Journal of Dairy Research

cambridge.org/dar

Modeling cow somatic cell count using sensor data as input to generalized additive models

Dorota Anglart[1,2], Charlotte Hallén-Sandgren[1], Patrik Waldmann[3], Martin Wiedemann[4] and Ulf Emanuelson[2]

[1]DeLaval International AB, PO Box 39, se-147 21, Tumba, Sweden; [2]Department of Clinical Sciences, Swedish University of Agricultural Sciences, PO Box 7054, se-750 07, Uppsala, Sweden; [3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, PO Box 7023, se-750 07, Uppsala, Sweden and [4]DeLaval GmbH, Wilhelm-Bergner-Str. 5, Glinde 21509, Germany

Research Article

Cite this article: Anglart D, Hallén-Sandgren C, Waldmann P, Wiedemann M and Emanuelson U (2020). Modeling cow somatic cell count using sensor data as input to generalized additive models. *Journal of Dairy Research* **87**, 282–289. https://doi.org/10.1017/S0022029920000692

Received: 19 September 2019
Revised: 6 March 2020
Accepted: 6 April 2020
First published online: 4 September 2020

Keywords:
Additive model; automatic milking rotary; somatic cell count; udder health

Author for correspondence:
Dorota Anglart,
Email: dorota.anglart@delaval.com

Abstract

This research paper presents a study investigating if sensor data from an automatic milking rotary could be used to model cow somatic cell count (composite milk SCC: CMSCC). CMSCC is valuable for udder health monitoring and individual cow udder health surveillance could be improved by predicting CMSCC between routine samplings. Data regularly recorded in the automatic milking rotary, in one German dairy herd, were collected for analysis. The cows (Holstein-Friesian, $n = 372$) were milked twice daily and sampled once weekly in afternoon milkings for 8 weeks for CMSCC. From the potential independent variables, including quarter conductivity, milk flow, blood in milk, kick-offs, not milked quarters and incomplete milkings, new variables that combined quarter data were created. Past period records, i.e. lags, of up to seven days before the actual CMSCC sampling event were added in the dataset to investigate if they were of use in modeling the cell count. Univariable generalized additive models (GAM) were used to screen the data to select potential independent variables. Furthermore, several multivariable GAM were fitted in order to compare the importance of the potential independent variables and to explore how the model performance would be affected by using data from various number of days before the CMSCC sampling event. The result of the model selection showed that the best explanation of CMSCC was provided by the model incorporating all significant variables from the variable screening for the seven preceding days, including the day of the CMSCC sampling event. However, using data from only three days before the CMSCC sampling event is suggested to be sufficient to model CMSCC. Variables combining conductivity quarter data, together with quarter conductivity, are suggested to be important in describing CMSCC. We conclude that CMSCC can be modeled with a high degree of explanation using the information routinely recorded by the milking robot.

Somatic cell count (SCC) has long been a common and valuable method for monitoring udder health in dairy herds (Sharma *et al.*, 2011) and could also be a tool for identifying intramammary infections in individual cows (International Dairy Federation, 2013). To monitor SCC levels, farmers usually sample their cows at composite level according to the testing procedures of their local milk testing organization, normally recommended to be conducted once a month, or to use on-site farm tests. The California mastitis test is probably the most commonly applied cow-side test used to indicate the SCC at the quarter level. It is cheap and rapid but not very precise or accurate (Schukken *et al.*, 2003; International Dairy Federation, 2013). A more precise method is fluoro-opto-electronic instruments in which cells are fluoresced and counted using flow cytometry (Kitchen, 1981; International Dairy Federation, 2013). This is also the only standardized method for determining SCC (International Dairy Federation, 2013). The method could either be used in a standalone device or integrated in the milking system, providing the farmer with SCC values of individual cows' milk after every milking. Frequent sampling for SCC could improve the monitoring of individual cows by detecting deviations from the individuals' normal patterns or rapidly elevated SCC levels as well as indicating recovery from elevated SCC. Furthermore, daily variations in SCC can affect the monthly sampling results, and additional input of SCC values in connection with the sampling day could reduce the risk of udder health misclassification (Quist *et al.*, 2008). However, more frequent sampling of individual cows will increase costs or workload for the farmer in systems where integrated sampling devices are not possible, so it would be advantageous if the SCC could be predicted based on information that is continuously and automatically recorded. To our knowledge, only two previous studies have attempted to model SCC patterns using sensor data such as conductivity or milking duration (Sitowska *et al.*, 2017; Ebrahimie *et al.*, 2018), and only one using routinely recorded sensor

© The Author(s), 2020. Published by Cambridge University Press on behalf of Hannah Dairy Research Foundation. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hannah Dairy Research Foundation

CAMBRIDGE UNIVERSITY PRESS

data (Sitowska *et al*., 2017). However, combining quarter information from routinely recorded milking data to describe cow level somatic cell count (composite milk SCC: CMSCC) has not earlier been explored.

The objective of this study was to investigate if existing sensor data from an automatic milking rotary (AMR) could be used to model CMSCC. Additionally, the ambition was to screen which variables were the best possible independent variables in modeling CMSCC, and over what time period by using past-period variables. The outcome of this study could potentially be used for imputing missing SCC values or as supplementary information for the farmer between routine SCC measurements.

