frontiers | Frontiers in Animal Science

TYPE Methods
PUBLISHED 16 December 2024
DOI 10.3389/fanim.2024.1444948

OPEN ACCESS

EDITED BY
Robert John Tempelman,
Michigan State University, United States

REVIEWED BY
Severiano Silva,
Universidade de Trás-os-Montes e Alto, Portugal
Svenja Woudstra,
University of Copenhagen, Denmark

*CORRESPONDENCE
Katharina Schodl
schodl@zuchtdata.at

RECEIVED 06 June 2024
ACCEPTED 27 November 2024
PUBLISHED 16 December 2024

CITATION
Schodl K, Stygar A, Steininger F and Egger-Danner C (2024) Sensor data cleaning for applications in dairy herd management and breeding.
Front. Anim. Sci. 5:1444948.
doi: 10.3389/fanim.2024.1444948

Sensor data cleaning for applications in dairy herd management and breeding

Katharina Schodl[1]*, Anna Stygar[2], Franz Steininger[1] and Christa Egger-Danner[1] on behalf of the D4Dairy-Consortium

[1]ZuchtData EDV-Dienstleistungen GmbH, Vienna, Austria, [2]Bioeconomy and Environment, Natural Resources Institute Finland (Luke), Helsinki, Finland

Data cleaning is a core process when it comes to using data from dairy sensor technologies. This article presents guidelines for sensor data cleaning with a specific focus on dairy herd management and breeding applications. Prior to any data cleaning steps, context and purpose of the data use must be considered. Recommendations for data cleaning are provided in five distinct steps: 1) validate the data merging process, 2) get to know the data, 3) check completeness of the data, 4) evaluate the plausibility of sensor measures and detect outliers, and 5) check for technology related noise. Whenever necessary, the recommendations are supported by examples of different sensor types (bolus, accelerometer) collected in an international project (D4Dairy) or supported by relevant literature. To ensure quality and reproducibility, data users are required to document their approach throughout the process. The target group for these guidelines are professionals involved in the process of collecting, managing, and analyzing sensor data from dairy herds. Providing guidelines for data cleaning could help to ensure that the data used for analysis is accurate, consistent, and reliable, ultimately leading to more informed management decisions and better breeding outcomes for dairy herds.

KEYWORDS

sensor data, dairy cow, data cleaning, activity sensor, rumination sensor

1 Introduction

Advancing technologies on dairy farms have unlocked the potential for data-driven decision support systems to improve herd management. Sensor technologies in automatic milking systems (AMS) or on wearable devices such as collars, ear tags or rumen boluses, allow the (continuous) recording of milk properties (e.g. milk amount, milk contents, electrical conductivity, etc.) or physiological and behavioral variables (e.g. rumination, activity, temperature, etc.). Based on changes in behavioral patterns or physiological parameters farmers should be alerted to animals in need of treatment or management

procedures. Many manufacturers already offer herd management products that use these technologies to detect cows in heat, cows about to calve, or cows that may need attention due to health problems (Caja et al., 2016; Mayo et al., 2019). However, the results of these predictions and alerts also rely on additional data to reach their full potential. This implies that farmers enter data such as calving or insemination dates into the system and the sensor device must be linked to the individual animal by a unique ID to ensure correct data assignment. In addition to the benefits of sensor technology to its target customer - the farmer - the data from these devices also offers a great opportunity to be used for research or other applications such as phenotyping for routine genetic evaluation or the development of new decision support tools for farmers. Dairy herd improvement (DHI) and breeding organizations have built up extensive databases over the past decades containing data on management practices, health records, conformation, etc. Integration of these different data sources should be the key to unlocking the full potential of precision livestock farming (PLF) technologies and providing farmers with decision support tools not only for specific management aspects, but also for herd health management and breeding decisions.

Data availability in the dairy sector is increasing as the number of farms with sensor or AMS technology continues to grow (e.g. Steeneveld et al., 2015). As sensor data have the potential to be used to provide information on regulatory compliance (Stygar et al., 2022) and consumer desired welfare attributes (Stygar et al., 2023) any failure in data cleaning may have not only technical but also legal consequences (Stöger et al., 2021). Therefore, a properly designed, implemented, and documented data cleaning procedure is becoming a necessity in the dairy industry. Guidelines for data cleaning in bioinformatics and health informatics (e.g. Van den Broeck et al., 2005; Chicco et al., 2022) and data cleaning protocols for the use of various data for routine genetic evaluation and tools provided by DHI and breeding organizations (see e.g. ICAR, 2022a) are well established. However, to the best of the authors' knowledge, there are no recommendations for data cleaning in dairy herd management and breeding applications when it comes to high resolution data from commercial sensor technologies. When discussing sensor data in the context of this paper, we are specifically referring to information collected by wearable sensor devices attached to individual animals within one or more farms or data coming from AMS. These devices are supplied and maintained by different companies and data sharing may be facilitated via interfaces or other means. Generally, the data obtained from sensors contain a lot of noise and usually have undergone prior processing to transform raw sensor measurements into variables such as activity levels or specific behaviors such as ruminating or eating (Schmeling et al., 2021). These variables are derived using specialized algorithms developed by the respective company and are not disclosed for intellectual property reasons. Consequently, the recipient of the data remains unaware of the specific procedures involved in data cleaning, imputation, pre-processing, and analysis, as these details are undisclosed by the data provider.

Before any further use, sensor data must be carefully inspected and evaluated to ensure that results are not biased by erroneous data due to sensor malfunction or measurement error Teh et al (2020) have delineated various types of errors inherent in sensor data that require attention in the context of data quality assurance. These include outliers, missing data, bias, drift, noise, constant value, uncertainty, and stuck-at-zero conditions. Given that users are typically not provided any details on the methodologies employed by the data provider to address these quality issues, it becomes imperative for the data user to seek relevant information from the provider. Consequently, data users must scrutinize the data for potential sources of error upon receipt, considering the aforementioned types of data anomalies outlined by (Teh et al., 2020). In addition to information on the cleaning and pre-processing that has taken place prior to data delivery, some use cases may require knowledge of whether and how the data from these sensors has been validated. Often, sensor systems available on the market are not validated at all or are validated only for one or a few very specific purposes (Knight, 2020; Stygar et al., 2021). Data processing and software algorithms may emphasize some behaviors more than others to generate the most reliable alarms, for example emphasizing mounting activity for heat detection (Elischer et al., 2013).

