**Choosing when and how long to sample: Estimating the distribution of behavioral states in social data on a sampling budget**

Colin M. Lynch 1, Michaela Starkey2, Stephen C. Pratt1, Ioulia Bespalova1, Theodore P. Pavlic1, 2, 3, 4, Jennifer H. Fewell1

1School of Life Science, Arizona State University, Tempe, AZ, USA 85287

2School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA 85287

3School of Complex Adaptive Systems, Arizona State University, Tempe, AZ, USA 85287

4School of Sustainability, Arizona State University, Tempe, AZ, USA 85287

Corresponding author:

Colin Lynch1

[cmlynch2@asu.edu](mailto:cmlynch2@asu.edu)

(435) 881-8878

916 S Pioneer, Mesa, AZ, 85204

Word count: 11,055

**Highlights: 3 to 5 bullet points (maximum 85 characters including spaces and each bullet point should be on a separate line)**

* **Ethologists use continuous or instantaneous sampling to categorize behavior.**
* **Here we examine the benefits of blending these two sampling techniques.**
* **Instantaneous sampling can categorize behavior over a long period of time.**
* **Continuous sampling gives a detailed description of behavior.**
* **Many small continuous samples optimally trades-off different behavioral metrics.**

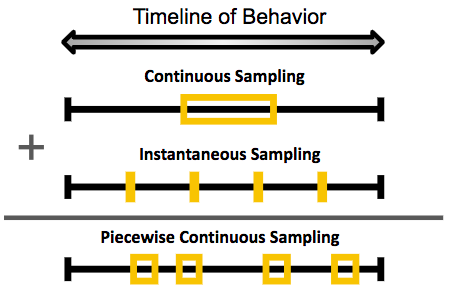
**Abstract (300 words).** Capturing qualitative features of animal behavior often requires manual tracking. Continuous sampling can capture many salient features of behavior, but can be time consuming if many individuals need to be recorded over long time periods. Instantaneous sampling can be less costly and be performed over longer time periods, but it can also miss many relevant features of behavior. We therefore explore the gradient in between these two extreme sampling techniques by continuously sampling multiple intervals of behavior which are randomly scattered throughout a time series. We first develop a protocol for estimating the number of instances necessary to reconstruct an animal’s behavior. We then divide this sample size into equally-sized segments which are randomly distributed across the time series. For each sample, we calculate two types of error. Independent error measures how poorly a sample estimates a feature of animal behavior that should be independent of the time at which the behavior occurs: the proportion of time spent in various states. Dependent error measures how poorly a sample estimates a feature of animal behavior that is dependent on the time at which the behavior occurs: average bout length for each state. We can then find the number of segments which minimizes both types of error simultaneously. Finally, we measure which behavioral attributes affect the optimal number of segments. Subsampling strategies were tested with recordings of harvester ant behavior and were validated with simulations. We find that an intermediate number of intervals optimally traded-off both types of error, and that this finding was robust to changes in underlying ant behavior, implying that this method could be used for other model organisms as well.

**Keywords:** Bout length, instantaneous sampling, continuous sampling, division of labor, *Pogonomyrmex californicus*, sample size, sampling technique, behavior

**Introduction**

All scientific datasets are finite due to monetary restrictions and limited time budgets. Some fields are capable of maximizing sample size for a time series by automating data collection with remote sensors (Ha et al., 2018; Kremer et al., 2017). However, although many phenomena can be recorded continuously, not all scientifically relevant data can be so easily collected without significant human assistance. While some behaviors such as activity level and velocity can be tracked via automated software (Birgiolas et al., 2017; Rice et al., 2020; Valentini et al., 2020), qualitative descriptions of discrete behaviors with an ethogram often require human observation ([Casarrubea](http://paperpile.com/b/zQgSv6/PIP2) et al., 2019). Attempts have been made to automatically categorize behavior by using tracking algorithms to infer the speed of these insects and their position relative to parts of the nest such as the entrance (Crall et al., 2018), but these approaches may not be able to capture subtle yet important behaviors and can misclassify actions. Manual tracking and behavioral coding are time consuming, especially if the relevant timescale is over hours or days, so we investigate optimal strategies for sub-sampling continuous behavior.

Two of the most used techniques for sampling animal behavior are instantaneous and continuous sampling. Continuous sampling is the continuous collection of all the behavior an animal does in a particular duration of time (Rose, 2000) and is considered the gold standard of behavioral sampling, as it simultaneously measures frequencies, durations, and the times at which behaviors are performed (Martin & Bateson, 1993). However, continuous sampling requires a great deal of manual labor when analyzing multiple individuals or if relevant timescales are long, and so frequent instantaneous or scan sampling techniques are sometimes used instead to approximate a continuous sample (Mitlöhner et al., 2001; Tacha et al., 1985). Instantaneous sampling can partially resolve this issue, as it reduces the amount of time necessary to analyze animal behavior (Hämäläinen et al., 2016). Instantaneous sampling is the recording of an individual’s behavior during preselected moments in time (Altmann, 1974). Instantaneous sampling is often used to minimize sampling bias (Mann, 1999) and to avoid systematic errors if intervals between samples are sufficiently short (Ary & Suen, 1983), although they are prone to random errors (Hämäläinen et al., 2016). Instantaneous sampling can capture the bout length of a behavior if the interval length is shorter than length of the bout (Lehner, 1992), however, bout length can be difficult to estimate without a preliminary dataset, and instantaneous/scan sampling misses momentary events or rare states (Lehner, 1992). To retain the benefits of both strategies, we explore the gradient between instantaneous and continuous sampling where multiple short intervals of behavior are recorded continuously as opposed to a single long interval (Fig. 1). That is, we adjust the number of intervals (I) to create an approximate continuum between instantaneous and continuous sampling. This type of sampling technique has not yet been named within ethology, so we call it *piecewise continuous sampling*. When I = 1, piecewise continuous sampling is equivalent to continuous sampling. As I increases and segments become smaller, then it approximates instantaneous sampling.

**

**Fig. 1. Illustration of piecewise continuous sampling versus continuous sampling.** Here, we make a remote record of a single animal’s behavior (e.g. video recording). Continuous sampling would require analyzing this video for an extended period of time, whereas instantaneous sampling would require analyzing the video for a single moment over regular intervals. Piecewise continuous sampling is a combination of these strategies, where there are multiple continuous segments spread throughout the recording. These segments, though, are randomly scattered rather than regularly spaced to avoid aliasing.

Each end of the spectrum of piecewise continuous sampling has different benefits. For I >> 1, this sampling technique can adopt the benefits of instantaneous sampling to avoid systematic bias and sample across a long period of time. This should allow strong estimates for independent metrics of animal behavior. Independent metrics are features of animal behavior that are independent of the time at which the sample is taken. For instance, local density of group members (the average distance between animals, Waters et al., 2010) can be measured with a single snapshot in time, and therefore randomized instantaneous sampling would be the best strategy to use to estimate local density across a long period of time. Conversely, when I is small and resembles continuous sampling, then the sampling strategy should be able to accurately estimate dependent metrics of animal behavior. Dependent metrics are features of behavior that are dependent on the time they are sampled. For instance, velocity requires at least two sequential snapshots of behavior. Intermediate values of I should allow for optimal trade-offs between these two biologically relevant features of animal behavior.

We study optimal sampling strategies using harvester ants (*Pogonomyrmex californicus*) as our model system. The question of choosing the correct intervals is particularly salient in social insects where potentially hundreds of individuals need to be tracked per colony (Hughes et al., 2008), and they make for a tractable system as all individuals can be contained in a single nest and their actions captured by a video recording (Holbrook et al., 2013). These videos can then be manually analyzed later and double checked for accuracy, and the environment can be tightly controlled in a laboratory setting. Additionally, as superorganisms, ants exhibit complex group-level dynamics (Holldobler & Wilson, 2009), and understanding these dynamics requires analyzing multiple aspects of their behavior. For instance, measuring division of labor requires measuring the proportion of time ants spend performing each task (Gorelick et al., 2004), but measuring the benefits of division of labor may also require counting the number of times ants switch tasks (Goldsby et al., 2012), so a sampling strategy in this context needs to consider both dependent and independent metrics of behavior.

To find an optimal piecewise sampling strategy, we first categorize the behavior of 9 ants for 3 hours on a second by second basis. We then use this data to estimate the minimal number of seconds necessary to reconstruct their behavior. Next, we simulate sampling for differing numbers of segments, calculating the errors associated with both dependent and independent behavioral metrics. Finally, we determined which behavioral attributes drive the optimal number of segments for each individual. These latter two steps were also validated with simulations of ant behavior.

