Correcting a bias in the computation of behavioral time budgets that are based on supervised learning

Yehezkel S. Resheff *[1], Hanna M. Bensch[2], Markus Zöttl[2], and Shay Rotics[2,3]

[1]Department of Computer Science, Holon Institute of Technology
[2]EEMiS, Department of Biology and Environmental Science, Linnaeus University, Kalmar, Sweden
[3]Department of Zooloy, University of Cambridge

Abstract

1. Supervised learning of behavioral modes from body-acceleration data has become a widely used research tool in Behavioral Ecology over the past decade. One of the primary usages of this tool is to estimate behavioral time budgets from the distribution of behaviors as predicted by the model. These serve as the key parameters to test predictions about the variation in animal behavior. In this paper we show that the widespread computation of behavioral time budgets is biased, due to ignoring the classification model confusion probabilities.

2. Next, we introduce *the confusion matrix correction for time budgets* – a simple correction method for adjusting the computed time budgets based on the model's confusion matrix.

3. Finally, we show that the proposed correction is able to eliminate the bias, both theoretically and empirically in a series of data simulations

*hezi.resheff@gmail.com

on body acceleration data of a fossorial rodent species (Damaraland mole-rat, *Fukomys damarensis*).

4. Our paper provides a simple implementation of *the confusion matrix correction for time budgets*, and we encourage researchers to use it to improve accuracy of behavioral time budget calculations.

Keywords— body-acceleration, bio-logging, behavioral time budget, biotelemetry, machine learning, animal behaviour

1 Introduction

The availability of affordable miniaturized bio-logger devices has revolutionized the field of behavioral ecology over the past decade [Kays et al., 2015]. Inertial measurement units, and especially accelerometers, provide information that can be translated to behavioral modes, typically using a supervised machine learning classification approach [Nathan et al., 2012, Resheff et al., 2014]. The detailed understanding of behavior and its location is key in the pursuit of questions at the heart of animal ecology [Nathan et al., 2008, Hays et al., 2016, Williams et al., 2020].

The process of inferring animal behavior from acceleration measurements using supervised machine learning requires, first, obtaining observations of animals fitted with the bio-logging devices to generate a training set of acceleration records coupled with known behaviors. These data are used to train machine learning models, that are then used to classify behavioral modes for body acceleration data of unobserved animals. Finally, the proportion of the classified behaviours, which are generally referred to as behavioral time budgets, are used to answer ecological questions about the distribution of behaviour across population in space and time (E.g. [Harel et al., 2016, Rotics et al., 2017, Chimienti et al., 2021, Weegman et al., 2017]).

Behavioral time budgets are commonly the key metric used for inferring animal behaviour and ecology based on body acceleration data. However, the regular practice of computing time budgets from the distribution of the classified behaviours does not consider the information regarding the classification model's accuracy. This informa-

tion includes the probability of classifying each behaviour incorrectly by confusing it with any of the other behaviours. The table of these probabilities is summarized in the model's 'confusion matrix'— a standard output of testing supervised machine learning accuracy using cross validation [Hastie et al., 2009]. For example, assuming we are interested in the 'running' behaviour, and the confusion matrix shows that in 10% of cases 'running' is wrongly classified by our model as 'walking' (whereas 'walking' is wrongly classified as 'running' in 5% of the cases), it would be important to adjust the calculated proportion of 'running' according to this information.

This problem has previously been discovered and studied in the field of machine learning, in a setting called *domain adaptation*, where the aim is to compute the distribution of classes in test data [Lipton et al., 2018] in order to train classifiers better suited for it. The authors found that simply counting classifier predictions leads to biased estimates of the distribution of classes in the test data, but a simple confusion-matrix based correction is enough to alleviate this problem. Following these results, we examined whether the computation of time budgets which ignores the classification model's confusion probabilities introduces a systematic bias, and it can be reduced by accounting for these probabilities.

Supervised machine learning models are optimized for the data distribution they are trained upon [Hastie et al., 2009]. If the distribution of behaviours in the training data differs considerably from the distribution of behaviours in the unobserved data that are to be classified by the supervised model, the classification accuracy is likely to drop. In such cases we hypothesize that the systematic bias of the time budget computation will increase, and its correction based on the model's confusion matrix will become even more important. A case of differing behavioral distributions between training and unobserved data may be fairly common in animal field studies. This is because the training data are usually collected under specific conditions under which observing the animal is more feasible (sometimes even in captivity, [Graf et al., 2015, Hammond et al., 2016, Clarke et al., 2021]), and which may not reflect the behavioral distribution when not observed. We therefore tested the time budget computation bias as well as its correction under data scenarios that simulate varying degrees of

difference between the behavioral distributions in the training and test data.

In this paper, we mathematically formulate and analyze the sources of bias in time budgets that are computed based on supervised machine learning models. Based on data simulations on acceleration records matched with known behaviours, collected in Damaraland mole-rats, we show that the standard time budget computation can be inaccurate, and that accounting for the confusion probabilities (the confusion matrix) substantially improves the accuracy of the computed time budgets. We demonstrate the implementation of the *confusion matrix correction for time budgets* and explore in which data situations it is particularly needed.

2 Estimating behavioral time-budgets

The standard method of computing time budgets as the distribution of classified behaviours introduces errors related to accuracy properties of the classifier. There are two sources of error when estimating the proportion of any specific behavior. Consider for instance the estimate of the proportion of Eating. Some of the samples where the correct behavior was Eating may be mistakenly classified as other behaviors (this is known as type II error; false negative). Conversely, some of the samples where in reality other behaviors took place may be wrongly classified as Eating (known as type I error; false positive). In case the two types of error happen to cancel each other out the estimation will be correct, whereas in any other case type I and type II errors will produce a systematic bias. This bias and method to correct for it were first formulated by Lipton et al. (2018) in the machine learning literature. We adapt the derivation here to elucidate the sources of the bias in a comprehensive way in the context of behavioral time budgets.