Against this background, this paper aims to address what needs to be considered when using sensor data from commercially available sensor systems. Starting with general aspects of data cleaning, it will be discussed how they apply to data from sensor systems and their specialties. Focus will be put on the discussion of sensor data cleaning for use in dairy herd management and breeding using examples from projects exploiting the potential of advanced data analysis for the improvement of animal health and welfare, nutrition, and product quality. Examples will be described from projects and studies using sensor data from different manufacturers in the course of the D4Dairy project (https://d4dairy.com). Based on the experience gained in these projects, general recommendations and conclusions will be drawn, which in turn will serve as a basis for guidelines for the use of sensor data developed by a joint initiative of ICAR (International Committee of Animal Recording) and IDF (International Dairy Federation).

2 Context and purpose

Decisions for data cleaning and processing depend very much on the purpose of data use and type of analysis. For example, requirements for applications related to herd management differ from those for breeding. If data are fed into a detection or classification algorithm for herd management, complete time series data are more important than it may be the case for some breeding purposes. Missing observations may result in a false or no alarm for a particular time slot (e.g. Stygar et al., 2023). In addition, most herd management applications are run in real time (e.g. Rustas et al., 2024), unlike breeding evaluation, which is performed at a specific point in time using appropriate types of data (e.g. Egger-Danner et al., 2012). Therefore, data cleaning for herd management applications may benefit more from automated approaches based on for instance principal component analysis, artificial neural networks or Bayesian networks (Teh et al., 2020).

Traits in livestock breeding must have a reasonably large genetic variability, be heritable and clearly defined (Shook, 1989). If traits can only be measured at high costs or have very low heritabilities, it may be necessary to define auxiliary traits with high genetic correlation to the target trait (Shook, 1989). While this approach makes perfect sense as a theoretical concept, the lack of a valid gold standard often makes it challenging to implement. Target traits may be too complex to be defined by a single value (e.g. resilience as described in Friggens et al., 2022), or there may be a variety of gold standards, making comparisons and trait development difficult [e.g. mastitis (Jensen et al., 2019)]. As the choice of the gold standard is crucial for the development of valid auxiliary traits, these must be chosen with caution and limitations have to be considered and disclosed. Due to the general lack in validation of sensor variables, i.e. if minutes spent ruminating as identified by a sensor system equal the actual duration of rumination of the cow, sensor data may for now be mainly considered for developing proxies for traits with low frequencies or heritabilities, such as functional traits. Furthermore, data from different sensor systems using different measuring devices and algorithms should be analyzed separately for as long as sensor measures are not objectively validated for measuring specific behaviors. Several studies demonstrated a lack of common trends across herds for predicting dairy cattle resilience (Adriaens et al., 2020), welfare (Stygar et al., 2023) and dry matter intake (Yilmaz Adkinson et al., 2024) when using sensor data. The prediction accuracy could be improved by the development of industry-wide standards and guidelines for sensor validation, which is an ongoing joint activity of ICAR and IDF (Egger-Danner et al., 2024).

Although sensor data are available continuously throughout the lactation or even the whole (productive) lifetime of a cow, not all data is necessary for trait definition and parameter estimation. In contrast, it may even be a problem regarding storage and computation time (Koltes et al., 2019). Carlström et al. (2013) investigated heritabilities of AMS-derived milkability traits during the whole lactation and concluded that standard errors of the heritability estimates did not increase when instead of the 330 only the first 100 observations per cow were used for analysis. Thus, for breeding purposes the importance of including a sufficiently large number of animals and covering the main factors of variation, as well as comprehensive pedigree information and precise definition of traits outweigh the amount of available data per animal as long as the genetic correlation with the desired (functional) trait is high enough.

3 Five steps to successful sensor data cleaning

As soon as purpose and context of sensor data use are clarified and made explicit, the actual data cleaning can be started. We identified six steps crucial for successful cleaning of sensor data, which are shown in the flow chart in Figure 1 and described in detail in the following.

3.1 Step 1: validate data merging process

Cows are officially registered with a unique official identification number on their ear tag, which should be used when merging data from multiple sources. Ideally, sensor data is already permanently assigned to this unique animal identifier. However, sometimes data are assigned to the ID of a sensor device and the assignment of the device to an animal ID is contained in a separate file. In this case, correct merging of animal ID data and sensor data must be performed by the data user. In doing so, the user must check for the possibility of more than one device being assigned to an animal. In case a sensor was broken or ran out of battery, an animal probably had to be equipped with a new device leading to multiple sensor devices sequentially recording data of the same animal. Vice versa, multiple animals may be equipped with the same device if farms own fewer devices than cows and they use the sensor only during specific periods in a cow's lactation for example for heat or calving detection. Obviously, the latter will not be the case for sensor systems using a rumen bolus. Thus, post-merging validation should be performed by running duplicate checks on sensor and animal ID assignments and obtaining information on the structure of data sets and handling of devices from the sensor company. Particularities may become apparent by employing dedicated algorithms to identify for example conspicuously short calving intervals or constant detection of heat events or other inconsistencies depending on the kind sensor information.

Furthermore, if sensors are attached for the first time to an animal the algorithm takes some time to learn before it works

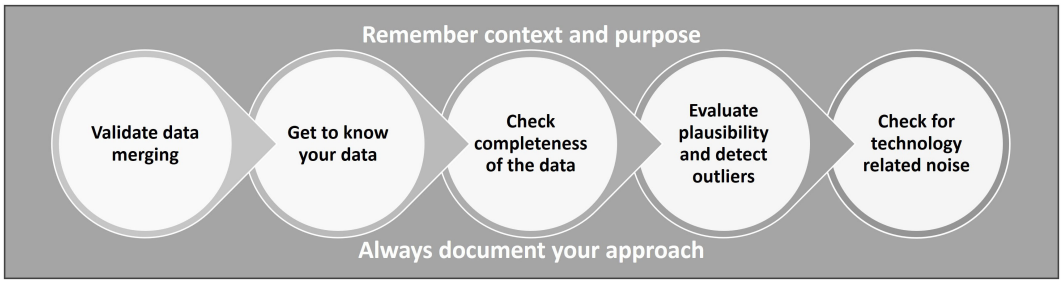

FIGURE 1
Flow chart of the sensor data cleaning process.

properly and sends meaningful alarms. These time windows are usually specified by the manufacturer and should not be used for data analysis. Other potential sources of error are differences in time zones or missing corrections for the daylight-saving time including winter and summer time. Sometimes variables are calculated from moving averages and thus, may be missing in the beginning and even be lagged compared to the other output variables. All these aspects are crucial to consider for integration with other data and should be clarified with the sensor company.