**Methods**

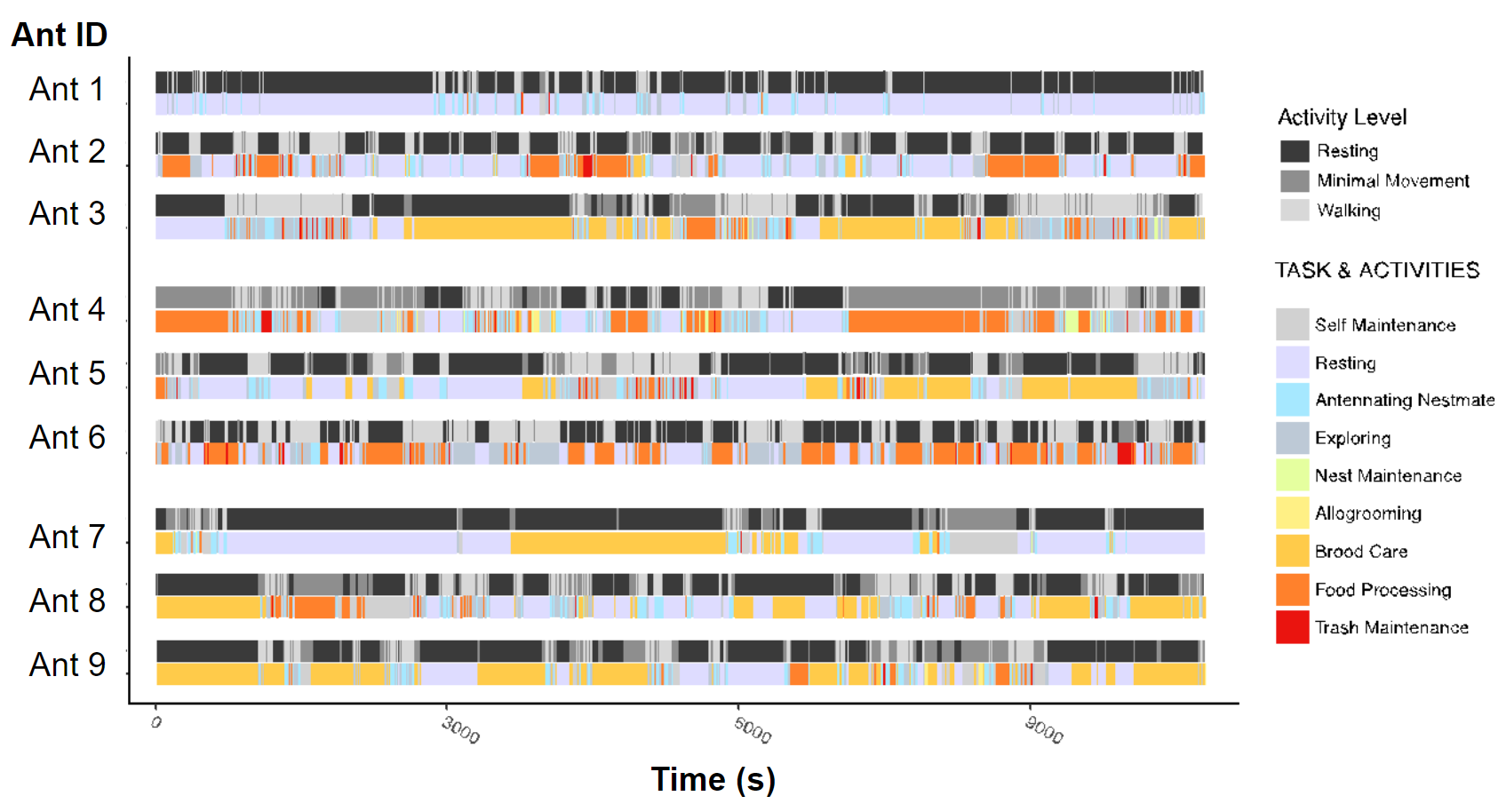
*Ant collection and colony maintenance & video analysis*

Newly mated P. californicus foundresses were collected on July 5, 2017, in Pine Valley, San Diego Co., California (lat 32°49′20″N, long 116°31′43″W, 1136 m elevation). Colonies were fed with *ad libitum* Kentucky bluegrass seeds, and provided frozen crickets or mealworms weekly. Colonies were also given water via a test tube plugged with cotton and were maintained at a consistent temperature of 30℃. Colonies were kept in a nest partitioned into a foraging arena and a brood chamber, where the total surface area was 242 cm2.

Nine ants were watched continuously for three hours (Fig. 2). Six were from a smaller colony (~30 workers, 2 queens) and three were from a larger one (~110 workers, 2 queens), though both were still considered incipient colonies as colonies in this species typically reach a size of 2,000 - 4,500 workers (Johnson, 2000). The workers that were followed were selected based on task, so as to capture a wide range of repertoires; three as they were performing brood care (interacting with brood), three as they were performing food processing (interacting with seeds or artificial diet), and three that were resting (immoble). Two from the small colony, and one from the larger, were selected for each task group. Switches between activities were manually coded using the program Cowlog (Version 3.0.2; Hänninen & Pastell, 2009), which allows logging of behaviors using keystrokes tied to video-based time-stamps. We categorized behaviors into 9 discrete tasks (Table 1).

| **Task** | **Description** |
| --- | --- |
| Brood Care | Carrying, manipulating, antennating, or resting on brood (larvae or live pupae) |
| Food Processing | Walking on food pile (intermittently antennating), carrying, manipulating, antennating, or resting on food (seeds, mealworms, or sacrificed pupa) |
| Trash Maintenance | Walking on the trash pile (intermittently antennating), carrying, manipulating, antennating, or resting on trash. |
| Antennating Nestmate | Antennating Nestmate (worker or queen) while standing, or while walking if antennations happen within 10 seconds of each other |
| Nest Maintenance | Digging, excavating/biting nest, dirt, or cotton |
| Resting | Does not include resting on trash, food, or brood |
| Self Maintenance | Self-grooming or being groomed |
| Exploring | Walking between tasks (not antennating nestmates for 10 sec), or trying to escape (climbing up the sides of the wall) |
| Allogrooming | Grooming nestmate (not being groomed by nestmate) |

**Table 1:** **Descriptions of each task used in video analysis.**



**Fig. 2:** **Visualization of the tasks each ant was performing every second.** Each row represents an ant’s task performance through time. Each ant was categorized into either what task they were performing at a given second (color) or what their activity level was (grayscale).

*Estimating minimum sample size for accurately reconstructing rare states: the 1% rule*

Here we outline two methods for estimating the sample size necessary for restoring a categorical time series. The first method finds the minimal number of trials necessary to draw the rarest state for an individual. The second method finds the minimal number of trials necessary to reconstruct the distribution of states for that same individual. As we will show later, these two methods give approximately the same sample size estimate per individual. These sample sizes will be used for comparing piecewise continuous sampling strategies.

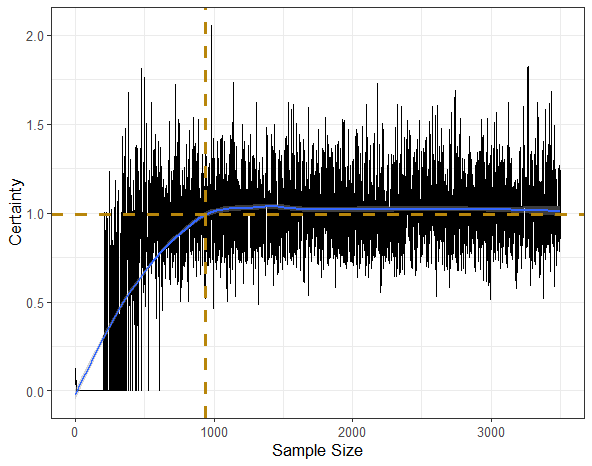
In order to reconstruct the probabilities that an individual performs all tasks, an experimenter must sample every task at least once in order to measure their frequencies of appearance. If the experimenter is sampling random instances of time, then the number of times s/he must sample is determined by the probability that they draw the rarest state, as chances are once they have drawn from this state, they have drawn from the other states as well. As there are only two outcomes, and we assume that each second drawn comes from a random second in the video, we consider each observation to be a Bernoulli trial. The probability of drawing from the rarest state, then, is given by a binomial random variable. Let pir be the smallest non-zero probability that an ant i performs the rarest task r and ni be the number of independent random draws for i. The expected value for a binomial distribution is E[X] = nipir. Setting E[X] = 1 and solving for ni yields equation 1:

![n_i = \frac{1}{p_{ir}}

](data:image/gif;base64,R0lGODlhPwApALMAAP///wAAAKqqqoiIiJiYmO7u7lRUVMzMzNzc3Lq6uhAQEERERCIiImZmZnZ2djIyMiH5BAEAAAAALAAAAAA/ACkAAAT+EMhJq7042/SO/mAIIsMReGKqpufqvlcLz69M36KN75nO/xQfECgc8opGXCCR/BUchsDDgWpar9isdhsSDAgFAsFQ5Y4SBwVBIliEGoa4fO4YCgAExmRgoAjWZhkNdRIGA1YBiYqLjI0SCkwSAQiBIggBEwIPAJFjGHBzoYRABJsSg3ipDAWVF18TBw4CrAAIpq0qDoC4Kau7vCANBJTAxcbHPF4DCGINyBUFswGAws8TTAkKE1Bsv8Z8EwbO1hMLd4/n5AABtAl6eH0AA7IEka0mtB2q0QAKB168BgjzQqzWJgQI3BRb4E2CLgkEDhWbdMEXAAP2cJn4NAyAQl4GHAKMShEBADs=) (1)

We call the sample size estimate from equation 1 the binomial estimate. This estimate can determine the sample size at which sample variance s2 converges on population variance σ2. When considering all possible states rather than just the rare state, the appropriate reference distribution for comparing a categorical time-series to a sample from that time-series is a 𝝌2 distribution. In this case, an experimenter can construct a contingency table comparing a subsample to the full dataset where the columns represent tasks, rows correspond to either the subsample or the full dataset, and each cell gives the number of observations a task was observed for a given ant. The variance of a 𝝌2 distribution is 2 times the degrees of freedom (σ2 = 2df), but at small sample sizes s2 > σ2. Here, we measure the degree to which the sample variance given by equation 1 converges onto population variance.

To measure sample variance at different sample sizes, we first iterate through all values of ni between 1 and 3,500. For each ni, we randomly sample with replacement from the time series that corresponds to ant i. We then calculate 𝝌2 for that sample (where the observed counts for each task are the counts from the sample and the expected counts is the probability that an ant performs each task times ni). We repeat this process 100 times to create an experimental 𝝌2 distribution which we can use to estimate s2. We also calculate σ2 for each ant (df = (row # - 1) × (column # - 1) = (2 - 1) × (#tasks ant i performs - 1), σ2 = 2df) so that we can calculate the ratio between population and sample variance (σ2/s2) which we call certainty (C). We calculate C for all ni = {1, 2, …, 3,500}. We use locally weighted smoothing (LOESS) to fit a smooth curve over the data (Fig. 3) and find the smallest value of ni where the fitted curve for C ≥ 0.99. This gives a strong estimate for sample size we can compare to the binomial estimate for sample size. We call this new estimate the certainty estimate. We repeat this process for each individual ant. As we have 9 ants, we end with 9 certainty estimates for sample size which we can compare to our binomial estimates. All simulations were run in RStudio and visualizations were made using the ggplot2 package (Wickham et al., 2016) and corrplot (Wei & Simko, 2017).