Materials and methods

Data collection

The data were collected from a German dairy farm with 372 Holstein-Friesian cows during an 8-week period in the summer of 2016. The cows were milked twice daily with a 13 h:11 h milking interval in an AMR (24 unit platform with 5 robotic arms, DeLaval International AB, Tumba, Sweden) and kept in a loose housing system. The cows were managed according to normal farm procedures and fed a total mixed ration. The daily average milk yield was 34 kg during the experimental period. Animal information such as days in milk (DIM) and lactation number (LN) was extracted from the herd management system, together with information from each milking during the 8 weeks (in total 39 587 records), henceforth referred to as milking data. Details regarding the milking data can be found in online Supplementary Table S1. Milking data at the quarter level comprised conductivity, blood in milk, milk yield, expected milk yield, mean and peak milk flow, cups kicked off during milking and incompletely or not-milked quarters. Milking data at the cow composite level comprised milking duration, milking unit number, mastitis detection index (MDi; an index representing the probability that the cow has mastitis calculated by incorporating different phases of conductivity during milking together with blood), and udder counters (UC; a counter increasing or decreasing after each milking depending on whether or not the MDi was above or below a pre-set threshold value).

Sampling for CMSCC analysis was done once weekly during the afternoon milking throughout the eight weeks. A milk sampler (DeLaval milk meter MM6) was attached to each milking unit to collect a representative sample from each cow. The samples were analyzed for CMSCC in a laboratory in Jena, Germany, using a Fossomatic 7, DC 600 system (ISO/IEC, 2005).

Data preparation

Cows were categorized into LN groups 1, 2, and ≥3. Cows not included in the weekly CMSCC sampling were removed from the milking data as were all milking events for cows during the first week of lactation. Mean and peak milk flow values classified as outliers according to boxplots (i.e. outside 1.5 × interquartile range above the upper and below the lower quartile) were removed. Quarter conductivity values considered not biologically plausible (i.e., below 3 mS/cm and above 10 mS/cm) were also removed. The CMSCC values were transformed to a log10 scale, henceforth referred to as log10CMSCC. Finally, log10CMSCC observations without a complete setup of independent variables were removed. In total, <1% of the milking data were removed due to cleaning (details in online Supplementary Table S1).

To analyze the dependent variable log10CMSCC at the composite level together with the independent variables at the quarter level, several new variables were created from the quarter variables of the milking data. The created variables were for instance variance between quarters, difference between quarters, lowest and highest value of a quarter or dichotomization of factor variables. The names of the created variables were given a suffix to indicate the transformation given. Details of all created variables can be found in online Supplementary Table S1. Past-period variables (lags) were created to evaluate how predictive some of the variables could be up to seven days (14 milking sessions) before the actual CMSCC sample. The lag of a variable was indicated by a suffix number corresponding to how many milking sessions before the CMSCC sampling event the variable was first recorded in the milking system, e.g. milking session_0 is corresponding to the day of CMSCC sample event. The milking data containing the created variables were merged with the CMSCC sample data. From this complete milking data, 2384 observations of 372 milkings cows with 934 potential independent variables (840 variables with day lags and 94 variables for milking session 0) were available. CMSCC observations that did not include a complete setup of independent variables for 14 milking sessions before the CMSCC sampling event were removed, leaving 319 cows with 1758 cow observations for variable screening.

Statistical analysis

Firstly, a variable scanning on all available data was performed to identify the independent variables suitable to incorporate for modeling CMSCC. Secondly, several models were fitted to investigate different variable setups as well as different variations of day lags, i.e., do models perform better using more or less data. The dependent variable for all statistical analyses was log10CMSCC. Initial regression analysis of variance for each quarter variable, showed that the interaction between quarter location (i.e., right front, left front, right rear, and left rear) and milk flow, conductivity, or milk yield, or the quarter location alone, was not significant. We therefore concluded that quarter location did not affect the relationship between the independent and dependent variable (log10CMSCC), so the potential quarter independent variables were used independently of quarter location.

To reduce the number of independent variables ($n = 934$) generalized additive models (GAM, Hastie and Tibshirani, 1990) were used to analyze the association of each potential independent variable from the milking data with log10CMSCC. Hence, log10CMSCC was set as dependent variable y and the potentially confounding variables LN (factor), DIM (linear variable), and Cow (random factor) were added to all models. The independent variables of interest, X, were then analyzed individually, either as factors or as smooths, i.e., non-parametric spline functions not forcing linearity between the independent variable and the dependent and allowing flexible estimation of the underlying patterns. In this way, 934 screening models were built according to:

$$y_i = \beta_0 + LN\beta_{LN} + DIM\beta_{DIM} + \alpha_{Cow} + f(X)_i + \varepsilon_i \qquad (1)$$

$$\alpha_{\text{Cow}} \sim N(0, \sigma^2_{\text{Cow}}), \quad \varepsilon_i \sim N(0, \sigma^2_{\varepsilon})$$

https://doi.org/10.1017/S0022029920000692 Published online by Cambridge University Press

Table 1. Possible independent variables used in model development, presented with definition of each variable and the corresponding past time-period records (milking session lag)