Table 1 provides a list of information that is important to request from sensor companies or data providers for a better understanding of the data and aspects that require special attention during the data cleaning process. It should be noted, however, that the provision of this information depends on the sensor companies' willingness to share their data and the type of data provided.

3.2 Step 2: get to know your data

Before starting with any further data cleaning steps, the nature of the data obtained must be described and any uncertainties about type and unit eliminated. For a better understanding of the data, it should be visualized. Whereas raw sensor data resembles sampling rate, usually Hz, the data which can be retrieved from interfaces is already processed by (proprietary) algorithms. Depending on the type of output, which is generated from the raw sensor signals, it can be expressed as durations (e.g. minutes of ruminating or time spent lying during a defined period), counts (e.g. number of drink cycles or rumination bouts) or as an index without any further specifications (e.g. activity levels). Thus, a proper description of the nature (including units) and the type (raw vs processed) of data should be asked from the sensor company together with the data (Bouchon et al., 2019). Knowing how values are expressed and what units are used is important for time series data, particularly regarding further aggregation steps. Rumination time, for example, may be expressed as minutes of rumination per hour or within the last 24 hours but still at an hourly rate. Furthermore, it is important to know if hourly measurements with a certain time stamp refer to values for the preceding or successive hour. Another important aspect is whether the data is a single shot measurement, such as a reticular temperature measurement at one point in time, or whether it is summarized, averaged, filtered, or otherwise processed by an algorithm, which can introduce a delay (Bouchon et al., 2019). One example of the latter is the activity level as provided in some of the sensor data we analyzed, which was based on the summary of the accelerometer data over the period of sampling intervals. Whereas time series data follows the pattern of the data frequency, events or alarms may be assigned to an exact timestamp within the time series intervals. Depending on the further use and aggregation of the data, different strategies have to be chosen for handling those differences in a timely resolution, i.e. how are events or alarms aligned with time series data or aggregated versions thereof. Plots such as scatter plots (e.g. Figure 2) or histograms containing statistical summaries or individual time series plots are recommended to gain valuable information about the nature of the data including distribution types, patterns, gaps, outliers or other problems in the data (Unwin, 2020).

3.3 Step 3: check completeness of data

When sampling and data rate, type, and unit of data are clarified, the next step is to check for missing data as well as duplicates. Completeness of data is easy to assess in time series data where data frequency is known, whereas it may not be that straightforward for datasets containing only alarms or other information that does not follow an expected frequency or pattern. In case of missing data, it is recommended to consult with the data provider for underlying reasons and recommendations on how to handle them. Reasons for missing data comprise data transmission problems due to low battery status of the sensor or bad connectivity, power failures on the farm or other technical malfunctioning (Borchers et al., 2017; Ren et al., 2021).

TABLE 1 Key information to obtain from sensor data providers.

Data cleaning step	Required information
Step 1: Validate data merging process	• Structure of data file(s) and correct assignment of animal ID to sensor device ID • Timestamps: time zone (e.g. UTC, etc.), correction for day light savings, timely alignment of variables, intervals • Alignment of measurements and alarms/events • Time the algorithm needs to learn before it starts to work properly, and data can be used
Step 2: Get to know your data	• Variable description: units, sampling and data rate, raw data or processed, single-shot measurement or aggregated, time span and aggregation method, timely alignment (e.g. for aggregation over sliding windows) • Handling of missing data: imputation, tolerance level for calculation of variables, appearance in provided data
Step 3: Check completeness of the data	• Missing values: NULL/NA values or extended intervals between data points • Reasons for missing values (e.g. data transmission problems, low battery, etc.) or duplicates
Step 4: Evaluate plausibility and detect outliers	• Outlier identification and handling: by algorithm prior to data provision or not • Biological reference values • Reasons for potential outliers in the data (e.g. failed measurements)
Step 5: Check for technology related noise	• Sensor drift and calibration: handled by the algorithm prior to data provision or still part of the data set • Information on updates of hard- or software (e.g. algorithm, output variables, etc.)