We can quantify the amount of estimation error in terms of the unknown correct time-budget and the predictor's confusion matrix. The *proportion of false negative* for a specific behavior is defined as the probability of the reality being the specific behavior, and the classifier predicting otherwise:

$$Pr\big(y = i \; and \; f(x) \neq i\big) \tag{1}$$

where x denotes an acceleration (ACC) sample of a corresponding behavior y, f is the classifier (see Appendix A for a full notation table). Using b_i to denote $Pr\big(y = i\big)$, the proportion of behavaior i in the data, equation (1) can equivalently be written as:

$$b_i \cdot Pr\big(f(x) \neq i \,|\, y = i\big) \tag{2}$$

Similarly, the *proportion of false positive* for the $i-th$ behavior is defined as the probability of the classifier predicting the $i-th$ behavior when the correct label for the sample is a different behavior:

$$Pr\big(y \neq i \; and \; f(x) = i\big) \tag{3}$$

as before, equation (3) can be written as:

$$(1 - b_i) \cdot Pr\big(f(x) = i \,|\, y \neq i\big) \tag{4}$$

and in total, the bias in the estimation of the proportion of time spent in the $i-th$ behavior, is the difference of the two:

$$\Delta_i = (1 - b_i) \cdot Pr\big(f(x) = i \,|\, y \neq i\big) - b_i \cdot Pr\big(f(x) \neq i \,|\, y = i\big) \tag{5}$$

We denote by o_i the observed proportion of time spent in the $i-th$ behavior as computed from the classified behaviors, we can express the expected bias in estimation for the $i-th$ behavior as:

$$b_i + \Delta_i = o_i \tag{6}$$

For each behavior $i \in \{1, ..., n\}$ there is a single linear equation (6). This gives a collection of n linear equations in n variables, the simultaneous solution of which provides the corrected time budget. In matrix form, this set of equations can be written as:

$$o = C^T b \tag{7}$$

where o is the vector $[o_1, ..., o_n]$ of observed time budget per behavior, C is the (row-normalized) confusion matrix (the $ij-$th element of C is the fraction of samples of behavior i in the validation data, that were classified as behavior j), and $b = [b_1, ..., b_n]$ is the unknown real time budget (see proof in Appendix B). Inverting C yields:

$$b = (C^T)^{-1} o \tag{8}$$

which gives a corrected time-budget. The intuitive way to interpret this relation is that we ask what the real time-budget must have been, so that together with the known confusion matrix for our classifier, we would get the computed time budget. This sheds light on some properties of the time-budget correction.

First, as expected, the estimate of any behavior that is perfectly classified, in terms of recall and precision, will not be changed at all by the correction. This is true because the associated Δ for this behavior will necessarily be 0 (Equation 5). Second, due to equation (5), behavioral classes of lower proportion and lower classification precision will tend to be over-estimated before the correction. Classes with high correct proportion and low recall will tend to be under-estimated.

For more information on statistical properties of the estimates produced by (8), and a broader discussion of label shift in machine learning, we refer the reader to [Lipton et al., 2018] where to the best of our knowledge this correction was first introduced.

3 Methods

3.1 Body acceleration data

We examined the adjustment of the time budgets according to the confusion matrix using data simulations (detailed below) on an empirical dataset of body-acceleration records of known behaviours. We obtained this dataset from 16 Damaraland mole-

Behavior	Eat	Dig	Rest	Sweep	Stand	Walk	total
count	2238	1807	745	729	662	410	6591

Table 1: Overall distribution of labels

rats (DMRs) that were collared with acceleration loggers (Technosmart LTD, Italy) for 1-3 weeks, and videotaped during this period to match the acceleration records with known behaviours. The collars were fitted under isoflurane anesthesia, with collar weight (2.8 (g)) being less than the 3% of the smallest collared animal used in this study (108 (g)). Acceleration was recorded by the loggers continuously at 25Hz in each of three perpendicular axes. The collaring and videotaping took place in a laboratory facility in the southern Kalahari (Kuruman River Reserve, South-Africa), wherein several groups of mole-rats are housed in a large system of tunnels that mimic their underground habitat [Zöttl et al., 2016, Houslay et al., 2020]. These tunnels are built of mostly transparent tubes, allowing to observe the DMRs behaviours (see [Zöttl et al., 2016], for details). We recorded 57, 10-minutes videos of the collared individuals and labelled the behaviours when they were clearly visible using the Boris software [Friard and Gamba, 2016]. The ACC data were then coupled with labelled behaviours and the analysis was conducted on 2-sec segments of acceleration records of a single behaviour (shorter behaviours were omitted). Only the most frequent behaviours were included in the analysis, which were: resting, eating, walking, digging, sweeping, and standing (See Table 1 for the behavioral distribution of the dataset collected). There were another 26 classes of behaviours, consisting in total 17% of the labelled behaviours, which were not included in the analysis in order to simplify our study which solely aimed to examine a methodological concept (rather than the DMRs biology). For additional validation, we repeated the main analysis with all the beahviours included, with the infrequent behaviors aggregated to an "Other" class. The results did not change qualitatively (see Appendix E). All research including the housing and collaring of the DMRs were done with approval of University of Pretoria Animal Ethics Committee (permits EC089-12, SOP-004-13, EC059-18).

