Applied Animal Behaviour Science 249 (2022) 105597

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

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

Tracking performance in poultry is affected by data cleaning method and housing system

Laura Candelotto *, Klara J. Grethen, Camille M. Montalcini, Michael J. Toscano, Yamenah Gómez

Center for Proper Housing: Poultry and Rabbits (ZTHZ), Division of Animal Welfare, VPH Institute, University of Bern, Burgerweg 22, 3052 Zollikofen, Switzerland

ARTICLE INFO

Keywords:
Individual behaviour
Sensor-based tracking
Validation
Data cleaning
Laying hens
Broiler breeders

ABSTRACT

Sensor-based behavioural observation methods improve our understanding of individual behaviour and welfare in large commercial groups, including poultry. Validating automatically generated data is essential to account for potential sources of error. Our study aimed to validate a sensor-based tracking system for broiler breeders (BB) and laying hens (LH) in commercially relevant housing systems. The BB study was conducted in 10 pens with 33 females and three males (Ross 308) per pen. Half of the pens contained a raised slatted area and two raised group nests (Raised), while in the remaining five pens, the nests and slats were on the floor (Floor). For the LH study, six pens with a commercial aviary were used, with 225 Dekalb White hens housed per pen (Aviary). Focal hens (BB, 10/pen; LH, 18/pen) were equipped with backpacks containing tracking devices that registered transitions between four (BB) or five (LH) resource-related zones covering all accessible areas within each housing system. The tracking data was compared against video observations for 20 focal BB on two days and 18 focal LH on three days (3×20 min/day). Three data cleaning methods tested with 30 values of a duration parameter were evaluated for reliability and stability with a cross-validation approach. Initial and post-cleaning performance were assessed with accuracy, precision, and sensitivity of recorded transitions and by calculating the reliability for two aspects of movement: total transitions (Lin's Concordance Correlation Coefficient) and locations (mean proportion of matching duration). A mixed model was applied to evaluate the duration of stay after false and true tracking registrations. Initial location reliability was high (> 0.949) in all housing systems, while reliability of total transitions was low (< 0.264), particularly in both BB housing systems (< 0.064). The cross-validation revealed suitable cleaning procedures for Aviary and Raised but not for Floor, thus Floor was not used for further analysis. Cleaning improved total transitions (> 0.832) while reliability of locations remained high (> 0.949) in Aviary and Raised. The duration between registrations was affected by housing system ($p < 0.001$) and was longer for true compared to false registrations ($p < 0.001$). Initial tracking performance varied between movement aspects and housing systems. The difference in duration between true and false registrations allowed for the application of simple yet effective data cleaning in Aviary and Raised, ensuring that the generated data better represented the animal's actual movement with reduced error associated with the tracking system.

1. Introduction

An important aspect of animal welfare research is behaviour, particularly at the individual level (Dawkins, 2003; Broom, 2010; Richter and Hintze, 2019). Behavioural observations are commonly done using video recordings, however this is time-consuming and prone to subjective interpretation (Catarinucci et al., 2014; Main et al., 2014) leading to low external validity. These problems are aggravated in poultry in cage-free housing systems, where animals have a high uniformity in appearance and are maintained in high densities leading to video recordings being challenging for individual level observations. Automated observation methods may offer solutions with benefits over human-based observations (Blokhuis et al., 2010; Rushen et al., 2012). Indeed, there is a growing interest in the use of sensors for animal health management (Neethirajan, 2017) and research publications on the use of technology to monitor poultry are increasing (Rowe et al., 2019). One possible technological application is the use of a tracking system measuring individual movement within a barn. The accuracy of a

* Corresponding author.
E-mail address: laura.candelotto@vetsuisse.unibe.ch (L. Candelotto).

https://doi.org/10.1016/j.applanim.2022.105597
Received 1 December 2021; Received in revised form 18 February 2022; Accepted 1 March 2022
Available online 5 March 2022

tracking system can be affected by a variety of factors including: positioning of sensors (e.g., horizontal vs. vertical orientation), distance to the signal emitting device, overlapping detection fields and presence of obstacles including other animals and barn interiors (Maselyne et al., 2014; Siegford et al., 2016; Triguero-Ocaña et al., 2019; van der Sluis et al., 2020). Metal, such as that used in housing construction, can be particularly problematic for a tracking system as it may weaken signals or act as a false antenna (DeVries et al., 2003; Siegford et al., 2016). Tracking poultry is especially challenging given the relatively small size of birds and complex housing environments, particularly in laying hens where aviaries allow birds to move within the three-dimensional space and contain large amounts of metal (Ellen et al., 2019).

Based on the raised concerns, it is crucial to validate the data generated by a tracking system (Rushen et al., 2012; Winckler, 2019) for example, by comparing the tracking data with video observations (Mendes et al., 2011; Maselyne et al., 2014). Compared to group-level assessments using techniques such as optical flow (Dawkins et al., 2012), a more detailed validation is of particular importance when investigating individual differences in behaviour. In the case of an erroneous tracking system, individual variation may originate from tracking tags performing differently or locations differing in error rates rather than actual behavioural differences.

In general, two types of error can affect the performance of a tracking system measuring the presence of an animal at a specific location or in an area (e.g. DeVries et al., 2003): the animal may be present though not recorded by the tracking system (False Negative) or the tracking system may record an animal's presence despite its absence (False Positive). False registrations (i.e., false recordings of an animal's location in the tracking system software) can be reduced by adjusting the hardware or software configuration, or by processing the generated data (e.g. Cornou et al., 2011; Van Der Sluis et al., 2019; Ren et al., 2020). Many data cleaning efforts focus on the reduction of False Positives using assumptions about behaviour, for example, by removing unlikely short registrations (Ledgerwood et al., 2010) or by defining a minimum duration for specific behaviours (Rufener et al., 2019). However, deleting registrations solely based on behavioural assumptions may risk the removal of true, meaningful behaviours (Rushen et al., 2012). It is therefore essential to not only know the features of the behaviour of interest but also the characteristics of erroneous recordings to determine the most suitable data cleaning method and duration parameter that balances removal of the largest number of erroneous recordings while retaining those that are true. As a final consideration, validations are typically done on a subsample of all recorded animals, days and times of day (Maselyne et al., 2014; Rufener et al., 2018; Zhuang et al., 2020). Data cleaning should thus be stable against variation in those subsampling factors in order to apply it to the full dataset.

Our study sought to validate a sensor-based tracking system in three different commercially relevant poultry housing systems, including two types for broiler breeders and one for laying hens. While broiler breeders normally live in a horizontally organised housing system, resources are organised in three dimensions for laying hens. We also aimed to validate simple data cleaning methods considering the stability of the cleaning method against variations in tracking tags, day and time of day. As our cleaning methods were based on a duration parameter, we further investigated the duration of stay after true vs. false transitions for each housing system.

2. Material and methods

2.1. Ethical approval

The study was approved by the Veterinary Office of the Canton of Bern in Switzerland (Broiler breeders; BB: BE9/19; Laying hens; LH: BE75/19). All applied procedures were in accordance with the Swiss regulations and guidelines for animal experiments.