**Fig. 3. Estimation of sample size for a single ant.** The black line shows the ratio between population and sample variance (certainty) for ant i across sample sizes (ni). The blue line represents the LOESS fit on the data. The shaded region around the blue line gives the 95% confidence intervals for the LOESS fit. The horizontal dashed line shows where certainty = 0.99 and the vertical dashed line shows the smallest sample size where the LOESS fit gives a certainty that is greater than or equal to 0.99. The location where the horizontal line meets the LOESS fit gives the certainty estimate for sample size.

*Simulating piecewise continuous sampling, estimating error for each sample, and measuring behavioral attributes*

Piecewise continuous sampling is defined as the continuous observation of animal behavior across distinct intervals of time which can either be periodically or randomly spaced throughout an observational period. Here, we assume that biologically relevant rhythms are not known a-priori, or that our 3 hour sampling window was not large enough to capture those rhythms (a circadian rhythm can have a period of 24 hours in ants, Fujioka et al., 2017), so we sample randomly to resolve biases introduced by aliasing (Wisecup, 1998). Here, we describe the process by which we simulate piecewise continuous sampling on the raw dataset, how we measure error for each subsample, and how we measure behaviors of the ants to see how individual attributes affect optimal outcomes.

We simulate piecewise scan sampling by randomly selecting non-overlapping segments from the raw dataset. The segments for each ant are independent of one another. The number of intervals (I) is drawn from the geometric sequence {1, 2, 4, …, 128} so that a wide array of strategies could be tested with a limited number of simulations. The length of any given segment is the sample size per ant (n) divided by the number of segments (I/n). This segment length is rounded to ensure it results in a whole number. We run simulations across different values of n to estimate the effect of sample size.

Our proxy for dependent metrics is average bout length for each task per individual. Let the vector **l**i contain the average bout lengths for each task for ant i. Our proxy for independent metrics is the proportion of time each individual spends on each task. Let the vector **p**i contain the probabilities that ant i performs each task. Error for both metrics is defined as the sum of squared residuals between an ant’s ‘true’ behavior (as determined by the full dataset illustrated in Fig. 2) and a subsample’s estimate for that behavior. In other words, if is the estimate for ant i’s bout length for each task given a subsample and is the estimate for the proportion of time the ant spends on each task, then error for bout length is while the error for proportion is where k is the index of different tasks.

Our multi-objective optimization problem is to minimize both types of error with a given value of I within the limited sample size range. As the weights of both objectives are equal (that is, we do not favor minimizing one type of error over the other), but the ranges of both errors are different, so we transform both types of error so that they are on the same scale. We do this by dividing bout length error by the average bout length error, which in turn is divided by average proportion error. We find the I that minimizes the sums of these errors across ants and find them for each ant individually.

After calculating error, we next measure 3 behavioral attributes about each ant which could strongly influence each type of error. We measure the average (arithmetic mean) bout length for each ant across tasks, the number of unique tasks an ant performed, and we measured how specialized each ant is, which we define as the normalized sum of differences between an ant’s behavior and that of a theoretical perfect generalist (Appendix 1).

Increasing the number of segments could increase the probability that ants become misidentified across intervals, which in itself could be another source of error. To test whether our experimental design lends itself to this type of bias, we performed an experiment to see whether error increases with increasing I. The result of this experiment was insignificant (see Appendix 3), so we do not include a penalty term for our final error calculation.

*Validating optimization results with discrete events simulations*

To test the universality of various sampling techniques and to test hypotheses regarding the relationships between variables, we supplement our analysis of these 9 ants with additional simulations of their behavior. The bout lengths for each task are often not geometrically distributed (Appendix 4). Therefore, the assumptions of many models such as a discrete-time Markov chain are violated. As a result, we instead perform a discrete events simulation where we draw ‘events’ (or bouts) from the original dataset in such a way that we create a wide array of biologically-constrained behaviors. We draw from the raw bout list rather than from fitted distributions as bout lengths seem to be distributed from discrete power laws for 5 of the 9 states. Drawing from these power distributions directly can result in biologically infeasible behavior, such as an ant performing only a single task for the entire duration of the simulation.

To run these discrete event simulations, we first assign the probability that a simulated ant would perform task k. We draw this probability from a truncated geometric distribution (Chattopadhyay et al., 2014), as this is a discrete distribution whose probability of success parameter (here denoted as d) roughly corresponds to an ant’s degree of specialization. The number of tasks (K) sets the upper bound of the distribution while 1 is the lower bound. When d = 0, the truncated geometric distribution is converted into a uniform distribution where the probability of starting each task is 1/K. If bout lengths were equal, then this process would create a perfect generalist. Similarly, when d = 1, the probability of starting the first task is 1 and the probability of starting other tasks is 0, which is similar to a perfect specialist. Intermediate values of d can also produce intermediate degrees of specialization values.

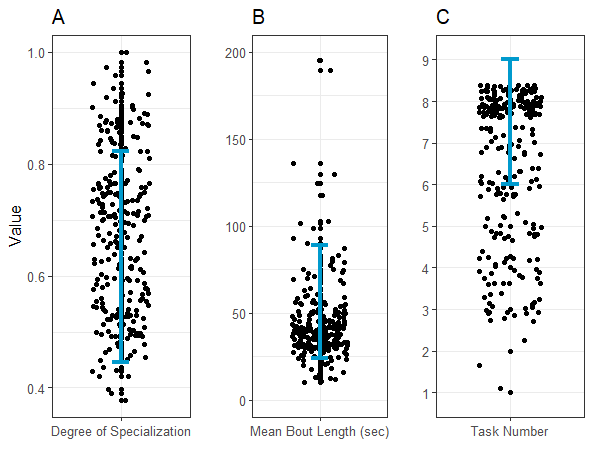
At the onset of each simulation, we shuffle the 9 states presented in Table 1 and are assigned new random indices x = {1, 2, …, K}. We then use equation 2 to set the probability that the ant will start k given its new index x:

\textnormal{Pr}(k) = \frac{(1-d)^{x-1}d}{1-(1-d)^{K-1}} (2)

We initialize ants by having them perform the task whose x = 1. Let **b**k be the vector that contains all the bout lengths that correspond to task k from the raw dataset. Once an ant is in state k, we randomly sample with replacement from **b**k. The ant remains in that task for as long as the bout drawn from **b**k. At the end of the bout, the ant must transition to a different task. This probability depends on those drawn from equation 2, however they are increased to compensate for the fact that the ant cannot transition back to the state it was in the previous timestep. Once the ant is in a new state, we draw the **b**k that corresponds to that state. This process repeats until the total length of the simulation = T, where T is the number of seconds each real ant was observed for (T = 11,041 seconds).

K = 9 for all simulations, but d ~ U(0, 1) every simulation. We run 250 simulations, and for each simulation we measure both types of error across I and sample sizes, the optimal value of I, mean bout length, number of unique tasks the virtual ant performed, and the ant’s degree of specialization.

This process reproduced behavior akin to that of the raw ants. The ranges of behavioral attributes for virtual ants overlapped that of the real ants (Fig. 4). The one exception is that none of the simulated ants performed all 9 tasks whereas 1 real ant performed all 9 tasks. In our statistical analyses, we subset the simulated data that falls within the biologically-determined range. Of the initial 250 simulations, 151 fall within all these ranges simultaneously. We used RStudio to run the simulations and the poweRlaw package to analyze power laws (Gillespie, 2015).

**

**Fig. 4. Output of model compared to real ranges for each behavioral attribute.** Each panel shows the values of different behavioral attributes for each simulated ant (black dots). The behavioral attribute is given on the x-axis of each panel. The blue bars represent the ranges for each of these values from real ants. Jitter added for visualization.

*Statistical analyses*

We have 3 sets of questions that we will address here. 1) How well do the binomial and certainty sample size estimates agree with one another? 2) How well do behavioral attributes predict optimal I? 3) How correlated are different behavioral attributes correlated with each other and different types of error?

1. We perform a simple linear regression between this certainty estimate for sample size and the binomial estimate for each real ant. We also calculate the 95% confidence interval for the slope to see if it overlaps 1, an indication of strong agreement between the two methods.
2. To explore the effects of individual attributes on the optimal sampling strategy, we performed a multiple linear regression between the optimal number of intervals (I) and the degree of specialization, mean bout length, and the number of tasks performed for both the real and simulated ants. Here the optimization occurred across the range of sample sizes, so we first find the average value of each metric across those sample sizes. After normalizing and summing error, we find the optimal I and then extract the values of each metric which correspond to that I. This process reduces the sample size to 151. To fit the best linear model, we compared all hierarchical interaction structures between the independent variables (degree of specialization, mean bout length, number of tasks performed) and the response variable (log-transformed optimal I) using AIC and BIC. For simulated ants, the model with no interactions among variables had both the lowest AIC and BIC, so we evaluated this model. While the independent variables did correlate slightly (variance inflation factors for degree of specialization, mean bout length, and number of tasks performed in this model are 1.312, 1.016, and 1.303 respectively), the degree to which they correlate is well below traditionally accepted thresholds for multicollinearity (O’brien, 2007). For real ants, the model with all possible interactions had the lowest AIC/BIC, but as none of the effects were significant (p-value from F-test > 0.05), we reduced the model to one with only main effects. Again the independent variables correlated slightly (variance inflation factors for degree of specialization, mean bout length, and number of tasks performed in this model are 2.675, 1.459, and 2.486 respectively), but are still below accepted thresholds of 4-10. All statistical tests were performed in R (R Core Team, 2021) using the stats, ggplot2, lmtest, fitdistrplus, corrplot, and mgcv packages.
3. Relationships between variables describing behavioral attributes (degree of specialization, mean bout length, number of tasks performed), variables describing sampling techniques (number of intervals I, sample size n), and variables describing the results of sampling (error for proportion of states and bout length) were explored using pairwise Spearman’s correlations, as many relationships were non-linear. A Bonferoni correction was made for multiple comparisons (α = 0.05/21 ≅ 0.0024). Here, the sample size is 151 simulations × 8 intervals × 250 sample sizes = 302,000.