3.1.1 Data processing

For each 2-sec acceleration record, 55 statistics were computed (e.g., Mean, Median, and Standard Deviation of each axis), and used as input to train the supervised machine learning models (See [Resheff et al., 2014] for a detailed manual of supervised learning of behavioral modes from sensor measurements). The models were trained to classify samples to one of the target behaviors (resting, eating, walking, digging, sweeping, and standing). In all experiments, the data were divided into three parts, designated train, validation, and test respectively. The size of each partition was reported for each experiment separately. The machine learning models (random forest with 250 trees, [Buitinck et al., 2013]) were trained using the train partition only. A confusion matrix was computed using the validation partition only. Time budget results were reported based on the test partition only.

3.2 Simulation experiments

The purpose of the following simulation experiments is twofold: first, to measure the amount of bias in the regular computation of behavioral time budgets (from the distribution of the classified behaviours) under various data scenarios; second, to quantify the ability of the confusion-matrix based correction described above to improve the accuracy of the calculated time budgets.

In the first set of simulations we examined the basic case where the training and test datasets have the same behavioral distribution. This reflects the ideal setting, in which the behavioral distribution during training is identical to that in the unobserved dataset. To do so, the entire data was evenly split at random into 3 equal sized partitions designated train, validation, and test, the classifier was trained on the train partition, the confusion matrix was generated based on the validation partition, and the regular and corrected time budgets were calculated on the test partition. To robustly collect statistics of estimation error, we repeated the process for 250 iterations with a different random split of data each time.

Next, we examined scenarios where a behaviour was represented disproportionately

in the training set versus the test set. This addresses the case when during observations for obtaining the training set, the animals were conducting some behaviours more or less frequently than when not observed. Keeping the test set uniform (100 samples of each behavior), we simulated cases where one of the behaviors was under or over-expressed (20 to 200 with increments of 10) in the training set while the others were held constant (at 60 samples each for all other behaviours). All the remaining data were assigned to the validation set. The process was repeated 10 times for each value of under or over expression, for each behavior, and the regular and corrected time budget tables were calculated.

Third, we examined a similar scenario to the above but this time keeping the training set distribution constant (60 samples per behaviour) and varying the extent of representation of a single behaviour in the test dataset (20-200 samples at increments of 10), while the others remained constant (100 samples per behaviour). All the remaining samples were assigned to the validation set. Again, the process was repeated 10 times for each value of expression, for each behavior, and the regular and corrected time budget tables were calculated.

4 Results

Train and test data with equal behavioral distributions. Our first, basic set of simulations with training and test sets of equal behavioral distributions showed that there is a bias in time budget estimates (Fig. 1, left column). For example, eating behaviour was estimated to constitute 22% of the total behaviour whereas its true proportion was 16.6%. The simulations also showed that on average the bias was eliminated completely when the 'confusion-matrix correction' is implemented to adjust the time budget estimates (Fig. 1, right column).

Train data with varying behavioral distributions. A series of simulations in which a single behaviour in the training set varied in its proportion (from under to over representation) showed that the time budget estimate (calculated on the test

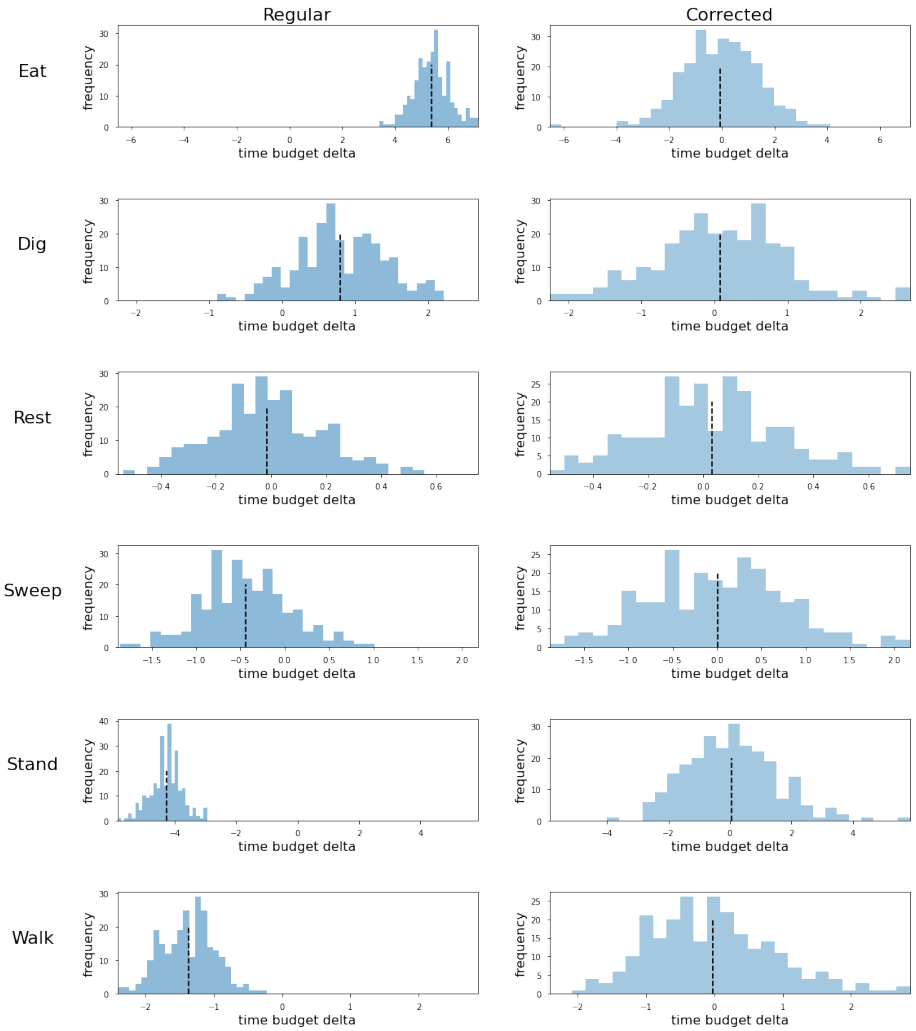

Figure 1: Distribution of deviation from correct time budget per behavior in 250 simulations for the regular time budgets (left column) and corrected time budgets (right column). Deviation is presented as the proportion (percentages) of the behaviour in the classified (annotated) behaviours minus its correct proportion (See Table 1.) Vertical dashed line represented the average of each distribution. Classifier performance (F1 mean $\pm$ std) per behavior across the 250 iterations: Dig 88.07 ± 0.92, Eat 85.5 ± 0.74, Rest 97.13 ± 0.66, Sweep 80.25 ± 1.75, Stand 56.41 ± 2.96, Walk 63.72 ± 3.15

