Computers and Electronics in Agriculture 218 (2024) 108729

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

DAIRY CHAOS: Data driven Approach Identifying daiRY Cows affected by HeAt lOad Stress

Marco Bovo *, Mattia Ceccarelli, Miki Agrusti, Daniele Torreggiani, Patrizia Tassinari

Department of Agricultural and Food Sciences, University of Bologna, Viale G. Fanin 48, Bologna, 40127, Italy

ARTICLE INFO

Keywords:
Dairy cow
Heat stress
Milk yield
Multifit
Animal welfare

ABSTRACT

In intensive farming systems the facilities have a central role on both animal welfare and animal production all this paving the way of researching new housing systems and management strategies for reducing the impacts. In particular, in the dairy cattle sector, the early detection of irregular productions is fundamental for animal health and safety. On the other hand, despite the growing interest concerning the modelling and forecasting daily production data, there is lack of studies devoted to identification of anomalous data. To this regard, in this work, a data driven approach for detecting milk production and behaviour anomalies is presented and applied to three farms selected as case study. The DAIRY CHAOS procedure proposed in this paper bases on two numerical algorithms having the scope of separately detect anomalies daily data for a single cow. Both the algorithms presented hereinafter have statistical foundations and take in input daily resting time, milk yield and climate data respectively recorded by pedometer worn by the cow, automatic milking robot and a thermo-hygrometer data logger installed in each barn. The first algorithm takes into consideration three indicators, namely Relative Yield Difference, Relative Laying time Difference and Cumulative Discomfort Index. An anomaly, i.e. a deviation from a normal value, is determined, for a single cow, for a specific day, if the three conditions assessing a noticeable deviation from the normal values of the three indicators above are contemporary verified. The second algorithm, by means of a multifit procedure, introduces the concept of reliability of robust statistics and provides statistically solid, since not affected by outlier values, milk yield and laying time trends for each animal. The application, in a production context, of the procedure proposed here can result extremely useful for the identification of animals suffering heat stress and therefore can become a support to the farmer's decisions for the mitigation of the heat stress effects and a more efficient management of the animals.

1. Introduction

In the dairy cattle sector, the cornerstones of sustainability can be recognised as milk production and milk quality, cow health and welfare, efficiency in the use of raw resources, and emissions reduction (Strpić et al., 2020). Animal welfare is strictly related to sustainability, due to the consequences in terms of milk quantity and quality, which affect the efficiency of the use of natural resources. For this purpose, a crucial point in the dairy cattle sector is the prevention and the proper management of the heat stress, as it is markedly jeopardising animal welfare in several countries in the Mediterranean area. The sudden increase in temperatures coupled with more frequent occurrences of extreme events observed in the recent years resulting from climate change, is having serious effects also on livestock production (Neves et al., 2022). Heat stress in dairy cows can be defined as the stress induced to the cattle when they are unable to dissipate heat without modifying the body thermal balance. Heat stress is usually related to environmental conditions, for example, in animals reared in environment characterised by high temperature, high humidity or exposed to strong solar radiation (Charlton et al., 2013; Herbut and Angrecka, 2017; Thornton et al., 2022; West, 2003; Bovo et al., 2022). In few cases, heat stress can be attributed to an internal heat overproduction by the animal (Bernabucci et al., 2014; Kadzere et al., 2002). Moreover, as widely described by literature, the heat stress effects will most likely continue to increase due to ongoing trend of temperatures (Ji et al., 2020; Burhans et al., 2022; Moore et al., 2023). In this scenario, several studies have shown how dairy cows subjected to adverse climatic conditions and heat waves, are often affected by stress (Cowley et al., 2015), with negative consequences not only on animals' health, but also on milk yield and milk quality and animal behaviour. Currently, heat stress is among the most investigated issues in the context of dairy cows, and is the subject for continuous analysis

* Corresponding author.
E-mail address: marco.bovo@unibo.it (M. Bovo).

https://doi.org/10.1016/j.compag.2024.108729
Received 14 August 2023; Received in revised form 11 January 2024; Accepted 6 February 2024
Available online 9 February 2024

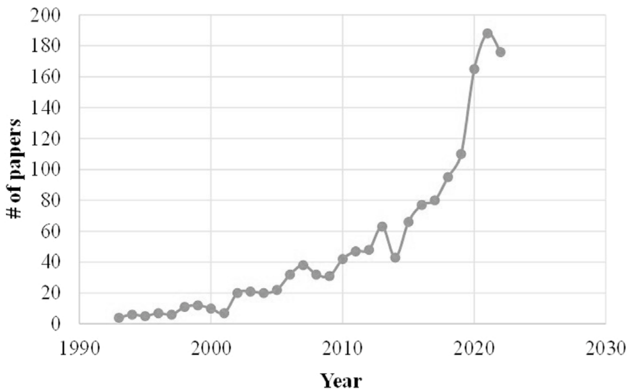


Fig. 1. Trend of the recent bibliography investigating the effects of the heat stress on dairy cows.
Source: Scopus (Elsevier, 2023).

and insights. As a further confirmation, Fig. 1 shows how the number of publications investigating the effects of the heat stress on dairy cows is constantly increasing in time (the trend covers a 30-year time span and represents the number of papers extracted by Elsevier (2023) by filtering for title, keywords and abstract).

On the other hand, the continuous increase and development of Precision Livestock Farming (PLF) techniques, representing an alternative approach for the management of the livestock aiming to increase both farm sustainability and animal welfare by the automatic real-time monitoring and controlling, is pushing the research towards the development of more complex numerical models. In fact, PLF paved the way for the collection and development of heterogeneous dataset.

Moreover, the recent increasingly widespread adoption of Automatic Milking System (AMS) technology and the availability of more precise climate data have facilitated the calibration of methods aimed at quantifying and predicting heat stress in a more accurate manner (Bonora et al., 2018; Benni et al., 2020). The availability of large dataset has provided the conditions for the application of advanced numerical techniques already applied in other fields of research.

The use of sensors that continuously collect punctual data regarding animal production, animal behaviour, animal welfare and the thermo-hygrometric conditions in the barn, allowed development of mathematical models capable on one hand of clustering animals based on their behaviour and characteristics and on the other, to quantitatively predict the daily milk yield as a function of the climatic conditions of the barn (Bovo et al., 2021).

As a matter of fact, several recent papers investigated the relation between environmental conditions and one or more animal-based indicators with the main objective of modelling short- or long-term effects of heat stress. In most of these works, the environmental conditions have been modelled by the Temperature–Humidity Index (THI) in the barn (Chamberlain et al., 2022) or indices derived by the THI, like the Heat Load Index (HLI), developed mainly for animals raised outdoors and considering also air velocity and solar radiation values (Lees et al., 2018). As far as the animal-based indicators are concerned, a large part of the research directly focused on milk yield and tried to establish numerical models for the assessment of the milk yield reduction or for the evaluation of the time lag between heat stress condition and production drop (Ekine-Dzivenu et al., 2020). On the other hand, heat stress is well correlated also to modification of the daily routine of the animals and the daily laying time, i.e., the number of hours in which the animal lays down, is a well recognised animal-based feature rather simple to measure and at the same time strongly influenced by heat stress. It is worth noticing that despite the growing interest concerning modelling and forecasting milk yield and animal behaviour data, there is a lack of studies investigating the identification of days with milk yield anomalies. The real time detection of milk yield anomalies could be of fundamental importance to establish animal health and welfare. To this regard, the present paper describes a data driven approach for detecting milk production and behavioural anomalies and applies it to three farms selected as representative case study. The DAIRY CHAOS (Data driven Approach Identifying daiRY Cows affected by HeAt lOad Stress) procedure proposed in the paper is based on two numerical algorithms, that have the objective of separately detect daily anomalies data for a dairy cow. Both algorithms use statistical foundations and take as input Daily Laying Time (DLT), Daily Milk Yield (DMY) and climate data. Data is recorded respectively by pedometer worn by the cow, automatic milking robot and a thermo-hygrometer data logger installed in each barn. The first algorithm takes into consideration three indicators, namely Relative Yield Difference (RYD), Relative Laying time Difference (RLD) and Cumulative Discomfort Index (CDI). An anomaly, i.e. a deviation from an ideal value, is determined, for a single cow, for a specific day, if the three conditions assessing a noticeable deviation from the expected values of RYD, RLD and CDY are simultaneously verified. The second algorithm introduces the concept of reliability of robust statistics and provides statistically solid trends of milk yield and laying time for an animal.

The paper is structured in the following way: Section 2, introduces the concept of ideal lactation curve, the concept of anomaly, the two algorithms used by the procedure and the main characteristics of the dataset adopted for the application of the method. Section 3 shows the main results of the method applied to the selected dataset and provides insights and details into the usefulness and reliability of the method in both theoretical and real applications. Section 4 summarises the main conclusions of the paper, highlights potential and criticality of the method and provides the indications that should guide future research and developments. For the sake of clarity, Table 1 shows acronym, type, unit and description of data and parameters used in the paper.

Table 1
Acronyms and labels.

Code	Type	Unit	Brief description
THI	Float	–	Temperature-Humidity Index
Datetime	Datetime	–	Date of report
Farm_id	String	–	Code of the farm
Animal	String	–	Animal ID
Parity	Integer	–	Number of lactations
DIM	Integer	–	Days In Milk
DMY	Float	kg/day	Daily Milk Yield
DLT	integer	h/day	Daily Laying Time
RYD	Float	–	Relative Yield Difference
RLD	Float	–	Relative Laying time Difference
CDI	Float	–	Cumulative Discomfort Index
MYT	Float	–	Multifit Yield Threshold
MLT	Float	–	Multifit Laying time Threshold

2. Materials and methods

This section is structured in the following way. First, the description of the main characteristics of the dataset used in the study is introduced. Then, a focus on the lactation curve model assumed in the work is reported. Moreover, the methodology used for the establishment of a cross-correlation analyses between DMY and THI and the concept of anomaly are described. The last section provides the details of the two algorithms which constitute the basis for the data-driven anomaly detection method proposed in this paper.