2.2. Tracking system

We used a custom-designed tracking system (® Gantner Solutions GmbH, Schruns, Austria) based on low frequency (LF) tracking and ultra-high frequency (UHF) communication (Fig. 1a) that recorded the time of an individual animal's movement between pre-defined areas (transitions). The tracking system consisted of three elements: readers, markers and tags. The markers were either used as point markers or connected to a cable in order to cover a larger area. Various resource-related areas (hereafter referred to as zones) could be determined by arranging those point markers and markers with cables (Fig. 1b and Fig. 1 c). Each marker sent modified (unique identifier for each marker) LF signals (125 KHz) at regular intervals ranging from 1.2 s to 1.4 s for BB and 1.1–1.4 s for LH. All focal hens wore a custom-designed backpack (14.5 cm × 13 cm; mass: 15.6 g) containing a tracking tag (5 cm × 4 cm; mass: 28.1 g). The backpacks consisted of a unique colour combination for visual identification of individual hens. Tags could detect the Received Signal Strength Indication (RSSI) of all nearby markers and an algorithm was then applied to process all received information (Pseudocode in supplementary S1). With every received signal, the algorithm considered all received signals from that tag within the preceding 4.5 s and recorded a new transition if the two highest RSSI were from the same marker and if the associated zone differed from the last registered one. Any change into a new zone as well as the respective date and time of the change was communicated to the reader via an UHF signal (868 MHz). The reader was connected to a computer which stored all of the above mentioned information into a CSV file. In BB, two sets of markers and cable loops connected to one reader were installed to cover five pens per set and in LH two separate tracking systems with separate readers were installed, each covering three pens.

All markers, connected to a cable (except for marker in the outdoor zone of LH), were supplied by a mains power source while point markers and tags were battery-powered. The latest communication of all markers and tags with the computer as well as the battery levels of all battery-powered devices were automatically recorded and regularly checked in order to replace low batteries and defect tags (i.e. absence of communication with the computer) when necessary.

The tracking system was calibrated in two steps with the help of a tag on a dummy. First, the best place of cables and markers was determined and finally the signal strengths of each marker and cable was adjusted either within the software or by changing the physical configuration of the associated cable. The calibration procedure was performed until the dummy recorded the correct zone in all zones, including critical places at the border between two zones.

2.3. Barn setup and Flocks

The BB study was conducted in 10 pens (4.3 m x 2.3 m) in an experimental barn in Zollikofen (Switzerland), with 33 female and three male BB (Ross 308) per pen. All pens contained four resource-related areas used as zones for the tracking system, namely the litter area, the slatted area and two group nests (Fig. 1b). In half of the pens the slatted area and the two nest boxes were raised (Raised) while the other half of the pens contained nest boxes on the floor with a small ramp leading to the entrance (Floor). A more detailed description of the barn setup can be found in van den Oever et al. (2021). Water was provided *ad libitum* via nipple drinkers. The amount of feed was adjusted every week relative to the birds' body mass according to the Aviagen management protocol and was provided via two feeder lines installed in each pen. Ten birds per pen were randomly selected as focal individuals and equipped with the custom-designed backpacks (with elastic bands around the wings) containing the tracking devices at 21 weeks of age (WoA).

The LH (Dekalb White) were kept in a semi-commercial layer barn (Zollikofen; Switzerland) previously described in Stratmann et al. (2015). In brief, the barn contained a three-tiered aviary system (Aviary; Bolegg Terrace, Krieger AG, Ruswil, Switzerland) split into 20 identical

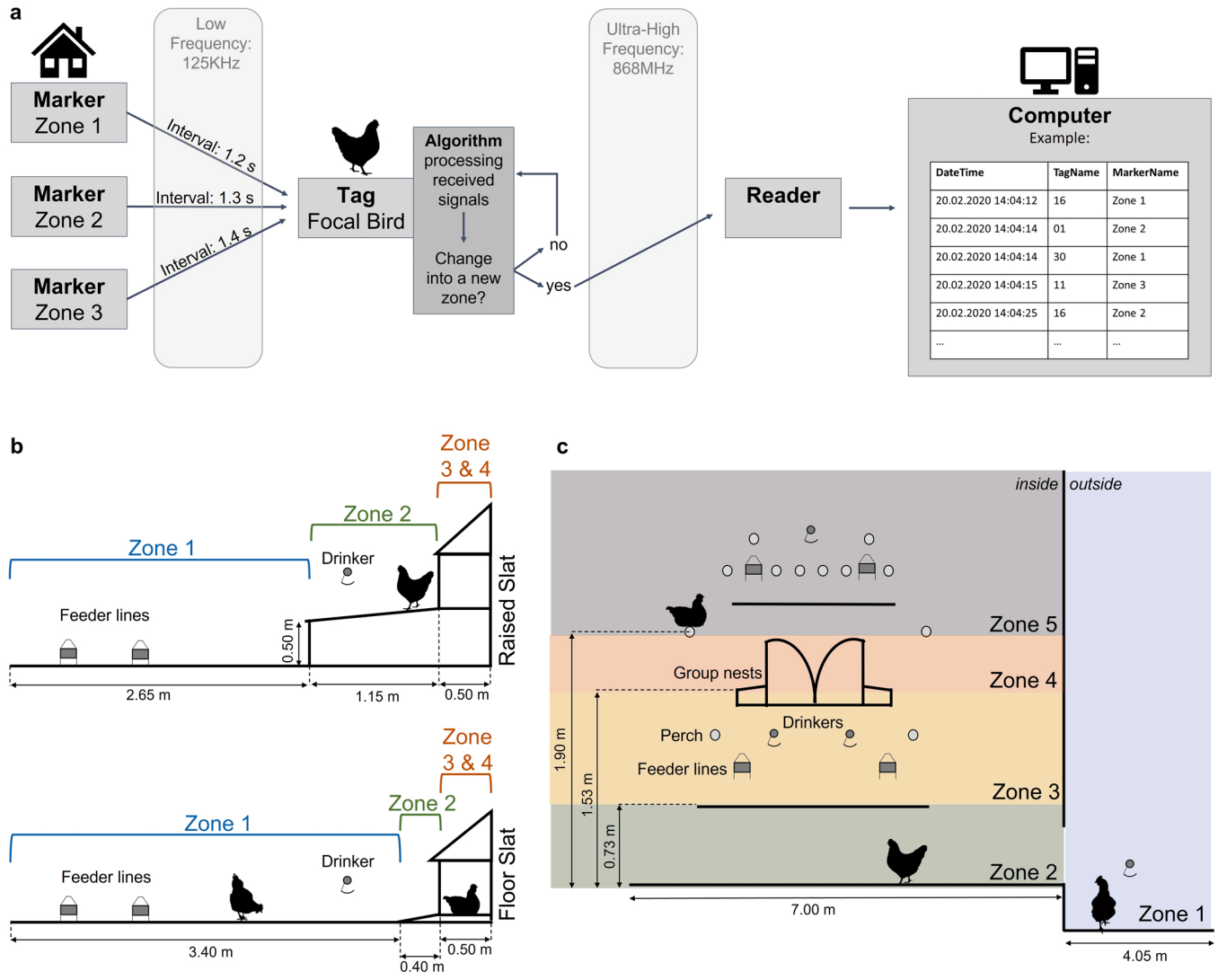

Fig. 1. a) Schematic representation of the general functioning of the used tracking system. The tracking system consisted of three principal technical elements: markers and their cables installed in each zone, tags attached to a backpack on hens and the readers sending the hen position to the computer. For this representation, three hypothetical zones were chosen as an example, though more zones were used for the actual tracking system. b) Side view of Broiler Breeder pens with raised slats and nest boxes (top) and small slats and nest boxes on the floor (bottom). The zones for the tracking system are given by the litter (zone 1), the slatted area (zone 2) and each of the two nest boxes (zone 3 & 4) with the left nest in each Pen as Zone 3 and the right nest as Zone 4. c) Side view of laying hen pens with the three-tiered aviary (zone 3, 4 and 5), the litter area (zone 2) and the wintergarden (zone 1) representing the 5 zones of the tracking system.