set) increased monotonically with the proportion of the behaviour in the training set (Fig 2). Thus, under or over representation in the training set was a source of bias in estimating the true proportion (time budget) of the behaviour in the test data. The range of estimation error was highly variable, depending on the behaviour, with, for example, estimates in the range of 5% - 30% for Stand (true value is always 16.66%), versus a range of roughly 16%- 18% for Rest (Fig 2). However, for all behaviours, the corrected time budget estimates (generated using the 'confusion matrix correction for time budgets') were uncorrelated with the behavior's proportion in the training set, showing that the correction eliminated the bias even in cases of large over (or under) expression of a behavior in the training data, and generally provided more accurate estimates than the regular, uncorrected time budget estimates (Fig 2).

Test data with varying behavioral distributions. A series of simulations in which a single behaviour in the test set varied in its proportion (from under to over representation) showed that the corrected time budget (using the confusion matrix correction) follows the true time budget more closely than the regular time budget, indicating the former is more accurate (Fig 3).

5 Discussion

When behaviours are classified from sensor measurement data using a supervised machine learning classifier, the straightforward approach of calculating behavioral time budgets is from the distribution of the classified behaviours. The drawback in this common approach is that it considers only the final output of the classification model - the classified behaviours, and neglects information regarding the rates of confusion between behaviours.

Our paper shows both theoretically and by using data simulations that the current standard method of computing time budgets is biased by the asymmetric confusion properties of the classifier. We show that this bias can be corrected by adjusting the time budget according to the confusion matrix of the classifier. We introduce

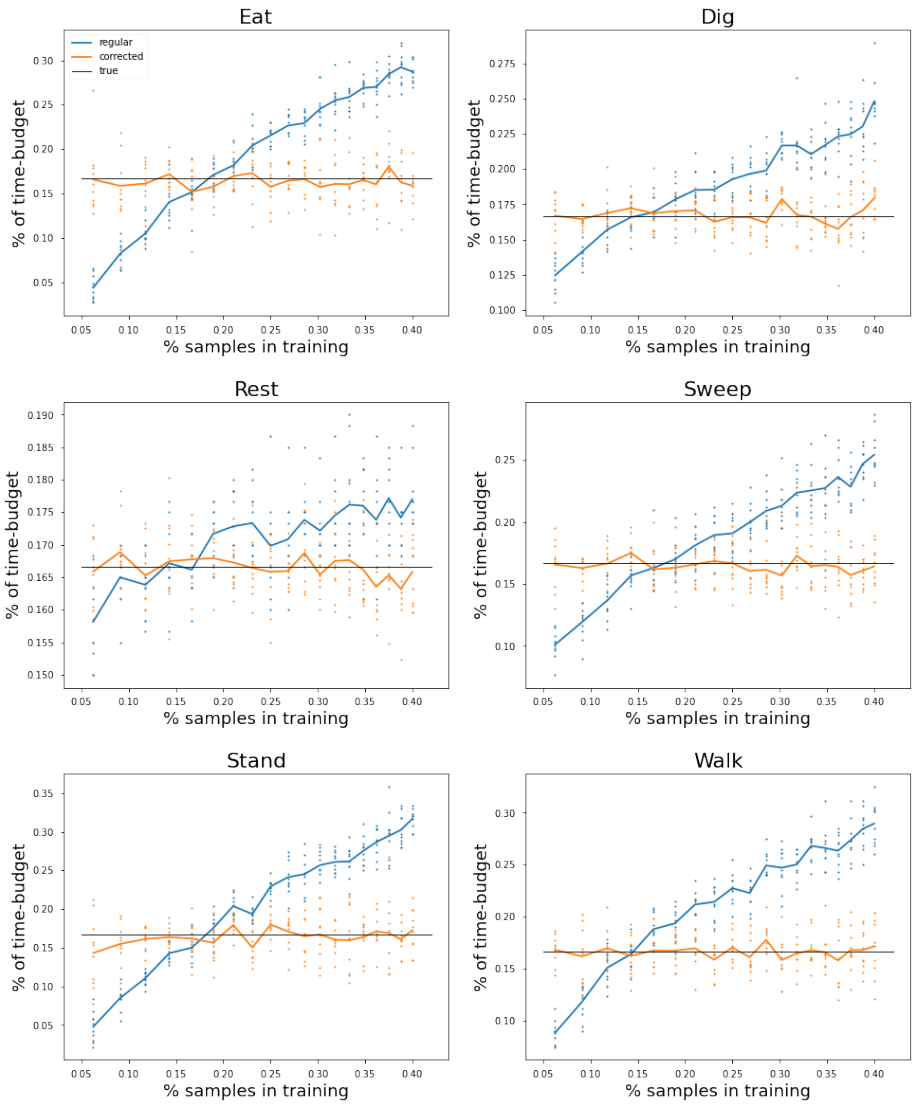

Figure 2: Effect of over or under expression of a single behavior in the training data on computed time budget in the wild for the same behavior. Test data is uniform. Blue - regular time budgets, Orange - corrected time budgets. Black line indicates the correct value. Dots indicate single trial results, solid lines are the averages.

