Behavioral Ecology

The official journal of the
ISBE
International Society for Behavioral Ecology

Behavioral Ecology (2020), XX(XX), 1–12. doi:10.1093/beheco/araa095

Vol. 31: 1379 - 1390 (2020)

Original Article

Comparing dominance hierarchy methods using a data-splitting approach with real-world data

Chloé Vilette,[a,b] Tyler Bonnell,[a,b] Peter Henzi,[a,b] and Louise Barrett[a,b]

[a]Department of Psychology, University of Lethbridge, 4401 University Drive Lethbridge, Alberta T1K 3M4, Canada and [b]Applied Behavioural Ecology and Ecosystems Research Unit, University of South Africa, Private Bag X6, Florida 1710, Republic of South Africa

Received 22 June 2020; revised 2 September 2020; editorial decision 6 September 2020; accepted 11 September 2020.

The development of numerical methods for inferring social ranks has resulted in an overwhelming array of options to choose from. Previous work has established the validity of these methods through the use of simulated datasets, by determining whether a given ranking method can accurately reproduce the dominance hierarchy known to exist in the data. Here, we offer a complementary approach that assesses the reliability of calculated dominance hierarchies by asking whether the calculated rank order produced by a given method accurately predicts the outcome of a subsequent contest between two opponents. Our method uses a data-splitting "training–testing" approach, and we demonstrate its application to real-world data from wild vervet monkeys (*Chlorocebus pygerythrus*) collected over 3 years. We assessed the reliability of seven methods plus six analytical variants. In our study system, all 13 methods tested performed well at predicting future aggressive outcomes, despite some differences in the inferred rank order produced. When we split the dataset with a 6-month training period and a variable testing dataset, all methods predicted aggressive outcomes correctly for the subsequent 10 months. Beyond this 10-month cut-off, the reliability of predictions decreased, reflecting shifts in the demographic composition of the group. We also demonstrate how a data-splitting approach provides researchers not only with a means of determining the most reliable method for their dataset but also allows them to assess how rank reliability changes among age–sex classes in a social group, and so tailor their choice of method to the specific attributes of their study system.

Key words: **data-splitting approach, dominance hierarchy, nonsequential approach, real-world data, reliability, sequential approach.**

INTRODUCTION

Dominance hierarchies are key to understanding social structure across many animal taxa. Recognition of their importance, and the need to represent them accurately, has driven the development of a variety of methods for inferring dominance hierarchies from observational data (reviewed in De Vries 1998; Bayly et al. 2006; Briffa et al. 2013). Given the array of options available, selecting the method that best fits a given dataset can thus prove challenging. So, how does one choose? One obvious possibility is to refer to the existing literature in order to assess which method is most commonly used for the particular study system or species on which one works, determine how and why such a choice was made, how it was justified, and then follow suit with one's own data. The flaw with this strategy, as we have discovered, is that there is wide variability in the methods used within and between study systems and species, and researchers rarely, if ever, provide any justification for their choice.

This is not to say that researchers have not assessed ranking methods in a systematic fashion. Indeed, there is a substantive literature that has focused on determining the validity of ranking methods. However, these studies do not provide all the resources needed to enable researchers to make a fully informed choice for their own datasets as they carry their own limitations, a point on which we now expand. In one set of studies, a number of ranking methods have been applied to an empirical dataset, and greater levels of agreement between methods have been taken to indicate that the methods are accurate and robust (e.g., De Vries 1998; Gammell et al. 2003; Neumann et al. 2011; Balasubramaniam et al. 2013). Sánchez-Tójar et al. (2018) argue, however, that similar results across methods could also mean they suffer from a common bias, rather than necessarily providing an index of the methods' robustness. As such, they do not necessarily offer guidance on which method should be selected: They could all be as bad as each other if common flaws cannot be identified.

Address correspondence to C. Vilette. E-mail: c.vilette@uleth.ca.

Consequently, Sánchez-Tójar et al. (2018) have recommended the use of simulated data, where the dominance hierarchy is created by, and thus known to, the researchers. Validity can then be assessed by correlating the hierarchy produced by different ranking methods to the known hierarchy. Simulations can therefore test for and identify flaws in ranking methods and give researchers confidence that a method is actually measuring what it claims to be measuring. However, simulation studies make the implicit assumption that their results will apply equally to real-world empirical datasets. This can be misleading because simulated data are much cleaner and less noisy than real-world data. The latter will always contain a certain amount of noise, as well as a degree of uncertainty with respect to the outcomes of agonistic interactions, and both of these contribute to the underlying structure of the dataset. Furthermore, while most ranking methods focus almost exclusively on dyadic interactions, this does not preclude the possibility that contest outcomes are influenced by the presence of other individuals, whether through tacit or overt support (Maestripieri and Higham 2010; Bissonnette et al. 2015). Again, this generates potential noise in real-world empirical datasets. Finally, in the real-world, one is forced to acknowledge that the true hierarchy can never be known. Thus, no matter how high the validity of a method tested on simulated data, it is impossible to determine whether an inferred hierarchy does, in fact, map onto the true hierarchy in an empirical real-world example.

Given this, we suggest that, in addition to simulation studies of validity, there is also value in assessing the reliability of different ranking methods when applied to real-world datasets. Specifically, one can use the hierarchy generated by a particular method to test whether it will correctly predict the outcome of future dyadic aggression between two opponents. This, we feel, is the closest one can get to determining if any given method produces reliable and, therefore, useful measures in the real world. Here, we offer a means by which researchers can compare different ranking methods and determine which offers the greatest reliability for their specific dataset. This will allow researchers to offer a clear justification for their choice of method, improve transparency and increase the rigor of behavioral research.

We base our argument for the value of reliability on the notion that dominance hierarchies reduce uncertainty about the outcomes of contests between group members (Beaulieu et al. 2014; Mendonça-Furtado et al. 2014), and the assumption that the state of the hierarchy at a given time will be predictive of future interactions (Rowell 1974; Hinde 1976; Drews 1993; Roney and Maestripieri 2003; Strauss and Holekamp 2019). Here "prediction" (Bernstein 1981) alludes to the confidence with which the statistical asymmetry in dyadic contests predicts the outcome of any subsequent conflict within the same dyad. Thus, it follows that, if the inferred relative rank position of two animals can predict the outcome of a later aggressive interaction, then we have good evidence to suggest that a method is reliable.

To estimate a method's reliability, we developed a "data-splitting" approach. Splitting a dataset into two distinct components—typically referred to as the training and testing datasets—is a common technique in predictive modeling and machine learning (Dupuy and Simon 2007; Liu et al. 2016; Liu and Cocea 2017; Kuhn and Johnson 2020). In machine learning, one of the main requirements is to build computational models with high predictive and generalization capabilities (Mitchell 1997). When an appropriate model for data is not completely known, the data themselves can be used to select the appropriate model using data splitting. Here, the training dataset is used to build a model (Faraway 1998). Once trained, the predictive power of the model can be assessed by running it on the testing dataset (Dupuy and Simon 2007; James et al. 2013; Liu and Cocea 2017; Oghaz et al. 2017; Ho 2012; Siva 2018). In our case, the model outputs are the dominance hierarchies (comprising each individual rank, rating, or score) produced by each ranking method. We then assess whether these outputs correctly predict the outcome of aggressive dyadic encounters in our testing dataset.

To demonstrate how data splitting can be used with a real-world dataset, we make use of a 3-year dataset of aggressive interactions in vervet monkeys (*Chlorocebus pygerythrus*), a gregarious primate species. We investigate the performance of seven alternative ranking methods. Each method's performance was assessed by determining whether individual ranks/ratings/scores obtained from the training dataset could successfully predict the outcome of the aggressive interactions that occurred in the testing dataset. Given the time period covered by our dataset and the possibility of large changes in rank structure over time, we were particularly interested in comparing the performance of ranking methods that are characterized as nonsequential (i.e., where interactions are aggregated over time), to those characterized as sequential (i.e., where the data are not aggregated and thus the sequencing of interactions is retained in the data). Given that, over time, changes in both demographic and ecological variables will give rise to changes in the dynamics of social groups, we predict that sequential approaches will perform better than nonsequential ones in our dataset (Neumann et al. 2011; Williamson et al. 2016; Goffe et al. 2018).