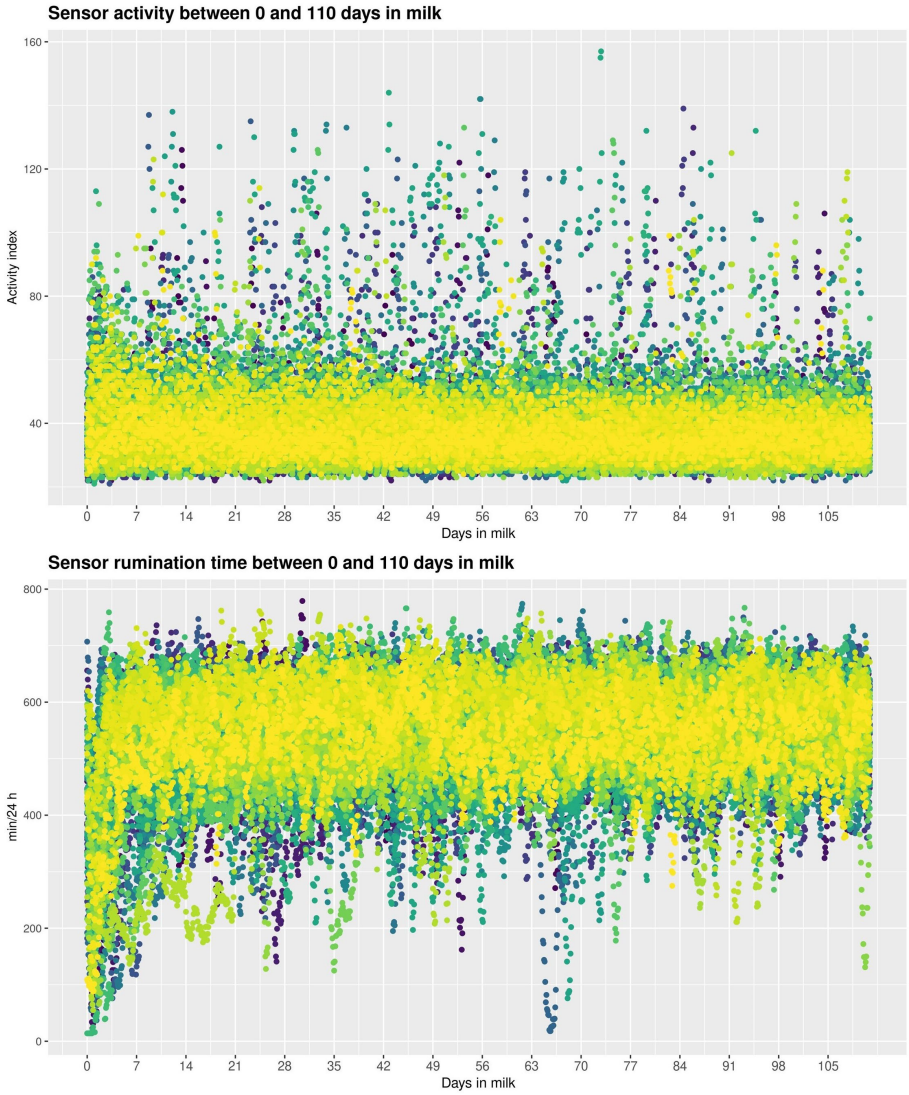

FIGURE 2
Scatterplots for bihourly sensor outputs for activity (top) and rumination time (bottom) over 110 days in milk for 63 lactations from 57 animals on one farm.

Missing data may occur as NULL/NA values or not appear in the data at all, i.e. intervals between subsequent data points are greater than the data rate. However, missing values may have already been imputed by the algorithm before output of sensor variables or due to the nature of the variable (e.g. moving averages or other types of aggregated data such as hourly output of mean rumination time over the last 24 hours) and thus may not be detectable in the data.

The decision about how to handle missing data should be based on the type and purpose of the analysis, further data processing, and the amount and distribution of missing data. For some types of analysis, it may be sufficient to set a threshold for the minimum number of records available per animal or to have observations over a specific period. As shown by Carlström et al. (2013), genetic parameter estimation for milkability traits from AMS data requires complete time series data; however, there was no difference in the accuracy of the parameter estimates when records between 8 and 330 days or only up to 100 days in milk were included. For herd management and disease predictions, the amount of time series data depends very much on the type (e.g. acute vs. chronic) and frequency of the disease, individual or herd level information as well as the quality of the algorithm (Hogeveen et al., 2021). If data are to be further aggregated, e.g. on a daily basis, averaging (mean or median values) will be less affected by missing values than summations (Mensching et al., 2020b). The latter is important for data such as (daily) milk yields recorded by automatic milking systems. Hogeveen et al (2001) suggested calculating kg milk per hour by dividing the milking yield of a milking event by the milking interval. They discarded records deviating more than 50% in hourly production from the average per hour production of the preceding and subsequent ten days and milking events of less than one kg milk in total. Flagging potentially flawed data from milk recordings using this approach is particularly important when using data from single milkings and information about milking intervals for health assessment to avoid bias. This approach was also applied in the D4Dairy project to identify gaps and potential outliers in the data as well as for the aggregation of 24-hour milk yields based on

timestamps of single milkings. Furthermore, depending on the underlying reasons for missing data, data may be missing completely at random or systematically, resulting in different distributions between missing and non-missing observations, leading to biased results (Bhaskaran and Smeeth, 2014). Imputation or interpolation of missing data may be considered necessary for some purposes or types of time series data analyses, which cannot handle missing values. Some analyses however may require complete datasets for producing unbiased results, which can be achieved by applying data imputation methods (You et al). However, by imputing values, particularly of larger blocks missing data, assumptions about the data and models must be made, which may or may not reflect the patterns of the 'real' data.

Opposed to missing values duplicates may also be found in the data. Complete duplicates, i.e. data points with equal timestamps and equal sensor values, should be removed. However, we also encountered data with multiple measurements per timestamp containing different values or – in the case of multivariate sensor data - values were split across rows for different variables. In the latter case rows with equal timestamps may be merged resulting in complete records whereas data with duplicate timestamps but differing values will have to be removed. Furthermore, data which was recorded in shorter intervals than the data frequency should be discarded from the data set. In their study on the effects of health disorders on sensor measurements, Siberski-Cooper et al (2023) even removed data three days prior to sensor failures.

3.4 Step 4: evaluate plausibility of sensor measures and detect outliers

Regarding outlier detection and their treatment different approaches exist, which can be broadly categorized into three principal approaches: those reliant on statistical parameters and the underlying distribution of the data, those employing thresholds grounded in biologically significant ranges, when applicable, and those identifying outliers in an iterative feedback modeling process applying e.g. machine learning algorithms or modeling approaches. Several investigations have delineated outliers based on statistical measures, such as the identification of values exceeding a specified number of standard deviations (SD) from the mean, as exemplified by approaches like removing values outside a confidence interval of ±4 SD from the mean (for daily milk yield and body weight, see Ouweltjes et al., 2021), ± 3 SD from the mean (for reticular body temperature, see Bewley et al., 2008) or standardized residuals outside ±3 SD (Mensching et al., 2020b). Alternatively, outliers were identified by assessing whether a value surpasses 1.5 times the interquartile distance from the lower or upper quartile (Mensching et al., 2020a). Moreover, employing smoothing techniques and analyzing deviations from expected curves can enhance the identification of potential outliers in the data.