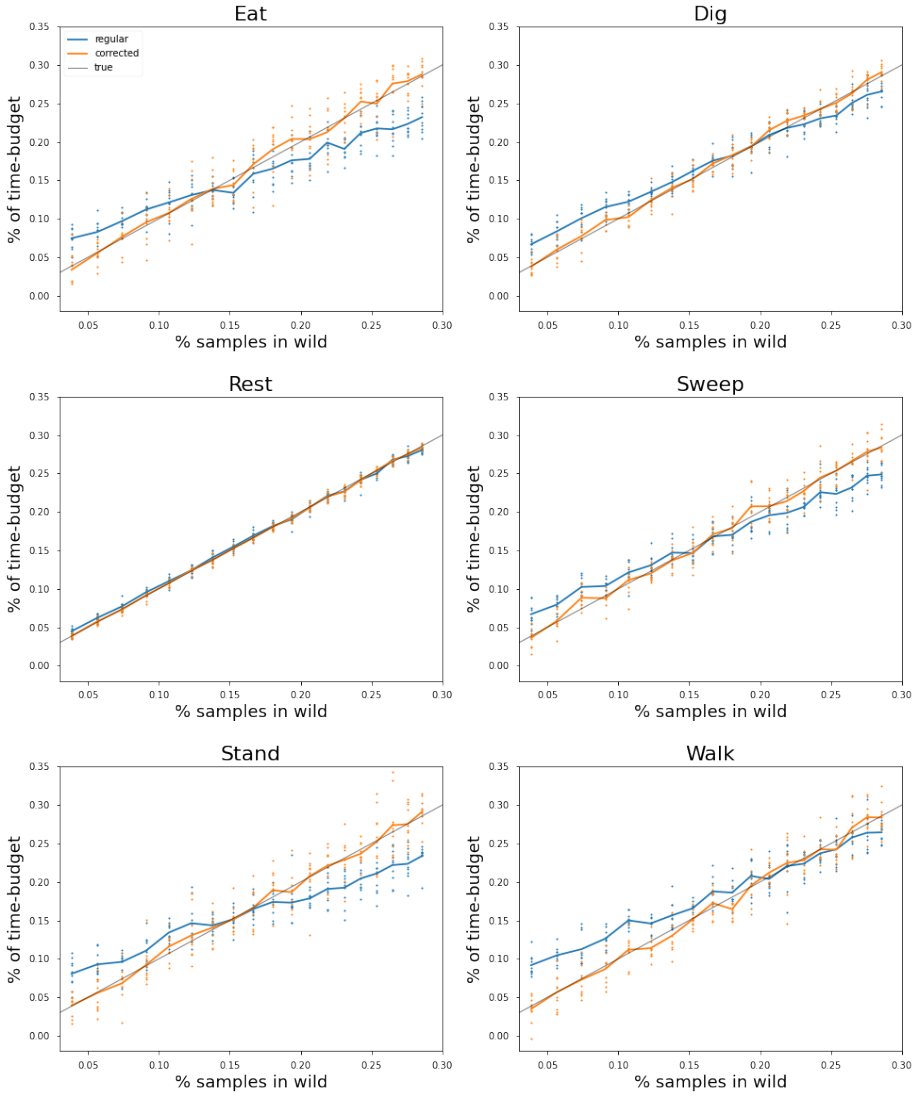

Figure 3: Effect of over or under expression of a single behavior in the test data on computed time budgets. Training data is uniform. Blue - regular time budgets, Orange - corrected time budgets. Black line indicates the correct value. Dots indicate single trial results, solid lines are the averages.

this correction following Lipton et al. (2018), and we call it the 'confusion matrix correction for time budgets'. The implementation of this method is simple, using the three-lines of code provided in appendix D. We demonstrate that using it improves the accuracy of time budgets (or of frequency of behaviours) that are derived from machine learning models.

In our first, basic series of simulations, where train and test distributions were identical (See Table 1 for the precise distribution), results showed varying degrees of time budget bias for the different behaviours. The bias was minor for behaviours with very high classification accuracy (e.g., rest, see for example a confusion matrix in Table 2, Appendix C), but other, less accurately-classified behaviours such as eat or stand were over or under estimated by up to 30% of their true proportion.

The bias increased when behaviours were over (or under) represented in the train data versus the test data (in which their proportion stayed fixed at 16.66%), with estimates biased as high as three times the true proportion of the behaviour. These simulation results showed a positive association between the behavior's representation in the training data and its estimated proportion, even though its true proportion stayed constant. Thus, the bias in time budget estimates increased with increasing disparity between the train and test data distributions. Consequently, one could reasonably obtain a wide range of behavioral time budgets for the same body acceleration data set, depending solely on the behavioral distribution in the data collected to train the classifier. This effect may have significant consequences for the validity of results obtained using the standard time budget estimation method without correction for this systematic bias.

In practice, we believe that a scenario of differing training and test behavioral distributions is common in wildlife research. Training data is usually confined to being collected when animal observation conditions are feasible or convenient. In some studies, it is collected from animals in captivity (e.g. [Graf et al., 2015, Hammond et al., 2016, Clarke et al., 2021]), in others only during more approachable life phases of the animal, such as only during breeding period in migrating birds (e.g. [Rotics et al., 2016]). Such training data is, therefore, unlikely to reflect precisely the distribution of behaviours in

the entire free-ranging data. Moreover, since research questions involve a comparison of time-budgets in different situations potentially having different budgets, it is not possible for training data to fit all the behavioral distributions.