pens (450 cm × 700 cm × 230 cm) with 225 hens each. For this study, six pens (three pens next to each other and separated by mesh wire) were used. Nipple drinkers, four feeder lines and group nests with nest balconies as well as 14 round metal perches per pen were integrated into the aviary. In addition, each pen contained a litter area covered with wood shavings and access to a fully enclosed and covered outdoor area ("Wintergarden"; 9.32 m^2). These areas represented one of the five tracking zones (Fig. 1c). In total, 18 birds per pen were selected as focal individuals and equipped with the backpacks and tracking devices at 19 WoA.

2.4. Video Validation

In each pen, one (BB) or three (LH) infrared-sensitive, wide-angled video cameras were installed (Samsung, South Korea) and connected to a MULTIEYE Network Video recorder (artec technologies AG, Diepholz, Germany) for validation.

For the BB study, focal individuals were tracked between June and August 2019, while LH were tracked between October 2019 and July 2020. For validation, two (BB; 26 and 31 WoA) or three (LH; 22, 25 and 35 WoA) observation days were selected. On each day, three observation intervals of 20 min each were chosen, representing the morning, midday and evening within the respective light cycles. Each transition between zones logged by the tracking system (Fig. 1) was verified on the video recordings by visually tracking the focal hens based on the unique colour combinations on their backpacks. Initial examinations of the tracking system revealed a maximum time lapse of 16 s (BB) and 24 s (LH) between the actual movement and the registered transition by the tracking system. As future investigations will focus on the location and movement patterns rather than the exact timing of a transition, we accepted a short time lapse of maximally 1 min between the tracking system and the video recordings during validation.

The videos of BB and LH were assessed by two and three different observers, respectively. To assess inter-observer reliability, approximately 20% of the full data set (BB: 8 birds x 3 observation intervals; LH: 10 birds x 3 observation intervals) was rated by all observers resulting in a fully-crossed design. The reliability was calculated as Cohen's Kappa for the BB and as the arithmetic mean of each pairwise Cohen's Kappa

for the LH (Hallgren, 2012). Both assessments revealed reliable video ratings with Kappa values above the required 0.8 (BB: 0.898, LH: 0.891) (Landis and Koch, 1977).

2.5. Data cleaning methods

Three data cleaning methods were developed and evaluated: Filtering, Binning and Sliding Bin (Table 1). All cleaning method codes were written in R (4.0.5; R Core Team, 2021) and are available on GitHub (https://github.com/vetsuisse-unibe/ZTHZ_data_cleaning). The main goal of the cleaning methods was to reduce False Positives (false registrations) while keeping True Positives (true registrations as validated by the video observations). All three cleaning methods were based on the duration of a stay in a new zone following a transition (duration parameter) as defined below.

- Filtering is a simple cleaning method, removing any zone with a duration of stay smaller than a given threshold duration.
- Binning calculates the duration of each zone within a given time window. Among all zones within that window, the zone with the longest total duration was chosen as the current zone for the bird for the full window.
- Sliding Bin is based on Binning but, instead of a fixed window, the bin is moving along a 1 s timeseries (Table 1). Within each window, the zone with the highest duration was chosen. If this zone was different from the former window, a change in zone was recorded. The timepoint of zone change was defined as the middle timepoint of the window (Odd window size: exact middle; even window size: middle rounded up) to avoid that the change in zone is shifted by half of the window size. Thus, in the resulting cleaned dataset, the minimal duration of a stay within a particular zone can be smaller than the chosen window size in contrast to Binning.

In the Binning and Sliding Bin, a tie in the durations of two or more zones within one window might occur. Occurrence of ties were resolved by using the last zone of those causing the tie within the window (Table 1). Thirty different values of the duration parameter (duration thresholds for Filtering and window sizes for Binning and Sliding Bin) were evaluated within each cleaning method ranging from 10 s to 300 s in 10 s intervals and applied to the tracking data of BB and LH.

2.6. Data processing

Two main features were extracted from the tracking data to describe animal movement: locations and total transitions. The location represents the current zone of a bird at any given time and was obtained as the duration spent within a zone. In the present study, whenever a zone registered by the tracking system was equal to the zone observed on the video, the matching duration was extracted and all those matching durations were summed up for each bird and each observation interval to

Table 1
The basic concept of the three tested cleaning methods. The following examples show two potential movement patterns as time series sequences (Example 1 represents a bird that did not move and Example 2 represents a bird with fast movements between zones) in the raw tracking data as well as the true location and transitions of the respective birds. Both examples show the zone of stay per second in a 20 s interval with changes in location being marked as bold and underlined (transitions). Raw logged zones contain not only the actual true changes in locations but also some wrong logged transitions. In this example, a duration parameter of 5 s was used for each of the three cleaning methods. Locations removed during filtering are marked as crossed out "e.g. ~~333~~". Bins are indicated with "[]". The resulting location of a bin is shown with a blank arrow and the overall cleaned data is shown by a filled arrow with the final cleaned timeseries sequence for the 20 s interval. As Sliding Bin would include information before and after the given example of the 20 s interval windows to clean the full 20 s, only the 16 s that do not include additional information of prior or post are given for the overall cleaned data.