Finally, we highlight the use of the "data-splitting" approach as an opportunity to quantify how reliability decays within particular age–sex classes of opponents. Focusing first on the whole group, we look at the overall trend in reliability across time. Due to the likelihood of demographic and ecological change mentioned above, we expected to see an overall decay in reliability. We then go on to investigate reliability at the adult sex-specific dyad level. In vervet monkeys, females are the philopatric sex and (often) inherit a rank position just below their mothers'(Fairbanks and McGuire 1985). Thus, we predicted that reliability in predicted outcomes for females would remain stable through time. In contrast, we anticipated a decay in reliability at the adult male dyad level due to migration between groups, which generates variation in male cohort composition, and hence greater rank instability. This analysis also allows us to determine whether some dyads are over-presented in the dataset, which could in turn have an impact on rank-order computation and, hence on a method's reliability when applied to group as a whole.

It is important to note that our aim here is not to determine the most reliable ranking method in any absolute sense. We also acknowledge that the use of a single real-world dataset to assess the reliability of a method holds its own problems. As such, we recognize the necessity of repeating these analyses on other populations and/or species to determine which patterns generalize, and which are highly specific to a given dataset. However, the goal of this study is to demonstrate the value of a training–testing approach that will enable researchers to identify the most reliable method for their particular dataset. That is, we present a "proof of concept" to illustrate that our approach has value and make no claims for the generality of findings with respect to the relative performance of each specific method. However, we consider the training–testing approach itself to be widely applicable precisely because it is not tied to any specific data requirements (e.g., no minimum amount of data required or specific length

of study period needed) or to any particular assumptions (e.g., regarding age–sex classes of individuals, linearity of the hierarchy, or the nature of the interactions included). Our method thus offers researchers a useful tool with which to conduct a convenient systematic reliability assessment of available methods. Adopting this approach will increase the reliability of the literature as a whole, by ensuring selected methods are offered with appropriate justification.

METHODS

Study site and subjects

Data used for these analyses were collected between January 2015 and December 2017 as part of a long-term field project at the Samara Private Game Reserve, South Africa (32o22′S, 24o52′E). We used data from one of our three study groups (RBM). All animals were fully habituated and individually recognizable. The study group occupied semi-arid riverine woodland (Pasternak et al. 2013), and group composition varied across the study period (Males: 20–6, Females: 13–8; Juveniles: 33–9; Infants: 11–2).

Behavioral data collection

Agonistic behaviors, identities of participants, and interaction outcomes were recorded ad libitum on all group members (i.e., across all sex and age categories). We wished to make use of the most diverse and complete dataset and chose to leave our dataset in its original form; hence, we retained agonistic encounters with juveniles and infants as well as those that involved coalitions (i.e., where one or more animal comes to the aid of another against a common opponent). Only unknown outcomes were discarded. We decided on this approach as we wanted our dataset to include uncertainty and noise to ensure we would not artificially increase the reliability of a given method by training and testing on a circumscribed and clearly determined array of interactions. By training the ranking methods with a noisy dataset, we can get a better sense of how well they can generate a reliable set of ranks without any form of prescreening of "acceptable" interactions. Agonistic behaviors included displacements, threats, chases, and bites. The visibility of the habitat, together with the modal presence of more than one observer (Henzi et al. 2013; McFarland et al. 2014), means it is unlikely that there was any systematic bias in the recording of agonism. We recorded 11,323 agonistic interactions between 66 individuals across the 36-month period. The initial training dataset comprised 8308 interactions, with the testing dataset accounting for the remaining 3031 interactions. For more details on the training dataset structure, see Supplementary Material S1.

Methods used to infer ranks and ratings

Among the tested ranking methods, it is possible to distinguish two families that differ in their overall approach (Table 1). The first family is based on the sequence in which interactions occur (which we refer to here as sequential methods). It includes the Elo-rating method (Elo 1978), as well as two of its variants: the Bayesian Inference (BI) approach (Goffe et al. 2018) and the modified Elo-rating (Newton-Fisher 2017).

The second relies on interaction matrices and comprises the David's score method (David 1987; Gammell et al. 2003), the Inconsistencies and Strength of Inconsistencies (I&SI) method (De Vries 1998), and the Percolation and Conductance (P&C) method (Fujii et al. 2015). Finally, the randomized Elo-rating

Table 1
Summary of the different methods tested in this study

Method	Outcome	Analytical details
I&SI	Ordinal rank order	Package: compete Function: isi13 with nTries = 450
David's score	Individual overall success	Packages: compete, steepness and EloRating Indices: Pij and Dij
Percolation and conductance	Rank order + dominance uncertainty	Package: Perc MaxLength= 2 and 4
Randomized Elo-rating	Individual overall success + the uncertainty around the estimates	Package: aniDom with n.rand = 1000
Original Elo-rating	Individual overall success	Package: EloRating Initial scores: 1000 and k = 100
Modified Elo-rating	Individual overall success	Newton-Fisher (2017) code Four categories of aggression intensity. Lowest starting at k = 200, with k increasing by 25 per aggression intensity
Bayesian inference	Individual overall success, start ratings, the Elo-rating winning/losing shift coefficient (k) + the uncertainty around the estimates	Goffe et al. (2018) code

(Sánchez-Tójar et al. 2018) was also included to the nonsequential methods. Despite being derived from a sequential approach, the changes implemented mean that this method shares many common features with nonsequential methods. For a general introduction to each method's background, see Supplementary Material S2.

As noted above, several different statistical packages and options are available for the David's score, thus we assessed 13 methods in all. Regarding the computation of David's scores, we used the functions offered within each package. These functions differed in their input and/or their way of dealing with draws (i.e., undecided interactions where there were no unambiguous winners and losers), hence potentially leading to differences in inferred scores.

We used R to conduct all rank-order estimations and subsequent analyses. Reliability was calculated using a custom package "rankReliability." This package provides researchers with the opportunity to estimate how reliable their inferred ranks are through time, while giving them the freedom to choose their preferred ranking method, the dataset of their choice (e.g., including juveniles/polyadic interactions, keeping/excluding draws, and so on), as well as how to split the data. The code can be found at https://github.com/tbonne/rankReliability.

While the I&SI and P&C approaches produce ordinal ranks and David's score gives cardinal scores, the rest of the methods produce ratings as outputs. For simplicity, we refer to ranks, scores, and ratings as outputs. The choices made regarding the analyses of each method are detailed in Supplementary Material S3.

Construction of training and testing datasets and comparison of methods

To assess the performance of each method we estimated 1) the average percentage of future interactions correctly predicted, 2) the amount of data required to make reliable predictions, and 3) the rate of decay in

prediction accuracy. The first measure of performance evaluates the overall ability of each method, while the second looks at the sensitivity to training sample size, and the third captures the temporal stability in each method's future predictions (i.e., does the accuracy of predicted outcomes decline through time and, if so, how fast).

Determining the method's average reliability: how do nonsequential and sequential methods compare?

Data-splitting ratios often vary across studies, making it difficult to offer uniform guidelines on how data should be partitioned. Although some authors recommend using 70% of the data as the training set and 30% as the testing set (Liu and Cocea 2017), others prefer a ratio of 75:25 (Oghaz et al. 2017) or 80:20 (Siva). Whatever the ratio chosen, there are two conditions that should be kept in mind when splitting the dataset: The training set must be large enough to estimate meaningful ranks, and the testing dataset must be long enough to estimate mean predictive performance. This excludes any extreme cuts, for example, 99/01.

