250[i]) * (2.5 - Dim[i]) * theta[29] +
    day250[i] * (
      (Dim[i] - 2.5) * theta[30] +
      pow((Dim[i] - 2.5), 2) * theta[31]))

## from High
beta[10, i] <-
```

```
   pstate[i, 2] * (
    theta[32] + u1[hd_id[i], 9] + par2[i] * theta[33] +
    (1 - day250[i]) * (2.5 - Dim[i]) * theta[34] +
    day250[i] * (
     (Dim[i] - 2.5) * theta[35] +
     pow((Dim[i] - 2.5), 2) * theta[36]))

## from dry (impossible)
beta[11, i] <- pstate[i, 3] * gamma

## from first (impossible)
beta[12, i] <- pstate[i, 4] * gamma

    }


# prior for theta
    for (k in 1:36) {
        theta[k] ~ dnorm(0, 0.01)
    }

# for transitions that are impossible
# exp(-2000) ~ 0
gamma <- -2000


# prior for random effects
for (j in 1:nhd) {
        u1[j, 1:9] ~ dmnorm(zero1[1:9], tau.u1[1:9, 1:9])
    }

for (i in 1:9) {
        zero1[i] <- 0
    }

tau.u1[1:9, 1:9] ~ dwish(R2[1:9, 1:9], 9)
sigma.u1[1:9, 1:9] <- inverse(tau.u1[, ])
```

```
}
```

B.2 R function to generate the transition curves

B.2.1 Function

```
PredProb <- function(Dim, parity, pstat, theta){

day100 <- ifelse(Dim < 101, 0, 1)
day250 <- ifelse(Dim < 251, 0, 1)
Dim    <- Dim / 100
par2   <- ifelse(parity == 1, 0, 1)
pstate <- rep(0, 4)
pstate[pstat] <- 1


BeTa  <- vector(mode="numeric", length=12)
probs <- vector(mode="numeric", length=4)

### to state 2
BeTa[1] <- pstate[1] * (theta[1] +
                        par2 * theta[2] +
                        (1-par2) * Dim * theta[3] +
                        (1-par2) * Dim^2 * theta[4] +
                        par2 * Dim * theta[5] +
                        par2 * Dim^2 * theta[6])

BeTa[2] <- pstate[2] * (theta[7] +
                        par2 * theta[8] +
                        (1-par2) * Dim * theta[9] +
                        (1-par2) * Dim^2 * theta[10] +
                        par2 * Dim * theta[11] +
                        par2 * Dim^2 * theta[12])

BeTa[3] <- pstate[3] * (par2 * (1-day100) * theta[13] +
                        day100 * -2000)
```

```
BeTa[4] <- pstate[4] * (theta[14] + par2 * -2000)


### to state 3

BeTa[5] <- pstate[1] * (theta[15] +
                        par2 * theta[16] +
                        (1 - day250) * (2.5 - Dim) * theta[17] +
                        day250 * (Dim - 2.5) * theta[18] +
                        day250 * (Dim - 2.5)^2 * theta[19])

BeTa[6] <- pstate[2] * (theta[20] +
                        par2 * theta[21] +
                        (1 - day250) * (2.5 - Dim) * theta[22] +
                        day250 * (Dim - 2.5) * theta[23] +
                        day250 * (Dim - 2.5)^2 * theta[24])

BeTa[7] <- pstate[3] * ((1-day100) * theta[25] +
                         day100 * theta[26])

BeTa[8] <- pstate[4] * -2000


### to state 4

BeTa[9] <- pstate[1] * (theta[27] +
                        par2 * theta[28] +
                        (1 - day250) * (2.5 - Dim) * theta[29] +
                        day250 * (Dim - 2.5) * theta[30] +
                        day250 * (Dim - 2.5)^2 * theta[31])

BeTa[10] <- pstate[2] * (theta[32] +
                         par2 * theta[33] +
                         (1 - day250) * (2.5 - Dim) * theta[34] +
                         day250 * (Dim - 2.5) * theta[35] +
                         day250 * (Dim - 2.5)^2 * theta[36])
```

```
### probabilities
probs[1] <- 1
probs[2] <- exp(BeTa[1] + BeTa[2] + BeTa[3] + BeTa[4])
probs[3] <- exp(BeTa[5] + BeTa[6] + BeTa[7] + BeTa[8])
probs[4] <- exp(BeTa[9] + BeTa[10] + BeTa[11] + BeTa[12])

probs <- probs / sum(probs)
}
```

B.2.2 Values for theta[ ]