2.1. Dataset description

The data used in the work were gathered from March 2020 to May 2022 in three farms (named CR04, MN05 and MN07) located in the Po Valley region, in northern Italy (see Fig. 2). Each farm is equipped with two AMSs Merlin (Fullwood Packo, England) that collect daily

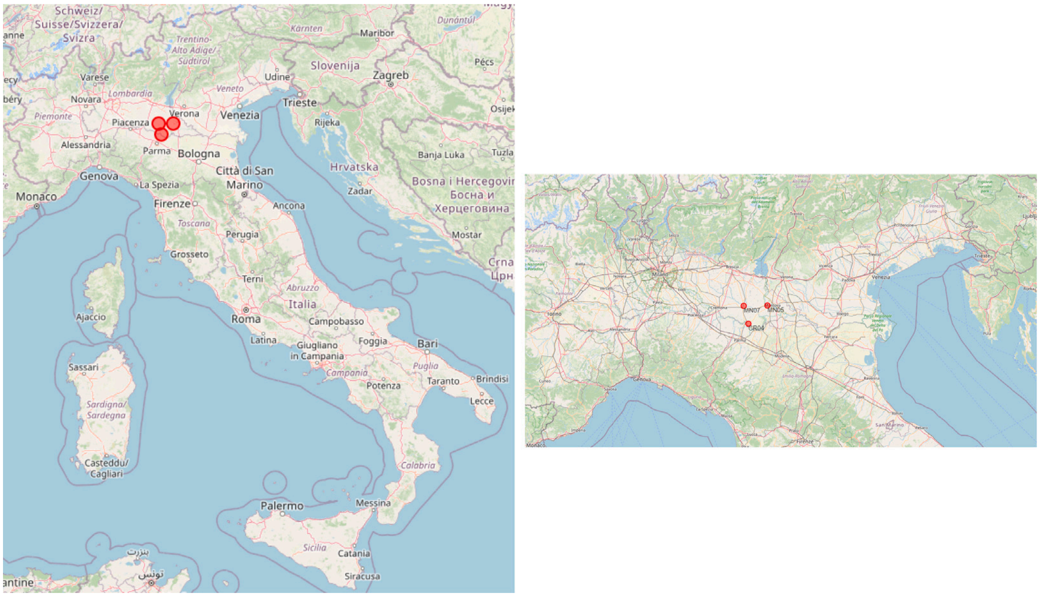


Fig. 2. Map of the position of the three barns indicated with three red dots: Italian territory (left) and regional territory (right).

data on milk yield and milk quality for each cow. The size of the three herds is similar. The three farms have about 120 milking cows in each farm. The daily laying time, i.e. the number of hours in which the animal lays down, is collected by a pedometer mounted on each animal and is automatically transferred, during the milking session of a single animal, to the software that manage the AMSs. In order to uniform the dataset length of the different cows, the lactation period was assume at the maximum of 305 days in milk (DIM). Therefore, for cows having a long lactation, only the first 305 days have been considered. Moreover, a lactation has been considered valid for the analyses only if it contains data at least for 250 days. In the two-year time span under consideration, CR04 provided data of 209 unique animals and 412 valid lactations, MN05 had data of 146 unique animals and 246 valid lactations, MN07 collected data of 291 unique animals for a total of 472 valid lactations. In total, the number of unique animals in the dataset is 646 and the number of valid lactations is 1130, with an average value of 1.75 lactation/cow.

A thermo-hygrometer data logger, PCE-HT71, with an accuracy of 3% for relative humidity and of 1 °C for temperature, has been positioned inside each barn approximately in the same position at the centre of the barn surface and at the same height of about 2 m from the pavement. The three thermo-hygrometer data loggers covered the same time span period from March 2020 to May 2022. The THI values adopted in the analyses was calculated as indicated by the National Research Council in Kelly and Bond (1971), from the data collected by the thermo-hygrometer:

$$THI = 1.8 \cdot T + 32 - [0.55 - 0.0055 \cdot rH \cdot (1.8 \cdot T - 26)] \tag{1}$$

where T is the air dry bulb temperature (°C) and rH is the air relative humidity (%). The daily average THI has been computed as the mean in 24 h.

A statistical description of the dataset is provided in Table 2 for each farm. The most common statistical parameters are reported for the features of interest for the analyses (DMY, parity number and DLT). Statistics for DIM are not reported since its value is uniformly distributed between 1 and 305 days for all the three herds.

2.2. Lactation curve modelling

The lactation curve, i.e. the relation between DMY and DIM is assumed in accordance with the Wood's model (Wood, 1967):

$$DMY(DIM) = a \cdot e^{-b \cdot DIM} \cdot DIM^{c} \tag{2}$$

Table 2
Statistical summary of the features DMY, parity number and DLT for the three farms.

	Min	25%	50%	75%	Max	Mean	Std
DMY $\left(\frac{kg}{day}\right)$							
CR04	0.9	30.5	36.7	43.8	76	37.0	9.7
MN05	0.9	35.4	41	48.7	79.2	41.8	10.6
MN07	1.5	33.6	40	48.3	82	41.1	11.1
Parity (–)							
CR04	1.0	1.0	2.0	4.0	8.0	2.7	1.6
MN05	1.0	1.0	2.0	3.0	8.0	2.0	1.2
MN07	1.0	1.0	2.0	3.0	8.0	2.3	1.4
DLT $\left(\frac{h}{day}\right)$							
CR04	0.0	8.0	10.0	11.0	19.0	9.9	2.2
MN05	0.0	10.0	12.0	14.0	23.0	11.7	2.8
MN07	0.0	8.0	10.0	11.0	19.0	9.6	2.4

where DMY is the daily milk yield at a lactation stage equal to the DIM value and a, b and c are three parameters that control the shape of the lactation curve: a is the scaling factor, that controls the production at beginning of lactation and peak production, b and c influence respectively the post peak behaviour and the final slope of the lactation curve (Silvestre et al., 2006). Data was stratified by parity and the model in Eq. (2) was fitted. As shown in Fig. 3, the main difference in lactation curves is between the model of primiparous and the models of multiparous cows. Primiparous cows (i.e. the red line) generally present a lower peak production with a more flat post-peak behaviour. All the other parities show a similar behaviour with a more evident production peak at about 50–60 days in milk.

Indeed, first parity cows show lower peak in production, but maintain an almost constant DMY until the end of the lactation period (i.e. 305 days). On the other hand, the curves of cows with parity greater than 1, have an higher peak DMY but a steeper decrease in the second stage of lactation (for DIM value between 50 and 200 days). In general, the total milk yield for a lactation increases after the first one. For these reasons, in order to obtain suitable characteristic lactation curves for the study, the dataset has been divided in two subset, analysed independently. A first subset with data of lactations with parity = 1 and a second subset containing data with parities ≥ 2. The parameters$a$, b and c are obtained by a non-linear least square minimisation, using the *lmfit* (Newville et al., 2014) library for Python. The results are shown in Fig. 4.

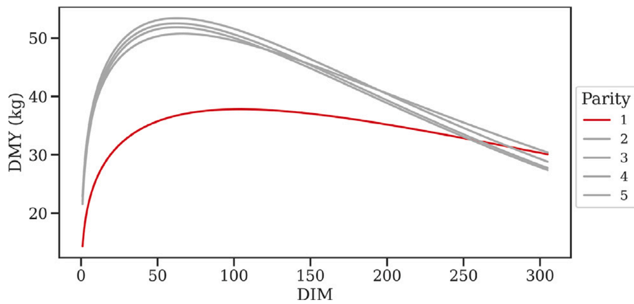


Fig. 3. Best fitting curves for the different parity numbers.

The final equations of the two best fit curves are:

$$DMY_1(DIM) = 14.9 \cdot e^{-0.0025 \cdot DIM} \cdot DIM^{0.255} \tag{3}$$

$$DMY_2(DIM) = 23.9 \cdot e^{-0.0039 \cdot DIM} \cdot DIM^{0.247} \tag{4}$$

respectively, for primiparous cows Eq. (3) and for multiparous cows Eq. (4).

2.3. Cross-correlation analysis

A further preliminary analysis has been carried out by implementing a cross-correlation was carried out between the time series of median DMY and median THI, in order to identify and quantify in an aggregated way the relationship between indoor climatic conditions in the barn (synthesised by the THI) and DMY of the herd. Indeed, a reduction of the milk production is one of the most recurring and evident effects induced by heat stress conditions (Wildridge et al., 2018). By means of the cross-correlation analysis, it is possible to quantify the herd's response time (in days) between high THI values and drop in DMY.

The time series of median DMY and median THI have been considered in the cross-correlation evaluation. Fig. 5 shows the trend of the time series of median DMY (black line) and median THI (green and red lines) for the whole dataset, which gather the three farms. The cross-correlation analysis has been computed between the time series of median THI with a forward-lagged version of median DMY, shifted by an increasing number of days $s \in [0, 100]$. The correlation has been estimated with the Pearson's correlation coefficient reported in (5):

$$\rho_s(DMY_s, THI) = \frac{cov(DMY_s, THI)}{\sigma_{DMY} \cdot \sigma_{THI}} \tag{5}$$

where ρ_s is computed with a forward-lag of s days for median DMY_s, cov is the covariance function and σ_x is the variance of the respective variable.

