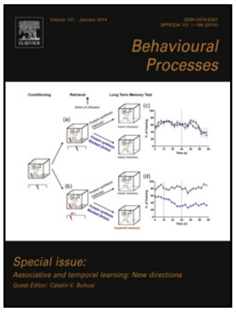
Accepted Manuscript



Title: Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle

Authors: Ruuska Salla, Hämäläinen Wilhelmiina, Kajava Sari, Mughal Mikaela, Matilainen Pekka, Mononen Jaakko

PII: S0376-6357(17)30114-6
DOI: https://doi.org/10.1016/j.beproc.2018.01.004
Reference: BEPROC 3578

To appear in: *Behavioural Processes*

Received date: 15-3-2017
Revised date: 5-1-2018
Accepted date: 8-1-2018

Please cite this article as: Salla, Ruuska, Wilhelmiina, Hämäläinen, Sari, Kajava, Mikaela, Mughal, Pekka, Matilainen, Jaakko, Mononen, Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle.Behavioural Processes https://doi.org/10.1016/j.beproc.2018.01.004

Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle

Ruuska Salla[a,1,2,] *Hämäläinen Wilhelmiina*[b], *Kajava Sari*[c], *Mughal Mikaela*[d], *Matilainen Pekka*[a] *and Mononen Jaakko*[c]

[a] University of Eastern Finland Department of Environmental and Biological Sciences
P.O.Box 162770211 KUOPIO, FINLAND
email: salla.ruuska@uef.fi; pekka.matilainen@uef.fi

[1] Corresponding author

[2] Present address: Savonia University of Applied Sciences P.O.Box 72 74101 Iisalmi, FINLAND
e-mail: salla.ruuska@savonia.fi

[b] Aalto University Department of Computer Science
P.O.Box 15400 00076 Aalto
email: wilhelmiina.hamalainen@aalto.fi

[c] Natural Resources Institute Finland Green Technology Halolantie 31 A, 71750 Maaninka, FINLAND
sari.kajava@luke.fi; jaakko.mononen@luke.fi

[d] University of Eastern Finland Department of Environmental and Biological Sciences
P.O.Box 111\ 80101 JOENSUU, FINLAND mikaela.mughal@uef.fi

Highlights

- Device validation with confusion matrices was evaluated empirically
- Assessments by confusion matrices, error indices and linear regression were compared
- Confusion probabilities help to analyse reasons for errors and their importance

Abstract

The aim of the present study was to evaluate empirically confusion matrices in device validation. We compared the confusion matrix method to linear regression and error indices in the validation of a device measuring feeding behaviour of dairy cattle. In addition, we studied how to extract additional information on classification errors with confusion probabilities. The data consisted of 12 h behaviour measurements from five dairy cows; feeding and other behaviour were detected simultaneously with a device and from video recordings. The resulting 216 000 pairs of classifications were used to construct confusion matrices and calculate performance measures. In addition, hourly durations of each behaviour were calculated and the accuracy of measurements was evaluated with linear regression and error indices. All three validation methods agreed when the behaviour was detected very accurately or inaccurately. Otherwise, in the intermediate cases, the confusion matrix method and error indices produced relatively concordant results, but the linear regression method often disagreed with them. Our study supports the use of confusion matrix analysis in validation since it is robust to any data distribution and type of relationship, it makes a stringent evaluation of validity, and it offers extra information on the type and sources of errors.

Keywords: *validation, confusion matrix, linear regression, error indices, confusion probabilities, feeding behaviour*

1. Introduction

The use of advanced technology for automated measurement of animal behaviour is increasing, and the accuracy of new devices has to be ensured in validation studies. However, there is no agreement on how validation should be done. In addition, many statistical techniques are sensitive to underlying assumptions and can produce misleading results, if the assumptions are not fulfilled. Therefore, it is always desirable to validate devices with several alternative methods and check their agreement.

The goal of validation is to guarantee that future measurements in similar situations are sufficiently accurate. When behaviour is measured, an important question is the time resolution of measurements. In one extreme, one may require that behaviour is classified accurately at every instant, while in other applications it suffices that hourly or daily durations are accurate. If consecutive classifications (i.e., results from every ‘instant’ or second) are available, then it is always possible to calculate durations of behaviour during longer time intervals and perform validation at any desired resolution.

Previous validation studies have mostly concentrated on validating numerical behaviour measurements like hourly durations spent on a certain behaviour. The strength of this relationship has been evaluated with correlation analysis (Büchel and Sundrum, 2014), linear regression (Chizzotti et al., 2015), a combination of these two methods (Elischer et al., 2013; Schirmann et al., 2009) or different kinds of error indices (Schirmann et al., 2009) like average, minimum, maximum and variance of error or error magnitude. When the underlying discrete classifications are available, an alternative approach is to validate classification accuracy with confusion matrices (Nielsen, 2013; DeVries et al., 2003). A confusion matrix presents information about how often a certain behaviour is detected correctly and how often it is classified as another behaviour. The classification accuracy is usually summarized by performance indicators like precision, sensitivity and specificity.

The aim of the present study was to evaluate the confusion matrix method in validation. For this purpose, we validated a halter device measuring eating, ruminating and drinking behaviour of cattle. The first objective was to compare the confusion matrix method with basic error indices and linear regression analysis in the validation of RWS. The main question was whether all three methods produce concordant assessments of the device accuracy. It was also hypothesized (in line with previous research, Miller-Cushon and DeVries, 2011) that the confusion matrix method could reveal more errors since it analyzes finer-grained measurements (behavior classifications on each second) while the latter two methods require coarser-grained measurements (hourly durations of each behaviour). The second objective was to analyse the confusion matrices in detail to find out reasons for errors. For this purpose, we introduce two types of confusion probabilities that extend the classical notions of precision and sensitivity and help in interpreting the matrix. In addition, we demonstrate how all available information on true behaviour can be utilized in the confusion matrix analysis even if the device measures only a few main classes. This gives valuable information on error-prone situations that should be targeted in further development of the device.