	Example 1	Example 2
true zones	**2** 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	**2** 2 2 2 2 2 **3** 3 3 3 3 3 3 **4** 4 4 4 4
raw logged zones	**2** 2 2 2 2 **3** 3 3 **2** 2 2 2 2 **3** 3 **2** 2 2 **3** 3	**2** 2 2 2 2 **5** 5 **3** 3 3 3 **5** 5 **3** 3 **4** 4 4 4 **5**
Filtering (5s)	2 2 2 2 2 ~~3 3 3~~ 2 2 2 2 2 ~~3 3~~ 2 2 2 ~~3 3~~ ⇩ **2** 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 ~~5 5~~ ~~3 3 3 3~~ ~~5 5~~ ~~3 3~~ ~~4 4 4 4~~ 5 ⇩ **2** 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Binning (5s)	[2 2 2 2 2] [3 3 3 2 2] [2 2 2 3 3] [2 2 2 3 3] ⇩ ⇩ ⇩ ⇩ 2 3 2 2 ⇩ **2** 2 2 2 2 **3** 3 3 3 3 **2** 2 2 2 2 2 2 2 2 2	[2 2 2 2 2] [5 5 3 3 3] [3 5 5 3 3] [4 4 4 4 5] ⇩ ⇩ ⇩ ⇩ 2 3 3 4 ⇩ **2** 2 2 2 2 **3** 3 3 3 3 3 3 3 3 3 **4** 4 4 4 4
Sliding Bin (5s)	[2 2 2 2 2] 3 3 3 2 2 2 2 2 3 3 2 2 2 3 3 ⇨ 2 2 [2 2 2 2 3] 3 3 2 2 2 2 2 3 3 2 2 2 3 3 ⇨ 2 2 2 [2 2 2 3 3] 3 2 2 2 2 2 3 3 2 2 2 3 3 ⇨ 2 2 2 2 [2 2 3 3 3] 2 2 2 2 2 3 3 2 2 2 3 3 ⇨ 3 … 2 2 2 2 2 3 3 3 2 2 2 2 2 3 3 [2 2 2 3 3] ⇨ 2 ⇩ **2** 2 2 **3** 3 3 **2** 2 2 2 2 2 2 2 2	[2 2 2 2 2] 5 5 3 3 3 3 5 5 3 3 4 4 4 4 5 ⇨ 2 2 [2 2 2 2 5] 5 3 3 3 3 5 5 3 3 4 4 4 4 5 ⇨ 2 2 2 [2 2 2 5 5] 3 3 3 3 5 5 3 3 4 4 4 4 5 ⇨ 2 2 2 2 [2 2 5 5 3] 3 3 3 5 5 3 3 4 4 4 4 5 ⇨ 5 … 2 2 2 2 2 5 5 3 3 3 3 5 5 3 3 [4 4 4 4 5] ⇨ 4 ⇩ **2** 2 2 **5** **3** 3 3 3 3 3 3 3 **4** 4 4 4

evaluate the reliability of locations. Transitions are the movements from one zone to another and were measured as the total number of all movements between any two zones per bird and observation interval. Locations and total transitions were statistically evaluated against video observations from the raw tracking data (initial performance) and the cleaned tracking data (post-cleaning performance) (Fig. 2). Hereafter, locations and total transitions refer to the above described matching duration and total number of transitions, respectively.

The cleaning methods were evaluated for their performance stability across observation intervals and tags using a cross-validation approach explained in Fig. 2. The cross-validation approach was chosen to prevent overfitting to the selected subset for the video validation and was applied separately for each cleaning method and each housing system: The full raw tracking data was split randomly in a train (BB: 40 datapoints/housing system; LH: 108 datapoints) and test (BB: 20 datapoints/housing system; LH: 54 datapoints) sample. Every data point represented one individual during one observation interval. With the train sample, a random subsample of 2/3 of all train-datapoints was produced during a run and in total 100 runs were performed (Fig. 2). Each train-datapoint appeared in several runs but each run used a unique randomly selected combination of tags and observation intervals. For every run, all 30 values of the duration parameter were applied and the reliability was calculated for each duration parameter value as well as the raw tracking data. Only those duration parameter values with a reliability above the required minimum of 0.8 (for total transitions) or 0.9 (for locations) were selected (see statistics Section 2.7). Duration parameter values that met the selection criterion of stability (reliable in at least 75 of 100 runs) were then considered in the determination of the duration parameter range. Finally, the determined duration parameter range within each cleaning method was validated by applying all values contained in the duration parameter range to the test sample and calculating the respective reliabilities.

Based on the cross-validation results, one cleaning method with one duration parameter value meeting the selection criterion was chosen as an example to demonstrate the post-cleaning performance for different housing systems.

To better understand the difference between true and false registrations by the tracking system, the duration of true transitions (duration of stay in a new zone before moving to another zone) observed in the videos and the duration of false transitions (registered by the tracking system but not observed in the video) was extracted from the raw tracking data.

2.7. Statistical analysis

All statistical evaluations were performed in R (4.1.0; R Core Team, 2021). To assess initial and post-cleaning performance (Fig. 2), a confusion matrix was created and the reliability of locations and total transitions was calculated before and after data cleaning. To measure transitions into any new zone, we extracted the following variables from the confusion matrix: the accuracy (the number of correct registrations vs. all registrations and observations), precision (the number of true registratered transitions vs. all registered transitions), and sensitivity (the number of true registered transitions vs. all observed transitions). A high sensitivity would imply that most observed transitions were registerd by the tracking system. However, a high sensitivity is not sufficient for good performance of the tracking system. If the precision would be low, the system would, in addition to the true transitions, register false transitions, i.e. transitons which never actually happened. Thus, a high sensitivity, as well as high precision are prerequisites for a reliable tracking system. For further understanding of zone-specific recording accuracy, the overall zone-specific accuracy as well as the zone-specific precision and sensitivity were extracted.

Locations and total transitions were further evaluated by calculating their reliability. We intended to not only detect over- or underestimations of transitions (given by the extracted variables of the confusion matrix) but also the "drift" (i.e., increasing bias with increasing true value) of the data from a line of perfect match (McBride, 2005). Therefore, Lin's Concordance Correlation Coefficient (CCC) was calculated for the reliability of total transitions, where values near + 1 indicate strong agreement (Lin, 1989; Chen and Barnhart, 2008; Akoglu, 2018). The CCC and its confidence interval was estimated by U-statistics to account for repeated measurements within each bird, the non-normal data distribution, and small sample size (King et al., 2007; Carrasco et al., 2013). For locations, the proportion of match was calculated by dividing the duration of matching zones (explained in Data processing, Section 2.6) by the total duration of each observation interval (20 min). Our goal was to achieve a mean proportion of match of at least 0.9 for locations, thus allowing for an average error of 2 min within a 20 min interval, and an estimated CCC of > 0.8 for total transitions based on existing literature, though the suggested threshold varies between authors and applications (McBride, 2005; Akoglu, 2018).