As reported in literature, THI can influence production with effects on milk yield on both long term and short term (Wildridge et al., 2018; Tao et al., 2018; M'Hamdi et al., 2021). To capture these effects, multiple cross-correlations were carried out with different thresholds for THI, starting from $THI \geq 30$ (i.e. including all data) to $THI \geq 70$, with a step of 1 THI. In Fig. 6 is shown the difference between two cross-correlation functions, i.e. $\rho_s(DMY_s, THI)$, obtained by using dataset filtered by THI. In the two cases the data considered were those higher than THI thresholds of 30 and 65, respectively. In the figure, the minimum correlation value and the corresponding time lag $\bar{s}$ have been reported.

On the one hand, the first function shows the long time effects of heat stress, which result in a lagged response of about 50 days. On the other hand, filtering by high THI values, highlights the short time response to heat stress, since the maximum negative correlation occurs 3–5 days after a day characterised by high THI. By collecting for each THI threshold the minimum correlation coefficients and the corresponding time lag $\bar{s}$, it is possible to find the relationship between time lag and THI and response time of DMY, which is plotted in Fig. 7. The time lag shows a progressively declining trend starting from 40–45 days up to a THI equal to 57, where the trend become rather flat around values of 4–5 days. From this analysis it is possible to conclude that maximum milk yield drop occurs about 4–5 days after the presence of high THI values in the barn, where high THI values are those higher than 57.

2.4. The DAIRY CHAOS approach

The DAIRY CHAOS procedure proposed in this work has the main goal to detect the arising of anomalous data in the production and behaviour trends of a dairy cow and induced by heat stress. The identification of the anomalous days follows the flowchart shown in Fig. 8.

Two different algorithms identify the potential anomalous daily data by following two independent paths. If both the algorithms identify the presence of an anomaly the proposed method classifies the day as anomalous for that cow. In the following, the two algorithms are deeply illustrated and commented.

2.4.1. Indicator-based algorithm

The first indicator-based algorithm used in the method, takes into consideration the three indicators listed below and already cited in the introduction section:

- Relative Yield Difference (RYD)
- Relative Laying time Difference (RLD)
- Cumulative Discomfort Index (CDI)

The RYD can be defined as the difference between relative differences in milk yield between the DMY of a single cow and the baseline model obtained for the whole herd. The baselines adopted in the study are the two best fit curves obtained in Section 2.2, for primiparous cows Eq. (3) and for multiparous cows Eq. (4). The RYD is then defined as:

$$RYD = \frac{\Delta DMY_t}{DMY_{t-1}} - \frac{\Delta DMYB_t}{DMYB_{t-1}} \tag{6}$$

where $\Delta DMY_t = DMY_{t-1} - DMY_t$ is the milk yield difference between day t and day $t-1$ for the specific animal, while $\Delta DMYB_t = DMYB_{t-1} - DMYB_t$ is the equivalent difference but calculated on the baseline curve (i.e. the best fit curve). RYD is a parameter measuring how different is the local trend of production curve of a specific animal compared with the baseline production curve.

The DLT parameter is conceptually similar to RYD but it is calculated on the data of daily laying time of cows. This measure is known to be related with heat stress (Hut et al., 2022) and is a reliable indicator of modification of the animal behaviour as a response to heat stress presence. In particular, a decrease in DLT is associated with increase in level of stress. The RLD parameter is defined as:

$$RLD = \frac{\Delta DLT_t}{DLT_{t-1}} - \frac{\Delta DLTB_t}{DLTB_{t-1}} \tag{7}$$

where: $\Delta DLT_t = DLT_{t-1} - DLT_t$ is the daily laying time difference between day t and day $t-1$ for the specific animal, while $\Delta DLTB_t = DLTB_{t-1} - DLTB_t$ is the equivalent difference but calculated on a baseline curve for laying time. In this case the best fit curve providing the relation between laying time and DIM has an almost linear trend. The following general equation has been assumed:

$$DLT(DIM) = d + e \cdot DIM \tag{8}$$

where: DLT (in hours) is the daily laying time, i.e. the time spent by a cow laying down, at a lactation stage equal to the DIM value and d and e are the two coefficients of the regression curve. The equations of the two best fit curves, obtained respectively for the primiparous and multiparous cows in the dataset, are:

$$DLT_1(DIM) = 8.83 + 0.0121 \cdot DIM \tag{9}$$

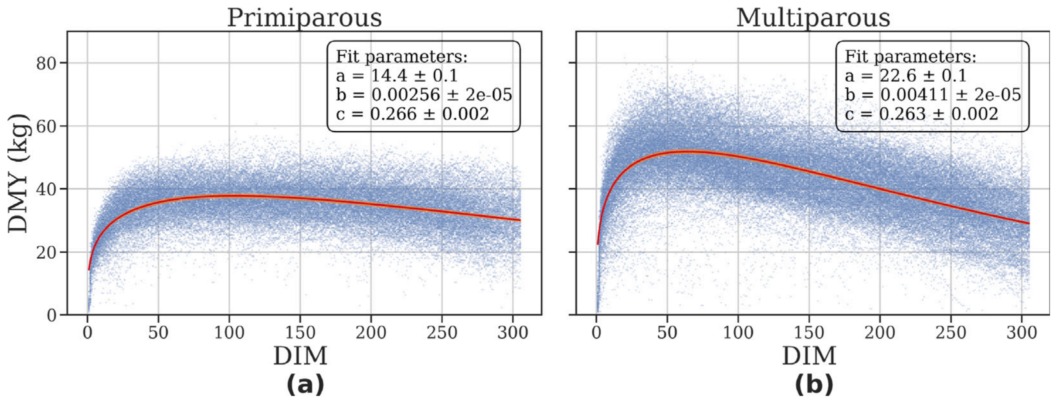


Fig. 4. Best fit curves (red) for **(a)** primiparous and **(b)** multiparous cows. The blue dots represent the milk yield data collected by the AMS of the three farms.

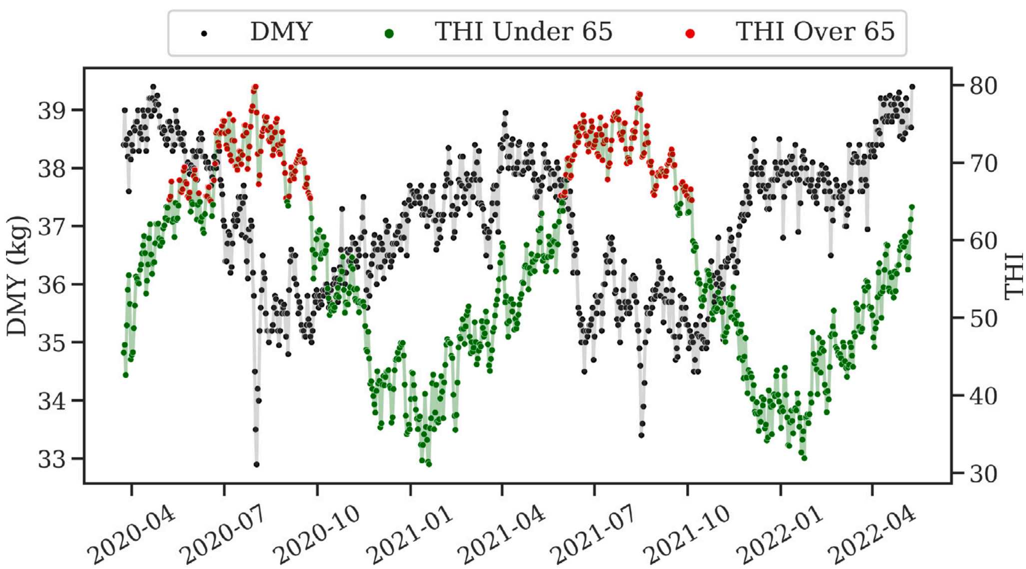


Fig. 5. Trends of the median DMY (black line) and median THI (green and red lines) for the whole dataset.

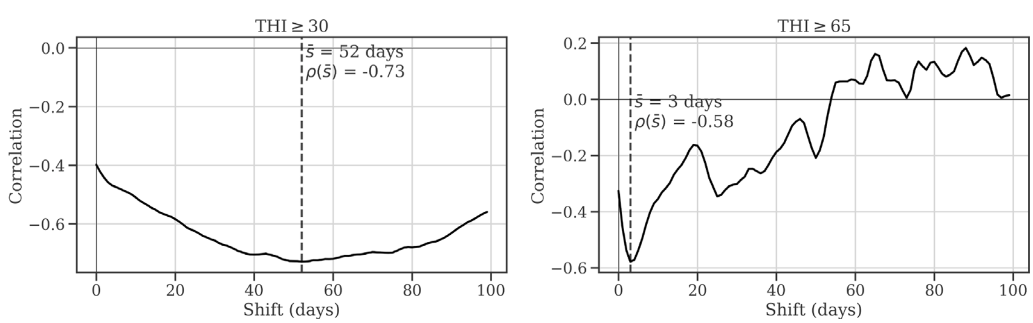


Fig. 6. Cross-correlation performed with data filtered by THI. On the left, the whole dataset is considered ($THI \geq 30$). On the right, only data with $THI \geq 65$ are considered.

$$DLT_2(DIM) = 9.12 + 0.0133 \cdot DIM \tag{10}$$

Finally, the CDI at time t (CDI_t) is obtained as the sum of the positive contributions in a moving window of N days of the terms ($THI_i - 72$) for $i = \{t - N, t - (N - 1), \dots, t - 1\}$. The formal definition is:

$$CDI_t = \sum_{j=t-N}^{i} THI_j - 72 \quad \forall \left(THI_j > 72 \right) \tag{11}$$

In this way, the CDI is a measure of the cumulative heat load in the barn at the day t and resulting as a sum of a defined period of N days. In this work N was assumed equal to 5, in agreement with the results in Section 2.3.

