An algorithm for quantifying and characterizing misleading trajectories in ecological processes

Christie A. Bahlai[1,2,*], Easton R. White[3,4], Julia D. Perrone[1,5], Sarah Cusser[2] and Kaitlin Stack Whitney[6]

1 Department of Biological Sciences and Environmental Science and Design Research Initiative, Kent State University

2 Kellogg Biological Station Long Term Ecological Research Site, Michigan State University

3 Department of Biology, University of Vermont

4 Gund Institute for Environment, University of Vermont

5 School of Information, Kent State University

6Department of Science and Technology Studies, Rochester Institute of Technology

* Corresponding author cbahlai@kent.edu

Abstract

A fundamental problem in ecology is understanding how to scale discoveries: from patterns we observe in the lab or the plot to the field or the region or bridging between short term observations to long term trends. At the core of these issues is the concept of trajectory—that is, when can we have reasonable assurance that we know where a system is going? In this paper, we describe a *non-random resampling* method to directly address the temporal aspects of scaling ecological observations by leveraging existing data. Findings from long-term research sites have been hugely influential in ecology because of their unprecedented longitudinal perspective, yet short-term studies more consistent with typical grant cycles and graduate programs are still the norm.

We directly address bridging the gap between the short-term and the long-term by developing an automated, systematic resampling approach: in short, we repeatedly 'sample' moving windows of data from existing long-term time series, and analyze these sampled data as if they represented the entire dataset. We then compile typical statistics used to describe the relationship in the sampled data, through repeated samplings, and then use these derived data to gain insights to the questions: 1) *how often are the trends observed in short-term data misleading, and* 2) *can we use characteristics of these trends to predict our likelihood of being misled?* We develop a systematic resampling approach, the 'bad-breakup' algorithm, and illustrate its utility with a case study of firefly observations produced at the Kellogg Biological Station Long-Term Ecological Research Site (KBS LTER). Through a variety of visualizations, summary statistics, and downstream analyses, we provide a standardized approach to evaluating the trajectory of a system, the amount of observation required to find a meaningful trajectory in similar systems, and a means of evaluating our confidence in our conclusions.

KEYWORDS: Population, time series, data mining, scaling, trajectory, firefly, lampyridae

Introduction

A fundamental problem in ecology is understanding how to scale discoveries: from patterns we observe in the lab or the plot to the field or the region, or bridging between short-term observations to long term trends and trajectories [1–3]. Shorter-term studies that are more consistent in length with typical grant cycles and graduate programs are still the norm but understanding where short term patterns fit, and how to interpret short-term patterns in the context of a system's trajectory remains an open question [4]. While long term studies are hugely influential in ecology, given that they require long-term access to research resources and infrastructure their unprecedented longitudinal perspective is not typical [5]. However, when available, these long-term data present a fundamental opportunity to bridge short and long-term trends through data mining. With long term data, we can systematically investigate the presence and prevalence of short-term trends and compare them to the long-term system trajectories these data document

The shape a time series can provide meaningful information about the properties of the system, the rules that govern its variability, and the trajectory that the system is taking [6]. The question of trajectory over time is central in ecology, particularly as related to how ecological systems on which humans depend are responding to disturbance or will behave under future climate or environmental conditions [7]. Trajectory is essential to our understanding of ecosystems, their management, and policy decisions, as we interact with our environment.

Ecological systems are inherently dynamic, and variations in the metrics humans collect about these systems can be driven by a variety of stochastic and deterministic processes, as well as by sampling error or other research-inducer effects [8]. Furthermore, short-term dynamics observed in an ecological system are not always indicative of the long-term trajectory of that system [9]. In population processes, for example, density-dependent deterministic mechanisms couple with environmental perturbations to produce highly variable population numbers during any given time slice [10]. Decoupling these processes can reveal the skeleton of a deterministic process interacting with external forces [11]. However, to disentangle these drivers from an empirical standpoint generally requires a substantial amount of data to be collected over time [12,13]. Indeed, in a recent study, White (2019) found that 72% of vertebrate population monitoring programs required at least a decade of observation before the overall trajectory of the population could be detected statistically. A recent study of trends in water bird populations found that short term trends were generally reflective of longer-termed patterns [4], but varied by the generation length of the organism under study. However, they found that, similar to the White (2019) study, greater than two decades of observations would be required to reliably detect a change of 1% per year. Conversely, a study of population viability modelling in snails determined that although longer time series were generally better for establishing the population's trajectory, diminishing returns in precision were observed after about 10-15 years of data were collected [15]. It is unclear how these findings can be generalized across organisms with differing lifespans and life histories, or other environmental processes.

Yet, it is not uncommon for a shorter-duration multi-year ecological study to extrapolate from its data, using the trends observed within their sampling window to draw conclusions about a system's apparent trajectory. For example, a study of British ladybeetle communities concluded that native ladybeetle species were in decline, as was total ladybeetle abundance, following the introduction of an invasive species [16]. Another found that the richness and abundance of seeds in a soil seed bank were in a recovery trajectory following a period of industrial pollution [17]. An adventive pest species was implicated in reducing carbon to nitrogen ratios, organic matter in soils of infested forests, thus substantially changing the ecosystem's function over time [18]. These examples, representing very different ecological domains, have a common element of a three year study duration. Patterns in publication (**Figure 1**) suggest that two to three-year studies dominate the ecology literature. Yet this three-year study duration, reflective of funding cycles or typical graduate program, may be fundamentally out of sync with the processes they aim to understand, from a temporal perspective [19].

A fundamental problem arises when shorter term studies apply statistical tools at time scales that are not matched with the underlying processes to make inferences about trajectory: not only may spurious trends be observed, but because only a portion of the underlying process variability is captured, a higher degree of statistical confidence in the result will be found. This concept is best illustrated with an example: in recent work, Bahlai and students examined a 12-year time series of fireflies at Kellogg Biological Station in southwestern Michigan [20] with two questions in mind:

When does firefly activity peak?
Are fireflies in decline?

The first question was practical in nature: humans are generally interested in fireflies, and we wished to create a model that would tell us when we could expect the most firefly activity. The second was driven by some concerns raised in the literature that fireflies were indeed in decline [21,22]. Yet, in this population, we found no evidence of decline over the 12 years (**Figure 2**): there was no significant linear relationship between average captures and year (p=0.32) in the larger time series (2004-2015), and, indeed, although the data were limited to capturing two cycles, there appeared to be evidence of a cyclical dynamic common to many populations near their carrying capacity (**Figure 2A**). However, we were compelled by the contrast we observed between the short-term pattern and long-term trends in this system. For example, if we had conducted the study over the four-year period from 2011-2014, we would have had dramatically different conclusions (**Figure 2B**). In this four-year period, we observed significant decline of 0.32+/-0.07 adults per trap per year (p<0.0001), and would likely have concluded that fireflies were indeed experiencing a sharp, consistent decline at our study site. Simply, with less data (even with a slightly longer sampling period than typical), we would have made the wrong conclusions, and we would have been confident in our wrong answer.