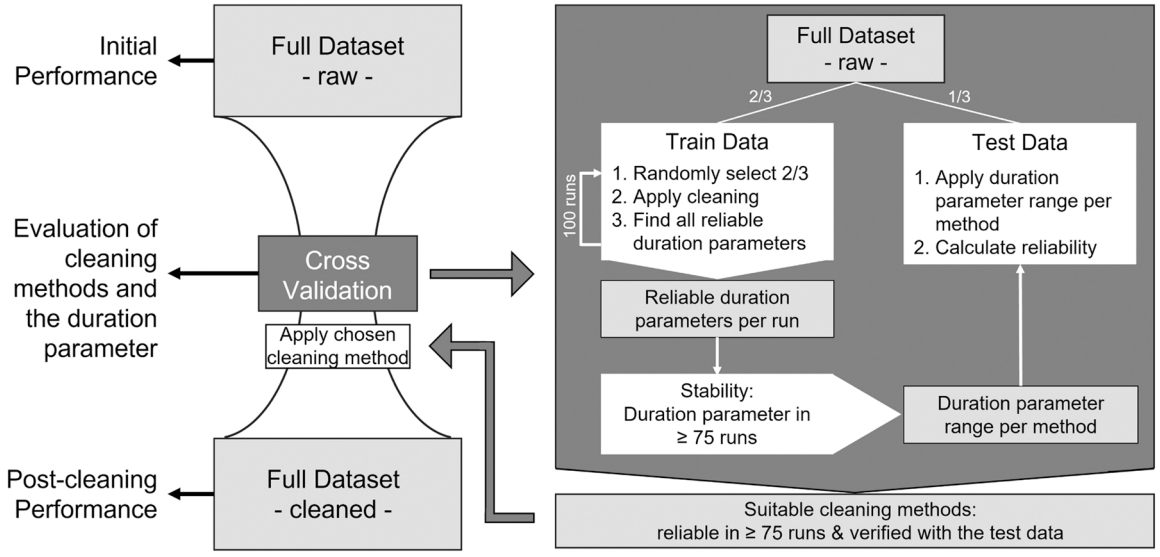

Fig. 2. Schematic description of data processing and analysis. The data processing and analysis consisted of three steps: Initial performance of the tracking system, the evaluation of cleaning methods and determination of suitable duration parameters and the post-cleaning performance. The process of the cross-validation is explained in detail in the dark grey box. White boxes represent data processing and light grey boxes show the resulting datasets.

For the duration of stay after a true transition compared to the duration of stay after a false transition, a negative binomial generalised linear mixed model with a log-link function was performed with duration between transitions as the response variable. Housing system and the type of transition (true vs false registrations) were included as fixed factors while individual and observation interval were included as crossed random factors. The residuals of the model were visually inspected for overdispersion and homoscedasticity. A stepwise model reduction was used to evaluate the significance of each fixed factor. The model estimates were backtransformed by exponentiation.

3. Results

3.1. Cross-Validation

For all 100 runs on locations at least one value of the duration parameter was producing reliable data (mean proportion of match $\geq$ 0.9) for all three housing systems and all three cleaning methods. A high stability ($\geq$75 runs) could be found for all cleaning methods and for several or all values of the duration parameter for Aviary (Filtering: 0 – 120 s; Binning: 0 – 100 s; Sliding Bin: 0 – 300 s), Raised (Filtering: 0 – 130 s; Binning: 0 – 120 s and 140 −170 s; Sliding Bin: 0 – 300 s) and Floor (Filtering: 0 – 100 s; Binning: 0 – 300 s; Sliding Bin: 0 – 300 s). According to the test sample, the stable values of the duration parameter of each cleaning method produced reliable locations (minimum mean proportion of match) for Aviary (Filtering – 0.956, Binning – 0.945, Sliding Bin – 0.952), Raised (Filtering – 0.946, Binning – 0.931, Sliding Bin – 0.956) and Floor (Filtering – 0.961, Binning – 0.972, Sliding Bin – 0.972).

For total transitions, reliability (CCC > 0.8) was achieved in all 100 runs for Aviary, but not in all 100 runs of Raised (Filtering: 84 runs; Binning: 88 runs; Sliding Bin: 74 runs) and Floor (Filtering: 74 runs; Binning: 94 runs; Sliding Bin: 57 runs). A high stability ($\geq$75 runs) was achieved by several values of the duration parameter for Filtering (20 – 40 s), Binning (30 – 80 and 100 s) and Sliding Bin (50 −100 s) in Aviary (Fig. 3). Only one value of the duration parameter was stable for Filtering (20 s) and Binning (60 s) in Raised and for Binning (20 s) in Floor (Fig. 3). Those values of the duration parameter producing stable results in total transitions were also reliable in the test sample for Aviary (Filtering CCC $\geq$ 0.847, Binning: CCC $\geq$ 0.839, Sliding Bin: CCC $\geq$ 0.840) and Raised (Filtering: CCC = 0.825, Binning: CCC = 0.835) but not for Floor (Binning: CCC = 0.355). For Aviary and Raised, values of the duration parameter producing stable and confirmed (by the test sample) results in total transitions are also producing stable and confirmed results in locations.

Based on the results of the cross-validation, the following cleaning method and value of the duration parameter was chosen as an example to demonstrate the post-cleaning performance: Filtering with a duration parameter of 30 s for Aviary and Filtering with a duration parameter of 20 s for Raised. Because results of Floor did not achieve the pre-defined selection criterion, post-cleaning performance was not assessed. All results of the cross-validation are summarised in the supplementary S2.

3.2. Initial and post-cleaning performance

For comparison, the results of initial and post-cleaning evaluation are presented in the same subsection, though consider two different datasets (raw full dataset of Aviary, Raised and Floor for initial performance and cleaned full dataset of Aviary and Raised for post-cleaning performance; Fig. 2).

In all three housing systems, high sensitivity and low precision was observed for initial performance, though precision was particularly low ($\leq$ 0.105) in both BB housing systems (Table 2). Based on the cross-validation, cleaning was successful for Aviary and Raised. The applied cleaning method greatly improved the precision of detecting transitions but lowered the sensitivity in Aviary and Raised (Table 2). The zone-specific overall accuracy was higher in LH compared to both BB housing systems and improved for both Aviary and Raised after cleaning (Table 2). The sensitivity and precision varied between different zones in all three housing systems before and after cleaning (values for all zones can be found in the supplementary S3). The lowest precision was found for the same zones in initial performance (Aviary: Zone 3 – 0.601; Raised: Zone 2 – 0.177; Floor: Zone 2 – 0.133) and post-cleaning performance (Aviary: Zone 3 – 0.795; Raised: Zone 2 – 0.659). The lowest

Table 2
The accuracy, sensitivity and precision of tracking system registrations. The table shows the initial performance ("Raw") and the post-cleaning performance ("Cleaned"; Filtering with a duration parameter of 30 s for Aviary and 20 s for Raised). As none of the tested (cross-validation) cleaning methods yielded reliable and stable results for Floor, no cleaned variables can be shown here.

	Aviary (LH)		Raised (BB)		Floor (BB)	
	Raw	Cleaned	Raw	Cleaned	Raw	Cleaned
Accuracy	0.694	0.807	0.506	0.567	0.523	–
Sensitivity	0.952	0.766	0.958	0.573	0.979	–
Precision	0.398	0.794	0.081	0.543	0.105	–
Zone-specific overall accuracy	0.757	0.843	0.558	0.763	0.571	–