```
theta <-
   structure(c(-3.0055, 0.33045, 0.16685, -0.00090055, 0.42405,
   -0.01302, -0.91145, 0.7252, 0.5777, -0.0447, 0.6296, -0.052355,
   -1.309, -1.559, -3.1975, 0.1522, -7.536, 2.533, -0.5641, -1.952,
   0.71515, -5.3585, 2.438, -0.5351, -5.3625, 14.555, -4.8865, 0.2999,
   0.12165, 1.5455, -0.2264, -2.45, 0.78135, -0.52115, 1.0085, -0.0919
   ), .Dim = c(1L, 36L), .Dimnames = list(NULL, c("theta[1]", "theta[2]",
   "theta[3]", "theta[4]", "theta[5]", "theta[6]", "theta[7]", "theta[8]",
   "theta[9]", "theta[10]", "theta[11]", "theta[12]", "theta[13]",
   "theta[14]", "theta[15]", "theta[16]", "theta[17]", "theta[18]",
   "theta[19]", "theta[20]", "theta[21]", "theta[22]", "theta[23]",
   "theta[24]", "theta[25]", "theta[26]", "theta[27]", "theta[28]",
   "theta[29]", "theta[30]", "theta[31]", "theta[32]", "theta[33]",
   "theta[34]", "theta[35]", "theta[36]")))
```

B.2.3 Prediction

The following line produces a matrix *pred* of probabilities of transition from *pstat* (set to 1) for cows of *parity* 1.

```
pred <- t(sapply(30:500,
   function(x) PredProb(Dim = x, parity = 1, pstat = 1, theta = theta)))
```

B.2.4 Variance-covariance matrix for the random effects

This matrix is provided for information. The above R code can easily be modified to incorporate the herd variability. The herd effects need to be sampled from a multivariate normal distribution with a vector of 0 for the means and the matrix provided below as the variance-covariance matrix.