Here, we chose to use the first 80% (2.1.2015–25.4. 2017) of our data to train the methods, with testing undertaken on the remaining 20% (26.4.2017–31.12. 2017; Figure 1a; Shah 2017). This ensured that we always had a training dataset with a sufficient number of observations to infer reliable outputs. We also excluded from analysis any animals that were present only during the testing phase of the dataset, but retained those present only in the training dataset, as the latter are able to provide information about their opponents. For each method, we calculated dominance hierarchies from our training dataset. As the estimated outputs are measured on different scales (e.g., ratings and scores), we converted these to ordinal ranks and used these new outputs for the rest of our study.

The first step in our analysis was to compare the ranking structure of each method. To do so, we visualized the data using a hierarchical clustering approach, which assembled results according to their similarity. Initially, each method was assigned to its own cluster. The algorithm then proceeded, joining the two most similar clusters at each stage and continuing until there was just one single cluster. In this way, methods that were most similar to each other were combined into branches, which were then fused higher up in the clustering process. Euclidean distance was used to measure the dissimilarity between each pair of methods. The hclust R function (stats package) was used to generate this hierarchical clustering.

Following this initial comparison, rank orders were then used to assess how well they matched the outcome of dyadic aggressive interactions in the testing dataset (matched = 1 or no match = 0). In other words, did the winner of the interaction in the testing dataset have a higher inferred rank than the loser? In the case of tied interactions in the testing dataset, the outcome was correctly predicted only if both individuals had been assigned the same ordinal rank. The proportion of correctly predicted outcomes was then translated into a percentage to determine which methods achieved better results than others. Finally, we applied a multilevel Bernoulli model to investigate how these correctly predicted outcomes (matched = 1, not matched = 0) varied across each ranking method. We then visualized the average percentage of correctly predicted outcomes for each method, along with the variance, using a violin plot.

Determining the optimal amount of data required for inferring reliable ranks

Our second aim was to estimate the amount of data needed to infer reliable outputs. To do so, we kept our testing dataset constant and modified the length of the training dataset. Specifically, we maintained the same end date for the training dataset, while varying its start date. Thus, as the training dataset decreased in size, only the most recent observations were included. Our original training

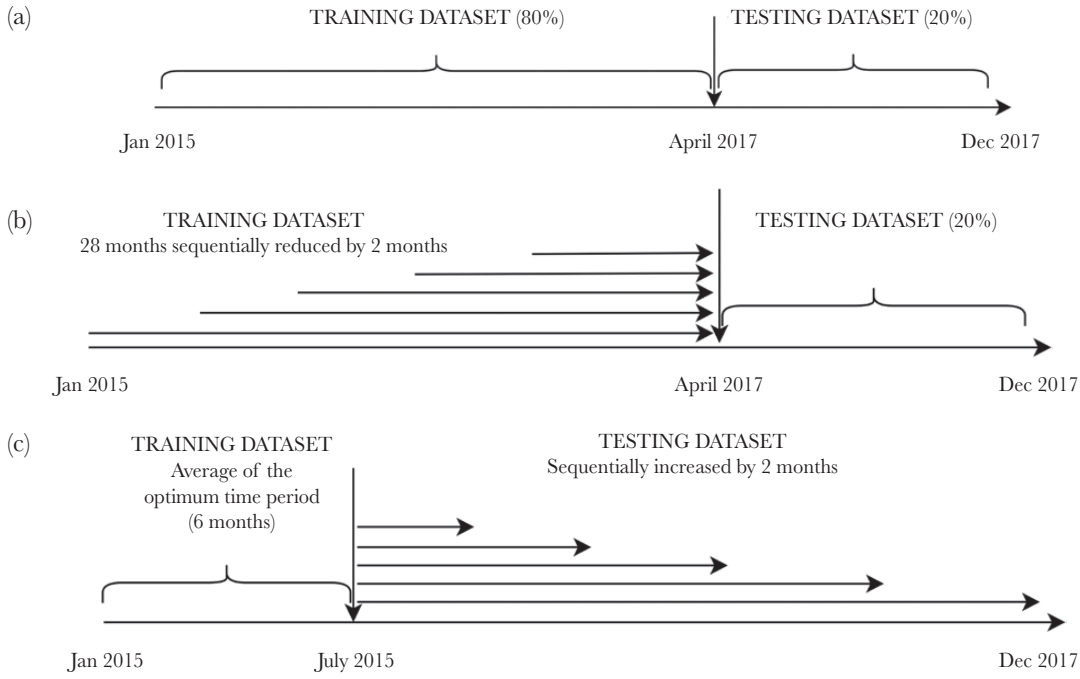

Figure 1

(a) Original approach used to assess methods' reliability. (b) Modified approach to assess the length of time period required for inferring reliable ranks. (c) Modified approach to assess the time period over which aggressive outcomes can be correctly predicted.

dataset comprised 28 months, which we reduced sequentially by 2 months, until only 2 months were left (i.e., we truncated the dataset starting from January 2015 toward April 2017; Figure 1b).

At each reduction in size, we computed the methods' output, converted them into ordinal ranks and assessed these against the interaction outcomes in the testing dataset. The same procedure outlined in the earlier section (see Determining the method's average reliability: how do nonsequential and sequential methods compare?) was used to calculate the percentage of correctly predicted outcomes. These percentages were then plotted in order to determine the amount of data needed to predict reliable ranks (i.e., ranks that were used to predict the outcome future interactions).

Determining the time period over which an inferred rank order can be used

To determine the length of time needed to correctly predict aggressive outcomes from obtained outputs, we performed the reverse procedure to that used above. That is, we gradually increased the size of our testing dataset and looked at its impact on the percentage of correctly predicted outcomes. Based on the results obtained in the previous analysis, we calculated the average optimal training dataset length across all methods. Using the average in this way meant that the training dataset could be kept constant, thus easing comparisons between the different methods. We used the remaining data as our testing dataset and systematically varied its length. We began with the 2-month period that followed directly from the training phase (July–September 2015) and then sequentially increased the testing dataset by 2 months until the 30-month limit was reached in December 2017 (Figure 1c).

Using the ordinal ranks inferred from the training dataset, we looked to see whether they matched the observed outcomes for each testing dataset. The percentage of correctly predicted ranks was plotted as a function of the testing dataset length to give us an insight into the rate of decay of each method's reliability (i.e., over what period can we use a given set of inferred ranks without any loss of reliability).

Using the testing dataset to probe reliability changes

To look at reliability changes in predicted aggressive outcomes, we made use of our testing dataset and the I&SI method, as the latter produced the most reliable outcomes for our dataset. We used the optimal training dataset length (i.e., 6 months) based on the results of our previous analysis (see Determining the optimal amount of data required for inferring reliable ranks). This ensured that we could compute reliable rank order while maximizing the size of the testing dataset (i.e., 30 months). Using the same approach as above, we looked to see whether the inferred ranks matched the observed outcomes in our testing dataset. We also determined whether the adult dyad participants were females or males for each observed outcome.

In this analysis, we first plotted the observed values of predicted outcomes at the group level to give a picture of the overall trend in reliability changes over the entire testing period. We then took our investigation down to the dyad level by plotting the observed values from the adult female–female and male–male dyads. We used the plotting function from the "rankReliability" package to plot the changes in outcome predictions over time at the group and sex-specific dyad level.

Ethical note

All protocols were noninvasive and adhered to the laws and guidelines of South Africa and Canada. Procedures were approved by the University of Lethbridge Animal Welfare Committee (Protocols 0702 and 1505).

RESULTS

Determining the method's average reliability: how do nonsequential and sequential methods compare?

Our dendogram identified the extent to which the methods provided similar estimates of rank order in our study group (Figure 2a). The output from the BI approach, the modified and the original Elo-rating cluster was the most different from the others, followed by the I&SI. The blue cluster, comprising all the David's scores methods, was the most similar in its outputs, followed by the cluster including the P&C and the randomized Elo-rating methods. Overall, the nonsequential methods produced a set of rank orders that were more similar to each other than to those produced by the sequential methods.