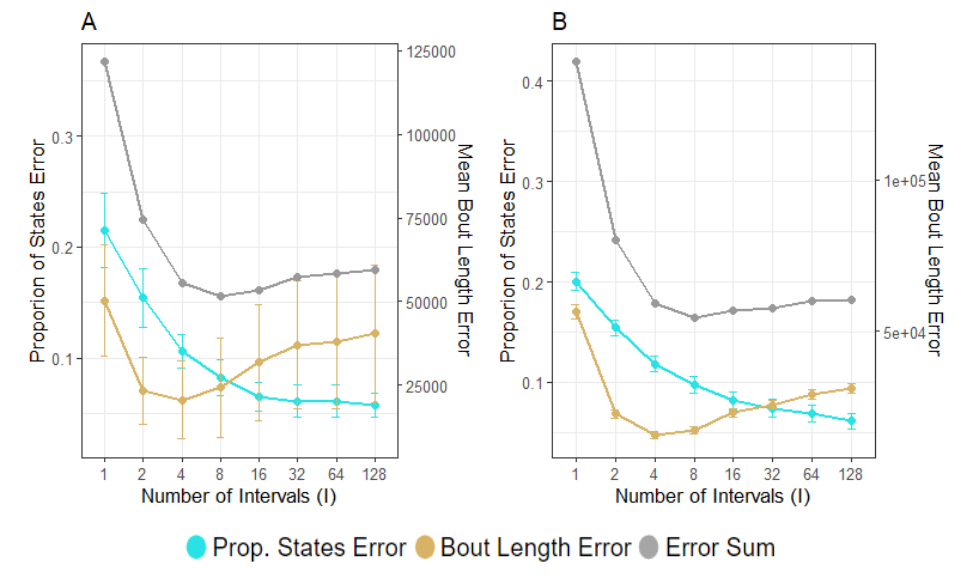
**Results**

*Agreement between binomial and certainty estimates for sample size*

If the certainty estimate for sample sizes matches the binomial estimate perfectly, then the slope of the regression line between them would be 1. While the slope is not exactly 1 (linear regression P < 0.001, slope = 1.12, adj. R2 = 0.973, df = 7), it is bound within the 95% confidence interval (CI [0.925 1.315]), indicating that it is not significantly different from 1. The range of binomial sample sizes across ants is between 56 and 2,208. The average is 660, and across simulations we iterate over n ∊ {600, 601, …, 700}.

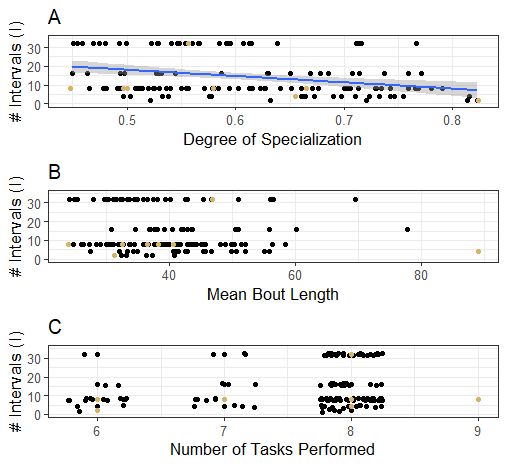
*Optimal I for both real and simulated ants*

The error curves for both real and simulated ants look very similar, each possessing a minimum at I = 8. This occurred because the rescaled error for bout length is saddle-shaped with its minimum at I = 4 while the rescaled error for the proportion of states continuously declines with increasing I (Fig. 5). The decline of the proportion of states error decreases with I though, and so it shifts the minima of the sum of rescaled errors from 4 to 8. Additionally, the mode for optimal I was 8 across all simulations, ants, and sample sizes.

****

**Fig. 5. Independent and dependent errors across interval numbers for simulations and raw dataset.** The left panel (A) shows ranges for the two types of error (probability of states error is teal, bout length error is gold) across sample sizes. Points show the mean error, and error bars are standard error. Note that these two errors are on different scales. The right panel shows error for simulated ants. The gray line shows the sum of these errors. Both panels display a minimum in the sum of errors at I = 8.

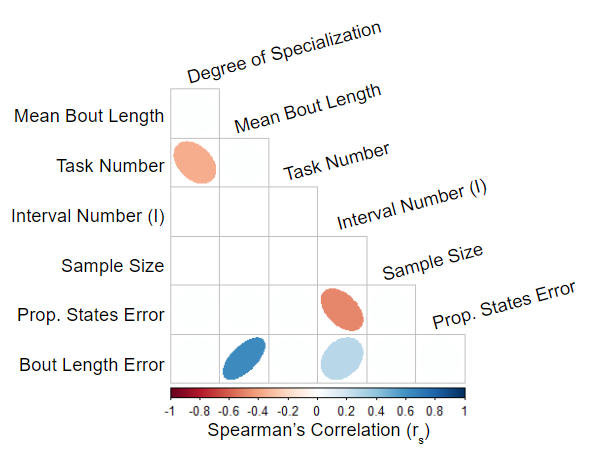
Despite the presence of this minima, the optimal I still varied across real and simulated ants. This variation only correlated slightly with measured behavioral attributes. Degree of specialization had a significant effect on optimal I in simulated ants (P for multiple linear regression model = 0.0682, P for task number > 0.05, task duration > 0.05, degree of specialization < 0.05). However, this is not the case for real ants (P for model and all effects > 0.05) and slopes for the degree of specialization effect for simulated ants is small (Fig. 6). Finally, the regression accounts for very little variation in the data (adj. R2 = 0.028), so it is questionable how much of an effect this behavioral attribute had on the optimal sampling strategy.

****

**Fig. 6. Effect of each behavioral attribute on the optimal number of intervals.** Each panel shows the effect that different behavioral attributes had on the optimal number of intervals (I). Block dots show simulated ants, and the regression lines (blue) are fit on these points in cases where the relationship between an attribute and optimal number of intervals was significantly correlated. Gold points signify real ants. Jitter was added to panel C for visualization.

*Effect of behavioral attributes on error and optimal sampling strategies*

While all pairwise comparisons among variables from simulated ants are significant (Spearman’s correlation P < 0.05/21, N = 151) except those between behavioral attributes (degree of specialization, number of tasks performed, mean bout length) and strategic variables (sample size, interval number), R2 for most comparisons are < 0.05 (Fig. 7). There is a negative correlation between the degree of specialization and bout number (rs = -0.369), a positive correlation between average bout length and bout length error (rs = 0.637), a negative correlation between interval number and the proportion of states error (rs = -0.48), and a positive correlation between interval number and bout length error (rs = 0.278).

**

**Fig. 7. Correlations between behavioral attributes, parameters governing sampling techniques, and error.** Each grid square gives the Spearman’s correlation between two variables as measured from the simulated ants. Empty cells signify relationships that are either insignificant (P > 0.05/21) or are close to 0 (R2 < 0.05). Color indicates Spearman’s correlation, and the orientation of the ellipse indicates a positive or negative correlation.

**Discussion**

The objective of this study was to assess the accuracy of different piecewise continuous sampling protocols and uncover what aspects of behavior could influence optimal sampling strategies. To this end, we recorded behavioral traces of 9 harvester ants and then simulated piecewise continuous sampling on this dataset, measuring independent and dependent error across the number of segments (I). We found that intermediate numbers of segments (4-16) tended to be the most accurate whereas sampling strategies that mimicked continuous sampling (I = 1) and instantaneous sampling (I >> 1) tended to do more poorly. We estimated sample sizes based on the rarity of states, and found that optimal values of I were not strongly predicted by various individuals' behavioral attributes. These results were validated with simulations which created a wide variety of differing-behaving individuals, indicating that this result is robust to differences among the individuals under observation.

First, we show that establishing the sample size necessary to reconstruct the rarest state of an animal is sufficient for reconstructing the distribution of states. Our binomial estimate for sample size results in a sample variance where 0.99σ2 ≤ s2 ≤ 1.01σ2, a reasonable property of any sample size estimate. If the probability of drawing the rarest state is known *a priori*, then equation 1 constitutes a simple back-of-the-envelope calculation for estimating sample size which we call the 1% rule. If this probability is not known, then it can be thought of as setting the rarest acceptable state for an analysis. This is similar in concept to how one sets power before performing a power analysis (Cohen, 1992).

Next, we found that piecewise continuous sampling can give a wide range of results depending on the interval number (I) selected. Small values of I (1, 2) tend to result in high independent and dependent errors whereas large values of I ( > 16) tend to minimize independent error at the cost of increasing dependent error. Intermediate values of I (4, 8, 16) tend to minimize dependent error while sufficiently reducing independent error such that the sum of the errors has a minimum at I = 8. However, the differences in error sum between I = 4, 8, 16 are small, so while 8 was the most common optimal value of I, many simulations also had minima at I = 4 and 16 (Fig. 5). Indeed, roughly 78% of all simulations had optima at either 4, 8, or 16. We therefore recommend a range of I values rather than a single value, although one should know that choosing I on the lower end of the range will minimize dependent error more while choosing values on the larger end of the spectrum tend to minimize independent error more.

This result was consistent across trials. The I = 8 minimum was found both for the raw dataset, the simulations, as well as a verification dataset where individuals were classified based on activity level rather than the task they were performing (Appendix 2). It was also found to be largely independent of individual behavioral attributes, suggesting that the range of minima between 4-16 could hold for a vast array of organisms. It should be noted, though, that this independence could be a feature of this particular dataset. While the number of tasks an ant performed was negatively correlated with its degree of specialization, it was not correlated with the average bout length of tasks, nor was the degree of specialization correlated with bout length. It is possible that in other datasets specialists spend more time on tasks, increasing their bout lengths and performing fewer tasks, and so the independent and dependent metric errors would track each more closely rather than being uncorrelated with each other. Therefore, minima may shift more dramatically between individuals than they are here.