```
sigma.u1 <-
   structure(c(0.3084, -0.01096, 0.21205, 0.04895, -0.063325, 0.089885,
   0.4533, -0.089325, 0.011505, -0.01096, 0.48005, -0.41545, 0.32595,
   0.26055, 0.049795, 0.1798, 0.2515, -0.00077065, 0.21205, -0.41545,
   0.97525, -0.3918, -0.4762, -0.01921, 0.2693, -0.55735, 0.11025,
   0.04895, 0.32595, -0.3918, 0.5414, 0.3065, -0.077725, 0.33975,
   0.3812, -0.13275, -0.063325, 0.26055, -0.4762, 0.3065, 0.4573,
   0.11705, 0.17505, 0.3603, 0.013185, 0.089885, 0.049795, -0.01921,
   -0.077725, 0.11705, 0.2673, 0.1698, -0.052465, 0.09662, 0.4533,
   0.1798, 0.2693, 0.33975, 0.17505, 0.1698, 2.622, -0.1627, 0.20925,
   -0.089325, 0.2515, -0.55735, 0.3812, 0.3603, -0.052465, -0.1627,
   0.67215, -0.10565, 0.011505, -0.00077065, 0.11025, -0.13275,
   0.013185, 0.09662, 0.20925, -0.10565, 0.25625), .Dim = c(9L,
   9L))
```

Appendix C

Parameter estimates of the linear models for milk yield and constituents as a function of days in milk and day of the year

Table C.1: Parameter estimates of the linear models for milk yield (mean and standard deviation) as a function of days in milk (1-60).

	Parity 1		Parity > 1	
Covariates[a]	Mean	Standard Deviation	Mean	Standard Deviation
Intercept	1.877e+01	5.111e+00	2.447e+01	6.276e+00
Dim	6.914e-01	3.622e-02	1.039e+00	1.305e-01
Dim^2	-2.045e-02	-3.249e-04	-4.038e-02	-2.496e-03
Dim^3	2.816e-04		8.507e-04	1.478e-05
Dim^4	-1.505e-06		-9.449e-06	
Dim^5			4.181e-08	

[a]Dim: days in milk

Table C.2: Parameter estimates of the linear models for the percentage of lactose and the percentage of butterfat (mean and standard deviation) as a function of days in milk (1-60).

	Lactose		Butterfat	
Covariates[a]	Mean	Standard Deviation	Mean	Standard Deviation
Intercept	4.042e+00	2.432e-01	4.995e+00	1.235e+00
Dim	7.351e-02	-4.079e-03	-1.083e-01	-8.824e-02
Dim^2	-4.073e-03	9.362e-05	4.243e-03	6.162e-03
Dim^3	1.201e-04	-6.908e-07	-9.357e-05	-2.218e-04
Dim^4	-1.945e-06		1.069e-06	4.220e-06
Dim^5	1.627e-08		-4.824e-09	-4.041e-08
Dim^6	-5.492e-11			1.535e-10

[a]Dim: days in milk

Table C.3: Parameter estimates of the linear models for the percentage of protein and the fat to protein ratio (mean and standard deviation) as a function of days in milk (1-60).

	Protein		Fat to Protein ratio	
Covariates[a]	Mean	Standard Deviation	Mean	Standard Deviation
Intercept	4.336e+00	4.827e-01	1.179e+00	3.181e-01
Dim	-1.521e-01	-3.846e-02	1.186e-02	-1.578e-02
Dim^2	7.777e-03	2.580e-03	-4.667e-04	1.344e-03
Dim^3	-2.278e-04	-8.921e-05	6.660e-06	-5.339e-05
Dim^4	3.886e-06	1.668e-06	-3.314e-08	1.062e-06
Dim^5	-3.539e-08	-1.599e-08		-1.037e-08
Dim^6	1.323e-10	6.155e-11		3.972e-11

[a]Dim: days in milk

Table C.4: Parameter estimates of the linear models for somatic cell count (mean and standard deviation) as a function of days in milk (1-60).

	Parity 1		Parity > 1	
Covariates[a]	Mean	Standard Deviation	Mean	Standard Deviation
Intercept	5.842e+00	1.168e+00	5.528e+00	1.277e+00
Dim	-1.956e-01	-1.677e-03	-1.536e-01	4.196e-03
Dim^2	9.253e-03		6.919e-03	-4.965e-05
Dim^3	-2.308e-04		-1.601e-04	
Dim^4	2.846e-06		1.857e-06	
Dim^5	-1.354e-08		-8.451e-09	

[a]Dim: days in milk

Table C.5: Parameter estimates of the linear models for milk yield (mean and standard deviation) as a function of days in milk (1-60) and day of the year (1-365).

	Parity 1		Parity > 1	
Covariates[a]	Mean	Standard Deviation	Mean	Standard Deviation
Intercept	1.877e+01	5.111e+00	2.535e+01	6.272e+00
Dim	6.913e-01	3.621e-02	8.267e-01	1.306e-01
Dim^2	-2.045e-02	-3.249e-04	-2.402e-02	-2.497e-03
Dim^3	2.816e-04		3.133e-04	1.479e-05
Dim^4	-1.505e-06		-1.611e-06	
cos(day)	-1.542e-01	6.552e-02	5.405e-01	3.264e-01
sin(day)	8.375e-02	2.872e-01	3.736e-01	3.210e-01
cos(2*day)	3.745e-01		3.183e-01	-7.454e-02
sin(2*day)	-1.232e-01		-4.579e-01	5.161e-02
cos(3*day)			2.212e-01	-1.872e-02
sin(3*day)				-8.995e-02
Dim*cos(day)			-1.523e-02	-2.367e-03
Dim*sin(day)	9.822e-03		1.746e-02	1.363e-03
Dim*cos(2*day)	-3.247e-03			
Dim*sin(2*day)	6.659e-03		9.050e-03	

[a]Dim: days in milk ; day: day of the year ; cos: *cosine* ; sin: *sine*

Table C.6: Parameter estimates of the linear models for the percentage of lactose and the percentage of butterfat (mean and standard deviation) as a function of days in milk (1-60) and day of the year (1-365).

	Lactose		Butterfat	
Covariates[a]	Mean	Standard Deviation	Mean	Standard Deviation
Intercept	4.074e+00	2.623e-01	4.894e+00	1.147e+00