It is because of this phenomenon of "highly-confident wrong answers" that long-term studies are so valued in the ecological community. Indeed, because biological systems are often defined by their variability, when studies are shown to be irreproducible, it is not necessarily due to poor research practice, but due to their inability to capture the full variability of the system within the limits of the study design [23,24]. Long-term ecological research provides insight into the inherent variability of natural systems [25], and insights are thus often only apparent after many years of study [26]. Beyond this, there are many other inherent benefits to long-term studies. Long-term studies are disproportionately represented in policy reports and in the ecological literature: studies involving long term observations are cited more often than studies of shorter duration [5]. Furthermore, long-term observational studies provide important baseline data: as the world itself changes, these data provide insight into how ecosystems function, instead of studying phenomena after they happen [27].

Although the importance of long-term studies is clear, empirical examinations of the converse are rare: just how frequently are we misled by short-term studies? Can we use knowledge generated by studying the relationship between short- and long-term studies to bridge our interpretations of short-term data to long-term processes? We use a synthetic, computational approach to develop a framework to address two hypotheses:

Shorter observation periods will increase the likelihood of observing misleading trends

Because exogenous forces are of greater influence at smaller spatial and temporal scales, we predict that short time periods will be more variable due to these processes, and conversely do not capture the full extent of natural variability [8,25], so they are more likely to result in "highly-confident wrong answers."

Statistical metrics often used as a proxy for 'confidence' in short-term trends (such as the p-value) will not be associated with an increased likelihood of capturing a time period consistent with long-term trends.

Following from the previous prediction, we predict that p-values will be inferior predictors of the 'correctness' of short-term trends in predicting longer term trajectory compared to other properties of the system. Better predictors may include statistical measures (slope, standard error), but trends are likely moderated by system specific predictors (e.g. site, data type).

In this study, we develop a suite of tools to directly address bridging this gap between the short-term and the long-term with an automated approach: in short, our algorithm repeatedly 'samples' sequential moving windows of data from existing long-term time series, and analyzes these sampled data as if they represented the entire dataset. The tool then compiles typical statistics used to describe the relationships in the sampled data, through repeated samplings, and then use these derived data to gain insights to the questions, *how often are the trends observed in short-term data misleading, and can we use characteristics of these trends to predict our likelihood of being misled?* Findings from this work will support the development of a deep understanding of temporal scaling in ecology, aiding in the interpretation of countless future short-term studies. Secondly, and more broadly, our findings have applicability across a variety of domains. Results from this approach will have the opportunity to guide science funding policy, experimental design and interpretation, and data archiving.

Materials and Methods

Develop 'bad-breakup' analysis algorithm

The bad-breakup algorithm breaks a time series dataset into all possible sequential subsets and then fits a linear model to each of these subsets and compiles the resulting summary statistics, allowing a user to identify and quantify spurious trends within their data. The algorithm is implemented as a series of functions written in R. The algorithm requires a user-inputted two variable data frame with a regular measurement interval as the first variable, and a response variable as the second variable. For the purpose of this study, we assume a yearly measurement interval and some integrative response metric (captures of organisms per trap, average reading, total yield). Data are first subjected to a standardization function which converts the response metric to a unitless Z-score to normalize the data and make it possible to compare datasets with responses of very different magnitudes, and to minimize the impact of measurement unit choice on the observed trends.

A function that fits a linear model to the data and produces a vector with the number of observations, the number of years in the study, and particular summary statistics of interest, namely, the slope of the relationship between the response variable and time, the standard error of this relationship, p-values for each of these statistics, and then R^2 and adjusted R^2. Although R^2 and p are not measures of statistical confidence per se, they are often used by ecologists in this way [28,29], and thus can be used as a means to approximate 'conclusions' that a researcher might make of the data. We use this fitting function within a moving window function that takes a provided data frame and iterates through it at all possible subsets and intervals, feeding each interval to the fitting function described above, and compiling the fit statistics for each into a single object.

The moving window function is defined as follows. Let $\boldsymbol{D}$ represent the complete dataset, with $\boldsymbol{D_{t,r}}$ representing a single observations of time $\boldsymbol{t}$ and response $\boldsymbol{r}$. Let $\boldsymbol{Y = (y_1, y_2, \ldots, y_n)}$ represent the set of unique values of $\boldsymbol{t}$ at for which observations are recorded, where $\boldsymbol{n}$ is the total number of unique values of $\boldsymbol{t}$. $\boldsymbol{D}$ is partitioned into sequential subsets of size $\boldsymbol{S = (3, 4, \ldots, n)}$ to create windows $\boldsymbol{w_{Y,S}}$ such that each window

$$\boldsymbol{w_{i,j} \subset D = \{ D_{t,r} \mid Y_i \leq t \leq Y_{i+Sj}, \forall\, Y_{i+Sj} \leq y_n \}}, \text{ and } \boldsymbol{w_{y1,n} = D}$$

For each $\boldsymbol{w_{i,j}}$, we apply the fitting function described above, and compile the resultant fit statistics for downstream analyses into a data frame. Then, we created functions that calculate several meta-statistics and produce visualizations of trends from the resultant data frame.

First, we defined the slope of the longest time series (i.e. the slope of the linear regression of the whole dataset, $\boldsymbol{D}$) as the proxy for the 'true' trajectory of the data (as it represents the best information available), along with the computed slope's standard deviation and standard error of the mean as measures of the 'true' variability of the set. Meta-statistics are computed based on comparison to these 'true' statistics.

For all meta-statistics based on frequentist assumptions, we used a set of frequently used 'significance' levels as defaults (i.e. an α=0.05 for line fit statistics) but also encoded the functions so that a user could change these default values easily through supplying a function with different arguments. For each relevant function, we allowed users to toggle via a function argument these meta-statistics based on the full set of windows tested, or only on the set of windows with statistically significant results, as defined above.