Variable	Definition	Milking session lag
MDi[a]	Mastitis detection index, based on different phases of conductivity and blood in milk	{0 1 2 3 4 5 6 7 8 9 10 11 12 13 14}
Peak flow.min[a]	Lowest milk peak flow value within cow and milking session	{0 1 2 11 12 13}
Conductivity.mean[a]	Arithmetic conductivity mean all quarters within cow and milking session	{0 4 9 11 14}
Conductivity.max[a]	Highest conductivity value within cow and milking session	{0 1 3 4 6 7 8 9 10 11 13 14}
Conductivity[a]	Quarter conductivity	{1 4}
Conductivity.diff[a]	Highest quarter conductivity value minus lowest value within cow and milking session	{0 1 2 3 4 5 6 7 8 9 10 11 12 13 14}
Conductivity.var[a]	Variance of conductivity between quarters within cow and milking session	{0 1 2 3 4 5 6 7 8 9 10 11 12 13 14}
Diff.milkings[a]	Quarter milk yield deviation from previous corresponding milking session at the quarter level	{0}
UC[b]	Udder Counter: If MDi > 1.4 counting up, if MDi < 1.4 counting down for last 10 milkings	{1 3 7 8 10 12}
Not milked.score[b]	Number of quarters with milk yield <0.2 kg (0–4)	{8}
Incomplete.score[b]	Number of quarters where milk yield <50% of expected yield if expected yield >1 kg and milk yield >3 kg (kg) was true (0–4)	{0 1 2 3 4 5 6 7 8 10 11 12 13 14}
Incomplete.score.7[a]	A 7- day rolling average of the number of incomplete.score	{0}

Milk session lag 0 is the day of sampling and 1 is one milking before the composite milk somatic cell count sampling event etc.
[a]Treated as smooth variable in all models.
[b]Treated as factor variable in all models.

where ***y*** is the independent variable, ***LN*** (factor), ***DIM*** (linear variable) and ***Cow*** (random factor) are potential confounding variables, ***f(X)*** are the non-parametric spline functions of the independent variables.

Of the 934 models, 905 converged, resulting in 268 independent variables for which $P < 0.001$, and thus kept for further analysis. A Bonferroni correction at the $P < 0.010$ level was performed, leaving 158 independent variables for further analysis. Multicollinearity within independent variables between lagged milking sessions was tested with a variance inflation test (VIF) and generalized VIF (GVIF) according to Fox and Monette (1992). Independent variables with VIF > 8 and independent variables specified as factors with a square root of $GVIF^{(1/(2*df))} > 8$ were removed from further analysis, as were factor variables with observations at one level only. Quarter-level variables were selected for analysis if at least three of the quarters within the same milking session had $P < 0.010$. All milk yield–associated variables (except milk yield.diff.milkings) were removed from further analysis, because milk yield was considered an intervening variable on the causal path between the other independent variables and the dependent variable. Thus, 1758 observations of 102 independent variables, in addition to LN, DIM, and Cow, from different time periods before the CMSCC sampling event were available for the model development (Table 1).

The final multivariable GAM were fitted with LN, DIM, and Cow as potentially confounding variables and log10CMSCC as the dependent variable, but with multiple potential independent variables in each model:

$$y_i = \beta_0 + LN\beta_{LN} + DIM\beta_{DIM} + \alpha_{Cow} + \sum_{j=1}^{p} f_j(X)_{ij} + \varepsilon_i \quad (2)$$

$$\alpha_{\text{Cow}} \sim N(0, \sigma^2_{\text{Cow}}), \quad \varepsilon_i \sim N(0, \sigma^2_{\varepsilon})$$

where ***y*** is the independent variable, ***LN*** (factor), ***DIM*** (linear variable) and ***Cow*** (random factor) are potential confounding variables, $\boldsymbol{f_j(X)}$ are the non-parametric spline functions of the independent variables.

The fitted models were model 1, which included all potential independent variables from the variable screening ($n = 102$), and model 2 ($n = 81$), which excluded all MDi and UC variables. Model 2 was developed to investigate the impact of using observed values of conductivity rather than values that had been derived by the MDi algorithm. To evaluate how well model 1 and model 2 would perform on milking data from restricted time periods before the CMSCC sampling event, 6 additional variations of model 1 and model 2 were fitted with potential independent variables from various time periods. This was done by removing milking data from days close to the CMSCC sampling event in some models and removing milking data from days further away in other models (e.g. both model variations restricted for 0–6 d, 0–3 d, 1–7 d, 1–5 d, 2–7 d and 2–5 before the CMSCC sample event). Most of the potential independent variables were fitted as numerical smooths in the models, while UC, Not.milked.score, and Incomplete.score were fitted as factor variables (Table 1). The smoothing parameter estimation method used in all model fits was restricted maximum likelihood (REML). The corrected Akaike information criterion (AIC), described by Wood *et al.* (2016), was used for model comparison. Models were also evaluated using the adjusted coefficient of determination (R^2_{adj}). All data cleaning and statistical analyses were performed in the program R using the 'mgvc' package for spline and GAM models (R Development Core Team, 2018).