There are other features of individual behavior that could also influence the optimal sampling technique. Given that the number of instances an individual can be observed for is n, then the length of each observation interval will be n/I. This sets a maximum observable limit on bout length, which could artificially influence the interpretation of the data. Additionally, nearly all behaviors fluctuate periodically as a function of time, either as a result of the day-night cycle (Emmer et al., 2018), seasonal changes (Leveau et al., 2018), or their own circadian rhythms (Bünning, 1973). If the period of biologically-relevant rhythms are known *a priori*, one can sample at the Nyquist sampling rate in order to avoid aliasing effects (Vaidyanathan, 2001; Brunton & Kutz, 2019). If one were to use piecewise continuous sampling with a fixed interval between sampled segments, then one could sample twice for every period of the highest frequency biological rhythm. However, if these rhythms are not known, or the rhythms account for only a small amount of the variation in the dataset, then one should space the piecewise segments randomly, which is what we do here. This has the additional benefit of removing the sensitivity of the collection protocol on starting point on accuracy that regular sampling strategies have (Hämäläinen et al., 2016).

While optimal segment numbers and lengths may differ across studies, the general finding that intermediate numbers of segments may be conserved across organisms for studies which look to analyze several aspects of behavior simultaneously. I = 1, 128 are the only solutions that are not pareto-efficient (Appendix 5), meaning that there are other strategies which can further minimize one or both types of errors simultaneously. All intermediate strategies are pareto-optimal, and can be used to trade-off different types of error if a study values one over the other more.

That said, these results are likely strongly dependent on the fact that the bout lengths for many tasks are drawn from power laws (much like locomotor velocities in the same species, Waters et al., 2010) rather than geometric or negative binomial distributions (Appendix 4). When bout lengths are distributed geometrically, then the error function for bout length as I increases is a concave saturating function with no local minima (Appendix 6). Thus, minimal error should occur when I = 1. However, when bout lengths are drawn from a power law, minimal error occurs around I = 4, as segments with smaller values of I tend to overlap abnormally long bouts which overestimate average bout length. When I is high, though, then average bout length is underestimated, and thus an intermediate I value minimizes overall error. Therefore, bout length distributions should be evaluated thoroughly before sampling strategies should be chosen.

As big data approaches become more common in animal behavior (Christin et al., 2019; Neethirajan, 2020) as well as approaches for testing multiple hypotheses (Betini et al., 2017), the need to collect more data is more prevalent than ever. Remote sensing and machine learning offer opportunities for data to be collected and categorized faster and more cheaply, but even these novel methods still require training datasets that are often manually collected (Wang, 2019). The need for classical sampling techniques may never disappear, though they may need to be adjusted so that they can account for multiple data types simultaneously. Piecewise continuous sampling is a promising adaptation of previous sampling techniques, one that can be invaluable to studies which look at multiple competing aspects of behavior.

**Acknowledgements**

We thank members of the Fewell and Pavlic labs for helpful discussion and their comments on the manuscript. We also want to thank Colin S. Gillespie for his assistance in using the poweRlaw R package.

**References**

Ary, D., & Suen, H. K.. **The use of momentary time sampling to assess both frequency and duration of**

**behavior**. Journal of Behavioral Assessment, 5(2) (1983), pp. 143-150 https://doi.org/10.1007/BF01321446

[Altmann, J. **Observational Study of Behavior: Sampling Methods**. Behaviour, (49) (1974), pp. 227–266](http://paperpile.com/b/zQgSv6/pZcp).

<https://doi.org/10.1163/156853974x00534>

Betini, G. S., Avgar, T., & Fryxell, J. M. **Why are we not evaluating multiple competing hypotheses in ecology**

**and evolution?** Royal Society Open Science, 4(1) (2017), pp. 160756. https://doi.org/10.1098/rsos.160756

[Birgiolas, J., Jernigan, C. M., Smith, B. H., & Crook, S. M](http://paperpile.com/b/zQgSv6/DmEc).[**SwarmSight: Measuring the temporal**](http://paperpile.com/b/zQgSv6/DmEc)

[**progression of animal group activity levels from natural-scene and**](http://paperpile.com/b/zQgSv6/DmEc)[**laboratory videos**. *Behavior*](http://paperpile.com/b/zQgSv6/DmEc)

[*Research Methods*, 49(2)](http://paperpile.com/b/zQgSv6/DmEc) (2017)[, pp. 576–587](http://paperpile.com/b/zQgSv6/DmEc). <https://doi.org/10.3758/s13428-016-0732-2>

Brunton, S. L., & Kutz, N. K. **Methods for data-driven multiscale model discovery for materials.**

Journal of Physics: Materials 2(4) (2019) pp. 044002. <https://doi.org/10.1088/2515-7639/ab291e>

Bünning, E, & Moser, I. (1973). **Light-induced phase shifts of circadian leaf movements of phaseolus:**

**Comparison with the effects of potassium and of ethyl alcohol.** Proceedings of the National Academy of

Sciences, 70(12) (1973), pp. 3387-3389. <https://doi.org/10.1073/pnas.70.12.3387>

Casarrubea, M., Aiello, S., Di Giovanni, G., Santangelo, A., Palacino, M., & Crescimanno, G. **Combining**

**Quantitative and Qualitative Data in the Study of Feeding Behavior in Male Wistar Rats.** Frontiers in

Psychology, (10) (2019). pp.881. doi:10.3389/fpsyg.2019.00881

Christin, S., Hervet, É., & Lecomte, N. **Applications for deep learning in ecology.** Methods in Ecology and

Evolution, 10(10) (2019), pp. 1632-1644. https://doi.org/10.1111/2041-210X.13256

Cohen, J. **Statistical power analysis**. Current directions in psychological science. 1(3) (1992), pp. 98-101.

https://doi.org/10.1111/1467-8721.ep10768783

Crall, J. D., Gravish, N., Mountcastle, A. M., Kocher, S. D., Oppenheimer, R. L., Pierce, N. E., & Combes, S. A.

**Spatial fidelity of workers predicts collective response to disturbance in a social insect.** Nature

communications, 9(1) (2018), pp. 1-13. DOI: 10.1038/s41467-018-03561

Delignette-Muller, M. L., & Dutang, C. **fitdistrplus: An R package for fitting distributions. Journal of**

**statistical software**. 64(4) (2015), pp. 1-34. 10.18637/jss.v064.i04

Emmer, K. M., Russart, K. L. G., Walker, W. H., Nelson, R. J., & DeVries, A. C. **Effects of light at night on**

**laboratory animals and research outcomes.** Behavioral Neuroscience, 132(4)(2018)*,* pp. 302–314.

doi:10.1037/bne0000252

Fujioka, H., Abe, M. S., Fuchikawa, T., Tsuji, K., Shimada, M., & Okada, Y. **Ant circadian activity**

**associated with brood care type.** Biology Letters, 13(2)(2017), pp. 20160743.

doi:10.1098/rsbl.2016.0743

Gillespie, C. S. **Fitting Heavy Tailed Distributions: ThepoweRlawPackage.** Journal of Statistical Software,

64(2)(2015)*.* doi:10.18637/jss.v064.i02

Goldsby, H. J., Dornhaus, A., Kerr, B., & Ofria, C. **Task-switching costs promote the evolution of division of**

**labor and shifts in individuality.** Proceedings of the National Academy of Sciences, 109(34)(2012), pp. 13686–13691. doi:10.1073/pnas.1202233109

Gorelick, R., Bertram, S. M., Killeen, P. R., & Fewell, J. H. **Normalized Mutual Entropy in Biology: Quantifying Division of Labor.** The American Naturalist, 164(5)(2004), pp. 677. doi:10.2307/3473178

Ha, J.-H., Kim, Y.-H., Im, H.-H., Kim, N.-Y., Sim, S., & Yoon, Y. **Error Correction of Meteorological Data**

**Obtained with Mini-AWSs Based on Machine Learning.** Advances in Meteorology, (2018) pp. 1–8.

doi:10.1155/2018/7210137

Hämäläinen, W., Ruuska, S., Kokkonen, T., Orkola, S., & Mononen, J. **Measuring behaviour accurately with**

**instantaneous sampling: A new tool for selecting appropriate sampling intervals.** Applied Animal

Behaviour Science, (180) (2016), pp. 166-173. https://doi.org/10.1016/j.applanim.2016.04.006

Hänninen, L., & Pastell, M. **CowLog: Open-source software for coding behaviors from digital video.**

Behavior Research Methods, 41(2) (2009), pp. 472–476. doi:10.3758/brm.41.2.472

Holbrook, C. T., Eriksson, T. H., Overson, R. P., Gadau, J., & Fewell, J. H. **Colony-size effects on task**

**organization in the harvester ant *Pogonomyrmex californicus*.** Insectes sociaux, 60(2) (2013), pp.

191-201. https://doi.org/10.1007/s00040-013-0282-x

Hughes, D. P., Pierce, N. E., & Boomsma, J. J. **Social insect symbionts: evolution in homeostatic fortresses.**

Trends in Ecology & Evolution, 23(12) (2008), pp. 672–677. doi:10.1016/j.tree.2008.07.011

Johnson, R. A. **Seed-harvester ants (Hymenoptera: Formicidae) of North America: an overview of ecology and**

**biogeography. Sociobiology.** 36 (1) (2000), pp. 89-122. doi: 10.11646/zootaxa.3646.3.1

Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K. S., & Igel, C. **Big Universe, Big Data: Machine**

**Learning and Image Analysis for Astronomy.**IEEE Intelligent Systems, 32(2)(2017), pp. 16–22.

doi:10.1109/mis.2017.40

Lehner, P. N. **Sampling Methods in Behavior Research.**Poultry Science, 71(4)(1992), pp. 643–649.

doi:10.3382/ps.0710643

Leveau, L. M., Isla, F. I., & Bellocq, M. I. **Predicting the seasonal dynamics of bird communities along an**

**urban-rural gradient using NDVI.** Landscape and Urban Planning, (177) (2018), pp. 103–113.

doi:10.1016/j.landurbplan.2018.04.007

Mann, J. **Behavioral Sampling methods for cetaceans: a review and critique**. Marine Mammal Science,

15(1) (1999), pp. 102–122. doi:10.1111/j.1748-7692.1999.tb00784.x

Martin, P., Bateson, P. **Measuring Behaviour: An Introductory Guide.** *Cambridge University Press*. (1993)

doi:10.1017/cbo9780511810893.003

Montgomery, D. C. **Introduction to statistical quality control.** *John Wiley & Sons*. (2020)

doi:10.1201/9781420038675.ch0

Mitlöhner, F. M., Morrow-Tesch, J. L., Wilson, S. C., Dailey, J. W., & McGlone, J. J. **Behavioral sampling**

**techniques for feedlot cattle**. Journal of Animal Science, 79(5)(2001).*,* 1189. doi:10.2527/2001.7951189x

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. **Response surface methodology: process and product**

**optimization using designed experiments.** John Wiley & Sons (2016).

Neethirajan, S. (2020). The role of sensors, big data and machine learning in modern animal farming. Sensing and

Bio-Sensing Research, 29, 100367.