We defined "stability time" as the number of time steps needed before a given proportion of slopes (default = 95%) observed in a window of that length are within a certain number of standard deviations (default = 1) of the true slope. We computed absolute range (minimum and maximum values) of slope across all windows, as well as relative range (minimum and maximum difference from the 'true' slope, computed as the slope($\boldsymbol{w_{i,j}}$) minus slope($\boldsymbol{D}$)). We also created functions that computed the proportion of windows examining a dataset would produce particular results. The proportion of statistically significant slopes produced by a given $\boldsymbol{D}$ measure the probability that a randomly selected window of time would produce a 'statistically significant' result. We defined the 'proportion wrong' as the proportion of windows producing statistics that would lead to conclusions differing from those observed for the 'true' trend (i.e. if the true trend was a positive slope, all windows suggesting a negative or non-significant slope were considered spurious, and so on). We provide functions to compute the proportion wrong for all windows in combination, for each window length, and in the set of windows with lengths less than stability time. In combination, these functions provide a standardized approach to asking the questions of how long a system must be observed to make consistent conclusions about its trajectory, and the likelihood of coming to misleading conclusions about a system if it is observed for less than that time period.

We created several visualization functions to enable a user to, for a given dataset $\boldsymbol{D}$**,** quickly interpret trends based on these meta-statistics, and compare trends in outputs across multiple datasets. A pyramid plot (**Figure 3**) uses the data frame of summary statistics from the fits of all windows. It plots the computed slope for each window on the x axis and the length of the window on the y-axis, resulting in a triangular or funnel shaped cloud of points. By default, point size is scaled by the R^2 of the response-by-time relationship within a given window and statistically significant points are demarcated by a circle, and non-significant points given by an 'X'. All points are given with lines indicating their respective standard error. A vertical dashed line indicates the slope of the longest time series, and two dotted vertical lines are plotted at one standard deviation from this value, allowing a user to visually identify the stability time, that is, the length of time required for the majority of windows to produce slopes within a certain interval of the true slope.

The "wrongness" plot (**Figure 4**) examines the same data from a summarized perspective- it plots the average R^2 value and proportion wrong on the y axis by number of years in a window on the x-axis, allowing a user to visualize the relationship between misleading results and the 'confidence' in them for a given ***D***. Finally, the "broken stick" plot (**Figure 5**) allows a user to visualize the raw time series from ***D*** simultaneously with some of the results of the bad-breakup algorithm. The z-scaled response metric (y-axis) is plotted by observation time (x-axis). The true slope of the entire dataset ***D*** is plotted as a solid black line. Then, best fit lines for each window of a user-specified length (default=3-time steps) are plotted, allowing a user to visualize the variation in trend at different points in the time series. Statistically significant slopes are given by dashed red lines, non-significant slopes are indicated by dotted lines. Finally, we created a function which layers and animates broken stick plots to visualize how window slopes change given increasing window length.

The R script was developed in RStudio Version 1.2.5033 "Orange Blossom" running R 3.6.2 "Dark and Stormy Night." The script, its development history and all code for case studies and figure generation are available on GitHub at https://github.com/BahlaiLab/bad_breakup_2

Case Study

We demonstrate the utility of the bad-breakup algorithm using the firefly study which inspired its development [20]. These data on firefly (beetles in the family Lampyridae, with those captured primarily thought to belong to *Photinus pyralis*) captures on insect sticky traps were collected 2004-2015 across 10 plant communities in southwestern Michigan. Complete sampling design and treatments descriptions are provided in Hermann et al (2016). For the purpose of this demonstration, we used the data collected at the perennial early secessional community plots, where fireflies were relatively abundant and complete data were available. Data were subjected to cleaning and quality control using scripts developed by Hermann et al (2016), and then compiled into a metric of total captures per trap, by year (N=12) and replicate (N=6), for a total of 72 observations.

The bad-breakup algorithm produced 55 unique windows (1 sequence of 12 years of data, 2 sequences of 11 years of data, … , 10 sequences of 3 years of data). The full 12 year, 72 observation dataset of the normalized response over time was found to have a non-significant slope (-0.01±0.03, p=0.70) and low R^2 value (0.002) suggesting there is unlikely a linear trend with time in these data (or, more accurately, we fail to reject the null hypothesis that there is no linear relationship between our response and time) (**Figure 3**). Values computed for the slopes across the various windows ranged ±1.2 units around the true slope. The algorithm found a stability time of 7 years, that is, once seven years of data were collected, slopes on >95% of windows tested were within one standard deviation of the slope of the longest series. Overall, nearly half (27/55) of the windows tested found a statistically significant slope, and thus there was nearly a 50% chance a shorter sample leading to a misleading conclusion. Although misleading slopes combined with significant p-values occurred for window lengths longer than 7 years (**Figure 4**), they were much more common with window lengths shorter than the stability time (68% of windows), yet these shorter windows were also more likely to be accompanied by a R^2> 0.1 (**Figure 4**). Although 3 of these 21 windows ≥7 years in length contained statistically significant trends, after stability time, relative slope ranged from -0.14 to 0.17 z-scaled units around the true slope (**Figure 5**).

Discussion

In this paper, we developed a method to directly address the temporal aspects of scaling ecological observations by leveraging existing data, particularly those produced by long-term studies, in the scaling of insights gained from shorter-term investigations. Scaling between the short-term study and the long-term trajectory of a system is a fundamental problem in ecology, and is essential to maximize the utility of observations made in shorter-term studies. Patterns observed in local scale, short-term ecology tend to be dominated by stochastic forces, making generalizations, extrapolations and predictions difficult at larger scales, yet are essential to capture fine-scale understanding of system dynamics [3,30].

The bad-breakup algorithm formalizes a framework for determining how long a system must be observed before conclusions about its general trends can be reached, and the prevalence of misleading results that occur prior to that time period. With our firefly case study, we found that trends observed prior to our ‘stability time’ of seven years had essentially even odds of being misleading: of three possible outcomes for each window (slope more negative than overall trend, slope more positive than overall trend, slope the same as overall trend), 2/3 of outcomes fell into the two former, and erroneous categories. In this case, no net linear trend was observed in the firefly population data (**Figure 3**), so future work should explicitly examine data with different structures to examine the relationship between time series shape and likelihood of erroneous conclusions at differing study lengths. Interestingly, we observed that in our case study, statistics commonly used as indicators of “strength” of relationship suggested more uncertainty, and less ‘confidence’ in results from windows of longer length: p- values, on average, went up, and R^2 values decreased on average as longer windows of the time series were examined (**Figure 4**). This finding shines an important light on the reliability of these statistical tools as indicators of model performance: although they provide measures of how well the data from a given window fit the selected model at that time, they also inflate our confidence in what is often an inappropriate model fit to a spurious or short-term trend. Future work must consider how process characteristics, data availability, and cultural precedent (i.e.: the history of use of a given approach in a scientific field) affect the selection and interpretation of these models.