Results

Descriptive statistics can be found in online Supplementary Table S2.

https://doi.org/10.1017/S0022029920000692 Published online by Cambridge University Press

GAM models: effects of variables

The results of the GAM fit of model 1 indicated that 24 of the independent variables were associated ($P < 0.050$) with CMSCC. The independent variable having the strongest statistical association with CMSCC was MDi at the milking session of the CMSCC sampling event (MDi_0, $P < 0.001$), followed by quarter conductivity one milking session before the CMSCC sampling event (conductivity.quarter_1) for three out of four quarters, i.e. right front, left rear ($P < 0.001$) and left front ($P = 0.011$). The results of the GAM fit of model 2 indicated that 17 of the independent variables were associated ($P < 0.050$) with CMSCC. The variables having the strongest statistical association with CMSCC were variance of conductivity between quarters (conductivity.var_1, $P < 0.001$), quarter conductivity (conductivity.quarter_1) for right front ($P < 0.001$) and the difference in conductivity between quarters (conductivity.diff_1, $P = 0.003$), all at one milking before the CMSCC sampling event. Furthermore, the maximum conductivity of the quarters during the same milking as the CMSCC sampling event (conductivity.max_0, $P < 0.001$) was among the variables having the strongest statistical association with CMSCC. See Table 2 for details regarding all variables with $P < 0.050$ in both models.

The overall results indicated that the independent variables closer to the CMSCC sampling event were least likely to be associated with the dependent variable due to chance alone. Most, 21 out of 31 of the significant variables, occurred within the 6 milking sessions closest to the CMSCC sampling event. Nonlinear relationships were found for several of the independent variables, such as MDi, variance in conductivity, difference in conductivity or maximum conductivity of a quarter, which is indicated by the effective degrees of freedom being > 1 (Table 2).

GAM models: independent variables plot interpretation

The independent variables having the strongest statistical association with CMSCC, are visualized by smooth plots, estimated by the screening models (Figs. 1, 2). The smooth plots are showing the partial effects between log10CMSCC and the independent variable. Since smooths are expressed as overall mean and centered, i.e. moving around zero, the plots are expressing the nonlinear pattern between the independent variable and the dependent variable, but does not give any information regarding the height of the smooth in actual CMSCC units.

The partial effects of quarter conductivity for all four quarters and CMSCC one milking session before the CMSCC sampling event can be found in Figure 1. The trend lines of the relationship between quarter conductivity and CMSCC differed between the four quarters. The three quarters that had a statistical association with CMSCC, i.e., right front, right rear and left front, had more similar trend lines, while the trend line for left rear ($P = 0.296$) looked unlike the others, flat and straight. The relationship between MDi and the dependent variable was positive and the trend line steeper between MDi = 1 and MDi = 2, after which the line flattens out (Fig. 2a). The relationship between conductivity.var_1 and CMSCC was nonlinear and mainly positive, although there was a small negative trend when the variance between quarters exceeded 0.4 (Fig. 2b). Similar relationship was found for difference in conductivity between quarters and CMSCC (Fig. 2c). The maximum conductivity, during the milking session at the CMSCC sampling event, showed a nonlinear and clear-cut positive relationship with CMSCC (Fig. 2d).

GAM models: model selection

The results of the model selection for all models are presented in Table 3. The lowest AIC value (246) was found for model 1, indicating that the best model was the model including all variables for seven days. Comparing model 1 with model 2, where MDi and UC were excluded, the difference in AIC was not large (246 *vs.* 270). The difference in R^2_{adj} between model 1 and model 2 was very small (0.80 *vs.* 0.79), indicating that both models explained the variance in the data well.

The model performance results including milking data from various time points before the CMSCC sampling event, showed a large range in AIC among the 12 models (AIC = 246–517). The range in R^2_{adj} between the 12 models was not very wide (R^2_{adj} = 0.80–0.76). Models including milking data from the same milking as the CMSCC sampling event (i.e., milking session 0) had the consistently lower AIC values, but did not distinguish themselves in how much variance in the data they explained (R^2_{adj} being very similar). According to AIC, using milking data from six days before the CMSCC sampling event gave the best model performance among the time-restricted models (model 1_0:6, AIC = 264). By excluding MDi and UC (model 2_0:6) AIC increased to 287. The difference in R^2_{adj} between the two models with data from six days was small (0.80 *vs.* 0.79). Models with milking data restricted to only three days before the CMSCC sampling event (model 1_0:3 and model 2_0:3) were very similar. The difference in AIC was minimal (281 *vs.* 285) as was the difference in R^2_{adj} (0.79 *vs.* 0.79). The models with the highest overall AIC values were all models that excluded milking data from the two days closest to the CMSCC sampling event (e.g. model 1_2:5), and again amplifying that using data closer to the CMSCC sampling event resulted in better model performance.

Discussion

The objective of this study was to use routinely recorded sensor data to model CMSCC. The results indicate that GAM are suitable for modeling of CMSCC relatively well (R^2_{adj} ranging from 0.76 to 0.80) by combining quarter and composite sensor information.

Using milking data from three or six days before the CMSCC sampling event did not affect the performance of the models much, as long as milking data from the same milking session as the CMSCC sampling event were included. Excluding milking data from the same milking session as the CMSCC sampling event had a considerable effect on the overall fit of both model 1 and model 2. This suggests that there is some important information in variables from the same milking that improves the explanation of CMSCC. Using all variables from all seven days before the CMSCC sampling event gave the best model fit (model 1). Also Hammer *et al.* (2012) found changes in several milking trait variables even seven days before an event, although they studied clinical mastitis where lagged variables potentially should have been less informative since clinical mastitis may be more of a sudden event than the CMSCC that we are modeling. However, since the model fit was not strongly affected by excluding the three days farthest from the CMSCC sampling event, we suggest that using all variables for seven days before the CMSCC sampling event is actually not necessary to describe CMSCC.