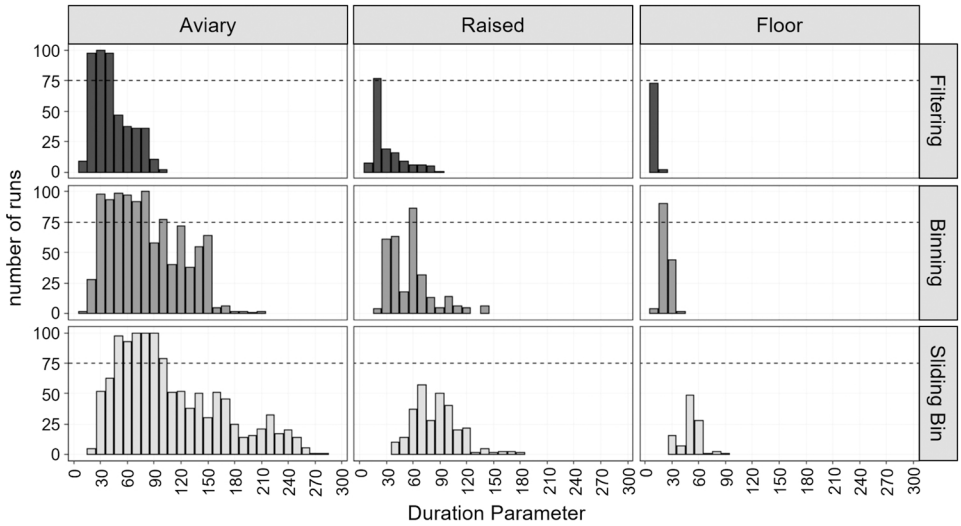

Fig. 3. The results of the cross-validation approach for total transitions. The y-axis shows the number of subsampling runs (total runs = 100) in which a given value of the duration parameter (x-axis) has produced stable total transitions for Filtering (black), Binning (dark grey) and Sliding Bin (light grey). Each run uses a different subsample containing 2/3 randomly drawn datapoints. Thus, each subsample has a unique combination of observation intervals (varying in day and timepoint) and tracking tags. To represent a high stability against variation in observation interval and individual, a minimum of 75 runs (dashed line) had to be achieved by a given value of the duration parameter.

sensitivity was found for Zone 2 in LH (0.675) and Zone 4 in BB (Raised: 0.500; Floor: 0.534) in initial performance and for Zone 4 in Aviary (0.661) and Zone 3 in Raised (0.375) in post-cleaning performance.

The initial performance of locations measured as mean proportion of match [95% confidence interval] was above the required minimum of 0.9 in all three housing systems (Aviary: 0.949 [0.923, 0.974]; Raised: 0.957 [0.939, 0.974]; Floor: 0.969 [0.954, 0.982]) and remained high in the post-cleaning performance (Aviary: 0.949 [0.924, 0.974]; Raised: 0.965 [0.947, 0.983]). However, the initial performance of total transitions measured as Lin's CCC [95% confidence interval] was below the required 0.8 in all three housing systems and praticularily low in both BB housing systems (Fig. 4; Aviary: 0.264 [0.143, 0.377]; Raised: 0.023 [−0.190, 0.234]; Floor: 0.064 [−0.229, 0.346]). The post-cleaning performance of total transitions (Lin's CCC) increased compared to initial performance and exceeded the minimum of 0.8 (Fig. 4; Aviary: 0.888 [0.841, 0.921]; Raised: 0.832 [0.768, 0.880]).

3.3. Durations of stay after true and false transitions (raw tracking data)

The duration of stay after a transition was affected by housing system ($p < 0.001$; $X^2 = 18.176$; $df = 2$), as well as by true vs. false transitions ($p < 0.001$; $X^2 = 1231.9$; $df = 1$). Individuals stayed longer (model estimate in seconds [95% confidence interval]; 225.8 s [179.9, 284.8], 119.1 s [67.2, 212.6] and 92.2 s [51.6, 166.4] in Aviary, Raised and Floor, respectively) in a zone after true transitions compared to false transitions (Fig. 5).

4. Discussion

While the initial performance for locations was adequate, total transitions showed a poor reliability and differed by housing system. Our results suggest that the tracking system was overestimating the number of transitions. For all transitions registered by the tracking system, less than 40% in LH and only 10% in both BB housing systems were verified in video recordings. Additionally, the low CCC values may indicate that a bird with high vs. a bird with low registered number of transitions may reflect tracking system errors rather than true behavioural differences. Therefore, ranked values are probably also not reliable.

The initial performance in measuring locations differed from the initial performance in total transitions, thus the validation of only one of the movement aspects would not have been sufficient. The difference in performance implies that it is necessary to validate sensory data for all

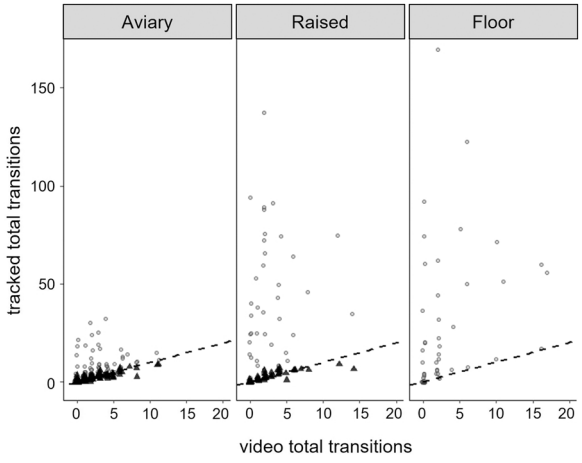

Fig. 4. The number of transitions for the tracking System (y-axis) and for the video observations (x-axis). The initial performance is shown in light grey circles and the post-cleaning performance in dark grey triangles. The dashed black line represents a perfect match (CCC = 1) between the tracking system and the video observations.

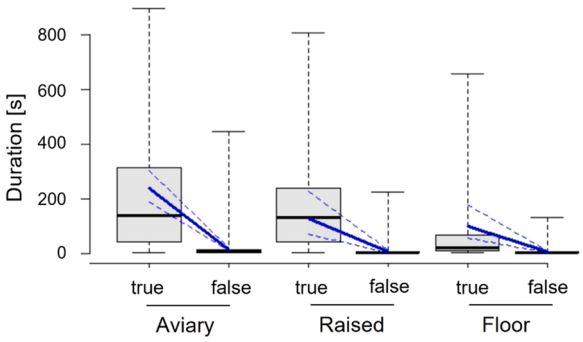

Fig. 5. The duration of stay before moving on to a next zone after a true or false transition. The boxplots represent the raw data with the blue lines showing the model estimates (solid line) and the 95%-confidence intervals (dashed lines).

distinct variables of interest. Moreover, the design of the housing system affected initial performance. In both BB housing systems, precision, CCC, and zone-specific overall accuracy were low compared to LH, which might be explained by the organisation of zones being mostly vertical in LH but horizontal in BB. Intuitively, the BB barn configuration appears simpler and previous studies have found high accuracy and correlation of automatically measured movement paths in floor pens (Rodenburg et al., 2017; van der Sluis et al., 2020). However, our tracking system measured transitions between zones rather than the movement path of an animal as in these previous efforts. This may suggest that a tracking system measuring an animals movement path may be advantageous for floor pens, whereas a tracking system using large, resource-related zones might be more suitable for commercial aviaries. Ultimately, the selection of tracking technology and the type of data generated will also depend on the research question. In areas of similar distance to marker cables, the signal strengths of markers may be similar causing the recorded zone to flicker back and forth between two zones in short intervals, rather than causing a long lasting false recorded zone. In the BB setups of the current study, the distance between zones, and thus marker cables, was smaller compared to LH and birds could be physically present between two zones. For LH, it would be difficult to be located between two zones due to the vertical configuration of zones, however tags may still receive similar signal strengths of two markers leading to erroneous data. Accordingly, our study found differences in zone-specific precision and sensitivity with particularly low initial precision in those zones being relatively small and close to neighbouring zones.