2. Materials and methods

2.1 Multiclass classification and confusion matrices

In this paper, we concentrate on a discrete multiclass classification task, where one should assign each observation into one of predefined classes $C_1,...,C_k$. Alternatively, one may perform probabilistic classification, where each observation is assigned a probability distribution describing probabilities of belonging into any of the given classes. Methods that perform the classification are called classifiers. Before new classifiers can be taken into use, their accuracy should be evaluated experimentally, by comparing the inferred classifications (measured by a new device or method) against the real classifications or gold standard classifications (measured by a benchmark method) with a sufficiently large test data set that is representative of expected future data.

Results of classifier validation studies are often presented as confusion matrices. A confusion matrix for k-class classification is a k x k contingency table whose cells [i,j] (i=1,...,k, j=1,...,k) present frequencies of observations with real class C_i and inferred class C_j. A binary confusion matrix is a special case when there are only two classes: C (positive class) and not-C (negative class). A k x k confusion matrix can always be represented as a set of k binary confusion matrices, one for each class C_i. An example is shown in Figure 1. In practice, explicit transformation into binary confusion matrices is not necessary, but it helps to demonstrate how different classification performance indicators are calculated from a multiclass confusion matrix.

In a binary confusion matrix (see Figure 1, on the right), observations classified correctly into the positive class are called true positives and observations classified correctly into the negative class are called true negatives. Instances of the positive class classified falsely as negative are called false negatives and instances of negative class classified falsely as positive are called false positives. Numbers of true positive, false positive, true negative and false negative observations are notated by TP, FP, TN and FN. From these frequencies, one can calculate classification performance indicators that reflect how the classifier performs in detecting the given class. The most common of such indicators are precision = TP/(TP+FP), sensitivity = TP/(TP+FN), specificity = TN/(TN+FP), and accuracy=(TP+TN)/(TP+TN+FP+FN).

In this paper, we also introduce a new way to analyze confusion matrices with two types of confusion probabilities. Both confusion probabilities, cp1 and cp2, can be calculated directly for each cell of a multiclass confusion matrix. The first confusion probability, cp1= P(real=B|measured=A), gives the proportion of instances classified as class A but actually belonging to class B. The second confusion probability, cp2=P(measured=B|real=A), gives the proportion of actual A instances that were classified as B. For any class A, its cp1 values in the same column reveal the main reasons for false positive errors and its cp2 values in the same row reveal the main reasons for false negative errors. When A=B, cp1 reduces to precision and cp2 to sensitivity.

Confusion probabilities can also be calculated for an extended confusion matrix where the main classes are divided into subclasses. Often the classifier can detect only the predefined main classes but the real classes in behavioural studies can be divided into detailed subclasses based on video observations. With confusion probabilities, one can use all available information and trace which subclasses were responsible for errors. Here, cp1 is especially useful because it shows the FP error distribution. Given B's subclasses, B_1,...,B_m, comparison of cp1=P(real=B_i|measured=A) shows which subclasses were responsible for A's false positive errors. On the other hand, given A's subclasses A_1,...,A_m, cp2=P(measured=A|real=A_i) shows which subclasses were detected most or least sensitively.

2.2 Data collection and formatting

The data was originally collected for a validation study (Ruuska et al., 2016) of a pressure sensor-based system (RumiWatch System, RWS, Itin + Hoch GmbH, Liestal, Switzerland) for measuring eating, ruminating, and drinking behaviour of cattle. By default, the RWS device reports hourly durations of behaviours and these estimates were used in the previous study. However, RWS offers also 10 Hz raw data consisting of ten behaviour classifications per second, and this data was used in the current study. The operating principles of RWS, as well as the experimental setup, collecting the RWS data, extracting correct behaviour classifications from video recordings (gold standard) and analysing the data with a linear regression method have been described in detail in the previous paper (Ruuska et al., 2016).

In the primary validation study six non-lactating dairy cows were housed in tie stalls and fed with grass silage (dry matter 22-26%) and water was provided *ad libitum*. Cows were equipped with RWS sensors and their behaviour was video recorded for 24 hours. Two trained observers monitored video recordings with continuous recording.

In this study, we used 12 h (43 200 s) observational and RWS measurement data from five animals in the original study (excluding one animal with missing RWS data). The RWS data consisted of classifications of the behaviour as eating, ruminating, drinking, or other behaviour (class 'other') with 10 Hz frequency for the whole 12 hours. Since the gold standard classifications were presented with 1 s resolution, the mode of 10 consecutive RWS classifications was calculated to represent the RWS's behavioural classification for that second. The observational data from video recordings was more detailed and the main classes for eating, ruminating, and other behaviour were divided into 32 subclasses according to body posture, feed type, and other activities which could possibly get confused with eating (like licking the feeding table, masticating feed, and grooming with the mouth) (Table 1). We note that in this study we used a strict definition of eating that excluded all related behaviour. The resulting 1 Hz resolution categorical data (43200 rows per animal) was used in the confusion matrix analysis. In addition, we created another version of the data, by calculating durations of eating, ruminating, drinking, and other behaviour in each hour. The resulting data consisted of 12 rows per animal, each row containing eight numerical variables, the real duration X and RWS estimate Y for each of the four behaviour classes. This numerical data was used for calculating error indices and fitting linear regression models.