For the indicator-based algorithm, an anomaly (or deviation by a typical trend) is identified, for a single cow and for a specific day (t) measured by the DIM value, if the following three conditions are respected:

- $RYD_t < \alpha$, with $\alpha < 0$ to spot a decrease of the milk yield
- $RLD_t < \beta$, with $\beta < 0$ to spot a decrease of the laying time
- $CDI_t > 0$

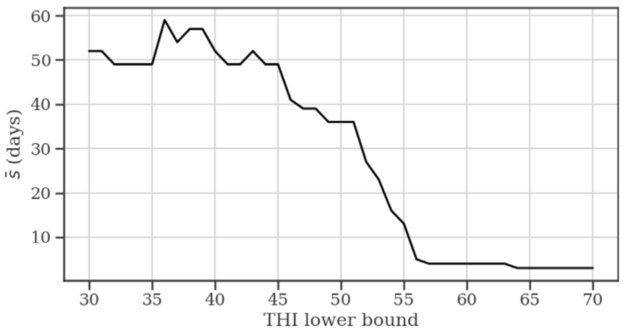


Fig. 7. Trend of the time-lag (shift) corresponding to the minimal correlation as a function of the THI threshold.

The parameters α and β can be assumed or tailored by the user, e.g. farmers, vets, technicians etc., adopting the detection method.

The occurrence of an anomaly for the generic animal is considered as a day in which the animal suffered from heat stress.

2.4.2. Multifit-based algorithm

The second algorithm implemented by the method make use of a multifit procedure (Fischler and Bolles, 1981) that identifies, for the single cow, robust baselines for DMY and DLT as a function of DIM. In fact, the multifit procedure is in general able to identify a more precise solution in a fitting-related problem. Moreover, by the procedure it is possible also to estimate the variance of the two parameters DMY and DLT along the lactation curve. The second algorithm is based on the concept of robust statistics and it uses the same information as the first algorithm. If the daily data do not respect conditions imposed on Multifit Yield Threshold (MYT), Multifit Laying time Threshold (MLT) and CDI, the algorithm recognises the presence of anomaly data.

In order to define the Eq. (2), the three parameters a, b, and c must be defined. In order to consider the Wood function that best fit the available data of a generic lactation of a specific animal, a least square regression procedure can be adopted. A more robust statistics can be obtained by randomly sampling the original dataset and producing a series of different fitting curves, one for each sample. In this way, a collection of Wood curves is obtained as:

$$DMY_k(DIM) = a_k \cdot e^{-b_k \cdot DIM} \cdot DIM^{c_k} \tag{12}$$

where: a_k, b_k, c_k are the parameters of the kth curve. The sampling and fitting process can be repeated N times, selecting each time a fixed fraction f of the original data. In the work the values $N = 500$ and f = 20% were assumed. The obtained group of curves can be used to select a reference curve and to define a measure of scattering of the parameters a, b and c among curve and curve.

In some cases, physically unacceptable curves were obtained, e.g., curves with negative or infinite values or unrealistic trends. For this reason, it was necessary to operate a selection of the meaningful curves characterised by an initial positive derivative. An example of the main results of the multifit process is shown in Fig. 9 for a representative lactation of a generic cow. In the figure, the grey lines represent the Wood curves obtained by the multifit procedure, the black line identify the Wood curve obtained by the fit procedure on the whole dataset represented by the red dots. Among the curves generated by the multifit procedure, the one selected because considered more reliable for the scope is the one that maximises the following quantity:

$$E_{Y,k} = (Y_k - Y)/Y \tag{13}$$

where: $E_{Y,k}$ is the relative difference on the total milk yield, i.e. the milk yield of the lactation duration, Y_k is the total milk yield of the kth curve and Y is the real total milk yield of the lactation as obtained by the sum of the real daily milk yield data recorded by the AMS.

In addition to the reference curve to use in the subsequent detection procedure, using the milk yield values obtained from the various multifit curves, it is possible to calculate the standard deviation σ_{DMY}(DIM) of the DMY, for each DIM of the lactation. In this way the algorithm provides for each DIM a reference value for daily milk yield, named $\overline{DMY}$, and the associated standard deviation, i.e. the couple of values $[\overline{DMY}(DIM); \sigma_{\overline{DMY}}(DIM)]$.

The same method was applied to the daily laying time, i.e. DLT. In this case, the multifit procedure provides the collection of N linear functions in the form:

$$DLT_j(DIM) = d_j + e_j \cdot DIM \tag{14}$$

where: d_j and e_j are the parameters of the jth curve. Among the curves generated by the multifit procedure, the one selected because considered the most reliable for the scope is the one that maximises the following quantity:

$$E_{L,j} = (L_j - L)/L \tag{15}$$

where: $E_{L,j}$ is the relative difference on the total laying time, i.e. the laying time in hours calculated along the lactation duration, L_j is the total laying time of the jth curve and L is the real total laying time during the lactation as obtained by the sum of the real daily laying time data recorded by the pedometers. So, at the end, the algorithm provides for each DIM a reference value for daily laying time, named $\overline{DLT}$, and the associated standard deviation, i.e. the couple of values $[\overline{DLT}(DIM); \sigma_{\overline{DLT}}(DIM)]$.

Fig. 10 shows the application of the multifit-based algorithm to the daily laying time. In the figure, the grey lines represent the trend obtained by the multifit procedure, the black line identify the trend obtained by the fit procedure on the whole dataset represented by the red dots.

For the multifit-based algorithm, an anomaly is identified, for a single cow and for a specific day (t) measured by the DIM value, if the following three conditions are respected:

- $DMY_t < \overline{DMY_t} - \gamma \cdot \sigma_{\overline{DMY},t}$, with $\gamma > 0$
- $DLT_t < \overline{DLT_t} - \delta \cdot \sigma_{\overline{DLT},t}$, with $\delta > 0$
- $CDI_t > 0$

Also in this case, the algorithm parameters γ and δ can be properly assumed by the user and can be personalised for the single cow as a function of the scatter and variability of its trends of milk yield and laying time along the lactation.

Finally, by following the two separate paths it is possible to evaluate, for every DIM, if the generic animal is suffering heat stress condition. This situation correspond to the case in which both the algorithms detected an anomaly for the animal. The two algorithms, based on completely different pipelines, are in this way combined in order to obtain a more robust and stable procedure for the detection of dairy cows affected by heat stress.

3. Results and discussion

3.1. Application of the multifit-based algorithm to three test-bed cases

The multifit procedure adopted by the multifit-based algorithm and presented in the previous section follows a pipeline derived by a fully robust statistics approach. The reader may wonder whether the application of this procedure is robust even in the case of applications to lactation curve data. In fact, to the knowledge of the authors, this is the first work ever that applied this statistical procedure to the daily milk yield data of dairy cows. Thus, in order to confirm the opportunity of applying the multifit procedure also to this type of data, this paragraph reports the main results obtained from the application

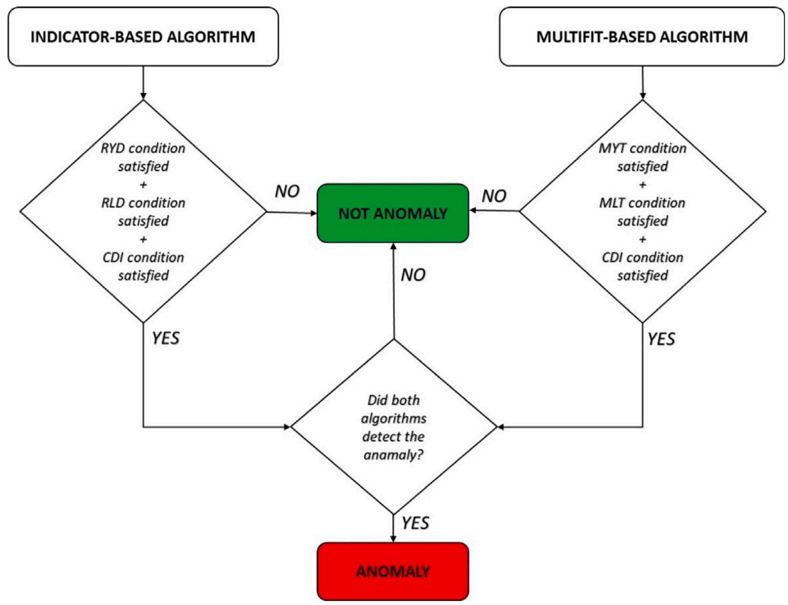


Fig. 8. Flowchart of the DAIRY CHAOS approach.