The independent variables with the strongest statistical association with the dependent variable, according to the model including all variables over the seven days before the CMSCC sampling

https://doi.org/10.1017/S0022029920000692 Published online by Cambridge University Press

Table 2. Independent variables ($P < 0.050$) for model 1 and model 2

	model 1		model 2	
Confounder		*P-value*		*P-value*
lactation 1 (intercept)		<0.001		<0.001
lactation 2		0.016		0.011
lactation 3		0.162		0.333
Cow		<0.001		<0.001
Days in milk		0.007		0.003
Independent variable	*edf*	*P-value*	*edf*	*P-value*
MDi_0	4.94	<0.001	–	–
conductivity[a]_1	1	<0.001	1	0.003
conductivity[b]_1	1	<0.001	1	<0.001
conductivity.diff_1	1	0.001	1	0.003
MDi_10	3.56	0.004	–	–
conductivity.max_0	3.74	0.005	4.43	<0.001
conductivity.max_1	2.40	0.005	1.99	0.003
milk yield.diff.milkings[b]	1	0.005	1.65	0.053
IncompleteScore_3:2[c]	*NA*	0.008	–	–
MDi_9	3.86	0.010	–	–
conductivity.var_1	4.56	0.011	5.64	<0.001
conductivity[d]_1	1	0.011	1	0.007
conductivity.mean_0	2.64	0.023	3.08	0.008
MDi_1	3.67	0.027	–	–
IncompleteScore_4:4[c]	*NA*	0.028	*NA*	0.173
IncompleteScore_4:2[c]	*NA*	0.031	*NA*	0.041
conductivity.diff _5	1	0.031	1	0.074
MDi_7	1	0.033	–	–
MDi_13	1	0.035	–	–
IncompleteScore_5:1[c]	*NA*	0.037	*NA*	0.099
UdderCounter_8:4	*NA*	0.038	–	–
MDi_12	1	0.039	–	–
UdderCounter_12	*NA*	0.045	–	–
IncompleteScore_13:1[c]	*NA*	0.049	*NA*	0.091
conductivity.var_3	1	0.080	1	0.014
conductivity.diff_9	3.44	0.584	1	0.017
conductivity.diff _3	1	0.053	1	0.004
conductivity.diff _0	1.15	0.454	4.41	0.030
conductivity[b]_4	2.27	0.119	4.59	0.044
conductivity.max_9	3.17	0.179	3.94	0.006
conductivity.diff_13	1	0.162	1	0.043

–, not used in model; *NA*, not applicable.
Effective degrees of freedom (edf) indicates the relationship (1 = linear, >1 nonlinear) with the dependent variable, i.e., composite milk somatic cell count. Suffix number of independent variable indicates number of milking sessions before the composite milk somatic cell count sampling event
[a]Left rear.
[b]Right front.
[c]Score level.
[d]Left front.

https://doi.org/10.1017/S0022029920000692 Published online by Cambridge University Press

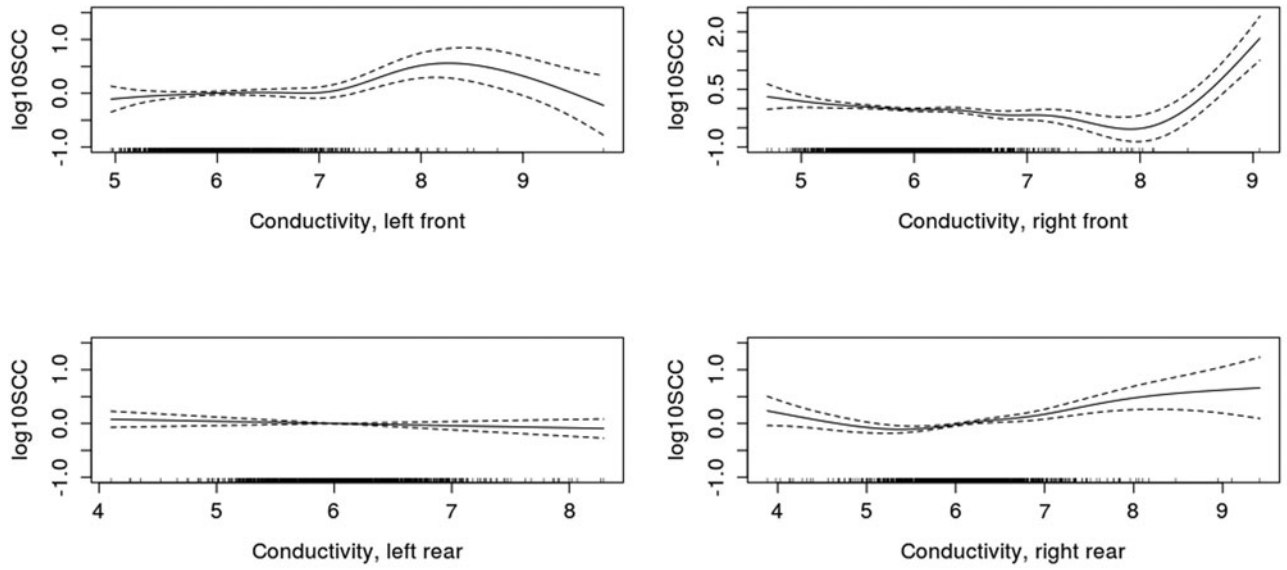


Fig. 1. The effect of quarter conductivity on composite milk somatic cell count one milking session before the composite milk somatic cell count sampling event, estimated by the screening model. The pointwise 95% confidence interval is shown by the dashed lines. The vertical lines on the *x*-axis show the individual quarter conductivity datapoints. The *y*-axis shows the composite milk somatic cell count transformed to a log10 scale (log10CMSCC). The smooths are expressed as deviations from the overall mean.