We visualized the average percentage of correctly predicted outcomes in the testing dataset in Figure 2b. While the dendogram indicates how similar the methods were in their outputs (rank order), Figure 2b shows the variance in the percentage of correct predictions produced by each method. In other words, they give us a sense of the "confidence" in the rank outputs produced (i.e., how effective were they at predicting future aggressive outcomes?). If we look at the red cluster (BI, the original and modified Elo-rating), for example, we see that these methods produced similar outputs (Figure 2a) and yet they differed in their reliability (Figure 2b): The BI approach had a higher percentage of correct predictions than the modified Elo-rating. Another intriguing pattern is that the P&C/randomized Elo-rating (pink cluster) and the I&SI (green cluster) differed in their outputs (Figure 2a), but the randomized Elo-rating method's reliability was more similar to the I&SI than it was to the P&C methods (Figure 2b).

The overall percentage of correctly predicted outcomes for each method is given in Table 2. These indicated that all methods did well in inferring reliable ranks (i.e., those that predicted future interaction outcomes). The BI method provided the best fit to the data, predicting 82.2% of aggressive outcomes, followed by the original Elo-rating method (81.0%). The two P&C variants (maxLength4 and maxLength2) produced an identical value of 80.6%, followed by the I&SI with 79.6%. The David's score obtained from the three different packages, and via the two different functions (Dij and Pij), were the lowest performing with values ranging from 76.4% to 79.3%, along with the modified and randomized Elo-rating with respectively a percentage of predicted outcomes of 79.2 and 79.1. The David's scores from the "EloRating" package and those from the "steepness" package (Pij and Dij function) gave the exact same percentage outcomes. Compared to Dij function, the Pij predicted a higher number of reliable outcomes across all three packages used. Moreover, the "compete" package had a higher efficiency than the "EloRating" and "steepness" packages (as the "EloRating" and "steepness" packages presented the same global percentage of reliability, as well as the same patterns throughout the rest of the analysis. We only use the "steepness" package in what follows from here). In general, and rather to our surprise, the family of sequential approaches was not more reliable with respect to predicting future aggressive outcomes. Taken together, these results showed that, despite these methods differing in their approach and the nature of their outputs, they all showed a high level of reliability when predicting the outcomes of future aggressive interactions.

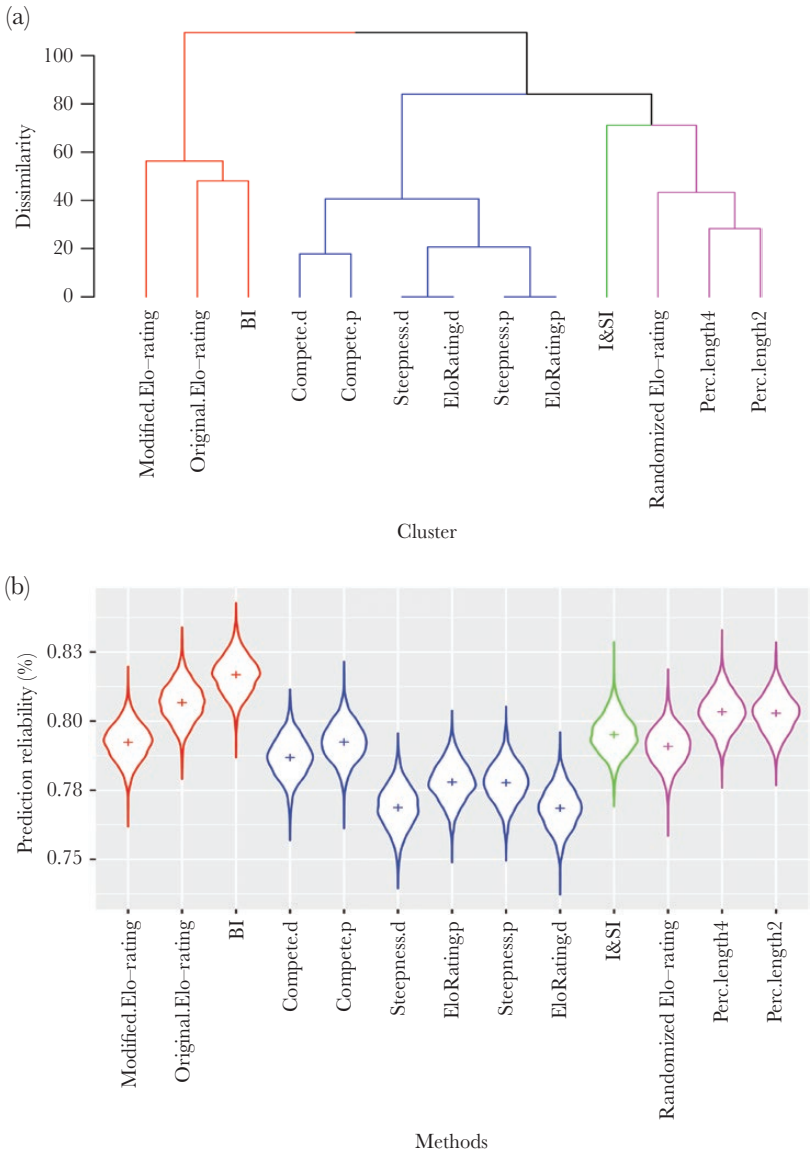

Figure 2

(a) Similarity between the rank orders produced by each method. The height of the split, on the vertical axis, indicates the similarity of rank order between two methods. The higher the split, the less similar the methods were in terms of their outputs. (b) Similarity of methods' reliability. Distribution of the percentage of correctly predicted outcomes across the methods used. Each cross represents the average estimate of the percentage of correctly predicted outcomes. Each color represents a dissimilarity cluster. Red, cluster represented the sequential approaches while pink, blue, and green ones were nonsequential methods. Blue corresponded to the David's score variants.

Determining the optimal amount of data required for inferring reliable ranks

We found that our original assumption was borne out: The reliability of predicted outcomes was not greatly affected by the length of the training dataset (Figures 3 and 4). The nonsequential approaches (Figure 3), however, did show more variation in the percentage of correctly predicted outcomes depending on the length of the training dataset. Despite this sensitivity, a maximum value for reliability could be found for each method. On average, these peaks occurred at 6 months, which we suggest represents the optimal length of time period needed to correctly predict future outcomes in this dataset. As the number of months in the training dataset increased, the nonsequential methods showed more of a decay in reliable prediction compared to the sequential methods (as one would expect). Moreover, the I&SI method displayed the highest percentage of correctly predicted outcomes when the training dataset spanned the period of 4–16 months.

Figure 4 shows the family of sequential approaches and shows that the length of the training dataset did not have any impact for two of the three methods: the original Elo-rating and the BI. In fact, both methods performed well regardless of the length of the training dataset. There is some evidence to suggest, however, that the original Elo-rating method and the BI approach produced slightly more reliable predictions once the training dataset exceeded 4 months and 10 months, respectively. As for the modified Elo-rating method, it shows a sensitivity to the amount of data in the training dataset. A first peak in reliable prediction appeared at training dataset lengths between four and 12 months. Beyond 12 months, reliability decreased and then stagnated as the training dataset length increased.

With the exception of P&C and randomized Elo-rating, all methods from the nonsequential approaches produced an optimal percentage of correct prediction with a 6-month training set. The sequential approaches reached saturation sooner at 4 months, although the BI approach showed a temporary decrease at 8 months. In our next analyses, we used a 6-month period for the training dataset as this represented the best compromise in terms of enabling comparison across all methods. Shortening the training set in this manner gave us a larger testing set of 30 months in total (2.5 years) to assess our third question.

Determining the time period over which an inferred rank order can be used

The percentage of correct predictions for each testing dataset length is plotted in Figure 5a–c. At first sight, all methods showed the same pattern. First, a decline in outcome predictability occurred at 4 months, which was then followed by a peak in prediction reliability, corresponding to a testing dataset of 8–10 months in length (Figure 5a). Past 10 months, the reliability of predicted outcomes showed a constant and slow decay.