O’brien, R. M. **A Caution Regarding Rules of Thumb for Variance Inflation Factors.** Quality & Quantity,

41(5) (2007), pp. 673–690. doi:10.1007/s11135-006-9018-6

R Core Team. **R: A language and environment for statistical computing. R Foundation for Statistical**

**Computing,** (2022) Vienna, Austria. URL [http://www.R-project.org/](https://www.r-project.org/)

Rice, L, Tate, S, Farynyk, D, Sun, J, Chism, G, Charbonneau, D, Fasciano, T, Dornhaus, A, Shin, M.

**ABCTracker: an easy-to-use, cloud-based application for tracking multiple objects.** (2020) arXiv preprint arXiv:2001.10072.

Rose, L. **Behavioral sampling in the field: continuous focal versus focal interval sampling.** Behaviour,

137(2) (2000), pp. 153–180. doi:10.1163/156853900502006

Tacha, T. C., Vohs, P. A., & Iverson, G. C. **A comparison of interval and continuous sampling methods for**

**behavioral observations.** Journal of Field Ornithology, 56(3)(1985), pp. 258-264.

Vaidyanathan, P. P. **Generalizations of the sampling theorem: Seven decades after Nyquist.**IEEE

Transactions on Circuits and Systems I: Fundamental Theory and Applications, 48(9)(2001), pp.

1094–1109. doi:10.1109/81.948437

Valentini, G., Mizumoto, N., Pratt, S. C., Pavlic, T. P., & Walker, S. I. **Revealing the structure of information**

**flows discriminates similar animal social behaviors.** Elife (9) (2020). doi: 10.7554/eLife.55395

Wang, G. (2019). Machine learning for inferring animal behavior from location and movement data. Ecological

informatics, 49, 69-76.

Waters, J. S., Holbrook, C. T., Fewell, J. H., & Harrison, J. F. **Allometric Scaling of Metabolism, Growth,**

**and Activity in Whole Colonies of the Seed‐Harvester Ant Pogonomyrmex californicus.** The

American Naturalist, 176(4) (2010), pp. 501–510. doi:10.1086/656266

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., & Zemla, J. **Package ‘corrplot’.** Statistician, 56(316) (2017).

Wichert, S., Fokianos, K., & Strimmer, K. **Identifying periodically expressed transcripts in microarray time**

**series data.** Bioinformatics, 20(1)(2003), pp. 5–20. doi:10.1093/bioinformatics/btg364

Wisecup, R. D. **Unambiguous signal recovery above the Nyquist using random‐sample‐interval imaging.**

Geophysics, 63(2) (1998), pp. 763–771. doi:10.1190/1.1444376

Yang, Z.-M., Huang, H., Cai, Z., & Qin, Y. **A theoretical framework for runtime analysis of ant colony**

**optimization.** *2010* I*nternational Conference on Machine Learning and Cybernetics*. (2010)

doi:10.1109/icmlc.2010.5580959

**Appendix 1: Normalized degree of specialization**

Division of labor statistics are used to calculate the degree to which individual identity predicts the task they perform (Holbrook et al., 2011). To arrive at this estimate, one must compute the Shannon diversity index for each individual ant and then average across individuals (Gorelick et al., 2004). This procedure can be used to estimate the degree to which individuals are specialized on tasks if the averaging step is skipped, however the computations are difficult to comprehend for scientists not trained in information theory Indeed the calculations are complicated enough that the original definitions for how individuals predict tasks and task predict individuals in Gorelick et al., 2004 were reversed (see Dornhaus et al., 2009). Here we derive a simpler equation that should be more tenable to biologists and provide sample R code for how the degree of specialization (D) can be calculated for an individual ant i.

Let the vector **p**i contain the probabilities that ant i performs each task k, where k ∊ {1, 2, …, K}. For example, if the number of observed tasks K = 3, then **p**i could equal [0.4 0.5 0.1] so that ∑ **p**i = 1. D can be understood as the difference between **p**i  and a perfect generalist who is equally likely to perform each task. Therefore, **p**G = [1/K 1/K … 1/K], a vector which contains K elements. Conversely, a perfect specialist will only perform one task, so **p**S = [1 0 … 0].

For interpretability, we calculate the degree of specialization as opposed to the degree of generalization. To do this, we define D as the sum of the absolute differences between **p**i and **p**G , or as equation 3:

D = \sum_{k=1}^{K} |p_{ik} - p_{Gk} |

 (3)

Alternatively, as **p**G contains the same value for all elements, this can be rewritten as:

D = \sum_{k=1}^{K} |p_{ik} - 1/K|

 (4)

This value, however, will produce a different range of values depending on the value of K. To ensure that the final value of D ranges between 0 and 1, we can normalize equation 3 with the formula D’ = (D-Dmin)/(Dmax-Dmin). First, we can solve for Dmax and Dmin. Let the vector **p**i contain the probabilities that ant i performs each task k, where k ∊ {1, 2, …, K}. **p**i for a perfect generalist = **p**G = [1/K 1/K … 1/K] while **p**i for a perfect specialist = **p**S = [1 0 … 0]. Dmin is trivial to calculate, as it is simply the difference between a perfect generalist and itself, so Dmin = ∑ |**p**G - **p**G| = 0. As the sum of each vector must equal 1, a perfect specialist is maximally different from a perfect generalist, so Dmax = ∑ |**p**S - **p**G|. To find a closed form solution to this equation, we first expand this summation:

\sum_{k=1}^K|\bold{p}_{Sk}-\bold{p}_{Gk}| = |1-\frac{1}{K}| + |0-\frac{1}{K}|+...+|0-\frac{1}{K}|

The first term in the expansion simplifies to (K-1)/K while all other terms simplify to 1/K. As there are K-1 1/K fractions, this equation simplifies to:

\frac{K-1}{K} + (K-1)\frac{1}{K} = \frac{2(K-1)}{K} = \frac{2K}{K} - \frac{2}{K} = 2 - \frac{2}{K}

Therefore, Dmax = 2 - 2/K. Plugging Dmin and Dmax into the normalization formula where x = D yields equation 5.

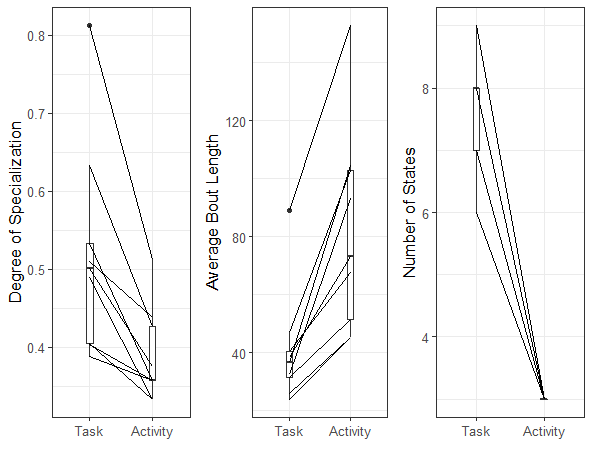
D' = \frac{\sum_{k=1}^{K} |p_{ik} - 1/K|}{2-2/K}

 (5)

**Appendix 2: Verification of sampling technique on activity data**

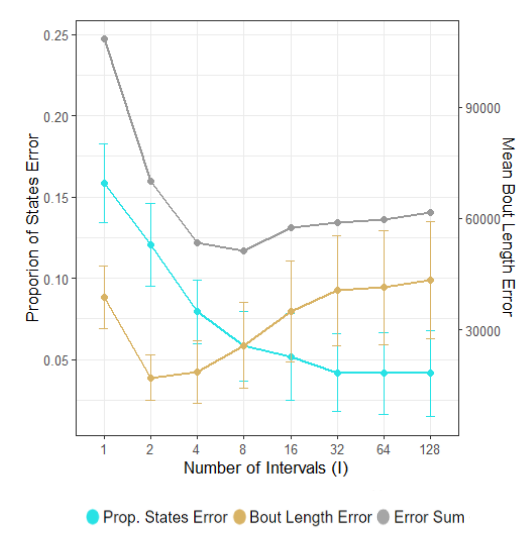
In addition to recording the task each of the 9 ants were performing each second, we recorded their activity level. The activity of each ant was evaluated in much the same way as their tasks were, except they could only be in three states, Resting, Minimal Movement, and Walking. These evaluations were made independently of the task they were performing, and thus represents a good test case for the optimal sampling strategy formulated in this paper.

Information derived from these datasets also significantly differ from each other. From each task or activity-based dataset, we calculate each ant’s degree of specialization, mean bout length, and the number of states they were in throughout the duration of the video (Fig. S1). We then performed a paired t-test comparing each of these attributes across datasets (Shapiro-wilks P-values > 0.05), bonferroni corrected P for all t-tests < 0.01.