2.3 Data analysis

In the data analysis, we analysed the categorical data with confusion matrices and the corresponding numerical data with linear regression and error indices. The analysis was done for all five individual data sets as well as for the pooled data (five data sets combined together).

First, we performed validation of the RWS system using only the four main behavioural classes with all three validation methods. In the confusion matrix analysis, the RWS classifications were compared to gold standard second by second. From each 4 x 4 confusion matrix, we calculated precision, sensitivity, specificity, and accuracy for all four behavioural classes. All calculations were done with AWK (GNU Awk implementation, Free Software Foundation) programs written by the authors. In the linear regression analysis, we determined the regression equations with the least squares method and calculated the coefficients of determination (R^2) with Microsoft Excel. In the error index analysis, we calculated the average error and average error magnitude defined as volume-weighted mean percentage error $\sum(x_i-y_i)/\sum x_i$ and mean absolute percentage error $\sum|x_i-y_i|/\sum x_i$ with AWK scripts written by the authors. The numerical data were also plotted with Microsoft Excel for visual inspection of linearity. The results are presented for all individual data sets as well as for the pooled data.

Second, we analyzed the confusion matrices further by calculating confusion probabilities. The analysis was done in two phases, proceeding from coarser level matrices (only four main classes) to more detailed analysis (using subclasses to trace errors). In the second phase, the main classes of real behaviour were divided into subclasses, until the sources of confusion were identified. For brevity, we present the confusion matrix together with confusion probabilities only for the pooled data.

3. Results

3.1 Comparison of the three validation methods

Comparison of the three validation methods is presented in Table 2 (pooled data) and Table 3 (individual data sets). The tables give error indices (average error and average error magnitude), the regression equations with coefficients of determination, and the classification performance indicators (precision, sensitivity, specificity and accuracy) for eating, ruminating, drinking and other behaviour. Scatterplots presenting relationships between the observed and measured hourly durations are shown in Figure 2 (pooled data) and as online material (individual data sets). In the qualitative characterization of the overall accuracy (excellent−poor), we emphasize average error magnitude, precision and sensitivity. In this context, error magnitude $\leq$ 10% can be considered an excellent result, 10−20% good, 30−40% moderate and >40% poor. Similarly, having both precision and sensitivity $\geq$ 90% indicates excellent accuracy while having either of them <60% indicates a poor classifier.

Eating was detected poorly, on average, but in two data sets (animals 79 and 3355) the performance was moderate. In all data sets, average errors were negative and precision was smaller than sensitivity, which means that the device overestimated the feeding time. Error magnitudes and classification performance measures were mostly in agreement but linear regression results were difficult to interpret. An obvious reason, detected from scatterplots, was that the relationship was far from linear except in two data sets (79, 3355).

Ruminating was detected reasonably well but there was individual variation ranging from excellent (animals 79, 102, 3355) to poor (154, 4293) accuracy. All three validation methods were in agreement in their assessments in spite of nonlinear relationships in two data sets (154, 4293).

Drinking was detected poorly, especially according to classification performance measures. There were no individual differences in error magnitudes or performance measures but R^2 was variable (0.00 − 0.62). Once again, error magnitudes and classification performance measures were in agreement. However, average error (-33%; -10 sec/h) gave an overly optimistic view of accuracy. None of the relationships were linear which made the linear regression results hard to interpret.

According to classification performance measures and error indices, other behaviour was detected moderately well in general, even better than eating. However, there was large individual variation ranging from poor (animals 154, 4293) to relatively good (79, 102 and 3355) accuracy. Linear regression disagreed with the other methods and in the pooled data, R^2 was smaller than in eating (0.33 vs. 0.58). However, linear regression agreed with other methods on the good accuracy in two sets (79 and 3355) where the relationship was linear. In all data sets, average errors were positive and precision was greater than sensitivity, which means that the device underestimated the duration of other behaviour.

3.2 Analysis of confusion probabilities

The confusion matrix of main classes is given in Table 4, together with confusion probabilities. The main diagonal cells give numbers of true positives and their cp1 and cp2 correspond to precision and sensitivity. The other cells show numbers of erroneous classifications with related confusion probabilities.

Eating was detected with good sensitivity (83.8%) but poor precision (44.9%) which means that the errors were mostly false positives. Cp1 values in the eating column reveal that over 55% of RWS eating classifications were false positives, mainly from class 'other' (46.8%), but also from ruminating (6.2%) and drinking (2.1%). Cp2 values in the eating row reveal that over 16% of eating was classified incorrectly as 'other' (8.6%) or ruminating (7.1%). Confusion probabilities of subclasses revealed that the main reason for false positives (21.5% of reported eating) was a single subclass of 'other', licking the feeding table or the base of the cubicle and gathering feed remainders. This behaviour is closely related to eating, and in a wider interpretation, it could be included into eating. Other reasons were standing still (16.6%), and, to a lesser extent, masticating feed (6.0%). It also turned out that the system was less sensitive for detecting eating feed other than silage or concentrates (sensitivity 64.5% vs. 84.2%), but this is a small subclass (1.9% of eating) with little practical importance.

Ruminating was detected relatively well since both precision and sensitivity were reasonable (79.5% and 77.8%, respectively). Analysis of cp1 reveals that over 20% of reported ruminating were false positives, mainly from class 'other' (16.1%) but also from eating (4.4%). Analysis of cp2 reveals that about 22% of actual ruminating was not detected. These false negatives were mostly classified as 'other' (15.1%), but also as eating (7.0%) and drinking (1.1%). Confusion probabilities of subclasses revealed that the system was more sensitivite to detect ruminating when the animal was lying on the right side (sternal recumbency) (sensitivity = 85.0%) than in other positions (standing or lying on the left side, sternal recumbency) (sensitivity = 74.5%).