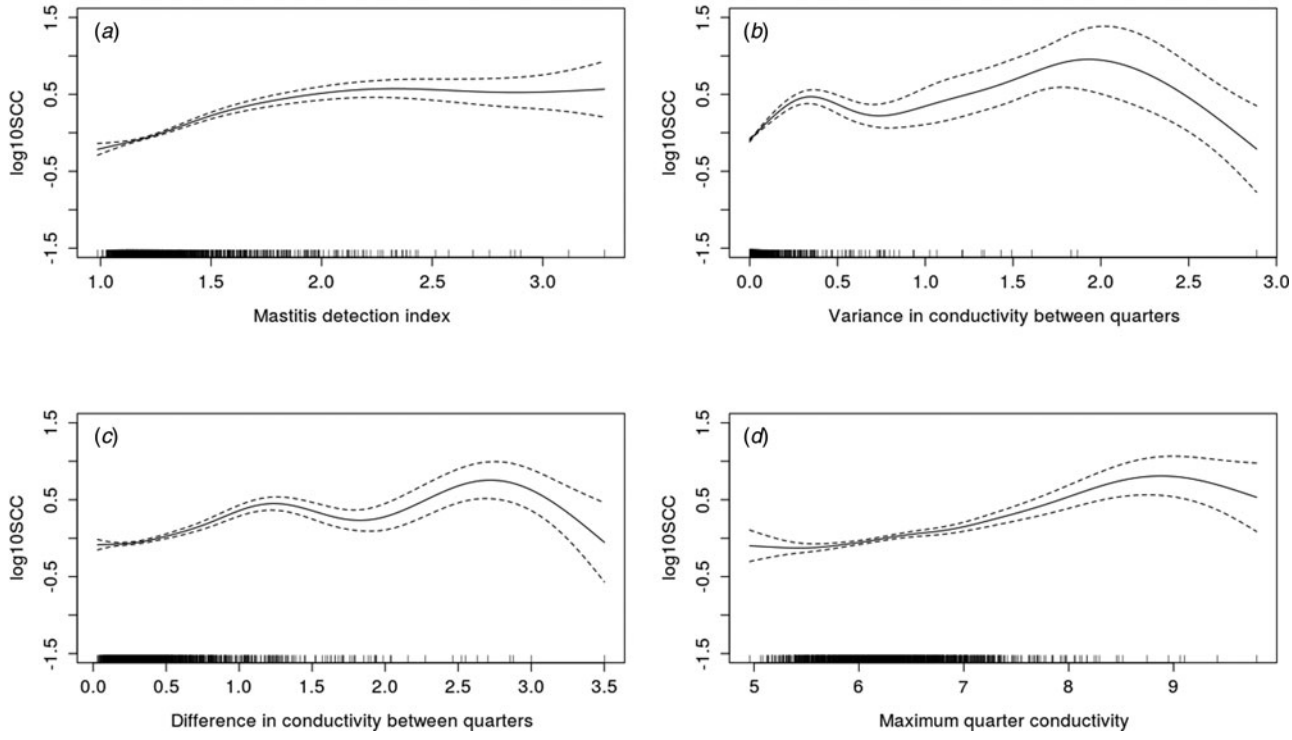


Fig. 2. The partial effects of the (a) mastitis detection index (MDi), (b) variance in conductivity between quarters (conductivity.var), (c) difference in conductivity between quarters (conductivity.diff), (d) maximum quarter conductivity (conductivity.max), one milking session before the composite milk somatic cell count sampling event, estimated by the screening model. The pointwise 95% confidence interval is shown by the dashed lines. The vertical lines on the *x*-axis show the individual MDi datapoints. The *y*-axis shows the composite milk somatic cell count.

event, was MDi at the same milking session as the CMSCC sampling event. The relationship between CMSCC and MDi was positive, indicating that higher MDi values (i.e. conductivity in different phases of the milking) were associated with higher CMSCC. However, there were few data points for which MDi > 2, which is probably why the line flattens for MDi > 2.5 (Fig. 2a). Contrary to our results, a previous study by Khatun *et al.* (2018) showed that MDi performed no better than did conductivity at the quarter level in a mastitis detection model, and MDi was excluded from their final model. The differences in

https://doi.org/10.1017/S0022029920000692 Published online by Cambridge University Press

Table 3. Performance of models using data from various time points before the composite milk somatic cell count (CMSCC) sampling event ranked according to lowest corrected Akaike information criterion (AIC)

Rank	Model	AIC	R^2_{adj}	*n* variables
1	model 1	246	0.80	102
2	model 1_0:6	264	0.80	91
3	model 2	270	0.79	81
4	model 1_0:3	281	0.79	53
5	model 2_0:3	285	0.79	44
6	model 2_0:6	287	0.79	72
7	model 1_1:5	321	0.79	66
8	model 2_1:5	336	0.78	51
9	model 1_1:7	339	0.79	91
10	model 2_1:7	364	0.78	71
11	model 2_2:5	490	0.76	39
12	model 1_2:5	504	0.76	49
13	model 2_2:7	513	0.76	57
14	model 1_2:7	517	0.76	74