The I&SI, as well as the P&C approach, stood out as the methods that led to the highest percentage of correctly predicted outcomes over the whole testing dataset's length (Figure 5a), followed by the BI approach. The remaining methods were clustered with a lower percentage of correctly predicted outcomes throughout the testing dataset's length.

Moving away from the general pattern, Figure 5a highlights the differences between nonsequential and sequential approaches. Again, the latter were no more reliable than nonsequential approaches. However, they distinguished themselves in the sense that all the tested methods were good at correctly predicting outcomes (i.e., they clustered in the centre of the range of performance), whereas the nonsequential methods showed a much wider range of variation.

In order to examine these patterns in more detail, we separated the nonsequential and sequential approaches to enable the similarities and differences—between and within each family—to be seen more easily. With regard to the nonsequential approaches (Figure 5b), both the I&SI and P&C methods displayed a pattern of fluctuation, whereas David's score showed a smoother curve with a constant decrease in reliability once past a testing dataset of 2 months. The "compete" package appeared to perform better than the "steepness" package; both packages produced similar curves. The randomized Elo-rating also presented fluctuations and was the method showing the lowest reliability through time. Finally, the I&SI and P&C methods displayed a higher percentage of correctly predicted outcomes compared to the sequential methods throughout the whole testing dataset length, except for P&C with maxLength 2 past 26 months of the testing dataset.

With respect to sequential approaches (Figure 5c), the BI and the original Elo-rating displayed the general pattern described above. The modified Elo-rating showed the same pattern in prediction reliability until 12 months, where its percentage of correctly predicted outcomes started increasing with the length of the testing dataset.

Table 2

Percentage of correctly predicted aggressive outcomes over an 8-month testing dataset for our vervet monkeys' troop (RBM)

Method	Package	Option	% prediction RBM
I&SI	Compete	isi13	79.6
David's scores	Compete	Dij	78.6
		Pij	79.3
	Steepness/ EloRating	Dij	76.4
		Pij	77.5
Percolance and conductance	Perc	maxLength4	80.6
		maxLength2	80.6
Elo-rating	EloRating	default	81.0
	Newton-Fisher's code	K variation	79.2
Bayesian inference	Goffe's code	Rstan	82.2
Randomized elo-rating	aniDom	n.rands = 1000	79.1

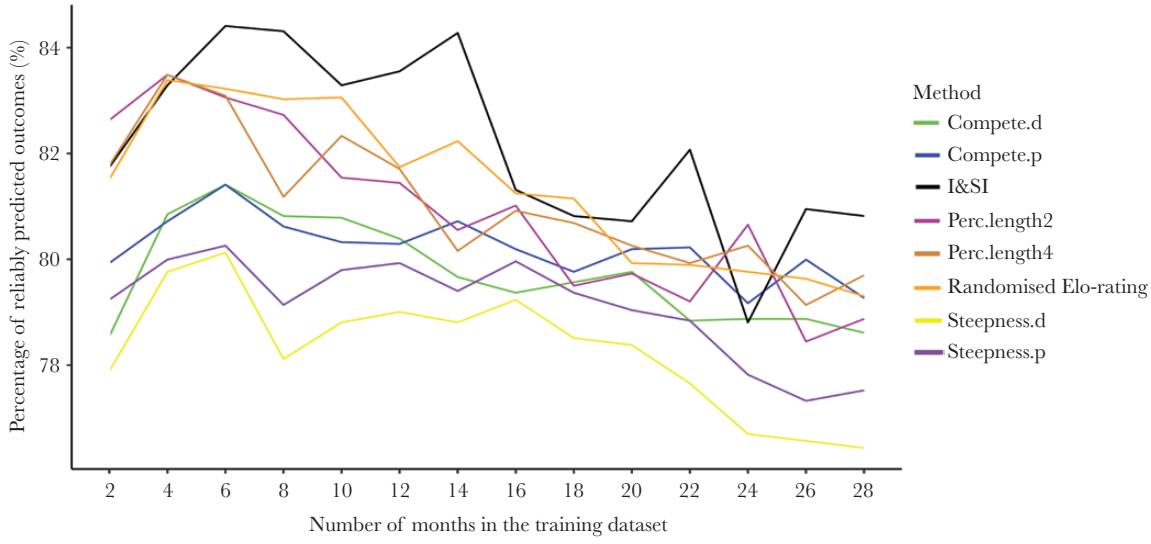

Figure 3

Variation of the percentage of outcome prediction with the nonsequential methods in function the number of months included in the training dataset.

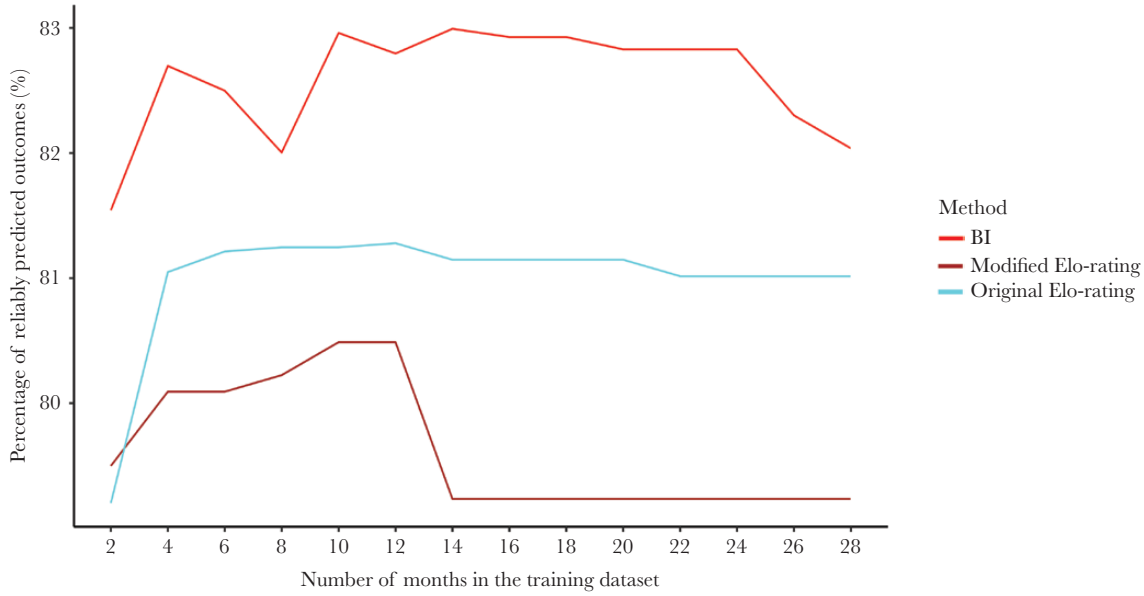

Figure 4
Variation of the percentage of outcome prediction with the sequential methods as a function the number of months included in the training dataset.