Drinking was detected poorly, with extremely low precision and sensitivity (5.6% and 7.4%, respectively). Analysis of cp1 reveals that over 94% of drinking classifications were false positives, mainly from class 'other' (86.8%) but sometimes also from eating (6.5%) or ruminating (1.1%). Analysis of cp2 reveals that over 92% of actual drinking was not detected. These false negatives were mostly classified as eating (66.4%) or 'other' (25.6%). Subclass analysis revealed that the main problem of false positive errors was standing (63% of reported drinking), followed by licking of the surroundings (10.4%), and, to a lesser extent, lying on the left side, masticating feed or eating (each about 4%).

Other behaviour (class 'other') was detected with high precision (87.2%) but relatively low sensitivity (66.5%), which means that the errors were mostly false negatives. Analysis of cp1 reveals that false positives (about 13%) originated mainly from ruminating (9.0%) or eating (3.2%). Analysis of cp2 reveals that over 43% of other behaviour was not detected, but was classified as eating (24.4%) or ruminating (7.2%) and sometimes even as drinking (1.9%). Subclass analysis revealed that false positives due to ruminating happened almost exclusively when the animal was either lying on the left side (sternal recumbency) or standing (8.5% of reported 'other'). An interesting observation was that ruminating on the right side (sternal recumbency) did not cause false positives (only 0.6% of reported 'other'), even though 32% of ruminating happened in this posture.

4. Discussion

The first objective of this study was to compare the confusion matrix method with basic error indices and linear regression analysis in the validation of RWS. The study revealed that different validation methods may sometimes produce discordant results. In general, the validation methods agreed when the results were clearly accurate (ruminating) or clearly inaccurate (drinking). Disagreement appeared only in intermediate cases, when the accuracy was poor to moderate (eating and other behaviour). In these cases, confusion matrix analysis and error magnitudes produced concordant results, but often linear regression disagreed with them.

Agreement on good accuracy is obvious, since error-free classifications mean error-free duration estimates and the linear equation coincides the identity line y=x with R^2=1. Agreement on inaccuracy is less certain. If the classifier performs no better than a random guess the hourly positive and negative errors (hourly FP and

FN) are randomly distributed. Still, it is possible that by chance the errors in each hour would cancel each other out (i.e., FP=FN) and hourly measurements would be accurate. However, it is very unlikely that such cancelling would happen in any larger extent.

The discrepancy between linear regression and the other two methods could not be explained by different time resolutions (second-based vs. hourly measurements), since error magnitudes were also calculated from hourly measurements. Instead, the main culprit was nonlinearity of data. Visual inspection of scatterplots revealed that in about half of the data sets the relationship was too nonlinear for reliable use of the linear regression method. The problem was aggravated by small sample sizes (12 observations per animal) and frequent outliers. This underscores that it is always crucial to plot the data and check linearity as well as homoscedasticity (constant variance) before applying linear regression. In the previous research, it has usually been assumed that the relationship between measured and reference values is linear or nothing has been said of the linearity, but sometimes graphical plots have been presented (e.g., Chizzotti et al., 2015, Elischer et al., 2013).

The effect of different time resolutions (second-based vs. hourly measurements) was best seen when error magnitudes were compared to the confusion matrix measures. In general, error magnitudes tended to give a more optimistic evaluation than the confusion matrix analysis because positive and negative errors (FP and FN) during the same hour could compensate each other. Usually, the compensations were relatively small, but in an extreme case (an hour in the animal 4293 data), the hourly error was only 1.1%, even if RWS detected only 67% of real eating and only 66% of measured eating were correct classifications. Average errors can hide even more errors, because they allow all positive and negative errors to cancel each other out. Therefore, average errors are alone unreliable measures of accuracy, but they allow easy detection of bias in hourly measurements (the sign shows over- or underestimation of durations).

We note that in this study we used hourly measurements for validation with linear regression and error indices. It is possible to calculate durations during shorter or longer time intervals which allows less or more errors to compensate each other. Therefore, the presented accuracy assessments with these two methods cannot be generalized to other time intervals. We also note that we used the RWS model from 2013 (firmware V01.13, data converter V0.7.0.0.; see Ruuska et al., 2016), and our results do not allow any conclusions on other models of RWS.

The second objective of this study was to analyze what extra information can be obtained with the confusion matrix method. Suggested confusion probabilities turned out to be a useful aid for utilizing all available information of real behaviour and detecting sources of errors. Especially, cp1 could reveal which subclasses were responsible for FP errors while cp2 revealed which subclasses were detected most or least sensitively. For further development of RWS, the most important findings were the reasons for poor detection of eating (certain eating related behaviour) and the effect of the animal's posture on accuracy. A possible explanation for the latter is that the measurements tended to be most accurate when the animal was in her most frequent ruminating posture and least accurate in the least frequent posture. Confusion probabilities also pointed out poor sensitivity for detecting eating feed other than eating silage or concentrates but this cannot be considered as a defect, since gnawing barn structures, feed troughs or halters could be rather classified as oral manipulation, although such a behavioural class was not used in the current study.