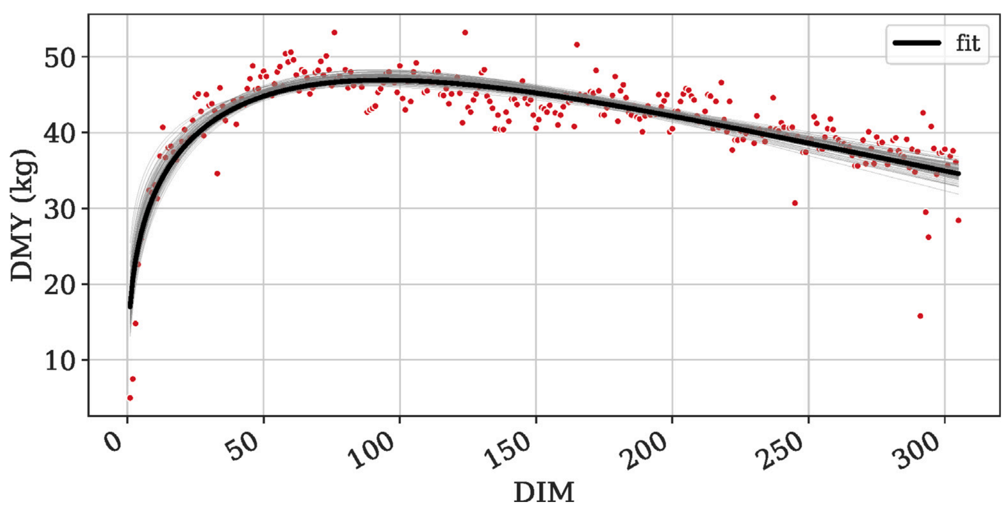


Fig. 9. Example of DMY curves obtained by the multifit procedure (grey curves) and comparison with the curve obtained by a fit procedure (black line) on the whole dataset (red dots) for a representative lactation.

of the multifit procedure to three examples believed significant because representative of the possible real cases that may arise.

(a) Short period anomaly In the first test-bed, the multifit procedure is applied to the case of a lactation curve affected by a production anomaly involving few days. This case could simulate the scenario in which the milk production of an animal has a sudden drop due to a few intense days characterised by high temperature and humidity values that induce evident heat stress effects. The objective of the test is to show how the adoption of a multifit procedure can improve the accuracy of the prediction of the reference curve if compared to a standard fitting procedure. For this, Fig. 11 shows the results of the multifit procedure applied to a Wood's curve with a short period anomaly. The figure shows with the red circles the original data affected by an anomaly of 5 days (between day and on the lactation period of 305 days in total), in grey the curves provided by the multifit procedure with in blue the selected one (obtained following the procedure described in Section 2.4.2) and, for comparison, in green the curve obtained by standard fit procedure on the whole dataset. By assuming the total milk yield provided by the red dots as reference value, the relative error calculated with the Eq. (13) is −0.018 and −0.001 respectively for standard fit and multifit procedure, so practically nullifying the error. The values reported above have been obtained adopting, in the multifit algorithm, the value of 500 as random samplings and a value equal to 20% for the sampling fraction. These values have been selected after a preliminary sensitivity analyses performed on the original dataset by considering for the number of samplings a range going from 50 to 1000 and for the sampling fraction a range going from 10% to 80%. The Fig. 12 reports the results of the sensitivity analyses and shows as the two selected values of 500 and 20% guarantee to obtain the highest value of E_Y for this type of anomaly.

(b) Long period anomaly

The second test-bed, similarly to the previous one, applies the multifit procedure to an ideal Wood's curve affected by a longer period of production anomaly. This case should simulate the scenario in which the milk production of an animal has a less marked drop but which lasts for a longer period due to several days at medium-high temperature and humidity values inducing long exposure to heat stress. For this case, the results are reported in Fig. 13. As before, by

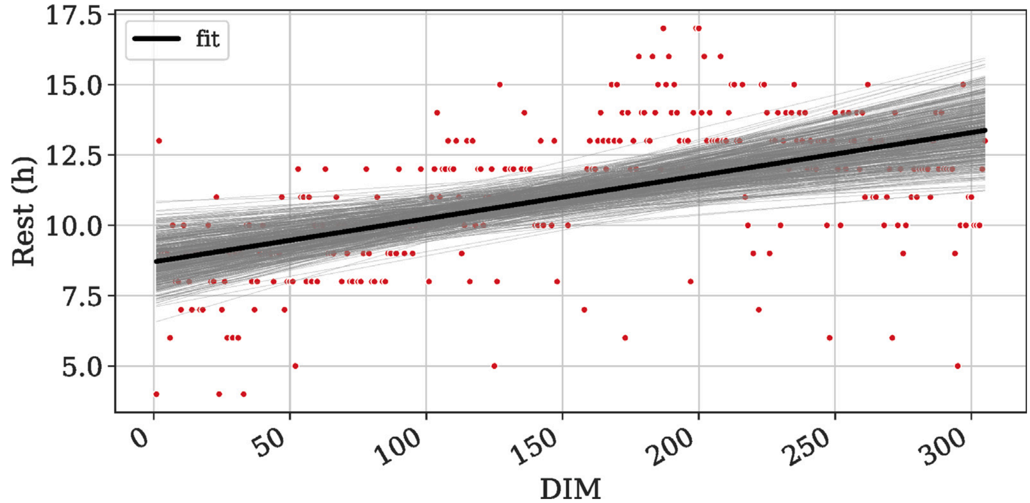


Fig. 10. Example of DLT curves obtained by the multifit procedure (grey lines) and comparison with the line obtained by a fit procedure (black line) on the whole dataset (red dots) for a representative lactation.

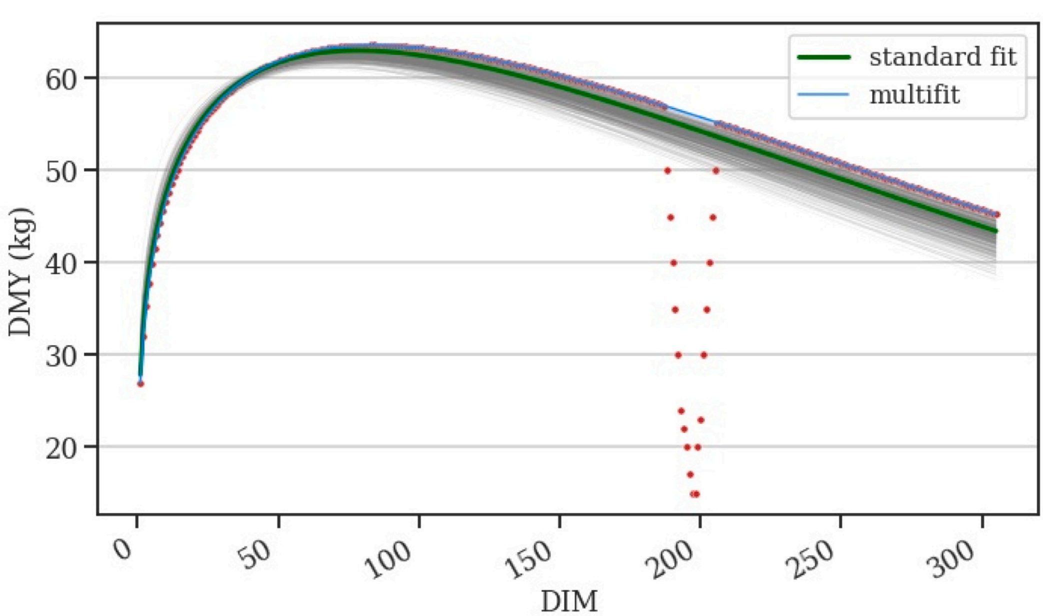
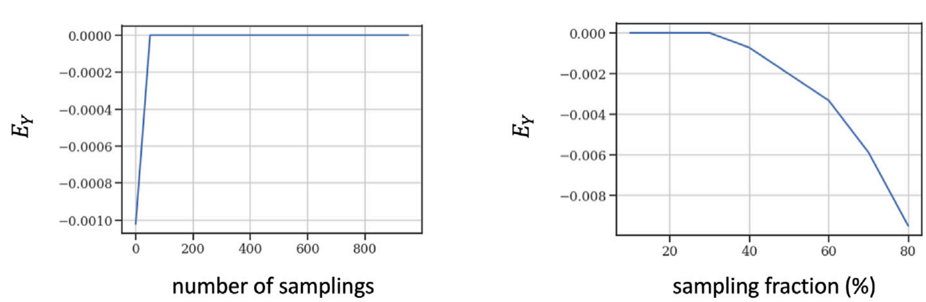


Fig. 11. Results obtained for the test-bed case with short period anomaly. Comparison between curves obtained by the multifit procedure (grey and cyan curves), the curve obtained by a standard fit procedure (green line) on the whole dataset (red dots) for a representative lactation.



(a) (b)

Fig. 12. Sensitivity analyses for short period anomaly case. Trend of E_Y as a function of (a) number of samplings and (b) sampling fraction.

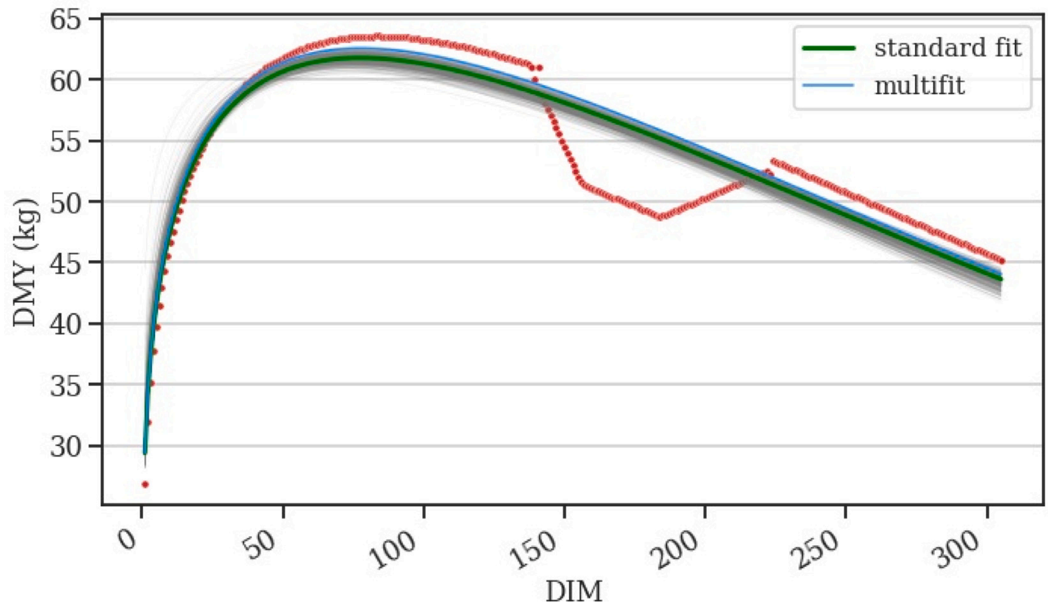


Fig. 13. Results obtained for the test-bed case long period anomaly. Comparison between curves obtained by the multifit procedure (grey and cyan curves), the curve obtained by a standard fit procedure (green line) on the whole dataset (red dots) for a representative lactation.