Dim	6.402e-02	-7.558e-03	-8.383e-02	-6.174e-02
Dim^2	-3.101e-03	2.791e-04	2.354e-03	3.445e-03
Dim^3	7.434e-05	-4.406e-06	-3.156e-05	-9.398e-05
Dim^4	-8.680e-07	2.481e-08	1.643e-07	1.210e-06
Dim^5	3.915e-09			-5.871e-09
cos(day)		-1.168e-02	8.175e-02	-3.924e-02
sin(day)	1.281e-02	5.016e-03	2.280e-02	2.104e-02
cos(2*day)			-4.253e-02	-4.845e-03
sin(2*day)	2.468e-03		-1.172e-02	-8.832e-03
cos(3*day)	-2.891e-03		-1.222e-02	5.615e-03
sin(3*day)	2.511e-03		-2.879e-02	-5.104e-03
Dim*cos(day)	3.079e-04	1.617e-04	1.623e-03	7.133e-04
Dim*sin(day)	6.328e-04	-1.717e-04	-3.631e-04	-3.686e-04

[a]Dim: days in milk ; day: day of the year ; cos: *cosine* ; sin: *sine*

Table C.7: Parameter estimates of the linear models for the percentage of protein and the fat to protein ratio (mean and standard deviation) as a function of days in milk (1-60) and day of the year (1-365).

	Protein		Fat to Protein ratio	
Covariates[a]	Mean	Standard Deviation	Mean	Standard Deviation
Intercept	4.259e+00	4.828e-01	1.179e+00	2.548e-01
Dim	-1.292e-01	-3.847e-02	1.186e-02	2.770e-04
Dim^2	5.435e-03	2.580e-03	-4.667e-04	-2.281e-05
Dim^3	-1.177e-04	-8.921e-05	6.660e-06	1.588e-07
Dim^4	1.291e-06	1.668e-06	-3.314e-08	
Dim^5	-5.623e-09	-1.599e-08		
Dim^6		6.155e-11		
cos(day)	-1.572e-02		3.016e-02	-1.152e-02
sin(day)	-3.648e-02		2.109e-02	1.028e-02
cos(2*day)	3.373e-03	-2.652e-03	-1.854e-02	-6.610e-03
sin(2*day)	-1.541e-02	-6.067e-03	5.601e-03	-4.807e-03
cos(3*day)	5.014e-03		-5.747e-03	1.912e-04
sin(3*day)	5.026e-03		-1.135e-02	-3.633e-03
Dim*cos(day)	9.299e-04	1.732e-04	2.226e-04	5.248e-05
Dim*sin(day)		-2.929e-05	-7.525e-05	-8.540e-05
Dim*cos(2*day)	-2.997e-04		1.962e-04	9.240e-05
Dim*sin(2*day)	-3.742e-04		6.098e-05	6.666e-05
Dim*cos(3*day)				5.385e-05
Dim*sin(3*day)				6.149e-05

[a]Dim: days in milk ; day: day of the year ; cos: *cosine* ; sin: *sine*

Table C.8: Parameter estimates of the linear models for somatic cell count (mean and standard deviation) as a function of days in milk (1-60) and day of the year (1-365).

	Parity 1		Parity > 1	
Covariates[a]	Mean	Standard Deviation	Mean	Standard Deviation
Intercept	5.558e+00	1.168e+00	5.350e+00	1.328e+00
Dim	-1.269e-01	-1.679e-03	-1.108e-01	4.849e-04
Dim^2	3.953e-03		3.611e-03	
Dim^3	-5.674e-05		-5.142e-05	
Dim^4	3.072e-07		2.727e-07	
cos(day)		6.503e-02	-5.478e-02	3.511e-02
sin(day)	1.479e-01	2.189e-02	1.484e-01	3.864e-02
cos(2*day)	2.102e-02			
sin(2*day)	1.192e-02			
Dim*cos(day)		-5.612e-04	-9.470e-04	-8.491e-04
Dim*sin(day)	-1.146e-03	8.862e-04	-2.492e-03	2.457e-04

[a]Dim: days in milk ; day: day of the year ; cos: *cosine* ; sin: *sine*

Appendix D

Examiners

My thanks go to D^r Theo Lam and D^r Wendela Wapenaar for having examined this work.

References

Abdallah, J. M., & McDaniel, B. T. 2000. Genetic parameters and trends of milk, fat, days open, and body weight after calving in North Carolina experimental herds. *J Dairy Sci*, **83**(6), 1364–1370.

Agresti, A. 2002. *Categorical Data Analysis (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Akers, R. M. 2002. *Lactation and the mammary gland*. Blackwell.

Albarrán-Portillo, B., & Pollott, G. E. 2008. Genetic Parameters Derived From Using a Biological Model of Lactation on Records of Commercial Dairy Cows. *J. Dairy Sci.*, **91**(9), 3639–3648.

Ali, A. K. A., & Shook, G. E. 1980. An Optimum Transformation for Somatic Cell Concentration in Milk. *J Dairy Sci*, **63**, 487–490.

Appuhamy, J. A. D. R. N., Cassell, B. G., & Cole, J. B. 2009. Phenotypic and genetic relationships of common health disorders with milk and fat yield persistencies from producer-recorded health data and test-day yields. *J. Dairy Sci.*, **92**(4), 1785–1795.

Arbel, R., Bigun, Y., Ezra, E., Sturman, H., & Hojman, D. 2001. The effect of extended calving intervals in high-yielding lactating cows on milk production and profitability. *J Dairy Sci*, **84**(3), 600–608.

Barbano, D. M., & Clark, J. L. 1989. Infrared Milk Analysis Challenges for the Future. *J. Dairy Sci.*, **72**(6), 1627–1636.

Bates, D., & Maechler, M. 2009. *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-31.

Bauman, D. E., Mather, I. H., Wall, R. J., & Lock, A. L. 2006. Major advances associated with the biosynthesis of milk. *J Dairy Sci*, **89**(4), 1235–1243.

Bauman, D.E., & Currie, W.B. 1980. Partitioning of Nutrients During Pregnancy and Lactation: A Review of Mechanisms Involving Homeostasis and Homeorhesis. *J. Dairy Sci.*, **63**(9), 1514–1529.