In medical science, there has been a long debate on appropriate validation methods for medical devices (e.g., Altman and Bland, 1983; Ludbrook, 2002; Zaki et al., 2012; Indrayan, 2013) and nearly all methods have been criticized. Pearson product-moment correlation has been considered inappropriate since it is unable to detect systematic error (bias); any linear relationship y=bx+a has perfect positive correlation (r=1), yet there may be substantial proportional (b≠1) or fixed (a≠0) bias (Zaki et al., 2012; Indrayan 2013; Hämäläinen et al., 2016). Linear regression gives more detailed information on the relationship between the reference and the measured value if the coefficient of determination ($R^2=r^2$), the slope and the intercept of the regression line are presented. Still, having slope=1, intercept=0 (no systematic error) and R^2 only slightly below 1 does not necessarily guarantee validity, as demonstrated by Indrayan (2013). In addition, the parameters of linear regression are difficult to interpret and there is no single estimate for random error or bias (Altman and Bland, 1983). Kappa coefficients are a family of statistics designed to assess inter-rater reliability both for nominal (Cohen's kappa coefficient) and ordinal or continuous (weighted kappa coefficients and intra-class correlation) data. In validation, they can produce counter-intuitive results, because they cannot detect bias and the coefficients are affected by the data distribution and variance between subjects (Zaki et al., 2012; Fay, 2005; Bland and Altman, 1990). Currently, the most popular validation method is so called Bland-Altman method or the limits of agreement (Altman and Bland, 1983). It includes a plot of differences x-y against (x+y)/2 (for checking bias) and then (in the absence of bias), analysis of the 95% confidence interval of differences. However, it has been shown that the Bland-Altman method may sometimes either under- or overestimate the bias and produce misleading results (Hopkins, 2004; Zaki et al, 2013).

Confusion matrices offer a viable alternative to validation when behaviour measurements are discrete classifications. Compared to linear regression and error indices, confusion matrix analysis has many advantages: it is robust to any data distribution and type of relationship; it makes a stringent evaluation of validity (with no chance for hiding errors); and it offers extra information on the type and sources of errors. Precision and sensitivity are easy to interpret and together they summarize classification performance on each class, but comparing overall performance between classes or data sets is more difficult. There are combination measures like F-score and AUC (area under the ROC curve; see e.g., Fawcett, 2006), but using a single measure loses always some information. In practice, we suggest to combine confusion matrices with other validation methods and make conclusions only if the methods are in agreement. For a comprehensive analysis, we suggest confusion probabilities that show the error distributions and help to detect the most serious errors and trace their reasons.

5. Conclusions

Validation methods are not always in agreement and, therefore, it is recommended that validation should be performed with several methods. If the assessments are concordant then conclusions can be made but otherwise more data should be collected for reliable validation. The current study also demonstrated that it is always important to check linearity before applying linear regression in validation. Confusion matrices are a robust validation method whenever discrete classifications are available. In addition, confusion probabilities offer extra information on the reasons for errors and their importance.

Acknowledgements

We would like to thank the technical staff of Natural Resources Institute Finland’s research station at Maaninka for their assistance with animal care. Lilli Frondelius, Dinara Kabylbekova and Sara Mykkänen are acknowledged for helping with data handling. The data collection was conducted in the Modern Cow Barn Technology project financed by the Regional Council of Pohjois-Savo (European Regional Development Fund, grant number A31185) and the Centre for Economic Development, Transport and the Environment of Pohjois-Savo (European Social Fund, grant number S11334). Salla Ruuska’s work was supported by the Olvi foundation and the Niemi foundation and Wilhelmiina Hämäläinen's work by the Academy of Finland (decision 307026).

References

Altman, D.G., Bland, J.M., 1983. Measurement in medicine: the analysis of method comparison studies. The Statistician. 32, 307–317. http://dx.doi.org/10.2307/2987937

Bland J.M., Altman D.G., 1990. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput. Biol Med. 20(5), 337-40. https://doi.org/10.1016/0010-4825(90)90013-F

Büchel, S., Sundrum, A., 2014. Technical note: Evaluation of a new system for measuring feeding behaviour of dairy cows. Comput. Electron. Agric. 108, 12–16. http://dx.doi.org/10.1016/j.compag.2014.06.010

Chizzotti, M.L., Machado, F.S., Valente, E.E.L., Pereira, L.G.R., Campos, M.M., Tomich, T.R., Coelho, S.G., Ribas, M.N., 2015. Technical note: Validation of a system monitoring individual feeding behavior and individual feed intake in dairy cattle. J. Dairy Sci. 98, 3438–3442. http://dx.doi.org/10.3168/jds.2014-8925

DeVries, T.J., von Keyserlingk, M.A.G., Weary, D.M., Beauchemin, K.A., 2003. Technical note: Validation of a system for monitoring feeding behaviour of dairy cows. J. Dairy Sci. 86, 3571–3574. http://dx.doi.org/10.3168/jds.S0022-0302(03)73962-9

Elischer, M.F., Arceo, M.E., Karcher, E.L., 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. http://dx.doi.org/10.3168/jds.2013-6790

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fay, M.P., 2005. Random marginal agreement coefficients: rethinking the adjustment for chance when measuring agreement. Biostatistics 6, 171–80. https://doi.org/10.1093/biostatistics/kxh027

Free Software Foundation. GNU Awk 4.0.1. https://www.gnu.org/software/gawk/manual/gawk.html

Hopkins, W.G., 2004. Bias in Bland-Altman but not Regression Validity Analyses. Sportscience 8, 42-46. http://sportsci.org/jour/04/wghbias.htm

Hämäläinen, W., Ruuska, S., Kokkonen, T., Orkola, S., Mononen, J., 2016. Measuring behaviour accurately with instantaneous sampling: A new tool for selecting appropriate sampling intervals. Appl. Anim. Behav. Sci. 180, 166–173. http://dx.doi.org/10.1016/j.applanim.2016.04.006

Indrayan, A., 2013. Clinical agreement in quantitative measurements – Limits of disagreement and the intraclass correlation. In: Methods of Clinical Epidemiology, pp. 17-27, ed. Doi, S.A.R., Williams, G.M. Springer-Verlag Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-37131-8