The bad-breakup algorithm has application beyond our single-population case study. In a recent study, Cusser et al [13] applied the algorithm to a thirty-year experiment comparing the sustainability and productivity attributes of an agricultural cropping system under several management regimes. In this system, due to high variability between treatments, 15 year observation periods were needed to detect consistent between-treatment differences in yield and soil water availability, and at least 1/5 of all windows examined resulted in spurious, statistically misleading trends (i.e. suggest the opposite relationship between management treatments). In another study, Christie, Stack Whitney et al (in prep) compiled 289 surveys of deer tick activity produced by public health departments and researchers primarily in the northeast and Midwest United States and subjected each set of observations to the bad breakup algorithm. They found none of the survey data reached stability time in less than 5 years, indicating that shorter term studies may be insufficient to infer long term population dynamics. Other work has focused on estimating the length of time series required to achieve high statistical power [14], the necessary frequency of monitoring [4], and studying data-poor fisheries [31]

The bad-breakup algorithm uses the longest available study duration as a proxy for 'truth' as its core assumption. However, long-term studies themselves are not immune to uncovering misleading trends. Methodology, site selection, and periods of disturbance following the initiation of a long-term study may inherently bias the apparent trajectory of a system [32]. For example, a 2002 study uncovered a significant multi-year cooling trend, from 1986-1999 in Antarctica's McMurdo Dry Valleys [33]. Yet, this study was initiated during an unusually warm year for the time period, essentially making it impossible for a 'statistically significant' increasing trend to be observed for many years without a series of record-high years: an unusual event in the first year of the study limited the possible outcomes of the statistical analysis. Recent years have seen temperatures stabilize and increase, and correspondingly, increasing stream flow and decreases of thickness of ice in glacial lake systems [34]. This highlights the importance not just of study duration, but of the selection of study starting and ending points: capturing an outlying data point or a high or low in a system's natural variability near the beginning or end of the study period will be highly influential on the statistical outcome, and thus the conclusions reached [32,35]. Understanding and characterizing these highly influential observations in the analysis process is essential to our interpretations of these ecological trajectories. Thus, it is important to consider these biasing factors when using long-term data in algorithms like the one presented herein.

The bad-breakup algorithm uses a linear model as its underlying structure, which is the simplest case of a relationship a response variable might take with time. However, many ecological processes are not linear with time and may be better described with non-linear approaches [4,11,36]. In the initial deployment of this algorithm, we created a tool for the simplest case that would be applicable under a whide variety of circumstances, but future iterations should consider multiple underlying model structures, as well as contingencies for unevenly spaced observations or missing data.

Conclusions and future directions

The ever-increasing availability of long-term data, fostered by the growth of technology that enables automated collection and sharing of data products, and the infrastructure availability and 'maturity' of projects like the US (and international) Long Term Ecological Research networks [37] and more recently, the National Ecological Observatory Network [38,39] present several key opportunities for new understanding of temporal processes in ecology. Not only can these data be used to observe long-term processes in their respective systems, these data can be used to contextualize the vast amount of data produced by shorter-term studies in our field. Ecology, until relatively recently, was a field defined by data scarcity: studies took place at local scales, over time periods manageable to small groups of researchers, and these shorter term studies remain the most common output in ecological research. Their work represents a huge human undertaking, however, and it is critical that we are able to interpret the insights these observations provide appropriately.

The bad-breakup algorithm provides a framework for understanding how ecological data produced by different domains behaves at different temporal scales. Thus, this tools can be used to synthesize data describing ecological processes, specifically examining how system properties (such as landscape, site, seasonality, lifespan in the case of organisms, management regimes, cycles in population trends) affect the likelihood of a spurious trend being observed. In future work, we will examine data of differing structures to identify the characteristics of observation periods that are more likely to produce misleading results, and conversely, the characteristics of time periods that are consistent with longer system trends. This framework will support ongoing research efforts to separate trends in ecological systems from natural variability and underlying processes, and provide critical insight into the scaling to temporal processes between short- and long-term experimental designs.

Acknowledgements

Data used in our firefly case study was collected on traditional Anishinaabe land where Hickory Corners, Michigan is currently located. The bad-breakup algorithm was initially inspired by conversations with John Andrew Gerrath, Ilya Gelfand, and Doug Landis and through feedback and refinements from G. Phillip Robertson, Scott Swinton, Elise Zipkin, Nick Haddad, and the rest of our colleagues at Kellogg Biological Station. Additionally, the algorithm has incorporated feedback from colleagues from the US-Long Term Ecological Research network throughout its development. A particular thanks to Sven Bohm for database curation. Infrastructure supporting this work was funded by the National Science Foundation Long-term Ecological Research Program (DEB 1832042) at the Kellogg Biological Station, and the algorithm was developed with funding from the National Science Foundation Directorate for Computer and Information Science and Engineering (OAC 1838807) to CB, JP and KSW.

References

1. Levin SA. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. Ecology. 1992;73: 1943–1967. doi:10.2307/1941447

2. Schneider DC. The Rise of the Concept of Scale in Ecology: The concept of scale is evolving from verbal expression to quantitative expression. BioScience. 2001;51: 545–553. doi:10.1641/0006-3568(2001)051[0545:TROTCO]2.0.CO;2

3. Chave J. The problem of pattern and scale in ecology: what have we learned in 20 years? Ecology Letters. 2013;16: 4–16. doi:10.1111/ele.12048

4. Wauchope HS, Amano T, Sutherland WJ, Johnston A. When can we trust population trends? A method for quantifying the effects of sampling interval and duration. Methods in Ecology and Evolution. 2019;10: 2067–2078. doi:10.1111/2041-210X.13302

5. Hughes BB, Beas-Luna R, Barner AK, Brewitt K, Brumbaugh DR, Cerny-Chipman EB, et al. Long-Term Studies Contribute Disproportionately to Ecology and Policy. BioScience. 2017;67: 271–281. doi:10.1093/biosci/biw185

6. Esling P, Agon C. Time-series data mining. ACM Comput Surv. 2012;45: 1–34.

7. Sutherland WJ, Freckleton RP, Godfray HCJ, Beissinger SR, Benton T, Cameron DD, et al. Identification of 100 fundamental ecological questions. J Ecol. 2013;101: 58–67. doi:10.1111/1365-2745.12025

8. Suding KN, Gross KL. The dynamic nature of ecological systems: multiple states and restoration trajectories. Foundations of restoration ecology. 2006; 190–209.