Adjusted *R*-squared (R^2_{adj}) and number (*n*) of variables in each model presented. Number after model name corresponds to closest day lag to CMSCC sampling event included in the model, while second number corresponds to farthest day lag from CMSCC sampling event included in the model (i.e., 0:6 is same day as CMSCC sampling event up to six days prior)

the results might be due to the nature of the different outcomes studied (SCC *vs.* clinical mastitis defined as veterinary treatment), model choice (nonlinear *vs.* linear model), difference in milking interval, or available data (one *vs.* two herds). Furthermore, Lusis *et al.* (2017) found the correlation between SCC and MDi to be poor; however, due to the small sample size and few observations in their study, the authors suggested that individual cows might have affected the results.

When MDi and UC were excluded (model 2), variance in conductivity between quarters had the strongest statistical association with CMSCC. This could indicate that independent variables combining quarter information are somewhat better at describing CMSCC than are independent variables from separate quarters. For example, variance in conductivity between quarters or MDi, had the strongest statistical relationship with CMSCC in both models. Most variables (15 of 17 variables with $P < 0.05$) in model 2 were conductivity related, i.e., variance, difference, mean or maximum conductivity. Previous studies have shown that including several different types of conductivity variables, such as variance or maximum conductivity values, in a mastitis prediction model, increased the specificity (Norberg *et al.*, 2004). The degree of explanation, of all models, could probably partly be explained by the inclusion of the different conductivity variables, describing different traits. The variables will probably contribute in different ways to the model and implies that adding several different types of quarter combined conductivity variables as nonlinear independent variables increases the level of explanation for CMSCC.

Conductivity at the quarter level was a statistically strongly associated independent variable in both models and stronger in model 2 than in model 1. Estimates from both models showed that not all four quarters were significantly related to CMSCC within the same milking session, e.g. right rear conductivity_1 in model 1 was $P = 0.497$ while remaining quarters conductivity_1 all had *P*-values < 0.050. However, the univariable screening model displayed a significant nonlinear relationship between CMSCC and conductivity for each quarter. Previous studies have shown that the relationship between SCC and conductivity seems to be positive (Hamann and Zecconi, 1998). This is in agreement with our findings, but the pattern is neither similar nor very clear for all quarters (Fig. 1). The varying results regarding the relationship between quarter conductivity and SCC may be because the independent variable and the dependent variable are measured at different levels (i.e., quarter *vs.* composite); also, as the CMSCC in this study was low (median 53 000 cells/ml), the trend would not be as clear as it might have been had the CMSCC been higher. Furthermore, the size of the dataset available might influence these results i.e., more data on each quarter could possibly equalize the differences between the quarters. The relationship between quarter conductivity and CMSCC has not previously been investigated using nonlinear modeling, which makes it impossible to compare the present and previous findings.

Quarter conductivity alone has been stated to be a poor predictor of clinical mastitis (Kamphuis *et al.*, 2008; Khatun *et al.*, 2018), and within-cow comparison of quarters is often recommended (Hamann and Zecconi, 1998). Our results suggest that conductivity variables at quarter level, such as quarter conductivity or maximum quarter conductivity, contributes in explaining the CMSCC to the same extent as conductivity variables from combined quarters, i.e. MDi or variance in conductivity between quarters. This, since quarter level conductivity variables had a strong statistical association with CMSCC in both univariable and multivariable models. Quarter conductivity observations close in time to the CMSCC sampling event (i.e., conductivity_1) showed a stronger statistical association with CMSCC than observations made several days before sampling. This is in line with the findings of Nielen *et al.* (1995) that the difference in conductivity between quarters was the largest at the milking when the clinical mastitis was observed compared with the two milkings before the observation.

High flow rates were not significantly associated with CMSCC in the univariable screening in contradiction to Sitowska *et al.* (2017) who found milk flow to be the second most important variable describing SCC i.e. higher milk flows being more associated with SCC > 80 000 cells/ml. The difference in the results might depend on what data were used as input to the decision tree, since even a small change in data can cause a large change in the final result of decision tree models (Gareth *et al.*, 2013). However, low peak flow rates (peakflow.min) were significant in the univariable screening model although not in the final models. This is somewhat in agreement with the findings of Hammer *et al.* (2012), who showed that low peak flows could be associated with clinical mastitis, but were not significant when used as an input variable in a multivariable model. Additionally, Ebrahimie *et al.* (2018) did not find peak flow to have any weight in either of their decision tree models predicting SCC.

Several of the most important variables describing CMSCC in the present study, such as DIM, LN, and conductivity were confirmed by the decision tree models for prediction of SCC by Sitowska *et al.* (2017) and Ebrahimie *et al.* (2018). The advantage of decision trees is that they are easily interpreted and explained although they often lack the accuracy of regression models. While the interpretation of GAM might be harder, GAM are flexible and can provide information regarding both linear and nonlinear relationships between the independent variables and the dependent variable (Hastie and Tibshirani, 1990). So far, GAM

https://doi.org/10.1017/S0022029920000692 Published online by Cambridge University Press

have not been widely used in mastitis research. Ankinakatte *et al.* (2013) performed one of the few studies comparing a GAM and a neural network for detecting mastitis. Their results imply that the performance of the GAM model was slightly better than that of the neural network depending on what input variables were used.