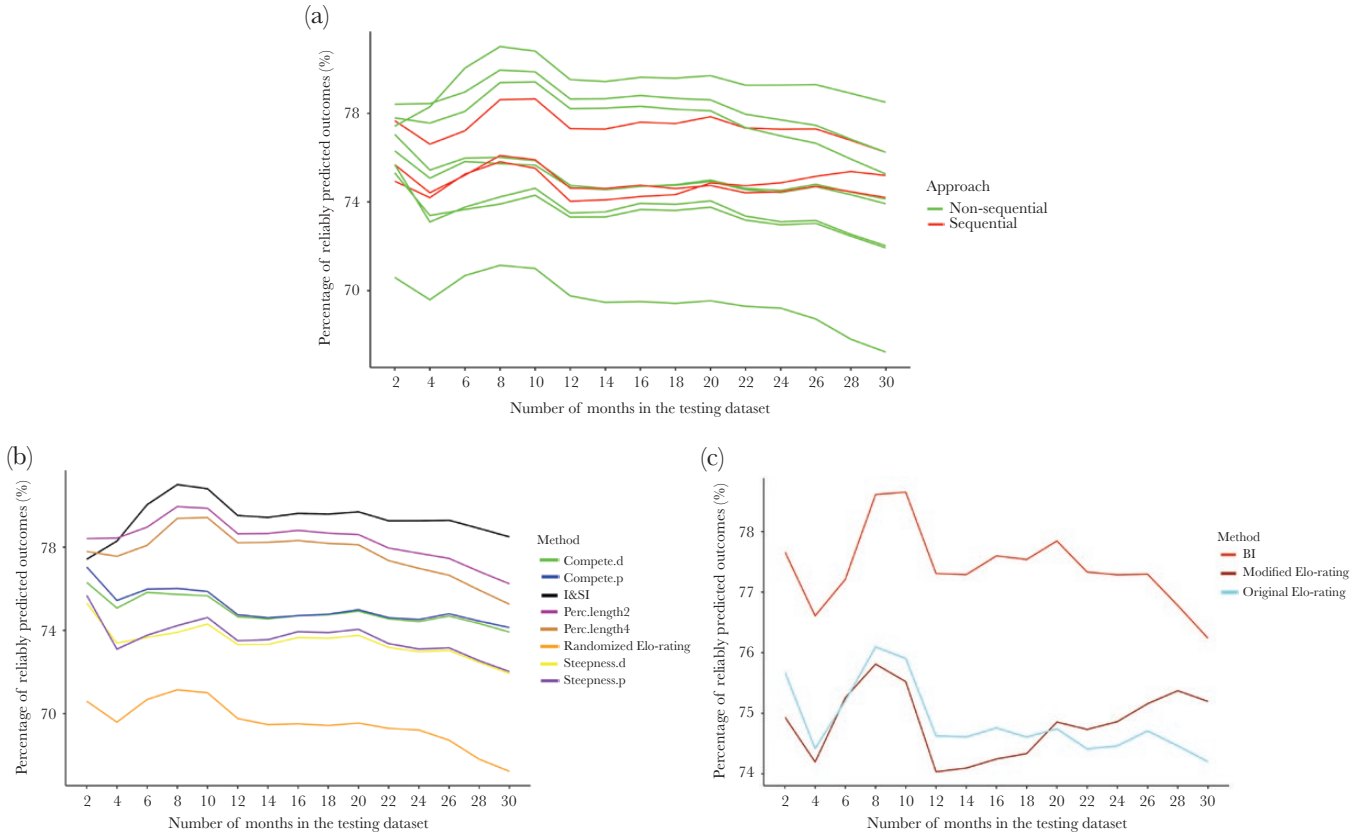

Figure 5
Variation of the percentage of correct predictions as a function of testing dataset length across (a) all methods, (b) sequential only, and (c) nonsequential approaches only.

Using the testing dataset to probe reliability changes

The reliability changes in the predicted outcomes of aggressive interactions over the 30-month testing dataset are plotted in Figure 6a–c. Looking at the global trend (Figure 6a), we found that our original assumption did not hold. Instead of a predicted decay in reliability of predicted outcomes, the overall reliability remained very stable across the 30 months.

In order to examine these reliability changes in more detail, we looked at the predicted outcome patterns at a finer scale: the adult

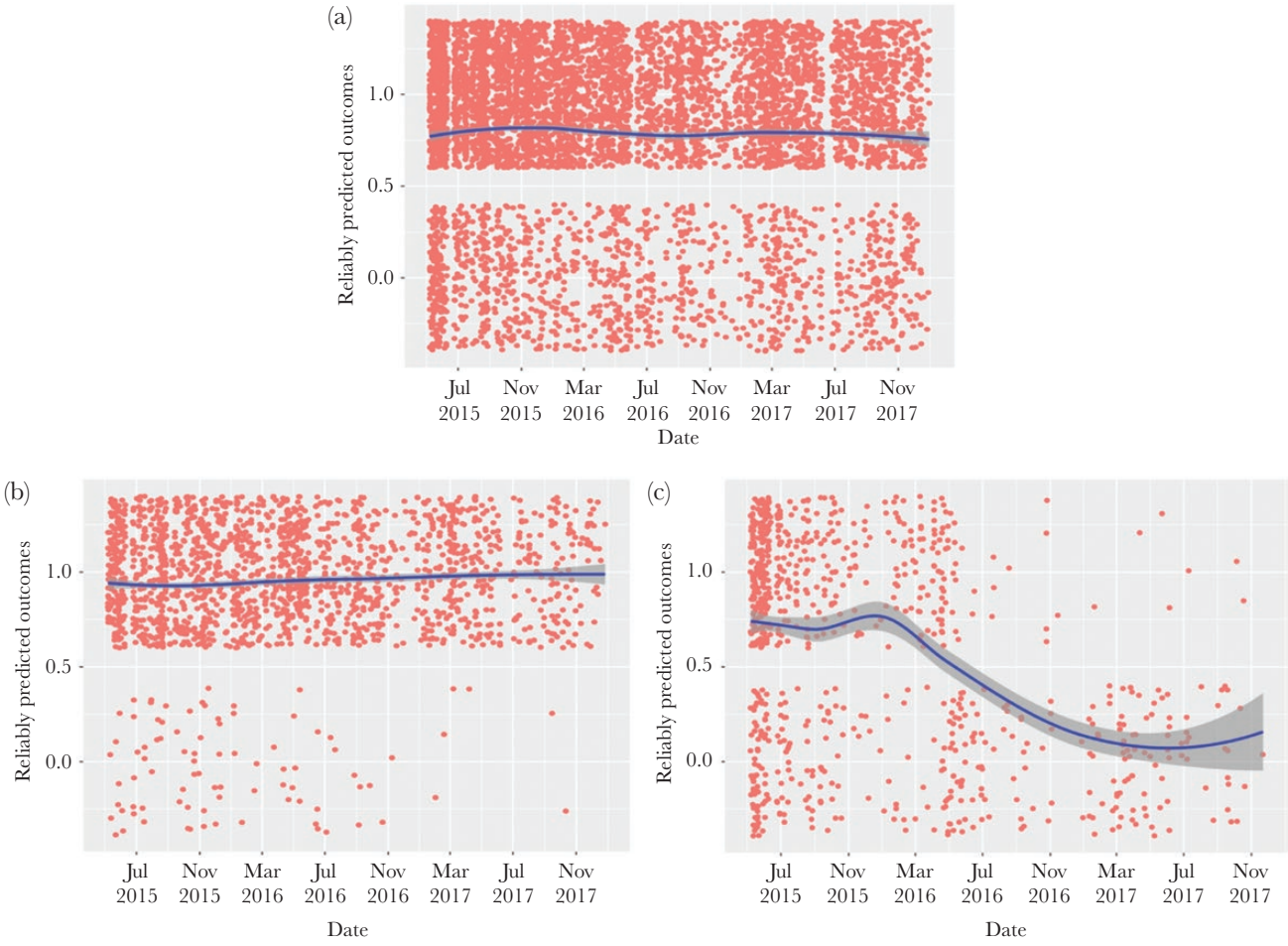

Figure 6
Variation of the fitted (line with 95% confidence interval) and observed (jittered points) values of outcome predictability (correct prediction = 1, not correct = 0) as a function of time at the (a) group level, (b) adult female dyad level, and (c) adult male dyad level.

sex-specific dyad level. We separated the adult female and adult male dyads to enable the patterns to be seen more easily (Figure 6b and c). With respect to the adult female dyads (Figure 6b), our assumption held: Reliability remained stable throughout the whole testing dataset. Regarding the adult male dyads, a stable pattern in the reliability was observed for the first 5 months, followed by a peak in over the next 3 months (Figure 6c). Past this peak, reliability showed a quick decline.