In our last series of simulations, in which behaviours were over (or under) represented in the test data versus the train data (in which their proportion stayed fixed at 16.66%), we found a 'regression to the mean' bias in the time budgets estimation. i.e., behaviors with small actual proportions are over-estimated, and conversely behaviors with large actual proportions are under-estimated, where the pivot point is around the proportion used in training data (16.66% for each behaviour). It is noteworthy that these simulation results with uniform training data show smaller overall estimation bias, compared to the simulations in which behaviours were over (or under) expressed in training data.

Our simulation results indicate that using the proposed '*confusion matrix correction for time budgets*' improves the time budget accuracy and on average eliminates the bias completely, regardless of the behavior's classification accuracy and the degree of disparity between the train and test data distributions.

Other methods for inferring animal behavior from acceleration measurements that do not rely on supervised learning include algorithms that characterize elements of movement such as turning points [Potts et al., 2018], and trajectory segments [Resheff, 2016]. These methods may also be susceptible to the bias arising from the confusion properties of the algorithm, and thus could benefit from the confusion matrix correction.

6 Conclusion

The current standard method for computation of behavioral time-budgets based on supervised learning of behavioral modes from acceleration data [Resheff et al., 2014, Nathan et al., 2012] ignores information about the confusion probabilities of specific behaviors and frequently leads to biased estimates of time budgets. This is especially the case for behaviours of lower classification accuracy, for small behavioral categories, and for behaviors that were over or under represented in training data. The corrected

time-budget estimates take the classifier's confusion matrix into account leading to more accurate results. These findings suggest that the *confusion matrix correction for time budgets* should generally be used whenever computing behavioral time-budgets. The correction should be applied on each time-budget computed, based on the specific unit of the analysis, i.e., per individual's time budget if individuals are being contrasted, or for example per individual and period if individual behaviour is compared between different periods (like summer and winter, or day and night).

7 Acknowledgements

We thank Prof Tim Clutton-Brock for his kind support in this research, and Susanne Siegmann and Dr Carin Bernardo for their essential contribution to the data collection. We thank the Wenner-Gren and the Blavatnik foundations for (non-parallel) stipends granted to S.R. The data collection was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 742808) and by the Crafoordska Stiftelsen (2018-2259 & 2020-0976). We are grateful to the Kalahari Research Trust and the Kalahari Meerkat Project for access to facilities in the Kuruman River Reserve, and to Prof Marta Manser for her contribution to the management of the reserve. We would like to thank the Northern Cape Department of Environment and Nature Conservation for permission to conduct the data collection. Finally, We would like to thank the reviewers and editor for their thoughtful comments and efforts towards improving our manuscript.

8 Author contributions

Y.S.R. and S.R. conceived the idea. H.M.B. and S.R. carried out the fieldwork with the help of M. Z.. Y.S.R conducted the analysis. Y.S.R and S.R. wrote the first draft of the manuscript and all authors substantially contributed to the revisions.

9 Conflict of interest

Nothing to declare.

References

[Buitinck et al., 2013] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

[Chimienti et al., 2021] Chimienti, M., van Beest, F. M., Beumer, L. T., Desforges, J.-P., Hansen, L. H., Stelvig, M., and Schmidt, N. M. (2021). Quantifying behavior and life-history events of an arctic ungulate from year-long continuous accelerometer data. *Ecosphere*, 12(6):e03565.

[Clarke et al., 2021] Clarke, T. M., Whitmarsh, S. K., Hounslow, J. L., Gleiss, A. C., Payne, N. L., and Huveneers, C. (2021). Using tri-axial accelerometer loggers to identify spawning behaviours of large pelagic fish. *Movement ecology*, 9(1):1–14.

[Friard and Gamba, 2016] Friard, O. and Gamba, M. (2016). Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in ecology and evolution*, 7(11):1325–1330.

[Graf et al., 2015] Graf, P. M., Wilson, R. P., Qasem, L., Hackländer, K., and Rosell, F. (2015). The use of acceleration to code for animal behaviours; a case study in free-ranging eurasian beavers castor fiber. *PloS one*, 10(8):e0136751.

[Hammond et al., 2016] Hammond, T. T., Springthorpe, D., Walsh, R. E., and Berg-Kirkpatrick, T. (2016). Using accelerometers to remotely and automatically characterize behavior in small animals. *Journal of Experimental Biology*, 219(11):1618–1624.

[Harel et al., 2016] Harel, R., Horvitz, N., and Nathan, R. (2016). Adult vultures outperform juveniles in challenging thermal soaring conditions. *Scientific reports*, 6(1):1–8.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learnin. *Cited on*, page 33.

[Hays et al., 2016] Hays, G. C., Ferreira, L. C., Sequeira, A. M., Meekan, M. G., Duarte, C. M., Bailey, H., Bailleul, F., Bowen, W. D., Caley, M. J., Costa, D. P., et al. (2016). Key questions in marine megafauna movement ecology. *Trends in ecology & evolution*, 31(6):463–475.

[Houslay et al., 2020] Houslay, T. M., Vullioud, P., Zöttl, M., and Clutton-Brock, T. H. (2020). Benefits of cooperation in captive damaraland mole-rats. *Behavioral Ecology*, 31(3):711–718.

[Kays et al., 2015] Kays, R., Crofoot, M. C., Jetz, W., and Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science*, 348(6240).

[Lipton et al., 2018] Lipton, Z., Wang, Y.-X., and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR.

[Nathan et al., 2008] Nathan, R., Getz, W. M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., and Smouse, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences*, 105(49):19052–19059.

[Nathan et al., 2012] Nathan, R., Spiegel, O., Fortmann-Roe, S., Harel, R., Wikelski, M., and Getz, W. M. (2012). Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures. *Journal of Experimental Biology*, 215(6):986–996.