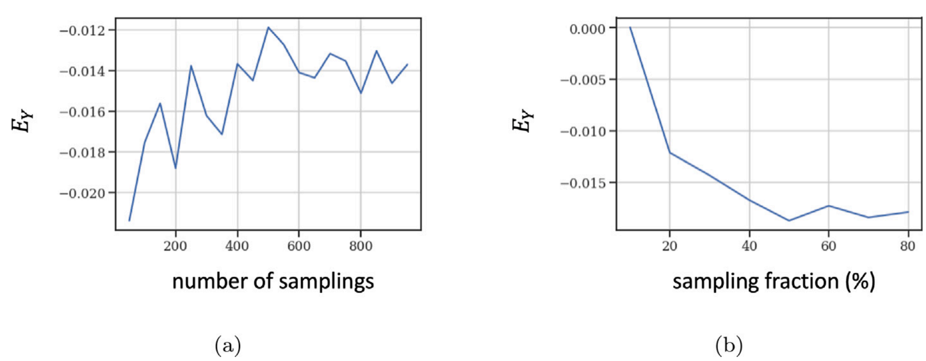


Fig. 14. Sensitivity analyses for long period anomaly case. Trend of E_Y as a function of (a) number of samplings and (b) sampling fraction.

assuming the total milk yield provided by the red dots as reference value, the relative error calculated with the Eq. (13) is −0.029 and −0.015 respectively for standard fit and multifit procedure, so halving the error and confirming also for this case the expected capabilities of the multifit procedure. Also for this case, the values reported above have been obtained adopting, in the multifit algorithm, the value of 500 as random samplings and a value equal to 20% for the sampling fraction. The results of the sensitivity analyses performed on this type of anomaly are summarised in Fig. 14 where it is showed that the two selected values of 500 and 20% guarantee to obtain highest value of E_Y also for this type of anomaly.

(c) Absence of anomaly

The last test-bed case considers an ideal production scenario not affected by anomaly but including the typical daily variations that normally occur during the whole lactation of a dairy cow. The results of this scenario are reported in Fig. 15. In this case multifit and standard fit procedures are expected to return a similar relative error value. And indeed it is, since the two relative errors are equal to −0.016 and +0.018 respectively for standard fit and multifit procedure. The results of this last case is not to be overlooked nor is it of little importance. In fact, a positive, rigorous and complete judgement on the robustness of the multifit procedure requires that it is reliable even in applications where anomalies are not present. Again, the results reported above have been obtained adopting, in the multifit algorithm, the value of 500 as random samplings and a value equal to 20% for the sampling fraction. The results of the sensitivity analyses performed for the case of absence of anomaly are summarised in Fig. 16. The outcomes of the sensitivity analyses shows that the value of E_Y is rather fluctuating and irregular with modification of the number of samplings and a clear trend is not identifiable. On the other hand, the analyses shows a more clear trend in the figure sampling fraction Vs. E_Y where, on the opposite respect the two previous cases, the E_Y value decreases by augmenting the sampling fraction value.

3.2. Application of the DAIRY chaos procedure to real data

In the present section, the application of the DAIRY CHAOS procedure to the real data collected in the three farms described in Section 2.1 is reported. The procedure, adopting the two different algorithms, has the scope of detecting the anomalous days for each milked cow for the monitored period. Then, in order to provide useful indication to the farmers, it is possible to obtain the number of animals that every day present production drop and behaviour deviating by the expected trend, effects that if in presence of high THI values can be attributed to heat stress condition. In Fig. 17 is showed an example of outcomes of the application of the indicator-based algorithm to the milk yield data of a representative cow. In this work α and β have been assumed equal to −0.05. The figure shows as every day the algorithm classifies the data as a normal milk yield data or an anomalous milk

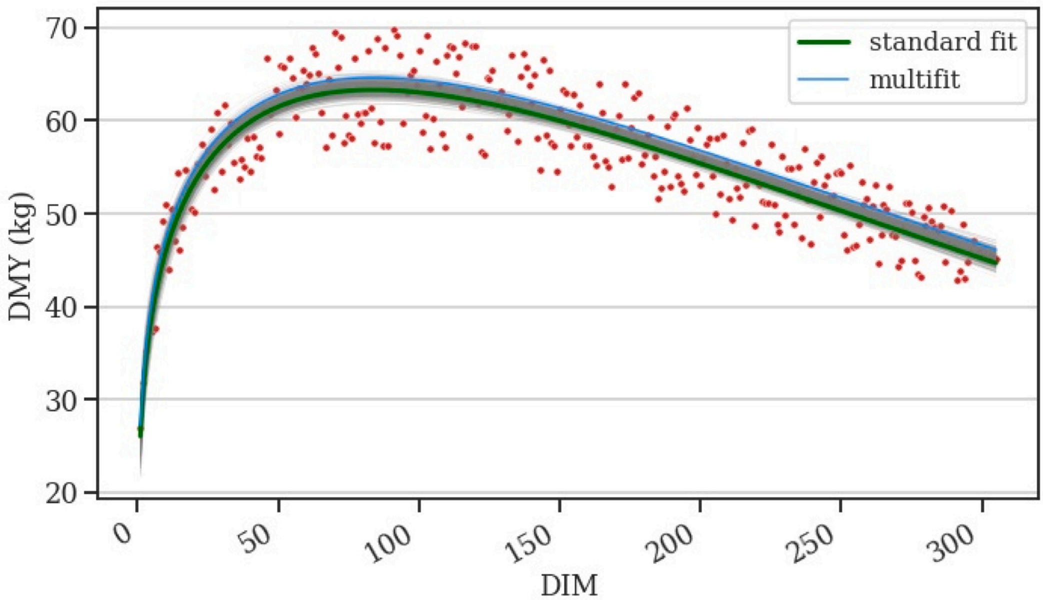


Fig. 15. Results obtained for the test-bed case with absence of anomaly. Comparison between curves obtained by the multifit procedure (grey and cyan curves), the curve obtained by a standard fit procedure (green line) on the whole dataset (red dots) for a representative lactation.

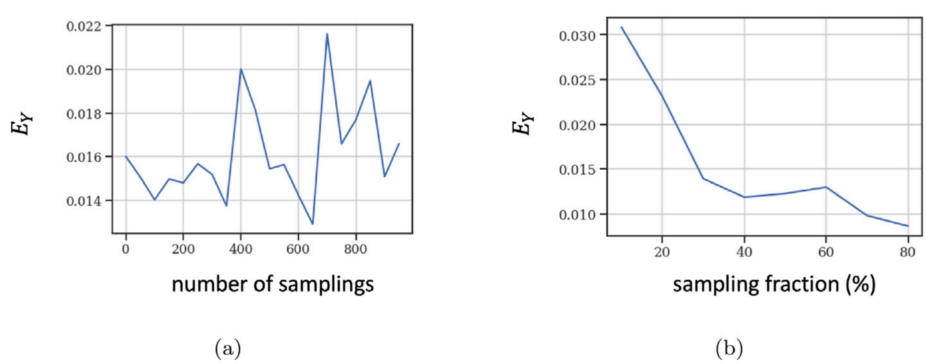


Fig. 16. Sensitivity analyses for the absence of anomaly case. Trend of E_Y as a function of (a) number of samplings and (b) sampling fraction.

yield data. A similar procedure is followed for the DLT of the cow and moreover, the calculation of the CDI is performed. Finally, the suspected anomalous days are identified if the three indicators RYD_t, RLD_t and CDI_t, respect, for the day t, the three conditions described in Section 2.4.1. Then, the DAIRY CHAOS procedure moves to the analysis of the dataset but approaching with the pipeline of the multifit-based algorithm. Fig. 18 shows the outcomes of the anomaly detection process for the multifit-based algorithm on the milk yield data of a representative lactation. The anomalous data are represented by the dark red coloured dots, whereas the normal ones are in light blue. The dashed line, represents the bounding curve and is obtained starting from the boundary condition reported in Section 2.4.2 using a value for γ equal to 3.0. Instead, Fig. 19 shows the identification of the anomalies in the time history of the DLT data of the same representative lactation above cited. Also for this case, the anomalous data are represented by the dark red coloured dots, whereas the normal ones are in light blue. For the definition of the bounding curve (see the grey dashed line) the value of δ equal to 3.0 has been adopted. After, the calculation of the daily CDI values, the anomalous days are identified when the three indicators DMY_t, DLT_t and CDI_t, respect, for the day t, the three conditions described in Section 2.4.2.