Bauman, D.E., & Griinari, J.M. 2003. Nutritional regulation of milk fat synthesis. *Annu Rev Nutr*, **23**, 203–227.

Bradley, A. J., & Green, M. J. 2001. Adaptation of Escherichia coli to the Bovine Mammary Gland. *J. Clin. Microbiol.*, **39**(5), 1845–1849.

Bradley, A. J., Leach, K. A., Breen, J. E., Green, L. E., & Green, M. J. 2007. Survey of the incidence and aetiology of mastitis on dairy farms in England and Wales. *Vet Rec*, **160**(8), 253–257.

Bradley, A.J., & Green, M.J. 2005. Use and interpretation of somatic cell count data in dairy cows. *In Practice*, **27**, 310–315.

Brolund, L. 1985. Cell counts in bovine milk. *Acta Vet. Scand.*, **Suppl. 80**, 5123.

Buckley, F., O'Sullivan, K., Mee, J. F., Evans, R. D., & Dillon, P. 2003. Relationships among milk yield, body condition, cow weight, and reproduction in spring-calved Holstein-Friesians. *J Dairy Sci*, **86**(7), 2308–2319.

Burvenich, C., Van Merris, V., Mehrzad, J., Diez-Fraile, A., & Duchateau, L. 2003. Severity of E. coli mastitis is mainly determined by cow factors. *Vet. Res.*, **34**(5), 521–564.

Butler, W. R., & Smith, R. D. 1989. Interrelationships between energy balance and postpartum reproductive function in dairy cattle. *J Dairy Sci*, **72**(3), 767–783.

Caccamo, M., Veerkamp, R. F., de Jong, G., Pool, M. H., Petriglieri, R., & Licitra, G. 2008. Variance components for test-day milk, fat, and protein yield, and somatic cell score for analyzing management information. *J Dairy Sci*, **91**(8), 3268–3276.

Capuco, A. V., Wood, D. L., Baldwin, R., Mcleod, K., & Paape, M. J. 2001. Mammary cell number, proliferation, and apoptosis during a bovine lactation: relation to milk production and effect of bST. *J Dairy Sci*, **84**(10), 2177–2187.

Capuco, A. V., Ellis, S. E., Hale, S. A., Long, E., Erdman, R. A., Zhao, X., & Paape, M. J. 2003. Lactation persistency: insights from mammary cell proliferation studies. *J Anim Sci*, **81 Suppl 3**, 18–31.

Cerbulis, J., & Farrell, H. M., Jr. 1975. Composition of Milks of Dairy Cattle. I. Protein, Lactose, and Fat Contents and Distribution of Protein Fraction. *J. Dairy Sci.*, **58**(6), 817–827.

Cleveland, W.S., & Devlin, S.J. 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, **83**(403), 596–610.

Cook, N. B., Bennett, T. B., Emery, K. M., & Nordlund, K. V. 2002. Monitoring nonlactating cow intramammary infection dynamics using DHI somatic cell count data. *J Dairy Sci*, **85**(5), 1119–1126.

Cook, N.B., Oetzel, G.R., & Nordlund, K.V. 2006. Modern techniques for monitoring high-producing dairy cows 1. Principles of herd-level diagnoses. *In Practice*, **28**, 510–515.

Council of the European Communities. 1992 (09). *Council Directive 92/46/EEC of 16 June 1992 laying down the health rules for the production and placing on the market of raw milk, heat-treated milk and milk-based products.* Official Journal L 268.

Cutullic, E., Delaby, L., Causeur, D., Michel, G., & Disenhaus, C. 2009. Hierarchy of factors affecting behavioural signs used for oestrus detection of Holstein and Normande dairy cows in a seasonal calving system. *Animal Reproduction Science*, **113**(1-4), 22 – 37.

DairyCo. 2009a (August). `http://www.mdcdatum.org.uk/index.html`.

DairyCo. 2009b. *Factors affecting milk supply.* Tech. rept.

de Haas, Y., Veerkamp, R. F., Barkema, H. W., Grohn, Y. T., & Schukken, Y. H. 2004. Associations Between Pathogen-Specific Cases of Clinical Mastitis and Somatic Cell Count Patterns. *J. Dairy Sci.*, **87**(1), 95–105.

de Haas, Y., Ouweltjes, W., ten Napel, J., Windig, J. J., & de Jong, G. 2008. Alternative somatic cell count traits as mastitis indicators for genetic selection. *J Dairy Sci*, **91**(6), 2501–2511.

de Vries, M. J., & Veerkamp, R. F. 2000. Energy balance of dairy cattle in relation to milk production variables and fertility. *J Dairy Sci*, **83**(1), 62–69.

Djabri, Belgacem, Bareille, Nathalie, Beaudeau, François, & Seegers, Henri. 2002. Quarter milk somatic cell count in infected dairy cows: a meta-analysis. *Vet. Res.*, **33**(4), 335–357.

Dohoo, I. R. 1993. An evaluation of the validity of individual cow somatic cell counts from cows in early lactation. *Prev Vet Med*, **16**, 103–110.