Ludbrook, J, 2002. Statistical techniques for comparing measurers and methods of measurement: a critical review. Clin Exp Pharmacol Physiol. 29(7):527-36. http://dx.doi.org/10.1046/j.1440-1681.2002.03686.x

Miller-Cushon, E.K., DeVries, T.J., 2011. Technical note: Validation of methodology for characterization of feeding behavior in dairy calves. J. Dairy Sci. 94, 6103–6110. https://doi.org/10.3168/jds.2011-4589

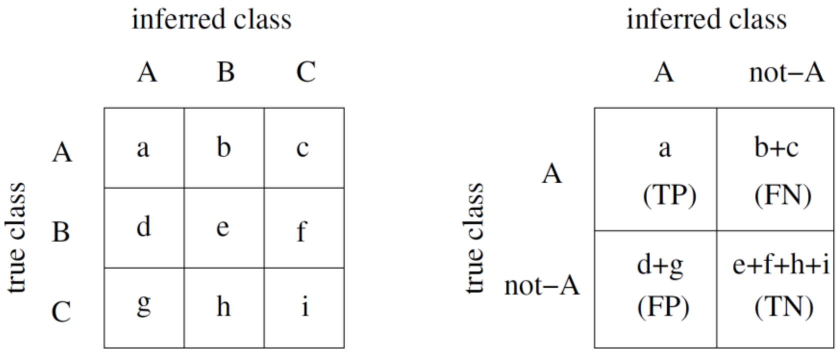
Nielsen, P.P., 2013. Automatic registration of grazing behaviour in dairy cows using 3D activity loggers. Appl. Anim. Behav. Sci. 148, 179–184. http://dx.doi.org/10.1016/j.applanim.2013.09.001

Ruuska, S., Kajava, S., Mughal, M., Zehner, N., Mononen, J., 2016. Validation of a pressure sensor-based system for measuring eating, rumination and drinking behaviour of dairy cattle. Appl. Anim. Behav. Sci. 174, 19–23. http://dx.doi.org/10.1016/j.applanim.2015.11.005

Schirmann, K., von Keyserlingk, M.A.G., Weary, D.M., Veira, D.M., Heuwieser, W., 2009. Technical note: Validation of a system for monitoring rumination in dairy cows. J. Dairy Sci. 92, 6052–6055. http://dx.doi.org/10.3168/jds.2009-2361

Zaki, R., Bulgiba, A., Ismail, R., Ismail, N.A., 2012. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PlosOne 7, e37908. http://dx.doi.org/10.1371/journal.pone.0037908

Zaki, R., Bulgiba, A., Ismail, N.A., 2013. Testing the agreement of medical instruments: Overestimation of bias in the Bland-Altman analysis. Preventative Medicine 57: S80-S82. http://dx.doi.org/10.1016/j.ypmed.2013.01.003



		inferred class		
		A	B	C
true class	A	a	b	c
	B	d	e	f
	C	g	h	i

		inferred class	
		A	not–A
true class	A	a (TP)	b+c (FN)
	not–A	d+g (FP)	e+f+h+i (TN)

Figure 1. An example of a 3x3 confusion matrix for classes A, B and C (left) and the corresponding binary confusion matrix for class A (right). TP=number of true positives, FP=number of false positives, TN=number of true negatives, FN=number of false negatives.