[Potts et al., 2018] Potts, J. R., Börger, L., Scantlebury, D. M., Bennett, N. C., Alagaili, A., and Wilson, R. P. (2018). Finding turning-points in ultra-high-resolution animal movement data. *Methods in Ecology and Evolution*, 9(10):2091–2101.

[Resheff, 2016] Resheff, Y. S. (2016). Online trajectory segmentation and summary with applications to visualization and retrieval. In *2016 IEEE international conference on big data (Big Data)*, pages 1832–1840. IEEE.

[Resheff et al., 2014] Resheff, Y. S., Rotics, S., Harel, R., Spiegel, O., and Nathan, R. (2014). Accelerater: a web application for supervised learning of behavioral modes from acceleration measurements. *Movement ecology*, 2(1):1–7.

[Rotics et al., 2016] Rotics, S., Kaatz, M., Resheff, Y. S., Turjeman, S. F., Zurell, D., Sapir, N., Eggers, U., Flack, A., Fiedler, W., Jeltsch, F., et al. (2016). The challenges of the first migration: movement and behaviour of juvenile vs. adult white storks with insights regarding juvenile mortality. *Journal of Animal Ecology*, 85(4):938–947.

[Rotics et al., 2017] Rotics, S., Turjeman, S., Kaatz, M., Resheff, Y. S., Zurell, D., Sapir, N., Eggers, U., Fiedler, W., Flack, A., Jeltsch, F., et al. (2017). Wintering in europe instead of africa enhances juvenile survival in a long-distance migrant. *Animal Behaviour*, 126:79–88.

[Weegman et al., 2017] Weegman, M. D., Bearhop, S., Hilton, G. M., Walsh, A. J., Griffin, L., Resheff, Y. S., Nathan, R., and David Fox, A. (2017). Using accelerometry to compare costs of extended migration in an arctic herbivore. *Current zoology*, 63(6):667–674.

[Williams et al., 2020] Williams, H. J., Taylor, L. A., Benhamou, S., Bijleveld, A. I., Clay, T. A., de Grissac, S., Demšar, U., English, H. M., Franconi, N., Gómez-Laich, A., et al. (2020). Optimizing the use of biologgers for movement ecology research. *Journal of Animal Ecology*, 89(1):186–206.

[Zöttl et al., 2016] Zöttl, M., Vullioud, P., Mendonça, R., Ticó, M. T., Gaynor, D., Mitchell, A., and Clutton-Brock, T. (2016). Differences in cooperative behavior among damaraland mole rats are consequences of an age-related polyethism. *Proceedings of the National Academy of Sciences*, 113(37):10382–10387.

10 Appendix A: Notation

	notation	meaning
1	x	Acceleration sample
2	y	Behavior label of an acceleration sample
3	f	classifier; $f(x)$ is the predicted behavior for sample x
4	b_i	Time budget for the $i-th$ behavior
5	o_i	Labelled (observed by classifier) time budget for the $i-th$ behavior
6	Δ_i	Estimation error of time budget for the $i-th$ behavior
7	C	Confusion matrix; C_{ij} is the fraction of samples of behavior i in the validation data, for which the classifier assigned behavior j

11 Appendix B: proofs

Proposition 1. *Let C denote the confusion matrix for a given classifier. The expected labeled time-budget o obeys: $o = C^T b$ where b is the correct time-budget.*

Proof. From equation (5) and (6):

$$o_i = b_i + \Delta_i \tag{9}$$

$$= b_i + (1 - b_i) \cdot Pr\big(f(x) = i) \,|\, y \neq i\big) - b_i \cdot Pr\big(f(x) \neq i) \,|\, y = i\big) \tag{10}$$

$$= b_i + \sum_{j \neq i} b_j \cdot Pr\big(f(x) = i) \,|\, y = j\big) - b_i \cdot Pr\big(f(x) \neq i) \,|\, y = i\big) \tag{11}$$

$$= b_i + \sum_{j \neq i} b_j \cdot C_{ji} - b_i \cdot (1 - C_{ii}) \tag{12}$$

$$= b_i + \sum_{j \neq i} b_j \cdot C_{ji} + b_i \cdot C_{ii} - b_i \cdot C_{ii} - b_i \cdot (1 - C_{ii}) \tag{13}$$

$$= b_i + \sum_{j} b_j \cdot C_{ji} - b_i \tag{14}$$

$$= \sum_{j} b_j \cdot C_{ji} \tag{15}$$

which by the definition of matrix multiplication is the $i - th$ element in $C^T b$. □

note – in the second transition we use the fact that on the one hand:

$\sum_{j \neq i} Pr\big(f(x) = i \cap y = j\big) = Pr\big(f(x) = i \cap y \neq i\big) = (1 - b_i) \cdot Pr\big(f(x) = i) \,|\, y \neq i\big)$

and on the other hand:

$\sum_{j \neq i} Pr\big(f(x) = i \cap y = j\big) = \sum_{j \neq i} Pr(y = j) \cdot Pr\big(f(x) = i) \,|\, y = j\big) = \sum_{j \neq i} b_j \cdot Pr\big(f(x) = i) \,|\, y = j\big)$

thus:

$(1 - b_i) \cdot Pr\big(f(x) = i) \,|\, y \neq i\big) = \sum_{j \neq i} b_j \cdot Pr\big(f(x) = i) \,|\, y = j\big)$.