Dohoo, I. R., & Morris, R. S. 1993. Somatic cell count patterns in Prince Edward Island dairy herds. *Preventive Veterinary Medicine*, **15**(1), 53 – 65.

Dohoo, I.R., & Leslie, K.E. 1991. Evaluation of changes in somatic cell counts as indicators of new intramammary infections. *Prev Vet Med*, **10**, 225–237.

Dopfer, D., Barkema, H. W., Lam, T.J.G.M., Schukken, Y. H., & Gaastra, W. 1999. Recurrent Clinical Mastitis Caused by Escherichia coli in Dairy Cows. *J. Dairy Sci.*, **82**(1), 80–85.

Duffield, T. F., Kelton, D. F., Leslie, K. E., Lissemore, K. D., & Lumsden, J. H. 1997. Use of test day milk fat and milk protein to detect subclinical ketosis in dairy cattle in Ontario. *Can Vet J*, **38**(11), 713–718.

Duffield, T. F., Lissemore, K. D., McBride, B. W., & Leslie, K. E. 2009. Impact of hyperketonemia in early lactation dairy cows on health and production. *J Dairy Sci*, **92**(2), 571–580.

Emanuelson, U., & Funke, H. 1991. Effect of Milk Yield on Relationship Between Bulk Milk Somatic Cell Count and Prevalence of Mastitis. *J. Dairy Sci.*, **74**(8), 2479–2483.

Fraser, Wilber J. 1933. The Formation of the American Dairy Science Association. *J. Dairy Sci.*, **16**(6), 583–586.

Gelman, A., & Hill, J. 2007. *Data analysis using regression and multilevel/hierarchical models*. Vol. Analytical methods for social research. New York: Cambridge University Press.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., & Mechelen, I. Van. 2000. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal Of The Royal Statistical Society Series C*, **49**(2), 247–268.

Gelman, Andrew, Carlin, John B., Stern, Hal S., & Rubin, Donald B. 2003. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC.

Green, L. E., Schukken, Y. H., & Green, M. J. 2006a. On distinguishing cause and consequence: do high somatic cell counts lead to lower milk yield or does high milk yield lead to lower somatic cell count? *Prev Vet Med*, **76**(1-2), 74–89.

Green, M. J., Green, L. E., Schukken, Y. H., Bradley, A. J., Peeler, E. J., Barkema, H. W., de Haas, Y., Collis, V. J., & Medley, G. F. 2004. Somatic cell count distributions during lactation predict clinical mastitis. *J Dairy Sci*, **87**(5), 1256–1264.

Green, M. J., Bradley, A. J., Newton, H., & Browne, W. J. 2006b. Seasonal variation of bulk milk somatic cell counts in UK dairy herds: investigations of the summer rise. *Prev Vet Med*, **74**(4), 293–308.

Green, M. J., Medley, Graham, & Browne, W. J. 2009. Use of posterior predictive assessments to evaluate model fit in multilevel logistic regression. *Veterinary Research*, **40**.

Grieve, D. G., Korver, S., Rijpkema, Y. S., & Hof, G. 1986. Relationship between milk composition and some nutritional parameters in early lactation. *Livestock Production Science*, **14**(3), 239 – 254.

Grohn, Y. T., Eicker, S. W., Ducrocq, V., & Hertl, J. A. 1998. Effect of Diseases on the Culling of Holstein Dairy Cows in New York State. *J. Dairy Sci.*, **81**(4), 966–978.

Haile-Mariam, M., Bowman, P. J., & Goddard, M. E. 2003. Genetic and environmental relationship among calving interval, survival, persistency of milk yield and somatic cell count in dairy cattle. *Livestock Production Science*, **80**(3), 189 – 200.

Hanigan, M. D., Crompton, L. A., Bequette, B. J., Mills, J. A. N., & France, J. 2002. Modelling Mammary Metabolism in the Dairy Cow to Predict Milk Constituent Yield, with Emphasis on Amino Acid Metabolism and Milk Protein Production: Model Evaluation. *Journal of Theoretical Biology*, **217**(3), 311 – 330.

Hawke, J.C., & Taylor, M.W. 2002. *Influence of nutritional factors on the yield, composition and physical properties of milk fat.* 2nd edn. Advanced dairy chemistry, vol. 2. Chapman &Hall. Chap. 2, pages 37–88.

Hennessy, D.A., & Roosen, J. 2003. A Cost-Based Model of Seasonal Production with Application to Milk Policy. *Journal of Agricultural Economics*, **54**(2), 285–312.

Heuer, C., Schukken, Y. H., & Dobbelaar, P. 1999. Postpartum body condition score and results from the first test day milk as predictors of disease, fertility, yield, and culling in commercial dairy herds. *J Dairy Sci*, **82**(2), 295–304.

Heuer, C., Straalen, W. M. Van, Schukken, Y. H., Dirkzwager, A., & Noordhuizen, J. P. T. M. 2000. Prediction of energy balance in a high yielding dairy herd in early lactation: model development and precision. *Livestock Production Science*, **65**(1-2), 91 – 105.