Establishing biologically meaningful thresholds based on domain knowledge presents another strategy in data plausibility checks and outlier identification within sensor data. Andreen et al (Andreen, 2020), for example, removed data with a total rumination time of less than 200 minutes per day because they suspected illness of the animal or technological malfunctioning of the sensor system. Additionally, they defined the ratio of time allocated to rumination relative to time spent eating should fall within the range of 0.5 and 10. Departures from this range triggered the suspicion of misinterpretation, such as rumination being erroneously identified as eating, or vice versa. Within the D4Dairy project, temperature data from a rumen bolus showed negative values, which is biologically impossible in living animals (Schodl et al., 2022). However, while it is easy to identify a particular range of measurements as implausible, it is difficult to set a precise threshold below or above which data can be flagged as a potential outlier or measurement error. Thus, exactly defining these thresholds based on domain knowledge alone is not straightforward and poses various challenges, primarily attributable to several inherent reasons. One notable challenge arises from the fact that variables, such as activity indices, are expressed as 'arbitrary' or dimensionless values, lacking a predefined reference spectrum for their biological significance. Unlike established and measurable physiological parameters, these values currently lack a universally acknowledged standard that delineates their biological relevance. Consequently, the absence of a well-established framework complicates the task of setting thresholds based on biological significance. In instances where extreme values are evident, manifesting as physiologically implausible outcomes (e.g. reticular temperature values approximating or falling below 0°C, as elucidated in Schodl et al., 2022), their identification and subsequent removal can be readily executed through the application of domain knowledge. However, the intricacy arises when attempting to define specific threshold values for continuous variables that appropriately discern between normal fluctuations and values indicative of abnormal biological states. Thus, it is crucial to exercise caution when implementing outlier removal methods, as overly stringent criteria may inadvertently lead to the exclusion of genuine outliers. Such genuine outliers could hold valuable information for predicting diseases or characterizing traits, such as resilience (Ouweltjes et al., 2021). Or as Knorr and Ng (1998) have aptly phrased it: "One person's noise is another person's signal". Including other variables into the outlier detection process may help prevent the unintentional removal of genuine outliers. Mensching et al (2020b) developed a multivariate plausibility assessment algorithm to differentiate between 'physiologically normal', 'physiologically extreme' and 'implausible' observations in simultaneously recorded data. They based their concept on the assumption that different measurable parameters are physiologically linked and that in the case of a disease or other disturbances more than one parameter is altered. Vice versa, if only one parameter shows extreme deviations it is most likely implausible and to be classified as an outlier (Mensching et al., 2020b). A similar approach is the integration with other farm data such as calving, insemination or health records to cross-check if deviations in sensor patterns may have occurred due to events such as calving or heat. Within the D4Dairy project these data were available from the Austrian central cattle database Rinderdatenverbund (RDV, https://www.rdv-gmbh.net/en.html)

and were used to flag e.g. data during heat or calving events or around clinical diagnoses before further analysis.

Particularly when confronted with time series data, which is characterized by inherent noise or missing values, advanced modeling techniques provide suitable approaches. Dynamic Linear Models (DLMs) incorporating Kalman filtering emerge as particularly valuable tools for the modeling and estimation of the hidden state of a system undergoing temporal evolution, particularly in the design of control systems. An illustrative example of the efficacy of such methods is found in the work of Stygar et al. (2017), where statistical analyses applied to historical data enabled the estimation of measurement errors and variances associated with both, animal and environmental factors. This information, in turn, serves as a crucial metric for assessing the control status of the current production process. For instance, it allows an evaluation of the potential impact on milk production should a farmer implement a modified feeding strategy within a herd utilizing an AMS. Within the framework of the D4Dairy project, an innovative statistical sensor data processing framework has been conceptualized. This framework capitalizes on the interplay between data quality and model robustness, utilizing co-dependency to identify performance issues in data-driven predictive models (Papst et al., 2021). Papst et al (2021) exemplified the effectiveness of this framework by revealing that shifts in the distribution or mean of input data significantly impacted the quality of predictive models. They introduced an indicator capable of detecting such distributional shifts in the test data. When the test data contains too many outliers compared to the training data, intervention steps such as re-training of the model were applied, which lead to a noteworthy improvement of up to 62% in accuracy compared to predictions without interventions (Papst et al., 2021). This underscores the utility of advanced statistical frameworks in addressing data quality challenges and optimizing the performance of predictive models in agricultural contexts. While outlier identification is one important aspect the question of how to manage them is another one (Basu and Meckesheimer, 2007). Instead of flagging or merely removing outliers and potential measurement errors, specialized techniques quantifying the expected degree of measurement error may be employed to replace outliers with imputed values (Basu and Meckesheimer, 2007).

3.5 Step 5: check for technology related noise

Technology related noise such as sensor drift, updates, calibration, and the installation of new devices have to be considered and may not be straightforward to detect depending on its velocity and dimension. To detect and correct for sensor measurement drift, i.e. deviations from its true value over time, various methods (e.g. sensor calibration) and algorithms were already proposed (Teh et al., 2020). However, if drift is happening very slow, algorithms may not be able to detect it (Giannoni et al., 2018). Therefore, any suspicious sensor reading should be checked against herd management records and data should only be removed if it is associated with sensor malfunction (e.g. negative values for reticular temperature measurements within the D4Dairy project Schodl et al., 2022). In a study by Stygar and Kristensen (2016), daily observations of pig body weight gain were systematically decreasing for all pigs in a pen leading to the conclusion of malfunctioning of scales. By cross-checking the management records the alleged malfunctioning turned out to be a shortage of soybean meal in the pigs' diet due to a forgotten order to the feed company. Moreover, instruments may be prone to measuring instabilities over time (e.g. mid-infrared mass spectrometers) and measuring a standardized reference sample (e.g. samples of the same batch of milk) on a regular basis may help to detect this issue (Grelet et al., 2021). Drift can also occur for timestamps if data are transmitted in the wrong time slot due to clock drifts in sensor devices leading to gradual delay in transmission (Leliveld et al., 2024). Depending on the type of data provided by the sensor company and the extent of pre-processing, drift may have been accounted for prior to data provision. As the detection and correction of sensor drift is essential for a functioning herd management software, it can be assumed that the company has developed strategies to deal with this issue. Ideally, this information should be retrieved from the sensor company.

Software that analyzes sensor data for decision support is not a static tool and is regularly updated to fix bugs or add new analyses and evaluations. In addition, the sensors themselves, or the algorithms that convert the sensor signal into data, are constantly being improved or calibrated in the event of sensor drift. As beneficial as this is for the user of the herd management software, it can be a critical issue when it comes to using sensor-generated data for other purposes. Values or patterns may change after calibration or installation of new sensors, leading to biased results. During the D4Dairy project, one sensor company updated their software and replaced activity data with feeding time (Figure 3) and time resolution of data was changed from 2-hour to 1-hour intervals. In this case, detection of the update was straightforward, which would not have been the case if the sensor variables had remained the same. Therefore, when a sensor company provides data, it may be helpful to ask for additional information about software updates, changes in the algorithms that translate the sensor signal into data, sensor replacements (e.g. by changing the sensor IDs on the same animal) and the reason for the replacement (e.g. new sensor of the same type or a new generation of sensor). Assuming that this information is not always available, the data should be checked for this issue. Indications could be the introduction of new variables, different temporal resolution of the data, sudden or persistent changes in scale. In the case of a software update, these changes apply to all animals on a farm and occur on the same date. If data from the same sensor system are available for more than one farm, the occurrence of such changes on other farms in a similar period may also be a good indicator for a software update. In the case of