Finally, after that the two algorithms have been applied separately to the daily data, the last step of the DAIRY CHAOS procedure crosses the two judgements on the single cow and if both algorithms have labelled the data of the generic day of the cow as anomalous, the day is identified as a day on which the animal suffered heat stress. This last step is reported in Fig. 20. The figure includes for the whole monitored period (from March 2020 to May 2022) the daily average THI values, the percentage of animals with anomalous data as resulting from the two algorithms applied separately (see green line for the indicator-based algorithm assessment and the orange line for the multifit-based algorithm) and moreover the percentage of cows identified as suffering heat stress when both the algorithms labelled the data of a cow as anomalous for the particular day (see blue line in the same figure). As can be seen in the figure, as expected, in correspondence of the periods with the highest THI values, both the algorithms anomalous data, with number of anomalies having peak values around 25% for both the algorithms. Then, only a restricted percentage of these cows are simultaneously detected by both algorithms as anomalous. So, the trend of the percentage of animals in heat stress has peak values ranging around 10% in the hottest days of the two years. Obviously, data on the cows labelled as suffering heat stress can be used as useful information for the daily management of the herd by the farmers. In fact, in this way, the farmer can take the most effective decisions in order to ensure the welfare and health of the animals by implementing targeted, timely and customisable actions up to the scale of the individual animal. For example, the farmer can operate regrouping the animals most sensitive to heat stress and treat them with more intense soaking and ventilation treatments. At the same time, this can make it possible to minimise production losses with the lowest resource needs, thus increasing the

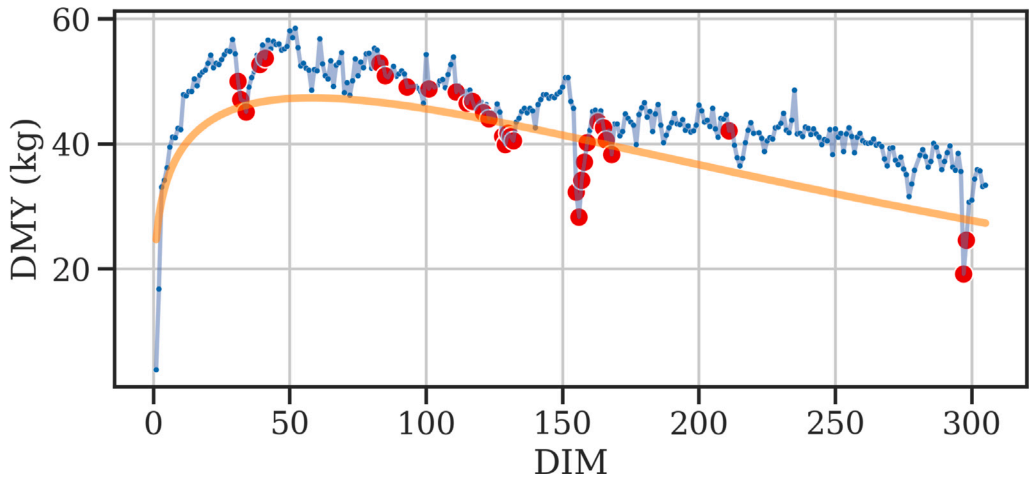


Fig. 17. Example of application of the indicator-based algorithm to the milk yield data of a representative cow. In blue the recorded DMY data are reported. The orange line represents the ideal trend whereas the red dots represent the day of production classified as anomalies by the algorithm.

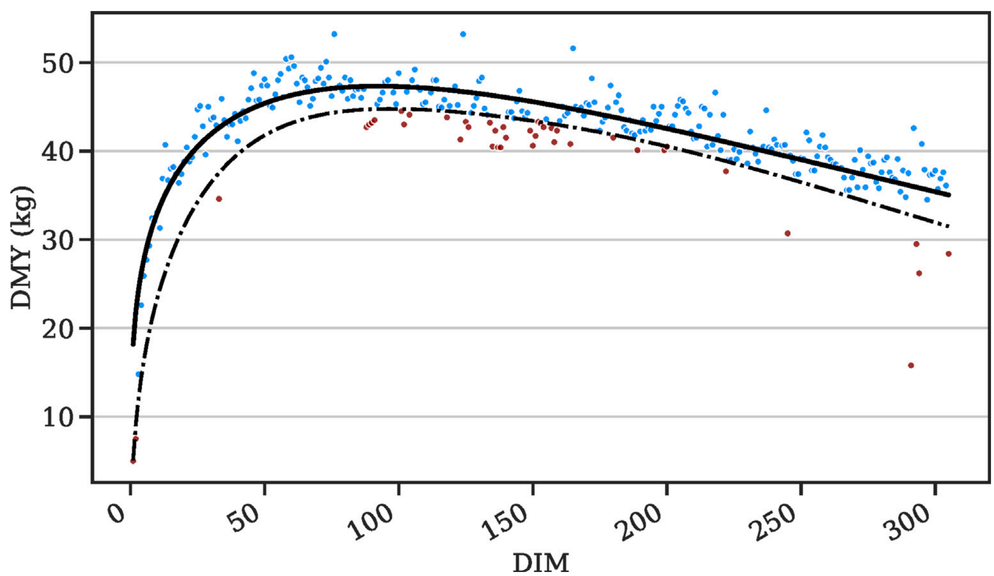


Fig. 18. Identification of the DMY anomalies with the multifit-based algorithm.

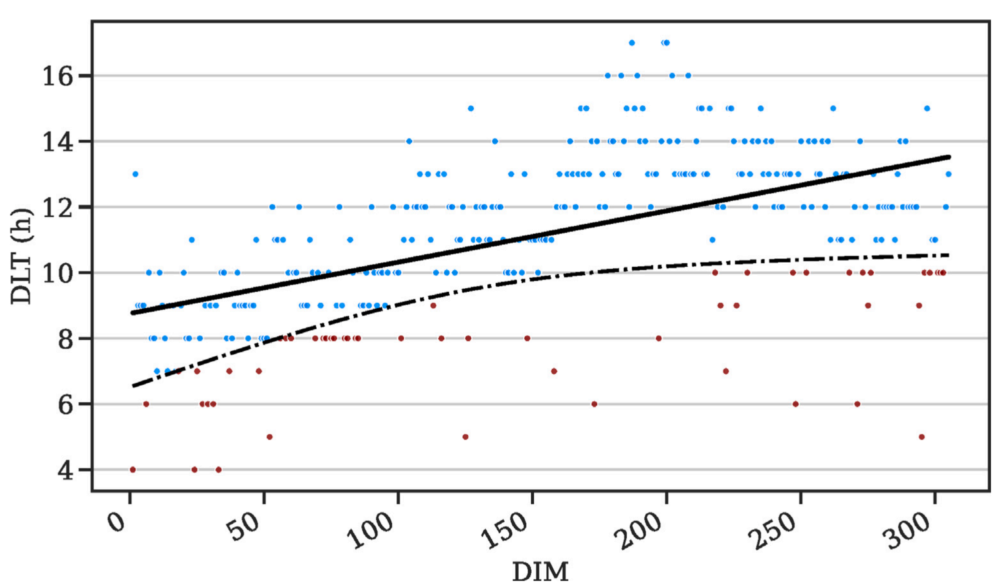


Fig. 19. Identification of the DLT anomalies with the multifit-based algorithm.

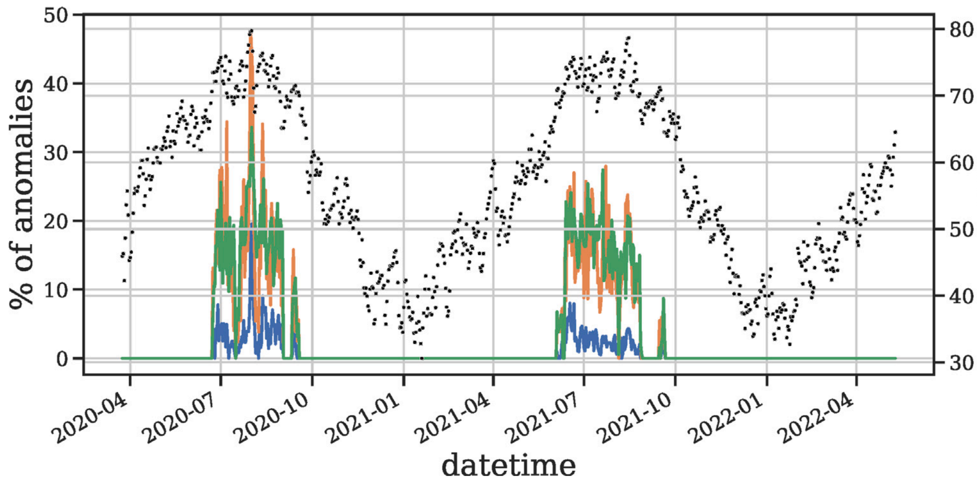


Fig. 20. Final response of the DAIRY CHAOS procedure.

profitability and the sustainability of the company. The present method could be easily integrated in a decision-support system since it needs of few data classes usually available in real time to the farmer.

4. Conclusions

In this paper a data driven approach for the identification of cows affected by heat stress is presented. The procedure, called DAIRY CHAOS, is based on two numerical algorithms, that separately detect daily anomalies that can be attribute to heat stress. Each one of the two algorithms take into consideration three indicators, two respectively related to milk yield and lying time and one based on climatic data in the barn. Following the procedure, an anomaly in the daily indicators identified by both the algorithms for a specific cow, labels the cow as suffering, for that specific day, heat stress. The originality of the approach mainly consists in the cross-use of two numerical strategies that start by different statistical approaches, are based on completely different pipelines but provide results that are then combined in order to obtain a more robust and stable procedure for the heat stress detection and at the level of single animal. Thanks to the available information, the farmers can take timely the most effective actions for example moving the animals most sensitive to heat stress in the most cool zones of the barn, treating them with more long soaking cycles and activating for them more intense ventilation treatments. The present data driven approach can be easily integrated in a decision-support system since it needs of few data classes usually available in real time to the farmer. The proposed approach represents one of the first attempt in this field and of consequence requires further studies, validation and application steps. These following steps will be mainly devoted to the calibration of the operating coefficients of the algorithms by comparison with recognised and already validated gold standard features or indicators of heat stress condition for the single animal.