9. Carey CC, Cottingham KL. Cross-scale Perspectives: Integrating Long-term and High-frequency Data into Our Understanding of Communities and Ecosystems. The Bulletin of the Ecological Society of America. 2016;97: 129–132. doi:10.1002/bes2.1205

10. Turchin P. Complex population dynamics: a theoretical/empirical synthesis. Princeton University Press; 2003.

11. Bahlai CA, Zipkin EF. The Dynamic Shift Detector: An algorithm to identify changes in parameter values governing populations. PLOS Computational Biology. 2020;16: e1007542. doi:10.1371/journal.pcbi.1007542

12. Higgins K, Hastings A, Sarvela JN, Botsford LW. Stochastic Dynamics and Deterministic Skeletons: Population Behavior of Dungeness Crab. Science. 1997;276: 1431. doi:10.1126/science.276.5317.1431

13. Cusser S, Bahlai C, Swinton SM, Robertson GP, Haddad NM. Long-term research avoids spurious and misleading trends in sustainability attributes of no-till. Global Change Biology. 2020;26: 3715–3725. doi:10.1111/gcb.15080

14. White ER. Minimum Time Required to Detect Population Trends: The Need for Long-Term Monitoring Programs. BioScience. 2019; biy144–biy144. doi:10.1093/biosci/biy144

15. Rueda-Cediel P, Anderson KE, Regan TJ, Franklin J, Regan HM. Combined Influences of Model Choice, Data Quality, and Data Quantity When Estimating Population Trends. PLOS ONE. 2015;10: e0132255. doi:10.1371/journal.pone.0132255

16. Brown PMJ, Frost R, Doberski J, Sparks TIM, Harrington R, Roy HE. Decline in native ladybirds in response to the arrival of Harmonia axyridis: early evidence from England. Ecological Entomology. 2011;36: 231–240. doi:10.1111/j.1365-2311.2011.01264.x

17. Wagner M, Heinrich W, Jetschke G. Seed bank assembly in an unmanaged ruderal grassland recovering from long-term exposure to industrial emissions. Acta Oecologica. 2006;30: 342–352. doi:10.1016/j.actao.2006.06.002

18. Orwig DA, Cobb RC, D'Amato AW, Kizlinski ML, Foster DR. Multi-year ecosystem response to hemlock woolly adelgid infestation in southern New England forests. Canadian Journal of Forest Research. 2008;38: 834–843.

19. Birkhead T. Stormy outlook for long-term ecology studies. Nature News. 2014;514: 405.

20. Hermann SL, Xue S, Rowe L, Davidson-Lowe E, Myers A, Eshchanov B, et al. Thermally moderated firefly activity is delayed by precipitation extremes. Royal Society open science. 2016;3: 160712.

21. Gardiner T. Glowing, glowing, gone? A history of glow-worm recording in the UK. Corby: British Nationalists' Association; 2009.

22. Chow AT, Chong J-H, Cook M, White D. Vanishing Fireflies: A Citizen-Science Project Promoting Scientific Inquiry and Environmental Stewardship. Science education and civic engagement. 2014;6: 23–31.

23. Jarvis MF, Williams M. Irreproducibility in Preclinical Biomedical Research: Perceptions, Uncertainties, and Knowledge Gaps. Trends in Pharmacological Sciences. 2016;37: 290–302. doi:10.1016/j.tips.2015.12.001

24. Voelkl B, Würbel H. Reproducibility Crisis: Are We Ignoring Reaction Norms? Trends in Pharmacological Sciences. 2016;37: 509–510. doi:10.1016/j.tips.2016.05.003

25. Lovett GM, Burns DA, Driscoll CT, Jenkins JC, Mitchell MJ, Rustad L, et al. Who needs environmental monitoring? Frontiers in Ecology and the Environment. 2007;5: 253–260.

26. Knapp AK, Smith MD, Hobbie SE, Collins SL, Fahey TJ, Hansen GJA, et al. Past, present, and future roles of long-term experiments in the LTER Network. Bioscience. 2012;62: 377–389. doi:10.1029/2008gb003336

27. Magurran AE, Baillie SR, Buckland ST, Dick JMcP, Elston DA, Scott EM, et al. Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. Trends in Ecology & Evolution. 2010;25: 574–582. doi:10.1016/j.tree.2010.06.016

28. Yoccoz NG. Use, Overuse, and Misuse of Significance Tests in Evolutionary Biology and Ecology. Bulletin of the Ecological Society of America. 1991;72: 106–111.

29. Nakagawa S, Cuthill I. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews. 2007;82: 591–605.

30. Willis KJ, Birks HJB. What Is Natural? The Need for a Long-Term Perspective in Biodiversity Conservation. Science. 2006;314: 1261. doi:10.1126/science.1122667

31. White ER, Bahlai CA. Experimenting with the past to improve environmental monitoring programs. EcoEvoRxiv. 2019 [cited 21 Nov 2019]. doi:10.32942/osf.io/cz5va

32. Fournier AMV, White ER, Heard SB. Site-selection bias and apparent population declines in long-term studies. Conservation Biology. 2019;33: 1370–1379. doi:10.1111/cobi.13371

33. Doran PT, Priscu JC, Lyons WB, Walsh JE, Fountain AG, McKnight DM, et al. Antarctic climate cooling and terrestrial ecosystem response. Nature. 2002;415: 517.

34. Gooseff MN, Barrett JE, Adams BJ, Doran PT, Fountain AG, Lyons WB, et al. Decadal ecosystem response to an anomalous melt season in a polar desert in Antarctica. Nature Ecology & Evolution. 2017;1: 1334–1338. doi:10.1038/s41559-017-0253-0

35. Chatterjee S, Hadi AS. Influential Observations, High Leverage Points, and Outliers in Linear Regression. Statist Sci. 1986;1: 379–393. doi:10.1214/ss/1177013622

36. Knape J. Decomposing trends in Swedish bird populations using generalized additive mixed models. Journal of Applied Ecology. 2016;53: 1852–1861. doi:10.1111/1365-2664.12720

37. Brunt JW, McCartney P, Baker K, Stafford SG. The future of ecoinformatics in long term ecological research. Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics: SCI. Citeseer; 2002. pp. 14–18.

38. Schimel D, Hargrove W, Hoffman F, MacMahon J. NEON: a hierarchically designed national ecological network. Frontiers in Ecology and the Environment. 2007;5: 59–59. doi:10.1890/1540-9295(2007)5[59:NAHDNE]2.0.CO;2

39. SanClements M, Lee RH, Ayres ED, Goodman K, Jones M, Durden D, et al. Collaborating with NEON. BioScience. 2020;70: 107–107. doi:10.1093/biosci/biaa005

Figures

Figure 1: Ecological studies are most often 3 years in duration. To gain an estimate of typical ecological study length in recent decades, we searched Google Scholar using the terms '"[X] year study" ecology', bounded for 1990-2019, where X= (1, 2, …, 20, 25, 30) and (one, two, …, twenty, twenty-five, thirty).