**Fig. S1. Measurements for each behavioral attribute for each ant across task-based and activity-based datasets.** Each panel compares a different behavioral attribute across the task-based dataset and the activity-based dataset. As all ants are present in both datasets, connected dots represent the same ant in either dataset. This demonstrates that division of labor metrics depend in part on how behavior is categorized. Here, allowing for fewer possible behaviors creates the appearance of more generalist ants.

We calculate bout length and proportion of states error for the activity-based dataset in the same way we calculated error for the task-based dataset. Summing these errors yields a minimum at I = 8 (Fig S2), which is identical to the task-based result. It should be noted though that the minimum for bout-length error shifted from I = 4 to I = 2 for this dataset, probably to compensate for increasing bout lengths in the activity dataset (Fig S1). Additionally, the mode I for this dataset is also 8.

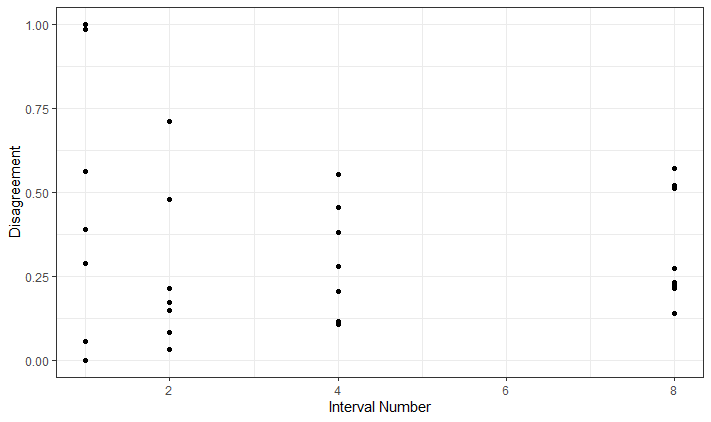


**Fig. S2. Independent and dependent error across intervals for the activity-based dataset.** Lines represent probability of states error is (teal) and bout length error (gold) across sample sizes. Points show the mean error, and error bars are standard error. The gray line shows the sum of these errors plus.

**Appendix 3: Penalty for increasing interval number**

To estimate the error introduced by increasing interval number (I), we performed an experiment where two researchers independently analyzed the same videos of 4 ants across 2 ant with 4 different interval levels (n = 4 ants × 2 colonies × 4 intervals = 32). We measured the degree to which the two researchers disagreed on what behaviors they were observing. A video of each colony was divided into 1, 2, 4, and 8 intervals, each of which were randomly dispersed throughout a 3 hour video of colony behavior. The length of these intervals combined was 700 seconds. 650 was the average estimated sample size for the 9 ants from this study, so we rounded up to the nearest hundredth. Researchers categorized what tasks the ants were performing every second. We used the task definitions of Table 1. The researchers are considered blocks, and the order in which each researcher analyzes a colony and with I intervals was randomized.

Disagreement is defined as the number of seconds the two researchers disagreed on a task divided by the total number of seconds (700). We performed a linear regression between the log of I and disagreement and found no relationship between the two variables (F1, 30 = 1.697, P = 0.2025, adj. R2 = 0.022, Fig. S3). While log-transforming I allows the residuals to be normally distributed (Shapiro-Wilk test P = 0.0989), the regression is still heteroscedastic (Breuch-Pagan test P < 0.01), indicating that disagreement is more variable with fewer intervals. This likely occurs because experimenters might not analyze the same ant. As the regression was insignificant, we do not include a penalty term for I in our final error measurement. However, increasing consistency with interval number could be an additional benefit of piecewise continuous and scan sampling.

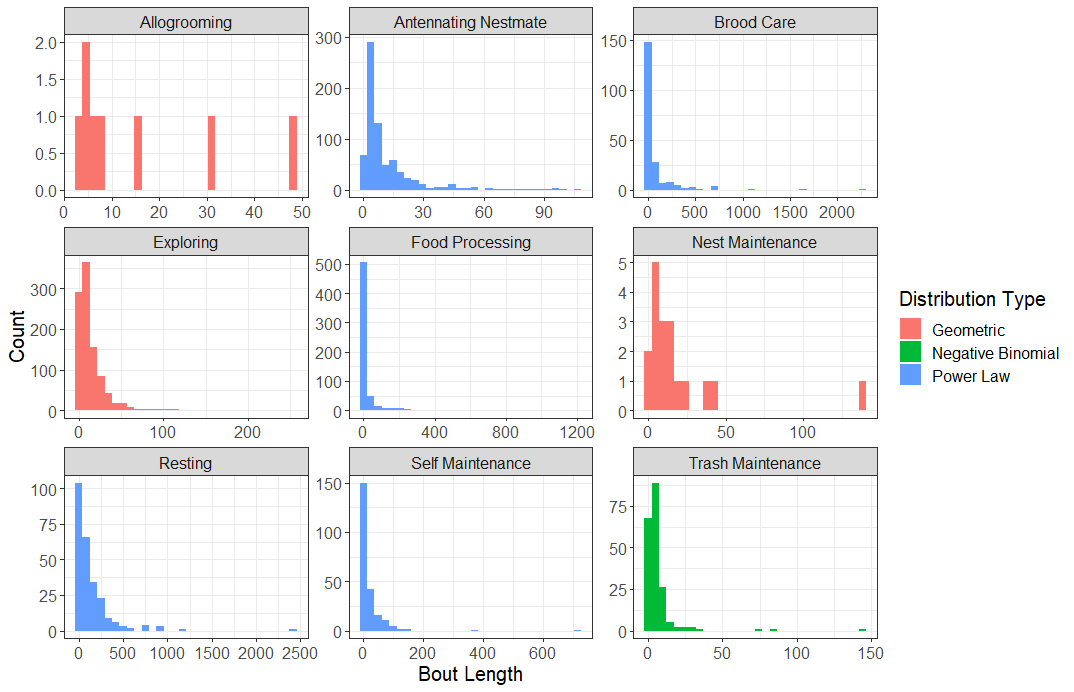


**Fig. S3. Error does not increase with interval number.** Each point represents the degree to which two researchers disagreed on the behavior of an individual ant across segments.

**Appendix 4: Fitting distributions to bout lengths per task**

We want to know whether bout lengths are geometrically distributed to see whether assumptions of Markovian models are met. To characterize the distributions of bout lengths for each task, we use maximum likelihood estimation from the fitdistrplus package (Delignette-Muller & Dutang, 2015) to fit the distributions of each task to 4 types of discrete distributions: negative binomial, geometric, hypergeometric, and poisson. We then performed a chi-squared goodness of fit test to which theoretical distributions were indistinguishable from the experimental distribution. If the p-value from multiple distributions was greater than 0.05, then we would choose the distribution with the lowest AIC and BIC scores. In cases where all 4 distributions were significantly different from the experimental distribution, we fit a power-law distribution with discrete maximum likelihood estimators (Clauset et al. 2009). We then performed a goodness of fit test via a bootstrapping procedure with the poweRlaw package (Gillespie, 2015) to see if the experimental distribution was significantly different from a power-law distribution.

4 of the 9 bout length distributions were indistinguishable from either a geometric or negative binomial distribution (Yang et al., 2010; Fig. S4; chi-squared goodness of fit p-values > 0.05). Of these 4, 3 were modeled more parsimoniously by a geometric distribution (lower AIC and BIC), whereas the last was best modeled by a negative binomial distribution. The remaining 5 distributions have extremely heavy tails, and as such they are indistinguishable from a power law distribution (goodness of fit test with bootstrapping procedure p-values > 0.05).

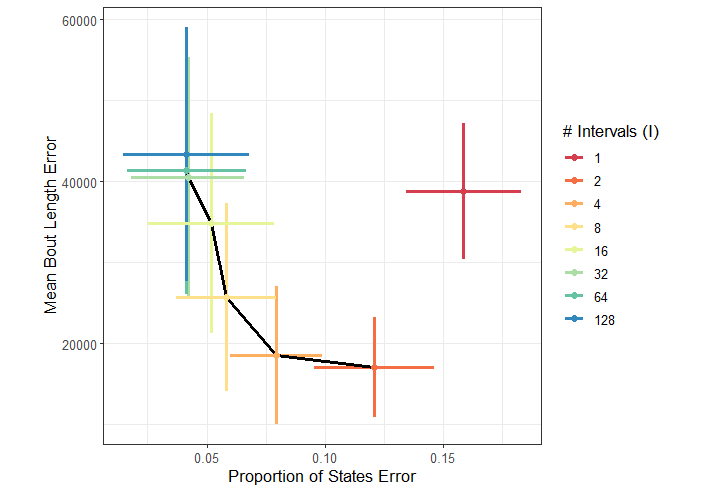


**Fig. S4:** **Distributions of bout lengths for different tasks.** Each panel shows the distribution of bout lengths for each type of task. Color indicates the type of theoretical distribution that best fits the experimental distribution.

**Appendix 5: Using interval number to explore pareto frontier in multi-objective space**

Previously, we assumed that minimizing both dependent and independent error was of equal importance, and thus we rescaled these errors so that we could find the minimum of the sum of these errors. However, there may be studies where it is more important to minimize one type of error over the other. In these cases, it may be more appropriate to find pareto-optimal solutions (Myers et al., 2016). A pareto-optimal solution is one where a criterion cannot be improved without worsening another criterion. In this case, a pareto-optimal solution is a sampling strategy where one type of error cannot be further minimized without increasing the other type of error. The pareto-frontier is the set of optimal solutions.

To find the pareto frontier, we calculate the average error rates for each sampling strategy across simulated and real ants and across sample sizes for each I where I is drawn from the geometric sequence {1, 2, 4, …, 128} (Fig. S5). We find that all solutions are pareto optimal except for I = 1, 128, which represent continuous sampling and our best approximation of instantaneous sampling, respectively. Increasing I can move an experimenter continuously up the frontier, decreasing independent error while increasing dependent error. This pattern holds until I = 128, which is barely off the frontier as its independent error is slightly higher than that of the independent error for I = 64. Conversely, when I = 1, both types of error are high, and thus should be avoided regardless of circumstance.