Funding

Part of this study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

CRediT authorship contribution statement

Marco Bovo: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Mattia Ceccarelli:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Miki Agrusti:** Formal analysis, Software, Supervision, Validation, Writing – original draft. **Daniele Torreggiani:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing. **Patrizia Tassinari:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

All authors approved the version of the manuscript to be published.

References

Benni, S., Pastell, M., Bonora, F., Tassinari, P., Torreggiani, D., 2020. A generalised additive model to characterise dairy cows' responses to heat stress. Animal 14 (2), 418–424. http://dx.doi.org/10.1017/S1751731119001721.

Bernabucci, U., Biffani, S., Buggiotti, L., Vitali, A., Lacetera, N., Nardone, A., 2014. The effects of heat stress in Italian holstein dairy cattle. J. Dairy Sci. 97, 471–486. http://dx.doi.org/10.3168/JDS.2013-6611.

Bonora, F., Pastell, M., Benni, S., Tassinari, P., Torreggiani, D., 2018. ICT monitoring and mathematical modelling of dairy cows performances in hot climate conditions: a study case in po valley (Italy). Agric. Eng. Int.: CIGR J. 20, 1–12.

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., Tassinari, P., 2021. Random forest modelling of milk yield of dairy cows under heat stress conditions. Animals 11 (5), 1305. http://dx.doi.org/10.3390/ANI11051305, URL: https://www.mdpi.com/2076-2615/11/5/1305/htmhttps://www.mdpi.com/2076-2615/11/5/1305.

Bovo, M., Santolini, E., Barbaresi, A., Tassinari, P., Torreggiani, D., 2022. Assessment of geometrical and seasonal effects on the natural ventilation of a pig barn using CFD simulations. Comput. Electron. Agric. 193, http://dx.doi.org/10.1016/j.compag.2021.106652, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122006577&doi=10.1016%2fj.compag.2021.106652&partnerID=40&md5=380cc280625d8a4ccff63a3c1e5b0340.

Burhans, W.S., Burhans, C.A.R., Baumgard, L.H., 2022. Invited review: Lethal heat stress: The putative pathophysiology of a deadly disorder in dairy cattle. J. Dairy Sci. 105, 3716–3735. http://dx.doi.org/10.3168/JDS.2021-21080.

Chamberlain, A., Powell, C., Arcier, E., Aldenhoven, N., 2022. The relationship between on-farm environmental conditions inside and outside cow sheds during the summer in England: can temperature humidity index be predicted from outside conditions? Animal - Open Space 1, 100019. http://dx.doi.org/10.1016/J.ANOPES.2022.100019.

Charlton, G.L., Rutter, S.M., East, M., Sinclair, L.A., 2013. The motivation of dairy cows for access to pasture. J. Dairy Sci. 96, 4387–4396. http://dx.doi.org/10.3168/JDS.2012-6421.

Cowley, F.C., Barber, D.G., Houlihan, A.V., Poppi, D.P., 2015. Immediate and residual effects of heat stress and restricted intake on milk protein and casein composition and energy metabolism. J. Dairy Sci. 98 (4), 2356–2368. http://dx.doi.org/10.3168/jds.2014-8442.

Ekine-Dzivenu, C.C., Mrode, R., Oyieng, E., Komwihangilo, D., Lyatuu, E., Msuta, G., Ojango, J.M., Okeyo, A.M., 2020. Evaluating the impact of heat stress as measured by temperature-humidity index (THI) on test-day milk yield of small holder dairy cattle in a sub-sahara african climate. Livest. Sci. 242, 104314. http://dx.doi.org/10.1016/J.LIVSCI.2020.104314.

Elsevier, 2023. Scopus. www.scopus.com, Accessed: 2023-03-03.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24 (6), http://dx.doi.org/10.1145/358669.358692.

Herbut, P., Angrecka, S., 2017. Relationship between THI level and dairy cows' behaviour during summer period. 17, 226–233 http://dx.doi.org/10.1080/1828051X.2017.1333892 URL: https://www.tandfonline.com/doi/abs/10.1080/1828051X.2017.1333892.

Hut, P., Scheurwater, J., Nielen, M., van den Broek, J., Hostens, M., 2022. Heat stress in a temperate climate leads to adapted sensor-based behavioral patterns of dairy cows. J. Dairy Sci. 105 (8), 6909–6922. http://dx.doi.org/10.3168/jds.2021-21756, URL: https://www.sciencedirect.com/science/article/pii/S0022030222003605.

Ji, B., Banhazi, T., Perano, K., Ghahramani, A., Bowtell, L., Wang, C., Li, B., 2020. A review of measuring, assessing and mitigating heat stress in dairy cattle. Biosyst. Eng. 199, 4–26. http://dx.doi.org/10.1016/J.BIOSYSTEMSENG.2020.07.009.

Kadzere, C.T., Murphy, M.R., Silanikove, N., Maltz, E., 2002. Heat stress in lactating dairy cows: a review. Livest. Prod. Sci. 77, 59–91. http://dx.doi.org/10.1016/S0301-6226(01)00330-X.

Kelly, C.F., Bond, T.E., 1971. Bioclimatic Factors and Their Measurement: A Guide to Environmental Research on Animals. National Academic Press, Washington DC.

Lees, J.C., Lees, A.M., Gaughan, J.B., Lees, J.C., Lees, A.M., Gaughan, J.B., 2018. Developing a heat load index for lactating dairy cows. Animal Prod. Sci. 58, 1387–1391. http://dx.doi.org/10.1071/AN17776, URL: https://www.publish.csiro.au/an/AN17776.

M'Hamdi, N., Darej, C., Attia, K., Znaidi, I.E.A., Khattab, R., Djelailia, H., Bouraoui, R., Taboubi, R., Marzouki, L., Ayadi, M., 2021. Modelling THI effects on milk production and lactation curve parameters of holstein dairy cows. J. Therm. Biol. 99, http://dx.doi.org/10.1016/J.JTHERBIO.2021.102917.

Moore, S.S., Costa, A., Penasa, M., Callegaro, S., Marchi, M.D., 2023. How heat stress conditions affect milk yield, composition, and price in Italian holstein herds. J. Dairy Sci. 106, http://dx.doi.org/10.3168/JDS.2022-22640, URL: https://pubmed.ncbi.nlm.nih.gov/37080787/.

Neves, S.F., Silva, M.C., Miranda, J.M., Stilwell, G., Cortez, P.P., 2022. Predictive models of dairy cow thermal state: A review from a technological perspective. Vet. Sci. 9, 416. http://dx.doi.org/10.3390/VETSCI9080416, URL: https://www.mdpi.com/2306-7381/9/8/416/htmhttps://www.mdpi.com/2306-7381/9/8/416.

Newville, M., Stensitzki, T., Allen, D.B., Ingargiola, A., 2014. LMFIT: Non-linear least-square minimization and curve-fitting for python. http://dx.doi.org/10.5281/zenodo.11813, URL: https://doi.org/10.5281/zenodo.11813.

Silvestre, A.M., Petim-Batista, F., Colaço, 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, 1813–1821. http://dx.doi.org/10.3168/jds.S0022-0302(06)72250-0, URL: http://www.journalofdairyscience.org/article/S0022030206722500/fulltexthttp://www.journalofdairyscience.org/article/S0022030206722500/abstracthttps://www.journalofdairyscience.org/article/S0022-0302(06)72250-0/abstract.

Strpić, K., Barbaresi, A., Tinti, F., Bovo, M., Benni, S., Torreggiani, D., Macini, P., Tassinari, P., 2020. Application of ground heat exchangers in cow barns to enhance milk cooling and water heating and storage. Energy Build. 224, 110213. http://dx.doi.org/10.1016/j.enbuild.2020.110213, URL: https://www.sciencedirect.com/science/article/pii/S0378778819332189.

Tao, S., Orellana, R.M., Weng, X., Marins, T.N., Dahl, G.E., Bernard, J.K., 2018. Symposium review: The influences of heat stress on bovine mammary gland function. J. Dairy Sci. 101, 5642–5654. http://dx.doi.org/10.3168/JDS.2017-13727, URL: https://pubmed.ncbi.nlm.nih.gov/29331468/.

Thornton, P., Nelson, G., Mayberry, D., Herrero, M., 2022. Impacts of heat stress on global cattle production during the 21st century: a modelling study. Lancet Planet. Health 6, e192–e201. http://dx.doi.org/10.1016/S2542-5196(22)00002-X.

West, J.W., 2003. Effects of heat-stress on production in dairy cattle. J. Dairy Sci. 86, 2131–2144. http://dx.doi.org/10.3168/JDS.S0022-0302(03)73803-X.

Wildridge, A.M., Thomson, P.C., Garcia, S.C., John, A.J., Jongman, E.C., Clark, C.E., Kerrisk, K.L., 2018. Short communication: The effect of temperature-humidity index on milk yield and milking frequency of dairy cows in pasture-based automatic milking systems. J. Dairy Sci. 101, 4479–4482. http://dx.doi.org/10.3168/jds.2017-13867, URL: http://www.journalofdairyscience.org/article/S0022030218301802/fulltexthttp://www.journalofdairyscience.org/article/S0022030218301802/abstracthttps://www.journalofdairyscience.org/article/S0022-0302(18)30180-2/abstract.

Wood, P.D., 1967. Algebraic model of the lactation curve in cattle. Nature 216, 164–165. http://dx.doi.org/10.1038/216164a0, URL: https://www.nature.com/articles/216164a0.