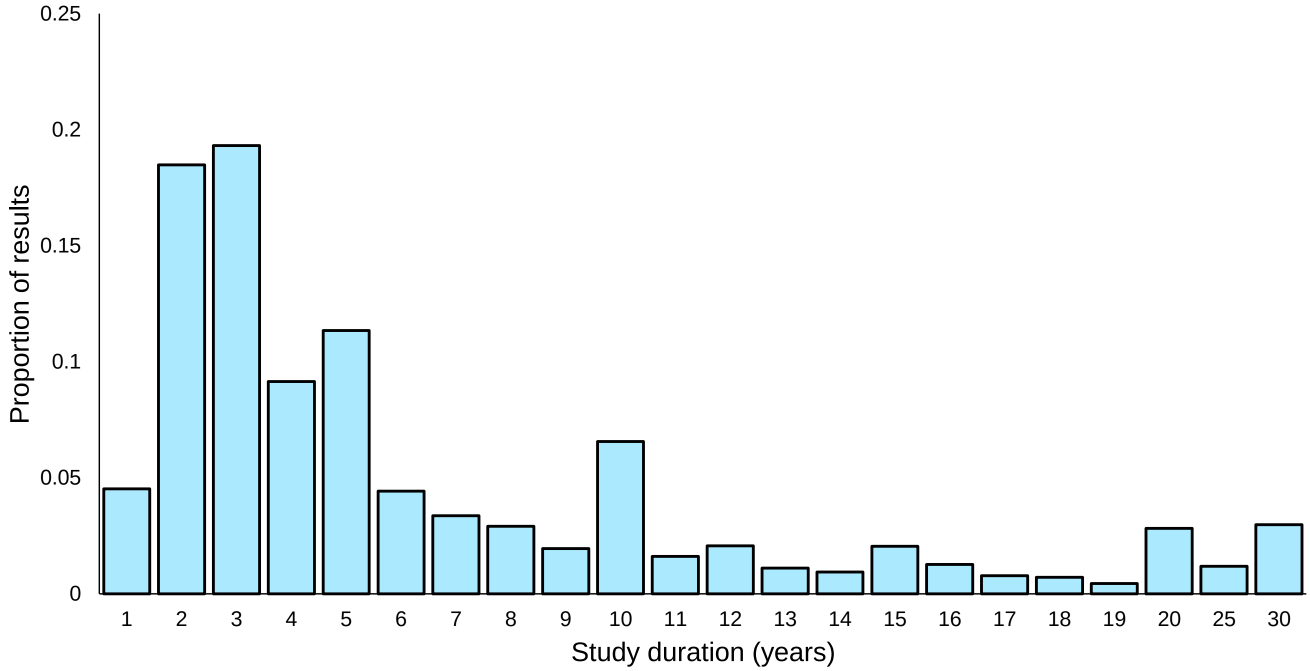
Figure 2: Same data, different observation periods, different conclusions. Firefly populations monitored in ten plant community treatments at Kellogg Biological Station in southwestern Michigan cycle over an approximately 6 year period (panel A). Yet, if sampling had only occurred over a 4 year period, we would conclude the population underwent a steep (and statistically significant) decline in the four years from 2011-2014 (panel B). Data and figures adapted from Hermann et al (2016).

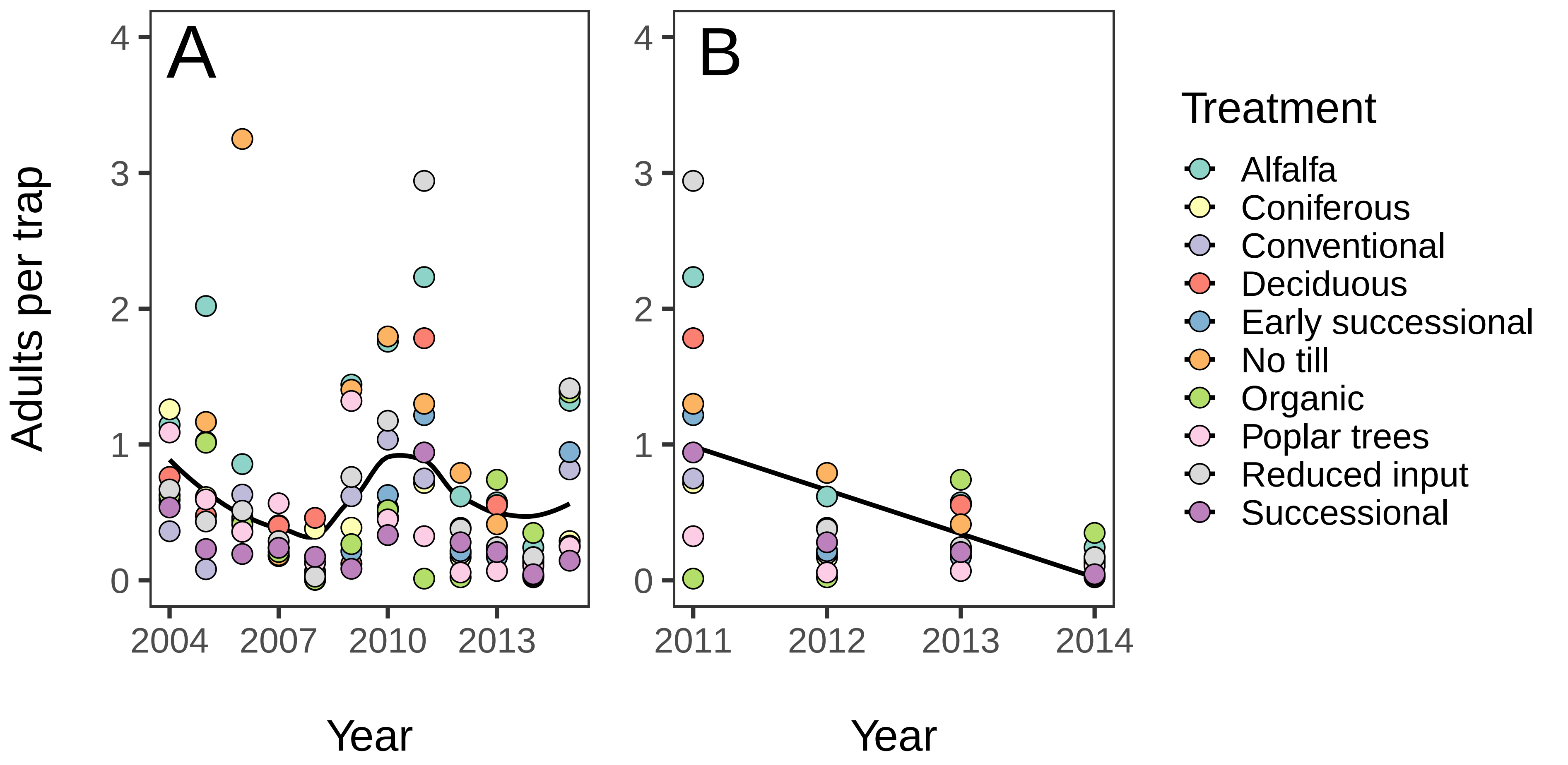
Figure 3: The pyramid plot gives a distribution of possible conclusions. Using the firefly data from the early sucessional plant community presented in Hermann et al (2016), we are able to compile 55 possible windows of three years or greater. On this plot, each point represents a window and its corresponding summary statistics for a linear relationship between the response variable (in this case, z-scaled population density of fireflies) and time. Point coordinates are defined by the slope and length of a window, and point size is scaled by the R^2 computed for that regression. The lines accpmanying each point represent standard error of the slope for each point. Statistically significant relationships (in this case α=0.05) are plotted as black circles, and non-significant slopes are plotted as red Xs. The vertical central dashed black line represents the slope of the complete time series (here with 12 years of data) and the vertical dotted grey lines are placed at one standard deviation in both the positive and negative direction from the 'true' slope.