**Fig. S5:** **Pareto optimal solutions across I.** Here we plot the average independent (proportion of states error) and dependent (mean bout length error) for each I (color), where the error bars represent standard error. The black line connects all points that lie along the pareto frontier.

**Appendix 6: Cause of U-Shaped Error Function for Bout Length**

The U-shape of the error function for bout length was unexpected, as longer sampling segments at small number of intervals (I) and a fixed sample size (n) should yield better estimates of bout length than shorter sampling segments. Thus, error should monotonically increase with I, approaching a limit where increasing I yields roughly the same error rate. Here, we show that this is the case when bout lengths are distributed geometrically, but the U-shape arises when one assumes a power law. To demonstrate this, we consider a simple case where there are only 2 tasks and 1 ant, so k ∈ {1, 2} and the index i for individuals is unnecessary. We calculate the error (E) associated with the bout length of task k = 1, so the expression simplifies to , where l1 is the average bout length for task 1 and is the bout length estimate for task 1. To get , we need to know the average estimate for bout length for each segment of length w = n/I.

First, we assume that bout lengths for both tasks 1 and 2 are drawn from a typical geometric distribution whose success probability is 1/l1 for task 1 and is 1/l2 for task 2. On average then, a time vector containing the states of the ant will periodically switch between states at this rate. For instance, if l1 = 2 and if l2 = 1, then the time vector will be [... 2 1 1 2 1 1 2 1 1 2 …]. To get , we average across all possible results a segment of length w would sample from this idealized vector. We first find the average number of instances of task 1 in a sample, and then we divide this by the expected number of bouts in that sample.

To get the average number of instances, we divide the total amount of instances of task 1 in a sample across all possibilities and divide this by the number of possibilities. There are l1 + l2 possible results, so this will be the denominator of the average. In the numerator, there are 3 groups of results that need to be considered. There is the possibility that the segment only overlaps task 2, and so will have an estimated bout length of 0. This can occur l2 - w + 1 times. There is also the possibility that the segment only overlaps task 1, in which case the estimated bout length will be w. This can occur l1 - w + 1 times. Finally, there is the case where the segment overlaps both task 1 and 2, in which case the estimated bout length will be 1, 2, …, w-1. This can occur 2(w-1) times. The sum 1 + 2 + 3 + … + (w-1) can be expressed as w(w-1)/2. The average number of instances in a sample is therefore:

\frac{0(l_2-w+1)+(w(w-1)/2)(2(w-1)) +w(l_1-w+1) }{l_1+l_2} = \frac{wl_1}{l_1+l_2}

The average number of bouts will be the number of occurrences where task 1 appears at least once in the sample divided by the number of possible samples. This is:

\frac{2(w-1)+(l_1-w+1)}{l_1+l_2} = \frac{w-1+l_1}{l_1+l_2}

To get the average bout length, we then divide the average number of instances by the average number of bouts:

\hat{l}_1 = \frac{wl_1}{l_1+l_2}/\frac{w-1+l_1}{l_1+l_2} = \frac{wl_1}{w-1+l_1}

However, when the average bout length is between 0 and 1, then this artificially inflates the bout length estimate. This means that we should only divide by the number of bouts when it is greater than 1. As w is a function of n and I, this inequality can be expressed as:

\frac{n/I-1+l_1}{l_1+l_2} > 1

Which can be further simplified and written in terms of I:

I > \frac{n}{l_2+1}

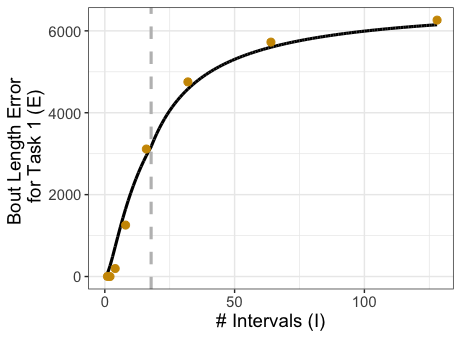
So:

\hat{l}_1 = \begin{cases} 
      \frac{wl_1}{l_1+l_2} & I \leq \frac{n}{l_2+1} \\
      \frac{wl_1}{w-1+l_1} & I > \frac{n}{l_2+1}
   \end{cases}

Plugging these estimates for bout length into the error function E and simplifying yields:

E = \begin{cases} 
      \frac{-l_1^2-l_1l_2+l_1w^2}{(l_1+l_2)^2} & I \leq \frac{n}{l_2+1} \\
      \frac{(-l_1^2+l_1)^2}{(w-1+l_1)^2} & I > \frac{n}{l_2+1}
   \end{cases} (6)

Despite being formulated in an ideal case, this formula fits simulated data quite well when the bout lengths of a behavioral vector are directly drawn from a geometric distribution for both task 1 and 2 (Fig. S6). In these simulations, a behavior vector would be created by first drawing from a geometric distribution whose expected value is 82 seconds, the mean length of brood care, which we take to be task 1. For instance, if the drawn value was 3, then the behavior vector would be [1 1 1]. Next, we would draw from a distribution whose expected value was 36 seconds, the average duration of all other tasks, which we take to be all other tasks. If the drawn value was 2, then this would be added to the previous behavior vector: [1 1 1 2 2]. Once the vector reached a length of 11,041 (the length of time spent observing real ants), we would then use piecewise continuous sampling to estimate the average bout length of brood care across values of I, where I ∈ {1, 2, 4, …, 128}. We repeated this process 1,000 times, and found the average error across simulations. Simulations were done with n = 660. The predictions from equation 6 fit these simulated values well, having an R2 of 0.995 (Fig. S6).



**Fig. S6:** **Simulated error vs closed form solution for error.** Error from both equation 6 (black line) as well as simulations (orange points) along the number of intervals I. The gray dashed line represents the threshold between using one type of error function over the other in equation 6. This threshold is n/(l2 + 1) = 660/(36 + 1) = 17.84.

From the graph, this error function appears to take on the predicted shape. However, this curve was produced from a single set of parameters, and could take on alternative forms under different conditions. To determine whether it monotonically increases to a limit, we first demonstrate that the limit of the function where I is high exists:

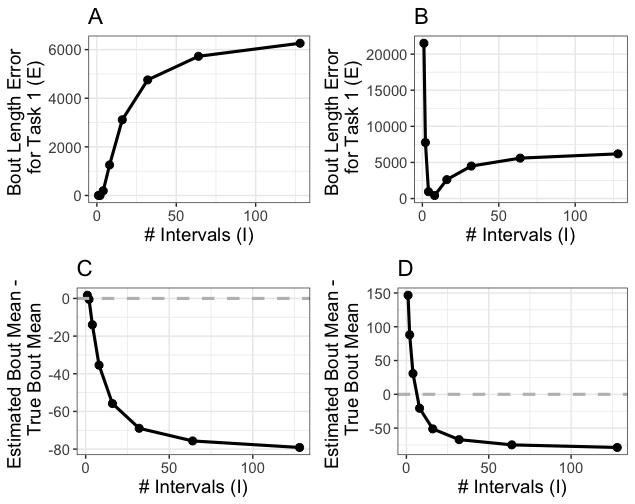
\lim_{I \to \infty} \frac{(-l_1^2+l_1)^2}{(n/I - 1+l_1)^2} = \frac{(-l_1^2+l_1)^2}{(l_1-1)^2}

and then find the derivative of the whole piecewise function with respect to I, which is always positive given the domain of I and the positive values of all constants:

\frac{dE}{dI} = \begin{cases} 
      \frac{2l^2_1n((l_1+l_2)I-n)}{(l_1+2)^2+I^3} & I \leq \frac{n}{l_2+1} \\
      \frac{2(l_1-1)^2l_1^2nI}{((l_1-1)I+n)^3} & I > \frac{n}{l_2+1}
   \end{cases} (7)

Thus, this function has the expected shape, which is distinct from the U-shaped curve in the main body of the text. One critical difference between the two curves is that the above analysis assumed a geometric distribution whereas the behavioral data largely follows a power law (Appendix 4). To see if this has an effect on the shape of the error function, we perform another set of simulations identical to the previous set, except instead of drawing from a geometric distribution, we drew from a power law that had been fitted to the geometric distribution (Gillespie, 2015). Here, we also calculate bout length error, but we also calculate the raw difference between the estimate and the true average bout length (Fig. S7).

Here, we can see that the errors from the geometric bout lengths derive from underestimating the true bout length. The error from the power law distributions, though, is U-shaped because piecewise continuous sampling initially overestimates the bout length at low values of I and then underestimates it at higher levels of I. This initial overestimation at low values of I likely occurs as a result of sampling from unusually long bouts that are present in the tails of the power distribution. High levels of I underestimate bout length for the same reasons it does for the geometric distribution. At the intermediate value of I, the segment length is roughly twice the length of the true bout length for brood care (660/4 = 165, 2\*82 = 164) balances these two phenomena to create a global minimum for error. It should also be noted that even though these simulations were done on a simpler model than those tested in the main body of the text, the global minimum is still the same (I = 4), so this result is strong validation that the U-shaped curve for the error function results from the underlying distribution of bout lengths, not the sampling strategy alone.



**Fig. S7:** **Bout length error for geometrically distributed and power distributed bout lengths.** Bout length error for both geometric (A) and power distributed (B) bout lengths as well as the difference between the estimated and true bout length means (C, D). In C, D, the horizontal dashed line represents a difference of 0.

**Author Contributions**

C. Lynch: Methodology, Investigation, Data Curation, Software, Analysis, Visualization, Modeling, Writing – Original Draft; I. Bespalova: Writing, Data curation, Data collection, Methodology, Investigation; M. Starkey: Conceptualization, Experimental Design, Modeling. T. Pavlic: Conceptualization, Methodology, Modeling, Writing - Review and Editing; S. Pratt: Conceptualization, Methodology; J. Fewell: Conceptualization, Methodology, Writing - Review and Editing, Project administration, Funding administration