Hintze, J.L., & Nelson, R.D. 1998. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, **52**(2), 181–184.

Hortet, P., & Seegers, H. 1998. Loss in milk yield and related composition changes resulting from clinical mastitis in dairy cows. *Prev Vet Med*, **37**(1-4), 1–20.

Jensen, R.G. 2002. The Composition of Bovine Milk Lipids: January 1995 to December 2000. *J. Dairy Sci.*, **85**(2), 295–350.

Jorritsma, R., Jorritsma, H., Schukken, Y. H., & Wentink, G. H. 2000. Relationships between fatty liver and fertility and some periparturient diseases in commercial Dutch dairy herds. *Theriogenology*, **54**(7), 1065–1074.

Kaylegian, K. E., Houghton, G. E., Lynch, J. M., Fleming, J. R., & Barbano, D. M. 2006. Calibration of Infrared Milk Analyzers: Modified Milk Versus Producer Milk. *J. Dairy Sci.*, **89**(8), 2817–2832.

Keefe, G.P. 1997. Streptococcus agalactiae mastitis: A review. *Can Vet J*, **38**, 429–437.

Kremer, W.D.J., Noordhuizen-Stassen, E. N., Grommers, F. J., Schukken, Y. H., Heeringa, R., Brand, A., & Burvenich, C. 1993. Severity of Experimental Escherichia coli Mastitis in Ketonemic and Nonketonemic Dairy Cows. *J Dairy Sci*, **76**.

Kristula, M. A., Reeves, M., Redlus, H., & Uhlinger, C. 1995. A preliminary investigation of the association between the first postpartum milk fat test and first insemination pregnancy rates. *Preventive Veterinary Medicine*, **23**(1-2), 95 – 100.

Kurtz, F.E. 1974. *The lipids of milk: composition and properties*. 2nd edn. Fundamentals of dairy chemsitry. Avi Publishing Co. Pages 125–219.

Lee, C. S., Wooding, F. B., & Kemp, P. 1980. Identification, properties, and differential counts of cell populations using electron microscopy of dry cows secretions, colostrum and milk from normal cows. *J Dairy Res*, **47**(1), 39–50.

Lievaart, J., Barkema, H.W., Hogeveen, H., & Kremer, W. 2009. Reliability of the bulk milk somatic cell count as an indication of average herd somatic cell count. *J Dairy Res*, **76**(4), 490–496.

Lievaart, J.J., Kremer, W.D.J, & Barkema, H.W. 2007. Short communication: Comparison of bulk milk, yield-corrected, and average somatic cell counts as parameters to summarize the subclinical mastitis situation in a dairy herd. *J Dairy Sci*, **90**(9), 4145–4148.

Linzell, J. L., & Peaker, M. 1971. Mechanism of milk secretion. *Physiol. Rev.*, **51**(3), 564–597.

Lucy, M. C. 2001. Reproductive loss in high-producing dairy cattle: where will it end? *J Dairy Sci*, **84**(6), 1277–1293.

Lukas, J. M., Hawkins, D. M., Kinsel, M. L., & Reneau, J. K. 2005. Bulk Tank Somatic Cell Counts Analyzed by Statistical Process Control Tools to Identify and Monitor Subclinical Mastitis Incidence. *J. Dairy Sci.*, **88**(11), 3944–3952.

Lunn, D.J., Thomas, A., Best, N., & Spiegelhalter, D. 2000. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**(4), 325–337.

McErlean, S. 1999. The price schedule required to smooth seasonal milk supply. *British Food Journal*, **101**(10), 785–796.

McKenna, D. 1994. Accuracy and precision of Fossomatic 250/300/360 series of somatic cell counters compared to Fossomatic 215: collaborative study. *J AOAC Int.*, **77**(4), 932–940.

Miller, N., Delbecchi, L., Petitclerc, D., Wagner, G. F., Talbot, B. G., & Lacasse, P. 2006. Effect of stage of lactation and parity on mammary gland cell renewal. *J Dairy Sci*, **89**(12), 4669–4677.

Miller, R. H., Paape, M. J., & Acton, J. C. 1986. Comparison of Milk Somatic Cell Counts by Coulter and Fossomatic Counters. *J. Dairy Sci.*, **69**(7), 1942–1946.

Mitchell, A., Bourn, D., Mawdsley, J., Wint, W., R. Clifton-Hadley, R., & Gilbert, M. 2005. Characteristics of cattle movements in Britain. An analysis of records from the Cattle Tracing System. *Animal Science*, **80**, 264–273.

Myung, In Jae. 2003. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, **47**(1), 90 – 100.

Nelder, J. A., & Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135**(3), 370–384.

NMR. 2008 (October). *Guide to Recording/Sampling.*

NMR. 2009. *History of NMR.* `http://www.nmr.co.uk/`.

Norman, H. D., Wright, J. R., Kuhn, M. T., Hubbard, S. M., Cole, J. B., & VanRaden, P. M. 2009. Genetic and environmental factors that affect gestation length in dairy cattle. *J Dairy Sci*, **92**(5), 2259–2269.

Olde Riekerink, R. G. M., Barkema, H. W., & Stryhn, H. 2007. The effect of season on somatic cell count and the incidence of clinical mastitis. *J Dairy Sci*, **90**(4), 1704–1715.