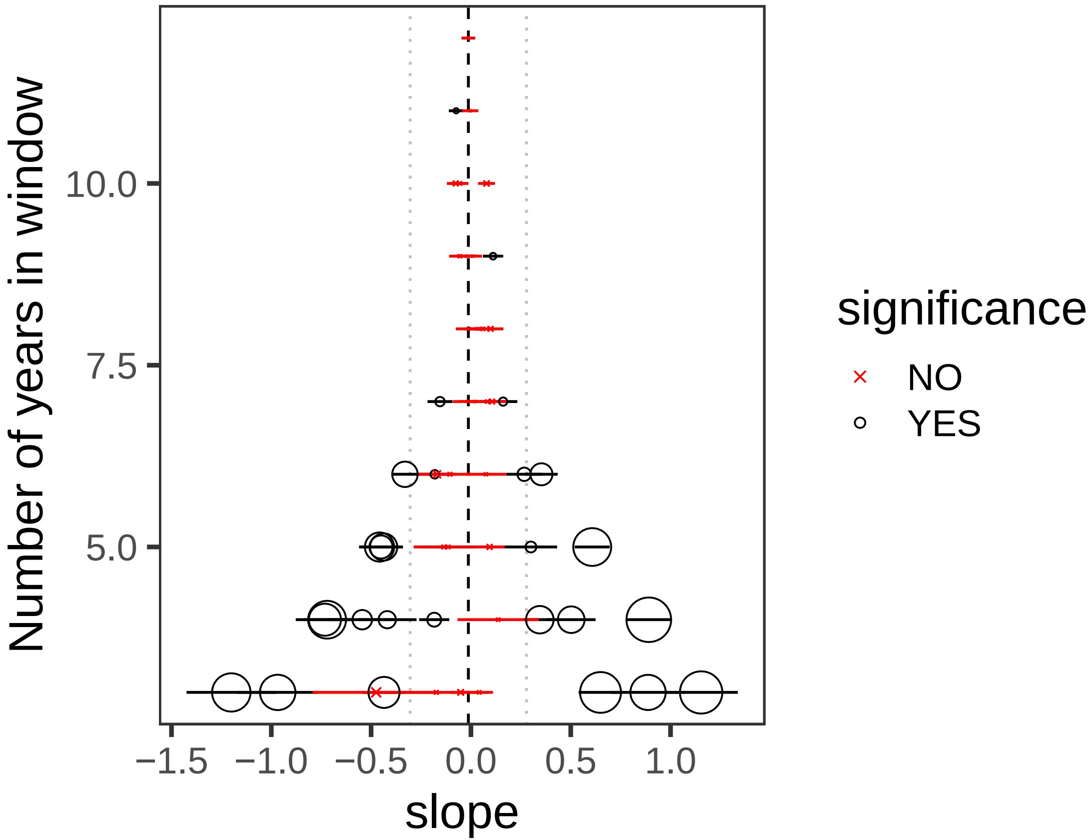
Figure 4: The 'wrongness plot' visualizes the relationship between the likelihood of a spurious conclusion and statistical proxies for 'confidence' in a relationship. Using the firefly data from the early sucessional plant community presented in Hermann et al (2016), we plot the proportion of windows where spurious slopes were observed by the length of window (black circular points with blue solid smoothing line), and the average R^2 value across windows of that length (orange triangular points with a dashed red smoothing line). The grey dotted vertical line is placed at the 'stability time' of 7 years, after which the slopes in 95% of the windows occur within one standard deviation of the 'true' slope.

Figure 5: The broken stick plot allows a user to visualize the magnitude of difference between the slopes produced at different window lengths. Using the firefly data from the early successional plant community from Hermann et al (2016), all of the nine panels presents the Z-scaled response of firefly density over time, and a solid black line indicates the linear regression of the full data series (the 'true'