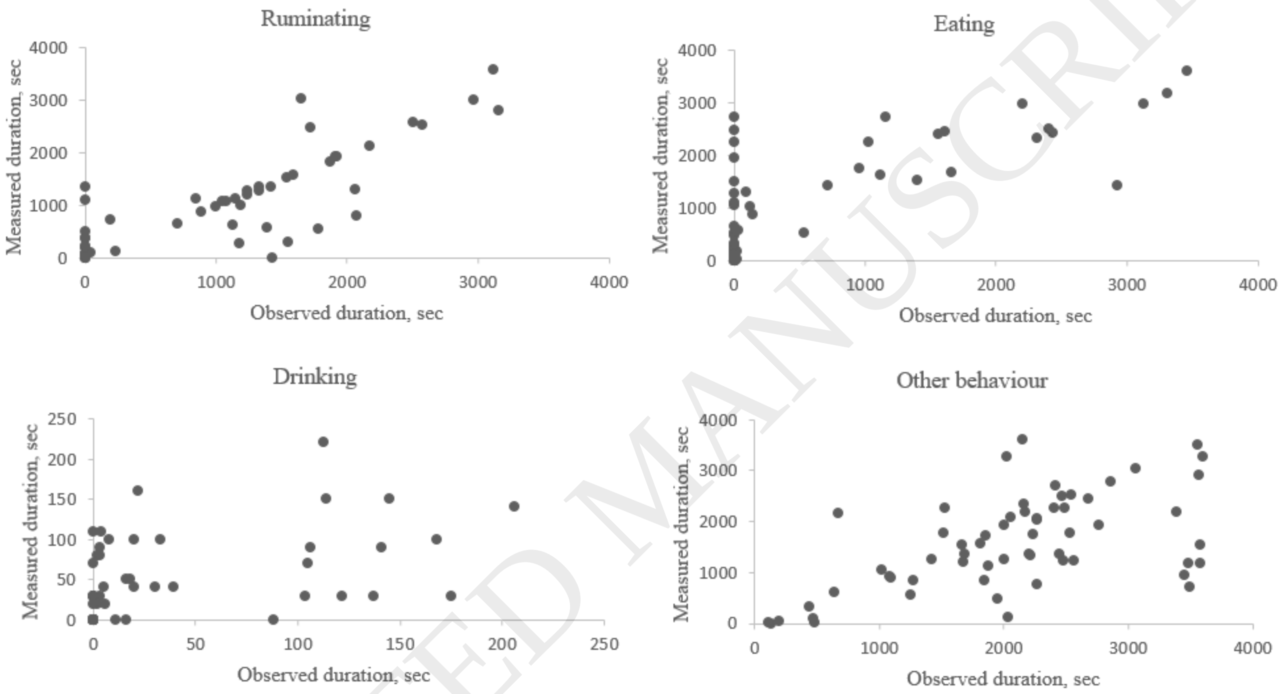
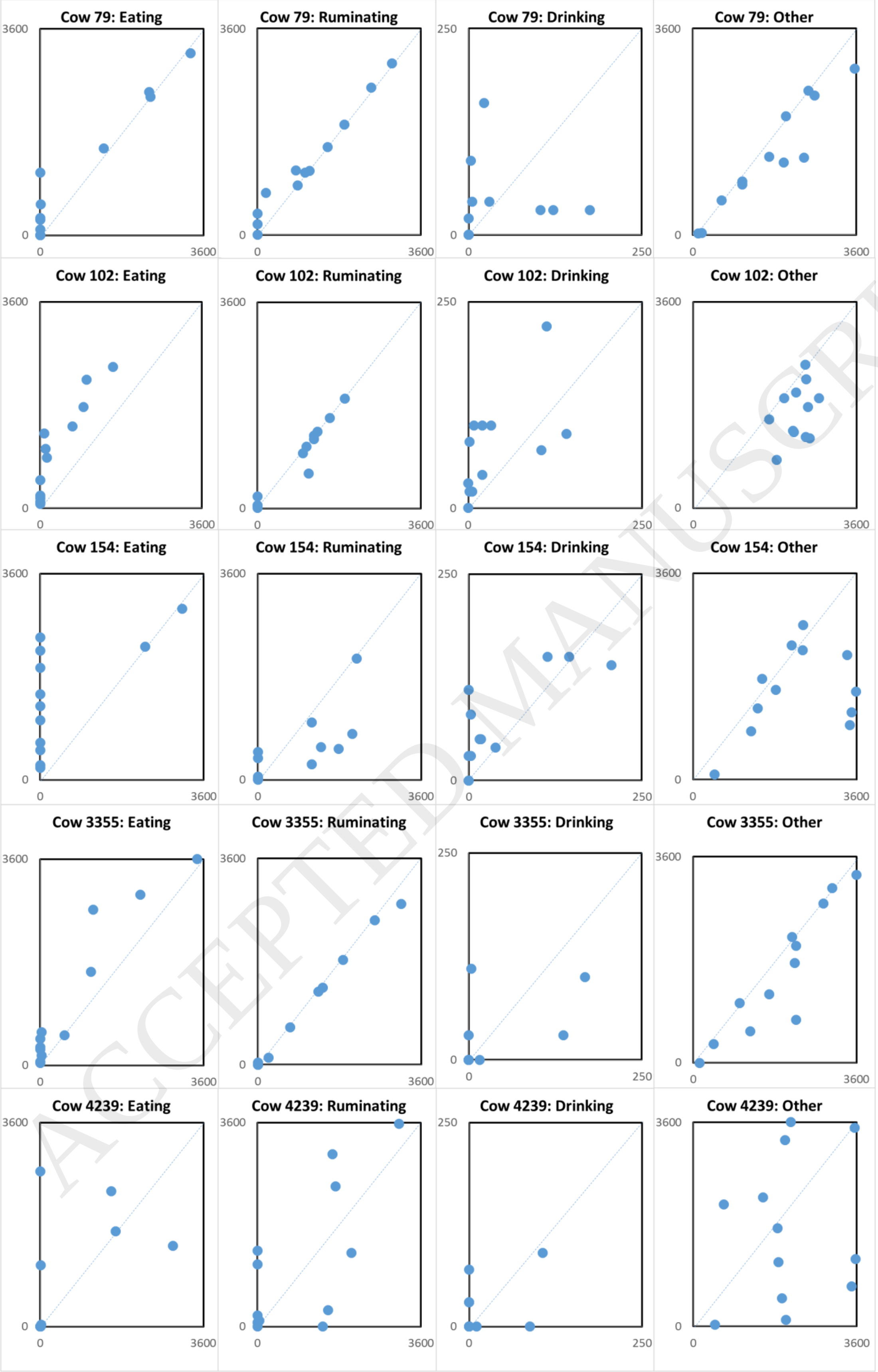


Figure 2. Scatterplots presenting relationships between observed and measured hourly durations of eating, ruminating, drinking, and other behaviour. Note that the drinking data are presented in a different scale.



Online material. Scatterplots presenting relationships between observed (video, x-axis) and measured (RWS, y-axis) hourly durations of eating, ruminating, drinking, and other behaviour for individual cows. There are 12 data points in each individual figure. The x-axes and y-axes range from 0 to 3600 seconds (one hour) for eating, ruminating and other behaviour, but only from 0 to 250 seconds for drinking. The diagonal lines illustrate the perfect linear relationship, i.e. y = x.

Table 1. The definitions of eating, ruminating, drinking and other behaviour and criteria for how subclasses were constructed.