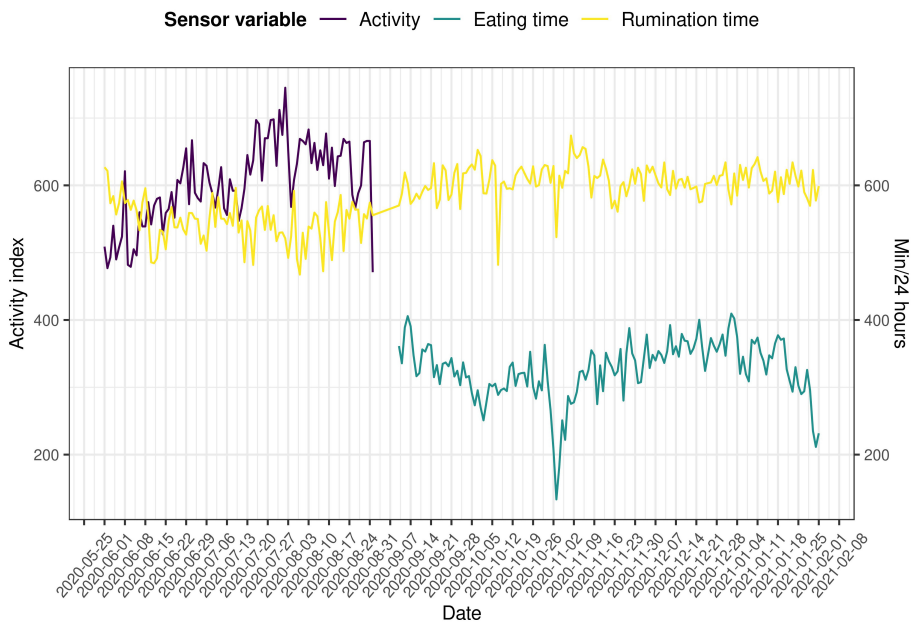

FIGURE 3
Example for sensor data from one animal for the replacement of activity data by eating time during a software update.

management changes that affect the behavior of the herd (e.g. animals are put out to pasture for a period of time, resulting in higher activity levels for all animals), care must be taken to interpret these changes accurately. If the above changes are only seen in individual animals at different times, this may indicate that the sensor was calibrated, or the device was replaced.

4 Document your approach

Efforts in data cleaning are largely underreported. For instance, only approximately 40% of medical studies provide a statement about data cleaning (Huebner et al., 2020). First and foremost, transparency and reproducibility in data processing and algorithm developments are essential for achieving progress in science (McKinney et al., 2020). Taking sensor data exploitation further in terms of the development of herd management or breeding tools or quality assurance within the dairy sector, proper documentation of data cleaning efforts may become a key factor in monitoring the effectiveness of these tools. Therefore, information on the number of missing observations or any efforts concerning imputation or interpolation of missing data would need to be provided if the data is shared through agricultural dataspaces. Information about uncertainty of obtained results is indispensable for the implementation of management strategies (Stygar et al., 2017); therefore, any deficiencies on data availability should be clearly communicated to the end users.

Number of days without observations, criteria for a plausibility check, number of removed outliers and information about data extrapolations are critical for transparency and reproducibility and as such should be reported. Therefore, all actions taken regarding data or issues which have been discovered in steps 1 to 6 should be described and this information should be provided to potential users of the data (e.g. scientists, other companies).

5 Conclusions

The use of data from wearable sensor devices on dairy cows has the potential to improve animal health and welfare on dairy farms on various levels in particular if this data is integrated with other farm data. However, depending on the information available on how these data are generated, to what extent they have been pre-processed and cleaned for erroneous measurements, sensor drift, noise or outliers, they have to be inspected for potentially biasing influences and cleaned accordingly. In this paper we highlighted important aspects of sensor data and presented suggestions on methods to detect potential sources of bias and approaches to data preparation based on our work in the D4Dairy project and scientific body of literature. Furthermore, we aimed to raise awareness of the importance of communication between data providers and recipients and which information may be important to retrieve from companies, if possible. These recommendations are targeted to all data analysts and professionals in the dairy sector, who are collecting and analyzing data for herd management or breeding applications. Standardizing data cleaning steps is crucial and one of the core functions of ICAR, who already published guidelines on e.g. cattle milk recording (ICAR, 2022b) or recording of direct health traits (ICAR, 2022a). Together with the IDF Standing committee on animal health and welfare, ICAR is currently developing guidelines regarding the use of sensor data in herd management and breeding to facilitate the use of sensor data in cattle breeding and management on a global level (Egger-Danner et al., 2024).

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data used in the current study are not publicly available due to privacy restrictions of the data provider and owner (LKV Austria Gemeinnützige GmbH, https://lkv.at/) and D4Dairy (https://d4dairy.com). Authors were provided with anonymized data according to an authorized material transfer agreement. Supplementary data may be available upon reasonable request from the corresponding author. Requests to access these datasets should be directed to egger-danner@zuchtdata.at.

Author contributions

KS: Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. AS: Conceptualization, Investigation, Methodology, Writing – review & editing. FS: Data curation, Writing – review & editing. CE-D: Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was conducted within the COMET-Project D4Dairy (Digitalization, Data integration, Detection and Decision support in Dairying, Project number: 872039; https://d4dairy.com/; accessed on 1 June 2022)) that is supported by BMK, BMDW, and the provinces of Lower Austria and Vienna in the framework of COMET-Competence Centers for Excellent Technologies. The COMET program is handled by the FFG (grant number 872039). Anna Stygar received financial support from the European Union's Horizon Europe Coordination and Support Action under grant agreement no. 101134866, (project Digi4Live, https://digi4live.eu/).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adriaens, I., Friggens, N. C., Ouweltjes, W., Scott, H., Aernouts, B., and Statham, J. (2020). Productive life span and resilience rank can be predicted from on-farm first-parity sensor time series but not using a common equation across farms. *J. Dairy Sci.* 103, 7155–7171. doi: 10.3168/jds.2019-17826

Andreen, D. M. (2020). Relationships between milk fat and rumination time recorded by commercial rumination sensing systems. *J. Dairy Sci.* 103, 8094–8104. doi: 10.3168/jds.2019-17900

Basu, S., and Meckesheimer, M. (2007). Automatic outlier detection for time series: an application to sensor data. *Knowl. Inf Syst.* 11, 137–154. doi: 10.1007/s10115-006-0026-6

Bewley, J. M., Einstein, M. E., Grott, M. W., and Schutz, M. M. (2008). Comparison of reticular and rectal core body temperatures in lactating dairy cows. *J. Dairy Sci.* 91, 4661–4672. doi: 10.3168/jds.2007-0835

Bhaskaran, K., and Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? *Int. J. Epidemiol.* 43, 1336–1339. doi: 10.1093/ije/dyu080

Borchers, M. R., Chang, Y. M., Proudfoot, K. L., Wadsworth, B. A., Stone, A. E., and Bewley, J. M. (2017). Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *J. Dairy Sci.* 100, 5664–5674. doi: 10.3168/jds.2016-11526

Bouchon, M., Bach, A., Meunier, B., Ternman, E., Van Reenen, E., Veissier, I., et al. (2019). *Guidelines for validation of sensor output*. Available online at: https://www.smartcow.eu/public-deliverables/ (Accessed October 4, 2023).