Proposition 2 (Alternative derivation of (7))**.** $o = C^T b$

Proof. $[C^T b]_i = \sum_j b_j \cdot Pr(f(x) = i \,|\, y = j) = \sum_j Pr(y = j) \cdot Pr(f(x) = i \,|\, y = j)$

$= \sum_j Pr(f(x) = i \,,\, y = j) = Pr(f(x) = i) = o_i$ □

12 Appendix C: confusion matrix

	Dig	Eat	Forward Loco	Rest	Stand	Sweep
Dig	1529	45	10	0	0	23
Eat	34	1970	10	0	16	8
Forward Loco	7	27	155	0	5	16
Rest	0	3	0	535	7	0
Stand	6	141	2	3	309	1
Sweep	28	19	4	0	2	476

Table 2: Confusion matrix. Training data: 200 samples of each behavior. Test data is all the remaining samples. Overall accuracy: 92.26%. Rows indicate observed labels. Columns indicate predicted labels.

13 Appendix D: code example

```python
import numpy as np
from sklearn.preprocessing import normalize


def compute_correction_time_budget(conf_mat, observed_budget):
    """
    :param conf_mat: numpy array (nxn)
        confusion matrix
    :param observed_budget: numpy array (n)
        observed time-budget
    :return: numpy array (n)
        corrected time-budget
    """

    # make confusion matrix row-normalized
    cm = normalize(conf_mat.astype(float), norm="l1")

    # correction
    return np.linalg.pinv(cm.T) @ observed_budget
```

14 Appendix E: Results with all behavioral classes

In the results reported in the paper (Section 4), only the most frequent behaviours were included in the analysis, which were: resting, eating, walking, digging, sweeping, and standing. There were another 26 classes of behaviours, consisting in total 17% of the labelled behaviours, which were not included in the analysis in order to simplify our study which solely aimed to examine a methodological concept (rather than the DMRs biology). Here we repeat the main results (Figure 1) without excluding the remaining behavioral classes. Instead, they are grouped and designated the "Other" label. The full distribution of samples is summarized in Table (3).

Behavior	Eat	Dig	**Other**	Rest	Sweep	Stand	Walk	total
count	2238	1807	**970**	745	729	662	410	7561

Table 3: Overall distribution of labels including "Over"

Results for this basic set of simulations with training and test sets of equal behavioral distributions showed here again that there is a bias in time budget estimates (Fig. 4, left column). For example, eating behaviour was estimated to constitute approximately 24% of the total behaviour whereas its true proportion was 14.3%. The simulations also showed that on average the bias was eliminated completely when the 'confusion-matrix correction' is implemented to adjust the time budget estimates (Fig 4, right column).

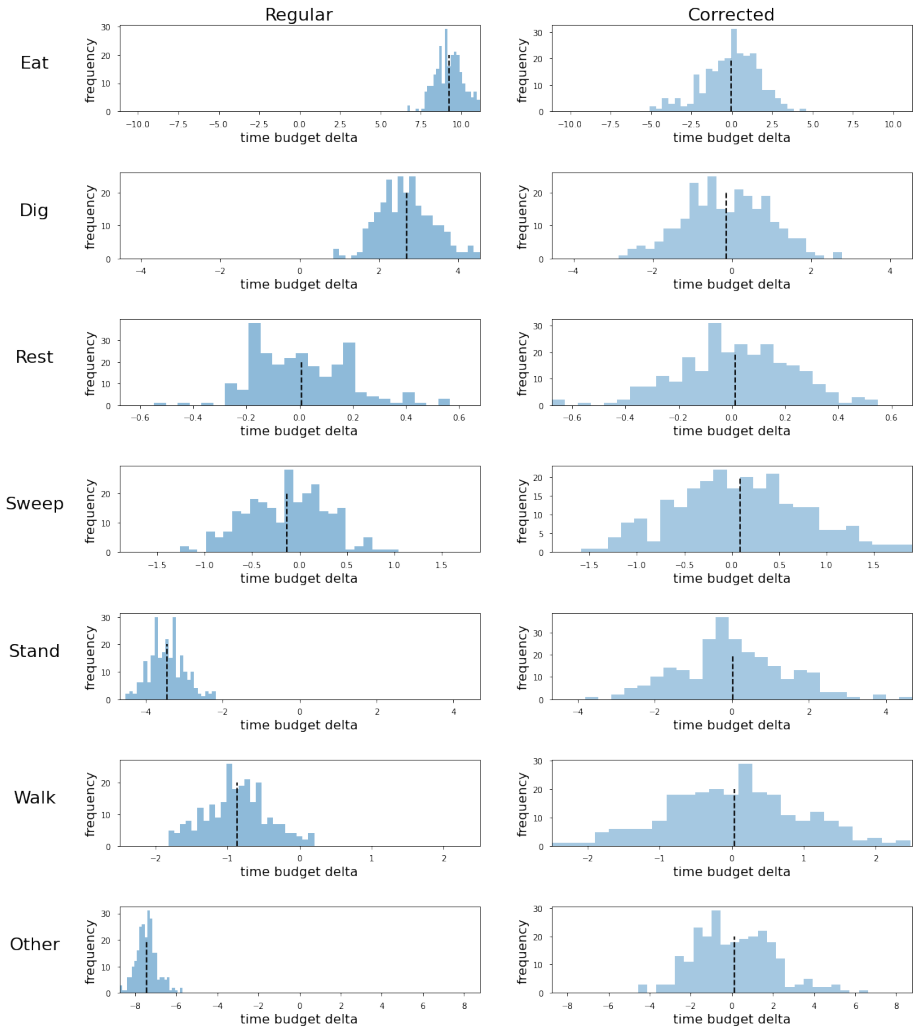

Figure 4: Distribution of deviation from correct time budget per behavior in 250 simulations for the regular time budgets (left column) and corrected time budgets (right column). Deviation is presented as the proportion (percentages) of the behaviour in the classified (annotated) behaviours minus its correct proportion. Vertical dashed line represented the average of each distribution.