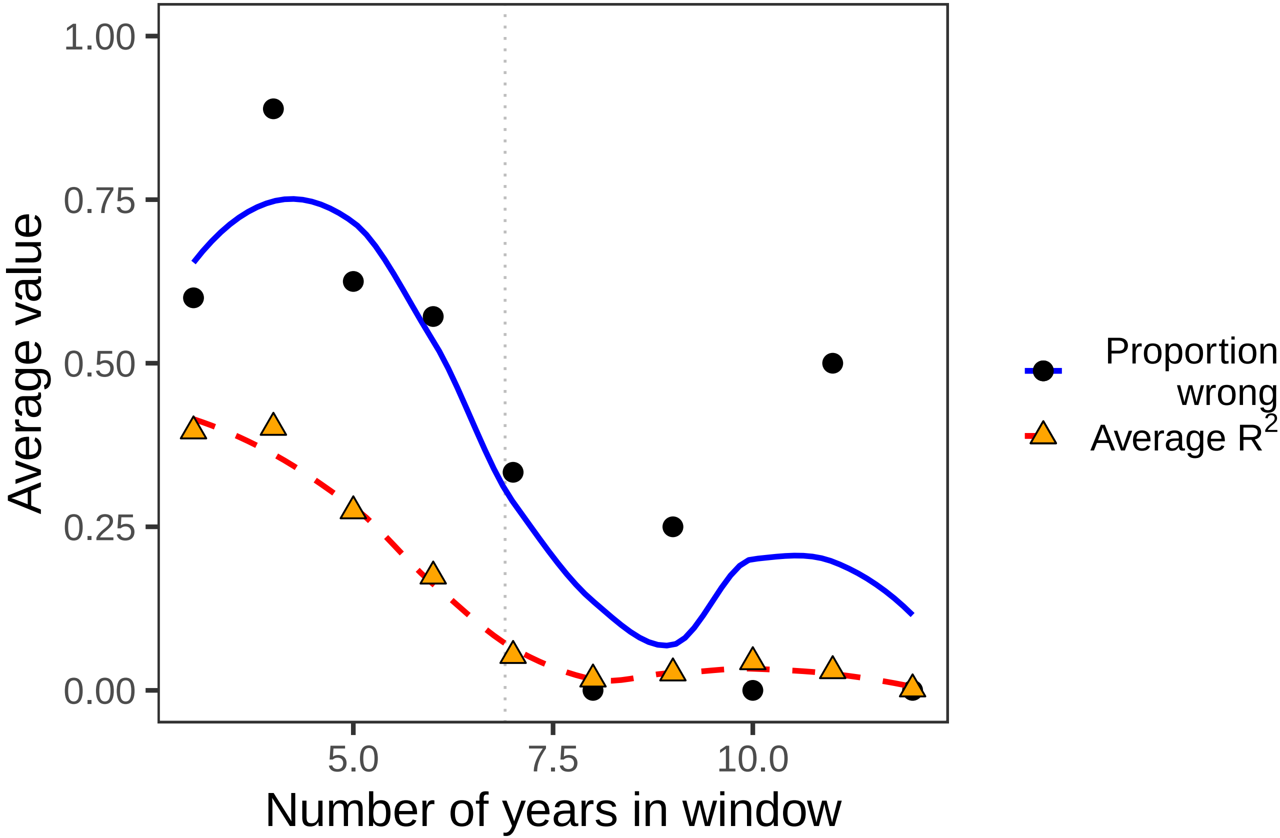
slope). The 95% confidence interval of this line is plotted in light blue. Within each panel, the linear regressions for each window of a given length are plotted: regressions with a statistically significant slope (at α=0.05) are given with red dashed lines, and non-significant regressions are plotted as grey dotted lines.


Proportion of results
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
25
30
Study duration (years)

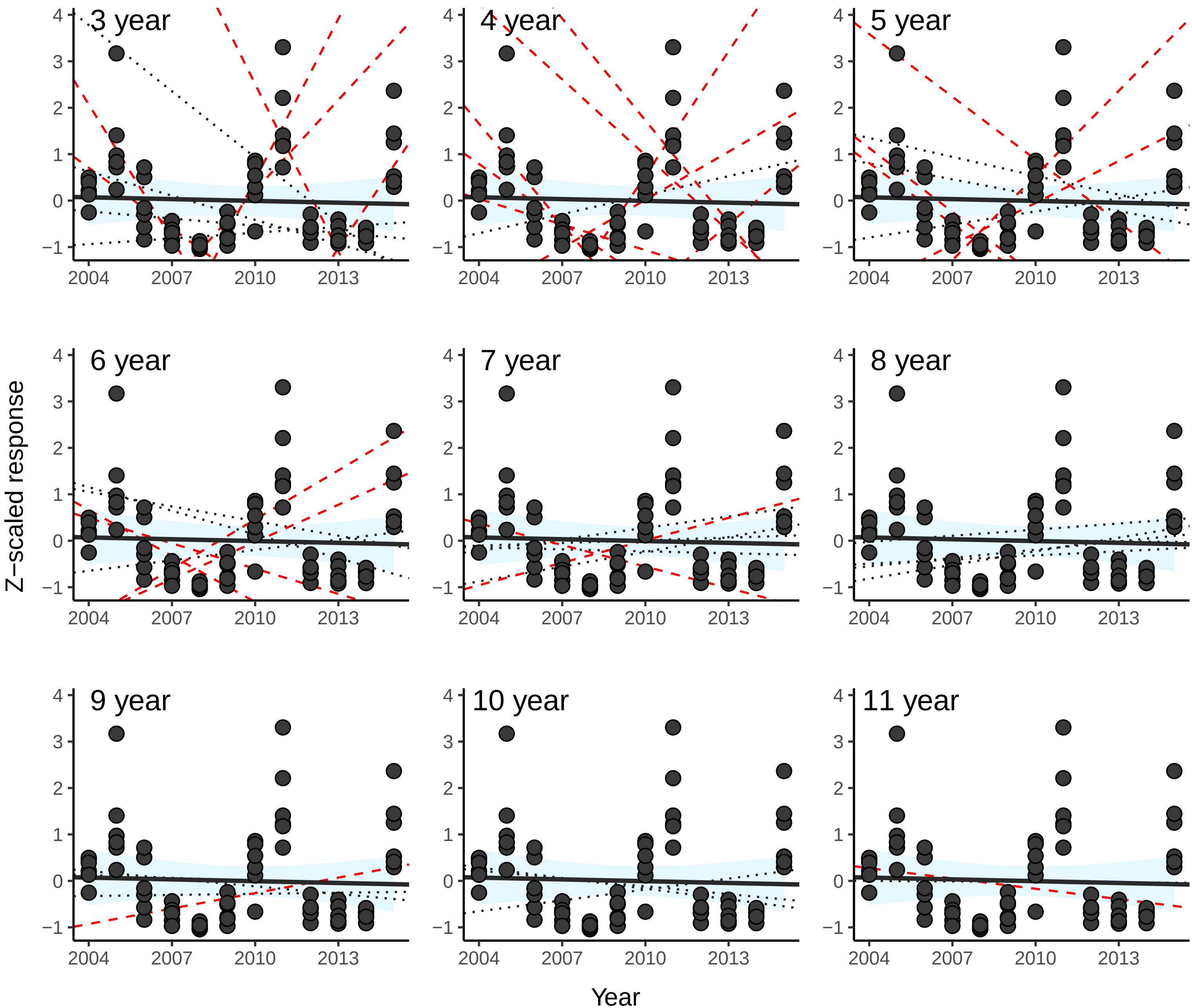

A
B
Adults per trap
0
1
2
3
4
2004
2007
2010
2013
2011
2012
2013
2014
Year
Year
Treatment
Alfalfa
Coniferous
Conventional
Deciduous
Early successional
No till
Organic
Poplar trees
Reduced input
Successional



Number of years in window
10.0
7.5
5.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
slope
significance
NO
YES



1.00
0.75
0.50
0.25
0.00
Average value
5.0
7.5
10.0
Number of years in window
Proportion wrong
Average R^2



3 year
4 year
5 year
6 year
7 year
8 year
9 year
10 year
11 year
Z−scaled response
Year
2004
2007
2010
2013
−1
0
1
2
3
4