Paape, M., Mehrzad, J., Zhao, X., Detilleux, J., & Burvenich, C. 2002. Defense of the Bovine Mammary Gland by Polymorphonuclear Neutrophil Leukocytes. *Journal of Mammary Gland Biology and Neoplasia*, **7**, 109–121.

Pantoja, J. C F, Hulland, C., & Ruegg, P. L. 2009. Dynamics of somatic cell counts and intramammary infections across the dry period. *Prev Vet Med*, **90**(1-2), 43–54.

Patton, J., Kenny, D. A., McNamara, S., Mee, J. F., O'Mara, F. P., Diskin, M. G., & Murphy, J. J. 2007. Relationships among milk production, energy balance, plasma analytes, and reproduction in Holstein-Friesian cows. *J Dairy Sci*, **90**(2), 649–658.

Podpecan, O., Mrkun, J., & Zrimsek, P. 2008. Diagnostic evaluation of fat to protein ratio in prolonged calving to conception interval using receiver operating characteristic analyses. *Reprod Domest Anim*, **43**(2), 249–254.

R Development Core Team. 2009. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rasbash, Jon, Steele, Fiona, Browne, William J., & Goldstein, Harvey. 2009. *A Users Guide to MLwiN Version 2.10*.

Robinson, Susan, & Christley, Rob. 2006. Identifying temporal variation in reported births, deaths and movements of cattle in Britain. *BMC Veterinary Research*, **2**(1), 11.

Santos, J. E. P., Cerri, R. L. A., Ballou, M. A., Higginbotham, G. E., & Kirk, J. H. 2004. Effect of timing of first clinical mastitis occurrence on lactational and reproductive performance of Holstein dairy cows. *Animal Reproduction Science*, **80**(1-2), 31 – 45.

Sarikaya, H., Schlamberger, G., Meyer, H. H D, & Bruckmaier, R. M. 2006. Leukocyte populations and mRNA expression of inflammatory factors in quarter milk fractions at different somatic cell score levels in dairy cows. *J Dairy Sci*, **89**(7), 2479–2486.

Schepers, A. J., Lam, T. J., Schukken, Y. H., Wilmink, J. B., & Hanekamp, W. J. 1997. Estimation of variance components for somatic cell counts to determine thresholds for uninfected quarters. *J Dairy Sci*, **80**(8), 1833–1840.

Schrick, F. N., Hockett, M. E., Saxton, A. M., Lewis, M. J., Dowlen, H. H., & Oliver, S. P. 2001. Influence of Subclinical Mastitis During Early Lactation on Reproductive Parameters. *J. Dairy Sci.*, **84**(6), 1407–1412.

Seegers, Henri, Fourichon, Christine, & Beaudeau, Franois. 2003. Production effects related to mastitis and mastitis economics in dairy cattle herds. *Vet Res*, **34**(5), 475–491.

Shennan, D. B., & Peaker, M. 2000. Transport of Milk Constituents by the Mammary Gland. *Physiol. Rev.*, **80**(3), 925–951.

Silvestre, A. M., Petim-Batista, F., & Colaco, J. 2006. The Accuracy of Seven Mathematical Functions in Modeling Dairy Cattle Lactation Curves Based on Test-Day Records From Varying Sample Schemes. *J. Dairy Sci.*, **89**(5), 1813–1821.

Silvestre, A.M., Martins, A.M., Santos, V.A., Ginja, M.M., & Colao, J.A. 2009. Lactation curves for milk, fat and protein in dairy cows: A full approach. *Livestock Science*, **122**(2-3), 308 – 313.

Spiegelhalter, S.D., Best, N.G., Carlin, B.P., & Linde, A.V.D. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.

Steele, F., Goldstein, H., & Browne, W.J. 2004. A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modeling*, **4**(2), 145–159.

Stoop, W. M., Bovenhuis, H., Heck, J. M. L., & van Arendonk, J. A. M. 2009. Effect of lactation stage and energy status on milk fat composition of Holstein-Friesian cows. *J. Dairy Sci.*, **92**(4), 1469–1478.

The Center for Dairy Information. 2009. *Breed Performance Statistics 1999 - 2008.* Tech. rept.

Therneau, T. 2009. *survival: Survival analysis, including penalised likelihood.* R package version 2.35-7 (original R port by T. Lumley).

Veerkamp, R. F., Oldenbroek, J. K., Van Der Gaast, H. J., & Werf, J.H.J. Van Der. 2000. Genetic Correlation Between Days Until Start of Luteal Activity and Milk Yield, Energy Balance, and Live Weights. *J. Dairy Sci.*, **83**(3), 577–583.

Weigel, K. A., Rekaya, R., Zwald, N. R., & Fikse, W. F. 2001. International genetic evaluation of dairy sires using a multiple-trait model with individual animal performance records. *J Dairy Sci*, **84**(12), 2789–2795.

Wilmink, J.B.M. 1987. Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. *Livestock Production Science*, **16**, 335–348.

Yang, M., & Goldstein, H. 2003. *Modelling Survival Data in MLwiN.*