Behaviour	Definition
Eating	Cow takes feed into its mouth, chews and swallows it
Subclasses of eating	Eating silage, concentrates or something else; body posture while eating silage (standing, lying, kneeling)
Ruminating	Cow regurgitates a bolus, chews and re-swallows it
Subclasses of ruminating	Body posture while ruminating: standing and lying (sternal recumbency, left/right side; lateral recumbency, left/right side)
Drinking	Cow puts its muzzle into the water bowl and swallows water
Other behaviour	Behaviour other than eating, ruminating or drinking
Subclasses of other behaviour	Licking the feeding table or cubicle mattress (standing, lying, kneeling), masticating feed (standing or lying), standing or moving, lying down and standing up, tongue rolling (standing or lying), auto-grooming (standing or lying) and allo-grooming, rubbing the halters, rubbing the head, lying (sternal recumbency, left/right side; lateral recumbency, left/right side), other behaviour

Table 2. Comparison of the three validation methods for RWS measuring eating, ruminating, drinking and other behaviour. The results are given for the pooled data (summed over all 5 cows). The compared methods were error indices (AE = average error, % and AEM = average error magnitude, %), regression analysis (RE = regression equation (y=bx+a) and R2 = coefficients of determination) and classification performance indicators (%; precision, sensitivity, specificity, accuracy).

	Eating	Ruminating	Drinking	Other
AE, AEM	-86.5, 97.1	2.1, 31.4	-32.5, 110.0	23.7, 33.0
RE	y = 0.80x + 612.7	y = 0.84x + 135.1	y = 0.46x + 28.5	y = 0.57x + 393.0
R^2	0.58	0.71	0.25	0.33
Precision	44.9	79.5	5.6	87.2
Sensitivity	83.8	77.8	7.4	66.5
Specificity	80.5	92.9	98.8	87.1
Accuracy	81.0	89.0	98.0	75.4

Table 3. Comparison of the three validation methods for RWS measuring eating, ruminating, drinking and other behaviour. The results are given for all five individual animals. The compared methods were error indices (AE = average error, % and AEM = average error magnitude, %), regression analysis (RE = regression equation (y=bx+a) and R^2 = coefficients of determination) and classification performance indicators (%; precision, sensitivity, specificity, accuracy).

	Eating	Ruminating	Drinking	Other
Cow 79				
AE, AEM	-24.5, 27.7	-11.0, 12.1	4.6, 130.4	18.7, 18.8
RE	y = 0.88x + 287.3	y = 0.91x + 220.5	y = 0.01x + 36.4	y = 0.84x – 47.8
R^2	0.93	0.97	0.00	0.90
Precision	68.7	88.7	2.0	97.8
Sensitivity	85.5	98.4	2.0	79.5
Specificity	88.9	94.6	99.0	98.4
Accuracy	88.2	95.7	98.0	89.6
Cow 102				
AE, AEM	-158.9, 158.9	2.4, 9.3	-93.3, 131.6	27.7, 27.9
RE	y = 1.39x + 468.9	y = 0.92x + 50.0	y = 0.68x + 47.1	y = 0.54 x + 419.4
R^2	0.83	0.94	0.35	0.13
Precision	35.4	93.3	6.8	98.1
Sensitivity	91.7	91.1	13.1	70.9
Specificity	79.6	97.9	98.1	97.5
Accuracy	80.9	96.2	97.2	80.6
Cow 154				
AE, AEM	-221.3, 226.4	36.2, 55.7	-52.6, 76.8	32.0, 37.1
RE	y = 0.54x + 1212.3	y = 0.49x + 118.8	y = 0.62x + 40.9	y = 0.27x + 939.3
R^2	0.37	0.55	0.62	0.14
Precision	25.8	81.8	7.3	86.9
Sensitivity	82.8	52.3	11.2	59.1
Specificity	65.7	96.6	98.2	84.6
Accuracy	67.8	86.5	97.1	68.4
Cow 3355				
AE, AEM	-56.2, 56.5	6.2, 7.4	16.9, 100.6	17.6, 17.9
RE	y = 1.08x + 341.0	y = 0.93x + 5.74	y = 0.37x + 12.6	y = 0.97x – 270.6
R^2	0.89	1.00	0.29	0.84
Precision	63.0	97.0	6.3	97.5
Sensitivity	98.4	91.0	5.2	80.3
Specificity	85.7	99.0	99.4	97.6
Accuracy	88.2	96.9	98.7	88.4
Cow 4293				
AE, AEM	-50.8, 99.9	-16.3, 76.1	-7.3, 119.5	20.0, 59.0
RE	y = 0.56x + 491.7	y = 0.81x + 337.6	y = 0.37x + 12.0	y = 0.23x + 1202.7
R^2	0.27	0.48	0.20	0.03
Precision	37.1	44.0	0.0	58.7
Sensitivity	55.9	51.2	0.0	46.9
Specificity	84.1	76.1	99.5	53.7
Accuracy	80.1	69.5	99.0	49.8

Table 4. The confusion matrix of eating (EAT), ruminating (RUM), drinking (DRI) and other behaviour (OTH). The matrix contrasts gold standard classifications (based on second by second continuous recording from videos) with RWS classifications for each second of the pooled data. Confusion probabilities cp1 and cp2 (%) are given in the parenthesis.

		RWS classification				
		EAT	RUM	DRI	OTH	∑
Gold standard	EAT	**28864** (44.9, 83.8)	2434 (4.4, 7.1)	170 (6.5, 0.5)	2977 (3.2, 8.6)	34445
	RUM	3959 (6.2, 7.0)	**43809** (79.5, 77.8)	30 (1.1, 0.1)	8504 (9.0, 15.1)	56302
	DRI	1319 (2.1, 66.4)	11 (0.0, 0.6)	**146** (5.6, 7.4)	509 (0.5, 25.6)	1985
	OTH	30088 (46.8, 24.4)	8869 (16.1, 7.2)	2284 (86.8, 1.9)	**82027** (87.2, 66.5)	123268
	∑	64230	55123	2630	94017	216 000