Indeed, the duration of stay after true transitions was longer compared to the duration of stay after false transitions in all three housing systems. All studied cleaning methods were based on a duration parameter (duration of stay after a transition, that we evaluated across a range of values) and were thus likely to be effective given the clear differentiation of true and false registrations. For demonstration of cleaning effectiveness, Filtering was applied for Aviary and Raised and showcased a considerable improvement of post-cleaning performance compared to the initial performance, particularily for Aviary. In Raised, a reliable CCC was achieved with Filtering, the accuracy and precision however remained insufficient. The effectiveness of cleaning was dependent on finding a suitable duration parameter that differentiatiated between false and true registrations. A duration parameter that is too short may remove real meaningful behaviour (Rushen et al., 2012), such as fast movements between zones during panics in poultry (Mills and Faure, 1990; Richards et al., 2012). The removal of true registrations increased the number of False Negatives in the cleaned dataset and thus decreased sensitivity.

In Filtering and Binning, short duration parameters led to reliable locations in all, but long duration parameters only in few runs, while in Sliding Bin all 30 values of the duration parameter maintained a high location reliability in almost all runs of the cross-validation. Filtering

seemed to be less sensitive to short, true locations compared to the other two cleaning methods, while Binning used fixed windows and thereby might have shifted the timing of true transitions by up to half of the duration parameter. In Aviary, fewer values of the duration parameter were suitable for Filtering compared to Binning and Sliding Bin, therefore a precise assessment of the most suitable duration parameter is especially important for Filtering. Overall, Filtering provides the simplest, yet effective cleaning though needs a precise assessment of a suitable duration parameter. The evaluations of cleaning procedures within the current study may benefit future efforts of similar tracking systems but would need to be validated with respect to the posed research question, the movement aspects of interest, the housing system, and the associated hardware configuration (e.g. by applying the validation procedures shown in our study).

In addition, the chosen duration parameter and cleaning method also must achieve high stability against variation in performance between sensors and timepoints, given that validations are typically done on a subsample of timepoints and/or individuals and thus sensors (e.g. Maselyne et al., 2014; Rufener et al., 2018; Zhuang et al., 2020). A low stability may risk the determination of a duration parameter being appropriate for the subsample used for validation though less so for the entire dataset. In the present study, some values of the duration parameter were indeed shown to only be suitable for a few specific combinations of sensors and timepoints. In this case, subsequent analyses of the full dataset may be strongly biased by variation in performance of the tracking tag. Stability seems particularly important when investigating individual movement, as a difference in registered transitions between individuals may be an artefact of tracking performance (such as variation in tag accuracy and precision) rather than an actual behavioural difference of the animals. In support of this possibility, varying sensor error distributions were also found by Ren et al. (2020). Further research on sensor technology used for individual behavioural assessments should consider quantifying the variance between sensors in the behavioural measure of interest. An extended video validation with a high number of individuals may give further insight as to the influence of the individual on sensor performance and consistency over time.

Despite the similar initial performance of both BB housing systems, a reliable and stable data cleaning method could only be found for Raised but not for Floor during the cross-validation procedure. The success of cleaning might be affected by the scale of difference between true and false registrations. The duration of stay after a true transition was shorter in Floor compared to Raised representing faster movement of BB between zones in pens with slats and nests on the floor compared to those with raised slats. True transitions and thus the distinction of true and false recordings was highest in Aviary, consequently several values of the duration parameter were effective to clean Aviary. Creating conditions, including physical barriers that slow down the movement leading to easier differentiation between true and false registrations, will improve ability to yield a more accurate dataset. For instance, the slower movement of LH between aviary tiers that required vertical transitions eased the process of finding an effective threshold for the duration parameter. For other investigators evaluating use of this technology, we recommend taking into account the speed of transitioning under normal circumstances (e.g., birds in an aviary will move slower between zone than in a single tier system) and whether artificially slowing birds would be suitable (e.g., introducing barriers). However, the latter depends on the posed research question as it may affect the behaviour of the birds.

Our study demonstrated a high relevance for validating all aspects of movement, not only in the raw output but also after applying data cleaning techniques to gain reliable data from automated behavioural tracking. The implications and benefits are particularly relevant when measuring individual movement in complex commercial housing systems as it highlights the need for differentiation between individual animal behaviour vs. sensor performance. The clear differentiation of false from true registrations allows the application of a relatively simple and easy to perform data cleaning method that nonetheless provides a drastic improvement in the reliability of total transitions. Despite these encouraging findings, it is important to validate the cleaning procedure before applying it to the full dataset to determine the most effective balance of elimination of false registrations against retention of true registrations. While the initial performance of the tracking system might be explained by the barn configuration, the ease of cleaning the data and generating accurate records relies on the distinction of true and false registrations and is thus influenced by the speed of birds when transitioning between zones. While the use of a tracking system may provide novel and highly valuable insight into individual birds' movement and resource usage within commercial settings and enables the link of bird movement to health and welfare, it is crucial to thoroughly validate the generated data and to evaluate the application of cleaning procedures before interpreting individual bird movement in the context of welfare.

Acknowledgements

This work was supported by the Swiss Federal Food Safety and Veterinary Office, Switherland [Project # 2.19.02]. The authors thank the Aviforum staff for taking care of the laying hens and broiler breeders as well as Alexander Kashev for his effort and support in developing and programming the data cleaning methods. We are grateful to Markus Schwab, Thomas Heinzel and Abdulsatar Abdel Rahman for their technical support as well as the specialised support from members of the Gantner Solutions GmbH (Austria) team including Michael Gantner. Furthermore, we would like to thank Doriana Sportelli, Nicola Knuchel, Mélina Chevalley, Patrick Hosmann and Mark Soltermann for their help in video coding, as well as Anne van den Oever for her support in the practical work..

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.applanim.2022.105597.

References

Akoglu, H., 2018. User's guide to correlation coefficients. Turk. J. Emerg. Med. 18 (3), 91–93.

Blokhuis, H.J., Veissier, I., Miele, M., Jones, B., 2010. The welfare quality® project and beyond: safeguarding farm animal well-being. Acta Agric Scand Sect A Anim Sci 129–140.

Broom, D.M., 2010. Animal welfare: an aspect of care, sustainability, and food quality required by the public. J. Vet. Med. Educ. 83–88.

Carrasco, J.L., Phillips, B.R., Puig-Martinez, J., King, T.S., Chinchilli, V.M., 2013. Estimation of the concordance correlation coefficient for repeated measures using SAS and R. Comput. Methods Programs Biomed. 293–304.

Catarinucci, L., Colella, R., Mainetti, L., Patrono, L., Pieretti, S., Secco, A., et al., 2014. An animal tracking system for behavior analysis using radio frequency identification. Lab Anim. 43 (9), 321–327.

Chen, C.C., Barnhart, H.X., 2008. Comparison of ICC and CCC for assessing agreement for data without and with replications. Comput Stat Data Anal 554–564.