DISCUSSION

We have presented a training–testing procedure, and associated code, that will allow researchers to determine the most reliable method for calculating dominance ranks for their particular dataset. We used data from our own long-term study of vervet monkeys to demonstrate the utility of the method. Overall, we found that all methods tested performed well at correctly predicting future aggressive outcomes in our dataset, that is, all methods were reliable. With respect to the impact of the length of the training dataset, all methods again displayed high reliability from the very start (i.e., with little to no training period), but all showed improvement as the length of the training period increased. With respect to the length of the testing dataset, we found that (with a fixed training period of 6 months), all methods could correctly predict aggressive outcomes for the subsequent 10 months. Finally, looking at changes in predicted outcomes over time, adult male dyads showed more variability than adult female dyads.

More specifically, our first analysis revealed that, despite some differences in obtained rank order, all methods succeeded at inferring reliable ranks. We suspect this may be because individuals whose ranks were inaccurately assigned were those that did not interact frequently; hence they did not appear often in the testing dataset, and so did not have an impact on the reliability of a given method. This finding goes against our prediction that sequential methods would perform better than nonsequential ones. Our prediction here was based on the assumption—built into the nonsequential family of methods—that all individuals represented in a matrix were coresident at some point (i.e., all had the chance to interact). With an original training dataset of 28 months length, we predicted that this assumption would most likely be violated and hence lower the performance of nonsequential methods relative to sequential ones. Contrary to our prediction, the nonsequential family performed well at producing reliable ranks, suggesting that violation of this assumption was of minor significance. In this study, we analyzed data from a species in which stable ranks tend to persist through time, and in which few rank reversals occur. This specific social dynamic may have enhanced the reliability of the nonsequential methods due to the large amounts of data included (i.e., more agonistic interactions to work with). If this were so, however, we would have expected to see poorer performance over

shorter time frames (investigating in the second part of the analysis), and this was not the case.

In our second analysis, we found that only a short training period was necessary to infer reliable ranks across all methods. The sequential methods, however, were less sensitive to the amount of data present in the training dataset, and hence showed a constant efficiency regardless of the length of training, compared to the nonsequential approaches. This is not unexpected given that sequential approaches track rate variations continuously and update the ratings after each interaction. From the nonsequential perspective, the combination of high overall reliability with some temporal fluctuation suggested that, in our study species, individual position in the rank order shows a form of "regression to the mean." That is, individuals may experience very mild shifts in relative rank position up or down the hierarchy across time, but nevertheless occupy more or less the same "absolute" position. This, in turn, suggests that rank changes may reflect the ecological and demographic contexts in which they occur, rather than pointing to genuine changes in inherent power. We should also highlight that, when the training dataset did not exceed 22 months, it was the nonsequential I&SI and P&C methods that produced the highest percentage of correctly predicted outcomes.

Finally, the training–testing procedure gave us greater insight into the working of the randomized Elo-rating method. When applied to a short training dataset (i.e., 6 months), the sequence of interactions clearly did not matter, and the method's reliability was high. As the training dataset increased up to 28 months, however, there was a decline in reliability, which reflects the fact that, over this period, demographic change was inevitable, and the order of interactions may well have begun to exert an influence on the structure of the hierarchy. Given this outcome with our data (and assuming this holds true across other datasets), this suggests that the randomized Elo-rating method will indeed prove useful in determining when interaction order matters in a given dataset, as originally suggested by Sánchez-Tójar et al. (2018).

We also found that, with the exception of the randomized Elo-rating, all methods correctly predicted aggressive outcomes for at least 10 months. Past this threshold, however, reliability in prediction decreased as the testing dataset increased in length. The decay was rather shallow, however, and there was still high predictability in aggressive outcomes, indicating overall rank stability during the period covered by our sample. This is not to say, however, that rank predictability did not fluctuate as the length of the training period increased, and it was apparent that the degree of fluctuation was dependent on the method used. The sequential methods produced more fluctuations than the nonsequential methods. This is probably because the former are likely to catch small shifts in rating position as they constantly update, whereas the nonsequential methods are more likely to produce a rank order that captures the overall social dynamic. The fluctuations observed suggest that rank shifts were occurring in the study group during particular periods, and so another advantage of using a training–testing procedure is that it provides researchers with a way to home in on periods of rank instability, which may prove useful when attempting to answer questions relating to the effects of dominance on various behaviors and in relation to ecological variables.

It is important to note here that the intention behind the data-splitting approach was to enable a better understanding of rank structure within a particular dataset, not to determine which method was absolutely the most reliable. That is, the specific results we have presented here may not generalize to other datasets. Indeed, differences may well be expected because other species and populations will vary in their frequency of agonistic interactions, the steepness of their hierarchy and the (a)symmetry in aggressive outcomes. Shizuka and McDonald (2015), for example, have shown that differences in dominance hierarchy structure across animals may be a consequence of the study design (e.g., how many animals to observe and how much interaction data to collect). Hence, our goal was to show that a training–testing approach can be applied to any dataset to determine the most reliable ranking method, and thus we consider this approach to be useful in and of itself. Having said this, it will be interesting to see whether any commonalities do, in fact, emerge across different datasets. It is therefore necessary to repeat these analyses on other populations and other species to determine what patterns might be more general, and which are highly specific to a given dataset. At present, we can say that the data-splitting approach allows researchers to assess which method will work best for their dataset, given the size of their sample, and the length of time over which the study was conducted.

For the purposes of comparison across methods, we converted all model outputs to ordinal ranks. Although we agree with Strauss and Holekamp (2019) that such conversion is useful for identifying hierarchy dynamics, we consider this to be a limitation of our study. In fact, we did not consider how the magnitude of rank differences might affect reliability, nor did we consider any uncertainty around rank calculations. These components may very well matter, especially in species where a linear rank order may not be representative of the social hierarchy. This point also serves to highlight the true advantage of methods like the BI, randomized Elo-rating and the P&C approach, which enable researchers to look at the uncertainty around ranks, and thus gain a more complete understanding of the social hierarchy.

We also acknowledge that the use of empirical data does not allow us to distinguish between the two sources of error that could explain differences in the methods' performance: 1) inadequacies of the method and 2) real biological change. Thus, to reiterate and emphasize the point made above, our findings are only valid with respect to our data and cannot be assumed to apply to other datasets. Within our dataset, however, we think it is safe to assume that variation in a method's reliability compared to others does, in fact, reflect something about the method itself. Given that we tested all methods on the same training/testing datasets, any potential biological changes within the dataset should have been detected by at least some of the methods. This, of course, is where the usefulness of simulated data comes into play, as simulation allows one to tease apart these two sources of error more effectively, as well as gaining some more general insights into each method (Sánchez-Tójar et al. 2018). Our suggestion here is that the most informative approach will involve comparing simulation studies of methods using constructed datasets with reliability studies of methods applied in real-world settings; in this way, we can determine whether methods that show high validity also show high reliability in real-world contexts. In fact, one of the latest studies to date (Strauss and Holekamp 2019) used both simulated and empirical data from a long-term field study of spotted hyenas (*Crocuta crocuta*) to assess the performance of the modified and unmodified methods in inferring longitudinal hierarchies.

Finally, using the testing dataset to investigate reliability changes in predicted outcome, at the group and dyad level, allowed us to get a better understanding of social dynamics. Adult male dyads displayed most variation, which was not detectable at the group level, while the adult female dyads remained stable across the entire study

period (30 months). Thus, the data-splitting approach can also help to achieve a better understanding of how dominance ranks vary within a given group over time in relation to factors like sex and age class. A multiscale approach can thus provide a more comprehensive perspective on the temporal dynamics in outcome predictions, and hence the social ranks, through time. In other words, we consider that data-splitting provides researchers with an excellent tool to probe the social dynamics of their study species in more depth, rather than simply offering a means of determining the most reliable ranking method. It would be also interesting (and possible) to look at the outcomes of aggressive encounters that do not match the ranks assigned to each participant, when both were extracted on the same day. This would give us a better idea of the true degree of outcome unpredictability, allowing us to assess whether uncertainty in rank assignment is due to the nature of the aggressive interaction itself or whether it reflects something about the context in which it takes place.