In conclusion, the present results indicate that GAM could be used for prediction models of CMSCC, since the CMSCC modeling results showed a relatively high degree of explanation using the information routinely recorded by the milking robot. Using milking data from the three days (i.e., six milking sessions) before the CMSCC sampling event also gave a high degree of explanation compared with that of the best model, which used milking data from seven days (i.e., 14 milking sessions). Variables combining quarter conductivity (e.g., MDi) and variance between quarters, but also quarter conductivity as such, are suggested to have the strongest associations with CMSCC and should preferably be combined when predicting CMSCC. Further research is needed to verify our results under other conditions i.e. other farms, voluntary milking or more cows, to evaluate whether GAM models can also accurately predict CMSCC using sensor data routinely collected by the milking robot.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0022029920000692

Acknowledgements. Financial support for this study was provided by Swedish Foundation of Strategic Research (SFF, Stockholm, Sweden). Data was provided by Delaval International. We would like to thank Staffan Betnér at Swedish University of Agricultural Sciences for statistical support, the farm staff and special thanks to Miriam Weber for collecting sample data.

References

Ankinakatte S, Norberg E, Løvendahl P, Edwards D and Højsgaard S (2013) Predicting mastitis in dairy cows using neural networks and generalized additive models: a comparison. *Computers and Electronics in Agriculture* **99**, 1–6.

Ebrahimie E, Ebrahimi F, Ebrahimi M, Tomlinson S and Petrovski KR (2018) Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. *Computers and Electronics in Agriculture* **147**, 6–11.

Fox J and Monette G (1992) Generalized collinearity diagnostics. *Journal of the American Statistical Association* **87**, 178–183.

Gareth James G, Witten D, Hastie T and Tibshirani R (2013) *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Hamann J and Zecconi A (1998) Evaluation of the electrical conductivity of milk as a mastitis indicator. *Bulletin-FIL-IDF (Belgium)* **334**, 5–22.

Hammer JF, Morton JM and Kerrisk KL (2012) Quarter-milking-, quarter-, udder- and lactation-level risk factors and indicators for clinical mastitis during lactation in pasture-fed dairy cows managed in an automatic milking system. *Australian Veterinary Journal* **90**, 167–174.

Hastie T and Tibshirani R (1990) *Generalized Additive Models*. New York: Chapman & Hall.

International Dairy Federation (2013) Guidelines for the use and interpretation of bovine milk somatic cell count. *Bulletin of the IDF* 466.

ISO/IEC (2005) General requirements for the competence of testing and calibration laboratories. 2nd ed. International Organization for Standardization/ International Electrotechnical Commission Committee on Conformity Assessment, Geneva, Switzerland. 17025:2005.

Kamphuis C, Sherlock R, Jago J, Mein G and Hogeveen H (2008) Automatic detection of clinical mastitis is improved by in-line monitoring of somatic cell count. *Journal of Dairy Science* **91**, 4560–4570.

Khatun M, Thomson PC, Kerrisk KL and García SC (2018) Development of a new clinical mastitis detection method for automatic milking systems. *Journal of Dairy Science* **101**, 9385–9295.

Kitchen BJ (1981) Review of the progress of dairy science: bovine mastitis: milk compositional changes and related diagnostic tests. *Journal of Dairy Research* **48**, 167–188.

Lusis I, Antane V and Laurs A (2017) Effectiveness of mastitis detection index for cow monitoring and abnormal milk detection in milking robots. *Engineering for Rural Development* **16**, 1383–1387.

Nielen M, Schukken Y, Brand A, Haring S and Ferwerda-Van Zonneveld R (1995) Comparison of analysis techniques for on-line detection of clinical mastitis. *Journal of Dairy Science* **78**, 1050–1061.

Norberg E, Hogeveen H, Korsgaard IR, Friggen NC, Sloth KHMN and Løvendahl P (2004) Electrical conductivity of milk: ability to predict mastitis status. *Journal of Dairy Science* **87**, 1099–1107.

Quist MA, LeBlanc SJ, Hand KJ, Lazenby D, Miglior F and Kelton DF (2008) Milking-to-milking variability for milk yield, fat and protein percentage, and somatic cell count. *Journal of Dairy Science* **91**, 3412–3423.

R Development Core Team (2018) R: A Language and Environment for Statistical Computing. Accessed December 12, 2018. http://www.r-project.org.

Schukken YH, Wilson DJ, Welcome F, Garrison-Tikofsky L and Gonzalez RN (2003) Monitoring udder health and milk quality using somatic cell counts. *Veterinary Research* **34**, 579–596.

Sharma N, Singh NK and Bhadwal MS (2011) Relationship of somatic cell count and mastitis: an overview. *Asian-Australasian Journal of Animal Sciences* **24**, 429–438.

Sitowska B, Piwczyński D, Aerts J, Kolenda M and Özkaya S (2017) Detection of high levels of somatic cells in milk on farms equippedwith an automatic milking system by decision trees technique. *Turkish Journal of Veterinary and Animal Sciences* **41**, 532–540.

Wood SN, Pya N and Säfken B (2016) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* **111**, 1548–1563.

https://doi.org/10.1017/S0022029920000692 Published online by Cambridge University Press