Cornou, C., Lundbye-Christensen, S., Kristensen, A.R., 2011. Modelling and monitoring sows' activity types in farrowing house using acceleration data. Comput. Electron. Agric. 316–324.

Dawkins, M.S., 2003. Behaviour as a tool in the assessment of animal welfare. Zoology 106 (4), 383–387.

Dawkins, M.S., Cain, R., Roberts, S.J., 2012. Optical flow, flock behaviour and chicken welfare. Anim. Behav. 84 (1), 219–223 (Available from). ⟨https://www.sciencedirect.com/science/article/pii/S0003347212002102⟩.

DeVries, T.J., von Keyserlingk, M.A.G., Weary, D.M., Beauchemin, K.A., 2003. Technical note: validation of a system for monitoring feeding behavior of dairy cows. J. Dairy Sci. 86 (11), 3571–3574.

Ellen, D.E., van der Sluis, M., Siegford, J., Guzhva, O., Toscano, J.M., Bennewitz, J., et al., 2019. Review of sensor technologies in animal breeding: phenotyping behaviors of laying hens to select against feather pecking. Anim.

Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant. Methods Psychol. 8 (1), 23–34.

King, T.S., Chinchilli, V.M., Carrasco, J.L., 2007. A repeated measures concordance correlation coefficient. Stat. Med. 26 (16), 3095–3113. https://doi.org/10.1002/sim.2778.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159.
Ledgerwood, D.N., Winckler, C., Tucker, C.B., 2010. Evaluation of data loggers, sampling intervals, and editing techniques for measuring the lying behavior of dairy cattle. J. Dairy Sci. 5129–5139.
Lin, L.I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 255–268. ⟨http://www.jstor.org/stable/2532051⟩.
Main, D.C.J., Mullan, S., Atkinson, C., Cooper, M., Wrathall, J.H.M., Blokhuis, H.J., 2014. Best practice framework for animal welfare certification schemes. Trends Food Sci. Technol. 127–136.
Maselyne, J., Saeys, W., De Ketelaere, B., Mertens, K., Vangeyte, J., Hessel, E.F., et al., 2014. Validation of a high frequency radio frequency identification (HF RFID) system for registering feeding patterns of growing-finishing pigs. Comput. Electron Agric. 102, 10–18.
McBride, G.B., 2005. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. NIWA client Rep HAM2005-062. Natl. Inst. Water Atmos. Res. Hamilt., N. Z. 62.
Mendes, E.D.M., Carstens, G.E., Tedeschi, L.O., Pinchak, W.E., Friend, T.H., 2011. Validation of a system for monitoring feeding behavior in beef cattle. J. Anim. Sci. 89 (9), 2904–2910.
Mills, A.D., Faure, J.M., 1990. Panic and hysteria in domestic fowl: a review. Soc. Stress Domest. Anim. Kluwer Academic Publishers, Dordrecht, pp. 248–272.
Neethirajan, S., 2017. Recent advances in wearable sensors for animal health management. Sens. Bio Sens. Res. 12, 15–29.
van den Oever, A.C.M., Candelotto, L., Kemp, B., Rodenburg, T.B., Bolhuis, J.E., Graat, E. A.M., et al., 2021. Influence of a raised slatted area in front of the nest on leg health, mating behaviour and floor eggs in broiler breeders. Animal 15 (2), 100109.
R Core Team, 2021R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ⟨https://www.r-project.org/⟩.
Ren, K., Karlsson, J., Liuska, M., Hartikainen, M., Hansen, I., Jørgensen, G.H.M., 2020. A sensor-fusion-system for tracking sheep location and behaviour. Int. J. Distrib. Sens. Networks.
Richards, G.J., Brown, S.N., Booth, F., Toscano, M.J., Wilkins, L.J., 2012. Panic in free-range laying hens. Vet. Rec. 519. https://doi.org/10.1136/vr.100685.
Richter, S.H., Hintze, S., 2019. From the individual to the population – and back again? Emphasising the role of the individual in animal welfare science. Appl. Anim. Behav. Sci. 212, 1–8.
Rodenburg, T.B., Bennewitz, J., De Haas, E.N., Košťál, L., Pichová, K., Piette, D., et al., 2017. The use of sensor technology and genomics to breed for laying hens that show less damaging behaviour. 8th Eur. Conf. Precis Livest. Farming 532–541. Nantes, France; 2017.
Rowe, E., Dawkins, M.S., Gebhardt-Henrich, S.G., 2019. A systematic review of precision livestock farming in the poultry sector: is technology focussed on improving bird welfare? Animals.
Rufener, C., Berezowski, J., Maximiano Sousa, F., Abreu, Y., Asher, L., Toscano, M.J., 2018. Finding hens in a haystack: Consistency of movement patterns within and across individual laying hens maintained in large groups. Scientific Reports 8 (1). https://doi.org/10.1038/s41598-018-29962-x.
Rufener, C., Abreu, Y., Asher, L., Berezowski, J.A., Maximiano Sousa, F., Stratmann, A., et al., 2019. Keel bone fractures are associated with individual mobility of laying hens in an aviary system. Appl. Anim. Behav. Sci.
Rushen, J., Chhapinal, N., De, Passille, M.A.B., 2012. Automated monitoring of behavioural-based animal welfare indicators. Anim. Welf. UFAW J. 21 (3), 339–350.
Siegford, J.M., Berezowski, J., Biswas, S.K., Daigle, C.L., Gebhardt-Henrich, S.G., Hernandez, C.E., et al., 2016. Assessing activity and location of individual laying hens in large groups using modern technology. Animals 6 (2).
Stratmann, A., Fröhlich, E.K.F., Gebhardt-Henrich, S.G., Harlander-Matauschek, A., Würbel, H., Toscano, M.J., 2015. Modification of aviary design reduces incidence of falls, collisions and keel bone damage in laying hens. Appl. Anim. Behav. Sci.
Triguero-Ocaña, R., Vicente, J., Acevedo, P., 2019. Performance of proximity loggers under controlled field conditions: an assessment from a wildlife ecological and epidemiological perspective. Anim. Biotelem. BioMed Cent. 7 (1), 1–9. https://doi.org/10.1186/s40317-019-0186-2.
van der Sluis, M., de Haas, Y., de Klerk, B., Rodenburg, T.B., Ellen, E.D., 2020. Assessing the activity of individual group-housed broilers throughout life using a passive radio frequency identification system—a validation study. Sensors.
Van Der Sluis, M., De Klerk, B., Ellen, E.D., De Haas, Y., Hijink, T., Rodenburg, T.B., 2019. Validation of an ultra-wideband tracking system for recording individual levels of activity in broilers. Animals 9 (8), 1–15.
Winckler, C., 2019. Assessing animal welfare at the farm level: do we care sufficiently about the individual. Anim. Welf. 28 (1), 77–82.
Zhuang, S., Maselyne, J., Van Nuffel, A., Vangeyte, J., Sonck, B., 2020. Tracking group housed sows with an ultra-wideband indoor positioning system: a feasibility study. Biosyst. Eng. 176–187.