In conclusion, a data-splitting approach gives researchers the power to tailor the selection of a dominance-ranking method to the particular nature of the dataset they are using. In addition, it provides insights into group dynamics, which can enable researchers to home in on regions of their dataset that will permit analyses into how and why rank shifts occur and discover the underlying causes of both rank stability and unpredictability across time.

SUPPLEMENTARY MATERIAL

Supplementary data are available at *Behavioral Ecology* online.

FUNDING

This work was supported by the Natural Sciences and Engineering Research Council of Canada Discovery grants to L.B. and P.H. and by the Canada Research Chairs program to L.B., T.B., and C.V.

We thank Mark and Sarah Tompkins for the permission to work at Samara, Kitty, and Richard Viljoen for their continued logistic support. We thank Damien Farine and Alfredo Sánchez-Tójar for constructive comments. We are also very grateful to the many research assistants who contributed to the database, as well as the monkeys (and George), without which only simulated data would be available.

Data availability: Analyses reported in this article can be reproduced using the data provided by Vilette et al. (2020).

Handling editor: John Quinn

REFERENCES

Balasubramaniam KN, Berman CM, De Marco A, Dittmar K, Majolo B, Ogawa H, Thierry B, De Vries H. 2013. Consistency of dominance rank order: a comparison of David's Scores with I&SI and Bayesian methods in macaques. Am J Primatol. 75:959–971.

Bayly KL, Evans CS, Taylor A. 2006. Measuring social structure: a comparison of eight dominance indices. Behav Processes. 73:1–12.

Beaulieu M, Mboumba S, Willaume E, Kappeler PM, Charpentier MJ. 2014. The oxidative cost of unstable social dominance. J Exp Biol. 217:2629–2632.

Bernstein IS. 1981. Dominance: the baby and the bathwater. Behav Brain Sci. 4:419–429.

Bissonnette A, Perry S, Barrett L, Mitani JC, Flinn M, Gavrilets S, de Waal FBM. 2015. Coalitions in theory and reality: a review of pertinent variables and processes. Behaviour. 152:1–56.

Briffa M, Hardy IC, Gammell MP, Jennings DJ, Clarke DD, Goubault M. 2013. Analysis of animal contest data. In: Hardy IC, Briffa M, editors. Animal contests. Cambridge: Cambridge University Press. p. 379.

David HA. 1987. Ranking from unbalanced paired-comparison data. Biometrika. 74:432–436.

De Vries H. 1998. Finding a dominance order most consistent with a linear hierarchy: a new procedure and review. Anim Behav. 55:827–843.

Drews C. 1993. The concept and definition of dominance in animal behaviour. Behaviour. 125:283–313.

Dupuy A, Simon RM. 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 99:147–157.

Elo AE. 1978. The rating of chess players, past and present. New York: Arco.

Fairbanks LA, McGuire MT. 1985. Relationships of vervet mothers with sons and daughters from one through three years of age. Anim Behav. 33:40–50.

Faraway JJ. 1998. Data splitting strategies for reducing the effect of model selection on inference. Comput Sci Stat. 30:332–341.

Fujii K, Jin J, Shev A, Beisner B, McCowan B, Fushing H. 2015. Perc: using percolation and conductance to find information flow certainty in a direct network. R Package Version 0.1. https://cran.r-project.org/web/packages/Perc/index.html.

Gammell MP, De Vries H, Jennings DJ, Carlin CM, Hayden TJ. 2003. David's score: a more appropriate dominance ranking method than clutton-brock *et al.*'s index. Anim Behav. 66:601–605.

Goffe AS, Fischer J, Sennhenn-Reulen H. 2018. Bayesian inference and simulation approaches improve the assessment of elo-ratings in the analysis of social behaviour. Methods Ecol Evol. 2018:2131–2144.

Henzi SP, Forshaw N, Boner R, Barrett L, Lusseau D. 2013. Scalar social dynamics in female vervet monkey cohorts. Philos Trans R Soc Lond B Biol Sci. 368:20120351.

Hinde RA. 1976. Interactions, relationships and social structure. Man. 11:1–17.

Ho R. (2012). Predictive analytics: evaluating model performance - Dzone performance. https://dzone.com/articles/predictive-analytics. Accessed 26 February 2020.

James G, Witten D, Hastie T, Tibshirani R. 2013. An introduction to statistical learning. Vol. 112, p. 18. New York: springer.

Kuhn M, Johnson K. 2020. Feature engineering and selection. New York, NY: Chapman and Hall/CRC.

Liu H, Cocea M. 2017. Semi-random partitioning of data into training and test sets in granular computing context. Granular Computing 2:357–386.

Liu H, Gegov A, Cocea M. 2016. Rule-based systems: a granular computing perspective. Granular Computing. 1:259–274.

Maestripieri D, Higham J. 2010. Revolutionary coalitions in male rhesus macaques. Behaviour. 147:1889–1908.

McFarland R, Barrett L, Boner R, Freeman NJ, Henzi SP. 2014. Behavioral flexibility of vervet monkeys in response to climatic and social variability. Am J Phys Anthropol. 154:357–364.

Mendonça-Furtado O, Edaes M, Palme R, Rodrigues A, Siqueira J, Izar P. 2014. Does hierarchy stability influence testosterone and cortisol levels of bearded capuchin monkeys (*Sapajus libidinosus*) adult males? A comparison between two wild groups. Behav Processes. 109 Pt A:79–88.

Mitchell TM. 1997. Machine learning. Burr Ridge, IL: McGraw Hill. 45:870–877.

Neumann C, Duboscq J, Dubuc C, Ginting A, Irwan AM, Agil M, Widdig A, Engelhardt A. 2011. Assessing dominance hierarchies: validation and advantages of progressive evaluation with elo-rating. Anim Behav. 82:911–921.

Newton-Fisher NE. 2017. Modeling social dominance: elo-ratings, prior history, and the intensity of aggression. Int J Primatol. 38:427–447.

Oghaz MM, Maarof MA, Rohani MF, Zainal A, Shaid SZM. 2017. A hybrid color space for skin recognition for real-time applications. J Comput Theor Nanosci. 14:1852–1861.

Pasternak G, Brown LR, Kienzle S, Fuller A, Barrett L, Henzi SP. 2013. Population ecology of vervet monkeys in a high latitude, semi-arid riparian Woodland. Koedoe. 55:01–09.

Roney JR, Maestripieri D. 2003 Social development and affiliation. In: Maestripieri D, editor. Primate psychology. Cambridge: Harvard University Press. pp. 171–204.

Rowell TE. 1974. The concept of social dominance. Behav Biol. 11:131–154.

Sánchez-Tójar A, Schroeder J, Farine DR. 2018. A practical guide for inferring reliable dominance hierarchies and estimating their uncertainty. J Anim Ecol. 87:594–608.

Shah T. 2017. About train, validation and test sets in machine learning. https://tarangshah.com/blog/2017-12-03/train-validation-and-test-sets/. Accessed 27 February 2020.

Shizuka D, McDonald DB. 2015. The network motif architecture of dominance hierarchies. J R Soc Interface. 12:20150080–20150080.
Siva C. 2018. Machine learning and pattern recognition. AI Zone journal. https://dzone.com/articles/machine-learning-and-pattern-recognition.
Strauss ED, Holekamp KE. 2019. Inferring longitudinal hierarchies: Framework and methods for studying the dynamics of dominance. J Anim Ecol. 88:521–536.
Vilette C, Bonnell TR, Henzi SP, Barrett L. 2020. Comparing dominance hierarchy methods using a data-splitting approach with real-world data. Behav Ecol. https://github.com/tbonne/rankReliability or doi:10.5061/dryad.612jm641s
Williamson CM, Lee W, Curley JP. 2016. Temporal dynamics of social hierarchy formation and maintenance in male mice. Anim Behav. 115:259–272.