Caja, G., Castro-Costa, A., and Knight, C. H. (2016). Engineering to support wellbeing of dairy animals. *J. Dairy Res.* 83, 136–147. doi: 10.1017/S0022029916000261

Carlström, C., Pettersson, G., Johansson, K., Strandberg, E., Stålhammar, H., and Philipsson, J. (2013). Feasibility of using automatic milking system data from commercial herds for genetic analysis of milkability. *J. Dairy Sci.* 96, 5324–5332. doi: 10.3168/jds.2012-6221

Chicco, D., Oneto, L., and Tavazzi, E. (2022). Eleven quick tips for data cleaning and feature engineering. *PLoS Comput. Biol.* 18, e1010718. doi: 10.1371/journal.pcbi.1010718

Egger-Danner, C., Fuerst-Waltl, B., Obritzhauser, W., Fuerst, C., Schwarzenbacher, H., Grassauer, B., et al. (2012). Recording of direct health traits in Austria—Experience report with emphasis on aspects of availability for breeding purposes. *J. Dairy Sci.* 95, 2765–2777. doi: 10.3168/jds.2011-4876

Egger-Danner, C., Klaas, I., Brito, L., Schodl, K., Bewley, J., Cabrera, V. E., et al. (2024). "Improving animal health and welfare by using sensor data in herd management and dairy cattle breeding – a joint initiative of ICAR and IDF," in *Precision Livestock Farming 2024 (Organising Committee of the 11th European Conference on Precision Livestock Farming (ECPLF)*, University of Bologna, Bologna, Italy), 56–63.

Elischer, M. F., Arceo, M. E., Karcher, E. L., and Siegford, J. M. (2013). Validating the accuracy of activity and rumination monitor data from dairy cows housed in a pasture-based automatic milking system. *J. Dairy Sci.* 96, 6412–6422. doi: 10.3168/jds.2013-6790

Friggens, N. C., Adriaens, I., Boré, R., Cozzi, G., Jurquet, J., Kamphuis, C., et al. (2022). Resilience: reference measures based on longer-term consequences are needed to unlock the potential of precision livestock farming technologies for quantifying this trait. *Peer Community J.* 2 (e38), 1–16. doi: 10.24072/pcjournal.136

Giannoni, F., Mancini, M., and Marinelli, F. (2018). Anomaly detection models for ioT time series data. *Preprint.* 1–10. doi: 10.48550/arXiv.1812.00890

Grelet, C., Dardenne, P., Soyeurt, H., Fernandez, J. A., Vanlierde, A., Stevens, F., et al. (2021). Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions. *Methods* 186, 97–111. doi: 10.1016/j.ymeth.2020.07.012

Hogeveen, H., Klaas, I. C., Dalen, G., Honig, H., Zecconi, A., Kelton, D. F., et al. (2021). Novel ways to use sensor data to improve mastitis management. *J. Dairy Sci.* 104, 11317–11332. doi: 10.3168/jds.2020-19097

Hogeveen, H., Ouweltjes, W., de Koning, C. J. A. M., and Stelwagen, K. (2001). Milking interval, milk production and milk flow-rate in an automatic milking system. *Livestock Production Sci.* 72, 157–167. doi: 10.1016/S0301-6226(01)00276-7

Huebner, M., Vach, W., le Cessie, S., Schmidt, C. O., Lusa, L., Cook, D., et al. (2020). Hidden analyses: a review of reporting practice and recommendations for more

transparent reporting of initial data analyses. *BMC Med. Res. Method.* 20, 61. doi: 10.1186/s12874-020-00942-y

ICAR (2022a). *Section 7 - Guidelines for Health, Female Fertility, Udder Health, Claw Health Traits, Lameness and Calving Traits in Bovine*. Available online at: https://www.icar.org/Guidelines/07-Bovine-Functional-Traits.pdf (Accessed February 21, 2024).

ICAR (2022b). *Section 2 - Guidelines for dairy cattle milk recording*. Available online at: https://www.icar.org/index.php/icar-recording-guidelines/ (Accessed April 19, 2024).

Jensen, D. B., van der Voort, M., Kamphuis, C., Athanasiadis, I. N., De Vries, A., and Hogeveen, H. (2019). "Comparison of data driven mastitis detection methods," in *Precision Livestock Farming '19* (European Association for Precision Livestock Farming, Cork, Ireland), 626–632.

Knight, C. H. (2020). Review: Sensor techniques in ruminants: more than fitness trackers. *Animal* 14, s187–s195. doi: 10.1017/S1751731119003276

Knorr, E. M., and Ng, R. T. (1998). "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the 24th VLDB Conference, New York* (San Francisco, CA, United States: Morgan Kaufmann Publishers Inc.), 392–403.

Koltes, J. E., Cole, J. B., Clemmens, R., Dilger, R. N., Kramer, L. M., Lunney, J. K., et al. (2019). A vision for development and utilization of high-throughput phenotyping and big data analytics in livestock. *Front. Genet.* 10. doi: 10.3389/fgene.2019.01197

Leliveld, L. M. C., Brandolese, C., Grotto, M., Marinucci, A., Fossati, N., Lovarelli, D., et al. (2024). Real-time automatic integrated monitoring of barn environment and dairy cattle behaviour: Technical implementation and evaluation on three commercial farms. *Comput. Electron. Agric.* 216, 108499. doi: 10.1016/j.compag.2023.108499

Mayo, L. M., Silvia, W. J., Ray, D. L., Jones, B. W., Stone, A. E., Tsai, I. C., et al. (2019). Automated estrous detection using multiple commercial precision dairy monitoring technologies in synchronized dairy cows. *J. Dairy Sci.* 102, 2645–2656. doi: 10.3168/jds.2018-14738

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94. doi: 10.1038/s41586-019-1799-6

Mensching, A., Bünemann, K., Meyer, U., von Soosten, D., Hummel, J., Schmitt, A. O., et al. (2020a). Modeling reticular and ventral ruminal pH of lactating dairy cows using ingestion and rumination behavior. *J. Dairy Sci.* 103, 7260–7275. doi: 10.3168/jds.2020-18195

Mensching, A., Zschiesche, M., Hummel, J., Schmitt, A. O., Grelet, C., and Sharifi, A. R. (2020b). An innovative concept for a multivariate plausibility assessment of simultaneously recorded data. *Animals* 10, 1412. doi: 10.3390/ani10081412

Ouweltjes, W., Spoelstra, M., Ducro, B., De Haas, Y., and Kamphuis, C. (2021). A data-driven prediction of lifetime resilience of dairy cows using commercial sensor data collected during first lactation. *J. Dairy Sci.* 104, 11759–11769. doi: 10.3168/jds.2021-20413

Papst, F., Schodl, K., and Saukh, O. (2021). "Exploring co-dependency of ioT data quality and model robustness in precision cattle farming," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. SenSys '21* (Association for Computing Machinery, New York, NY, USA), 433–438. doi: 10.1145/3485730.3493447

Ren, K., Nielsen, P. P., Alam, M., and Rönnegård, L. (2021). Where do we find missing data in a commercial real-time location system? Evidence from 2 dairy farms. *JDS Commun.* 2, 345–350. doi: 10.3168/jdsc.2020-0064

Rustas, B. O., Persson, Y., Ternman, E., Kristensen, A. R., Stygar, A. H., and Emanuelson, U. (2024). The evolutionary operation framework as a tool for herd-specific control of mastitis in dairy cows. *Livestock Sci.* 279, 105390. doi: 10.1016/j.livsci.2023.105390

Schmeling, L., Elmamooz, G., Hoang, P. T., Kozar, A., Nicklas, D., Sünkel, M., et al. (2021). Training and validating a machine learning model for the sensor-based monitoring of lying behavior in dairy cows on pasture and in the barn. *Animals* 11, 2660. doi: 10.3390/ani11092660

Schodl, K., Fuerst-Waltl, B., Schwarzenbacher, H., Steininger, F., Suntinger, M., Papst, F., et al. (2022). "Challenges of integration and validation of farm and sensor data for dairy herd management," in *ICAR Technical Series no. 26*. Eds. A. M. Christensen, D. Lefebvre, F. Miglior, R. Cantin, M. Burke and C. Mosconi (ICAR, Montréal, CANADA), 241–247.

Shook, G. E. (1989). Selection for disease resistance. *J. Dairy Sci.* 72, 1349–1362. doi: 10.3168/jds.S0022-0302(89)79242-0

Siberski-Cooper, C. J., Mayes, M. S., Gorden, P. J., Hayman, K., Hardie, L., Shonka-Martin, B. N., et al. (2023). The impact of health disorders on automated sensor measures and feed intake in lactating Holstein dairy cattle. *Front. Anim. Sci.* 3. doi: 10.3389/fanim.2022.1064205

Steeneveld, W., Hogeveen, H., and Oude Lansink, A. G. J. M. (2015). Economic consequences of investing in sensor systems on dairy farms. *Comput. Electron. Agric.* 119, 33–39. doi: 10.1016/j.compag.2015.10.006

Stöger, K., Schneeberger, D., Kieseberg, P., and Holzinger, A. (2021). Legal aspects of data cleansing in medical AI. *Comput. Law Secur. Rev.* 42, 105587. doi: 10.1016/j.clsr.2021.105587

Stygar, A. H., Frondelius, L., Berteselli, G. V., Gómez, Y., Canali, E., Niemi, J. K., et al. (2023). Measuring dairy cow welfare with real-time sensor-based data and farm records: a concept study. *animal* 17, 101023. doi: 10.1016/j.animal.2023.101023

Stygar, A. H., Gómez, Y., Berteselli, G. V., Costa, E. D., Canali, E., Niemi, J. K., et al. (2021). A systematic review on commercially available and validated sensor technologies for welfare assessment of dairy cattle. *Front. Veterinary Sci.* 8. doi: 10.3389/fvets.2021.634338

Stygar, A. H., Krampe, C., Llonch, P., and Niemi, J. K. (2022). How far are we from data-driven and animal-based welfare assessment? A critical analysis of european quality schemes. *Front. Anim. Sci.* 3. doi: 10.3389/fanim.2022.874260

Stygar, A. H., and Kristensen, A. R. (2016). Monitoring growth in finishers by weighing selected groups of pigs – A dynamic approach1. *J. Anim. Sci.* 94, 1255–1266. doi: 10.2527/jas.2015-9977

Stygar, A. H., Krogh, M. A., Kristensen, T., Østergaard, S., and Kristensen, A. R. (2017). Multivariate dynamic linear models for estimating the effect of experimental interventions in an evolutionary operations setup in dairy herds. *J. Dairy Sci.* 100, 5758–5773. doi: 10.3168/jds.2016-12251

Teh, H. Y., Kempa-Liehr, A. W., and Wang, K. I.-K. (2020). Sensor data quality: a systematic review. *J. Big Data* 7, 11. doi: 10.1186/s40537-020-0285-1

Unwin, A. (2020). Why is data visualization important? What is important in data visualization? *Harvard Data Sci. Rev.* 2 (1). doi: 10.1162/99608f92.8ae4d525

Van den Broeck, J., Cunningham, S. A., Eeckels, R., and Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med.* 2, e267. doi: 10.1371/journal.pmed.0020267

Yilmaz Adkinson, A., Abouhawwash, M., VandeHaar, M. J., Parker Gaddis, K. L., Burchard, J., Peñagaricano, F., et al. (2024). Assessing different cross-validation schemes for predicting novel traits using sensor data: An application to dry matter intake and residual feed intake using milk spectral data. *J. Dairy Sci.* 107, 8084–8099. doi: 10.3168/jds.2024-24701

You, J., Ellis, J. L., Adams, S., Sahar, M., Jacobs, M., and Tulpan, D. (2023). Comparison of imputation methods for missing production data of dairy cattle. *Animal.* 17, 100921. doi: 10.1016/j.animal.2023.100921