

A new tool for power analysis of fixed plot data: Using simulations and mixed effects models to evaluate forest metrics

KATHRYN M. MILLER^{1,†} AND BRIAN R. MITCHELL²

¹National Park Service, Northeast Temperate Network, Bar Harbor, Maine 04609 USA

²National Park Service, Northeast Temperate Network, Woodstock, Vermont 05091 USA

Citation: Miller, K. M., and B. R. Mitchell. 2014. A new tool for power analysis of fixed plot data: Using simulations and mixed effects models to evaluate forest metrics. *Ecosphere* 5(9):110. <http://dx.doi.org/10.1890/ES14-00113.1>

Abstract. Since 2006, the National Park Service's Northeast Temperate Network (NETN) has been monitoring forest health in 10 national park units in the northeastern U.S. using a protocol adapted from the U.S. Forest Service Forest Inventory and Analysis Program. To ensure current methods are appropriate for monitoring long-term trends in forest composition, structure and function, we performed a power analysis of key forest metrics using data collected in each park and covering two four-year survey cycles. We determined statistical power by repeatedly generating bootstrapped datasets with specified percent change between survey cycles in the value of each metric, and then testing whether a mixed effects model detected a significant change. We applied effect sizes ranging from a 50% decline to a 50% increase in 5% increments. Power analyses indicated that, for most key forest metrics, our monitoring program met the target of detecting a 40% change in a metric over a 12-year period with 80% power and a Type I error rate of 0.10. Power also tended to improve with subsequent repeated surveys. Native species richness and live-tree basal area metrics consistently performed well for all parks. Average percent cover of plant groups performed better than quadrat frequency. Regeneration metrics performed best in parks with low or high regeneration rates. Coarse woody debris volume, snag abundance, and invasive species richness did not meet the trend detection target in multiple parks. In most cases, metrics that failed to meet the trend detection target in one or several parks had high proportions of zeros and relatively low overall values for the respective park. In cases where high metric variability was the reason for poor trend detection, results indicated that post-stratifying can sometimes improve power. We developed the power analysis tool in R to be applicable for a range of data types, including proportional and count data, and for any number of sampling areas (e.g., parks) and sampling units (e.g., plots). Our approach represents one of the few tools available that can assess the power to detect change over time using mixed effects models.

Key words: forest health; long-term monitoring; mixed effects modeling; power analysis; simulation.

Received 9 April 2014; revised 8 July 2014; accepted 10 July 2014; final version received 18 August 2014; **published** 26 September 2014. Corresponding Editor: D. P. C. Peters.

Copyright: © 2014 Miller and Mitchell. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. <http://creativecommons.org/licenses/by/3.0/>

† **E-mail:** kathryn_miller@nps.gov

INTRODUCTION

In frequentist statistics, power is the probability of correctly rejecting the null hypothesis of “no change” when a change has in fact occurred. Power is influenced by sample size and variability,

measurement error, effect size, and the Type I error rate (probability of rejecting the null hypothesis when it is true; Cohen 1988), and it is important because it guides decisions on the sample size needed to detect a given effect. Power can be improved by reducing measure-

ment error and sample variability, and/or increasing the sample size or Type I error rate (Cohen 1988, Di Stefano 2001). Power to detect change is generally higher for large changes and lower for small changes (Di Stefano 2001). While knowing power is of interest for all ecological studies, it is particularly important to know for long-term monitoring programs, since a monitoring program with poor power will not be able to detect biologically meaningful changes in the resource being monitored (Urquhart et al. 1998, Legg and Nagy 2006). Power analyses are investigations of statistical power, and are conducted with estimates of sample variance, sample size, effect size, and desired Type I and Type II error levels (Zar 1999). A typical power analysis explores multiple sample sizes to determine the optimum number of samples, or investigates the ability to detect a given effect when error levels are adjusted. Power analyses can also be used to explore proposed changes in methods or stratification, which can affect sample variance. A properly implemented prospective power analysis can be a valuable tool during the design and early implementation of a monitoring program, because it can provide important information about the strength of evidence when no change is detected, and also show where improvements can be made in the monitoring program (Legg and Nagy 2006). That is, when the power to detect change in a metric is low, monitoring staff should consider improving measurement precision, increasing sample size, increasing the Type I error rate, or revising monitoring objectives and removing that metric from the monitoring protocol. Ideally, power analyses are performed during the initial stages of program development to ensure that sampling efforts are worthwhile from the beginning (Foster 2001, Di Stefano 2003).

Unfortunately, in many situations power analyses do not match the planned approach for statistical analysis after the data are collected (or as the data are collected in the case of long-term monitoring programs). Without matching the analytical methods, the power analysis is unlikely to be relevant to the study or monitoring program. Other problems include violation of key assumptions, insufficient data available to conduct the analysis, and inaccurate estimates of parameters used in the analysis (Morrison 2007).

For example, many monitoring programs are set up as fixed plots within a network of sites, and simple power analyses and existing software tools (such as MONITOR; Gibbs and Ene 2010) are not set up to handle this spatial and temporal structure. Analyzing these datasets requires sophisticated statistical tools, such as mixed effects modeling, to properly assess trends (Wang and Goonewardene 2004, Zuur et al. 2009). Therefore, to accurately assess power, we need a dataset that includes repeated measures, and we need to use a model structure and approach that closely mimics how we plan to analyze our long-term datasets. Unfortunately, there are no readily available tools to conduct a power analysis using mixed effects models. To fill this need, we developed a power analysis simulation tool in R that uses mixed effects models, and that can be applied to a range of data types (e.g., proportional and count data), for any number of sampling sites (e.g., parks) and sampling units (e.g., plots), and a range of repeated surveys.

In this paper, we demonstrate our simulation tool, which is broadly applicable for evaluating the power of long-term monitoring metrics (i.e., repeated measures data) collected at randomly located fixed sampling units that are sampled at fixed time intervals within a network of sites. The data used in the simulation can be from pilot data, from the first two rounds of data collection at all or a subset of plots, or they can be generated by the researcher's professional judgment. The simulation bootstraps populations (i.e., collections of randomly located plots) with site-specific means and variances (based on the data) and then systematically applies a range of linear changes over time. We use mixed effects models to analyze the power to detect change that occurs for the entire group of sites, and for each site (e.g., park) independently of the others. By using a model structure very similar to the models we plan to use for analyses of our long-term data and explicitly accounting for the spatial and temporal structure of our data, we have accounted for the most common pitfalls of power analyses (Morrison 2007).

We applied this new simulation approach to forest health monitoring data collected by the Northeast Temperate Inventory and Monitoring Network (NETN) of the National Park Service

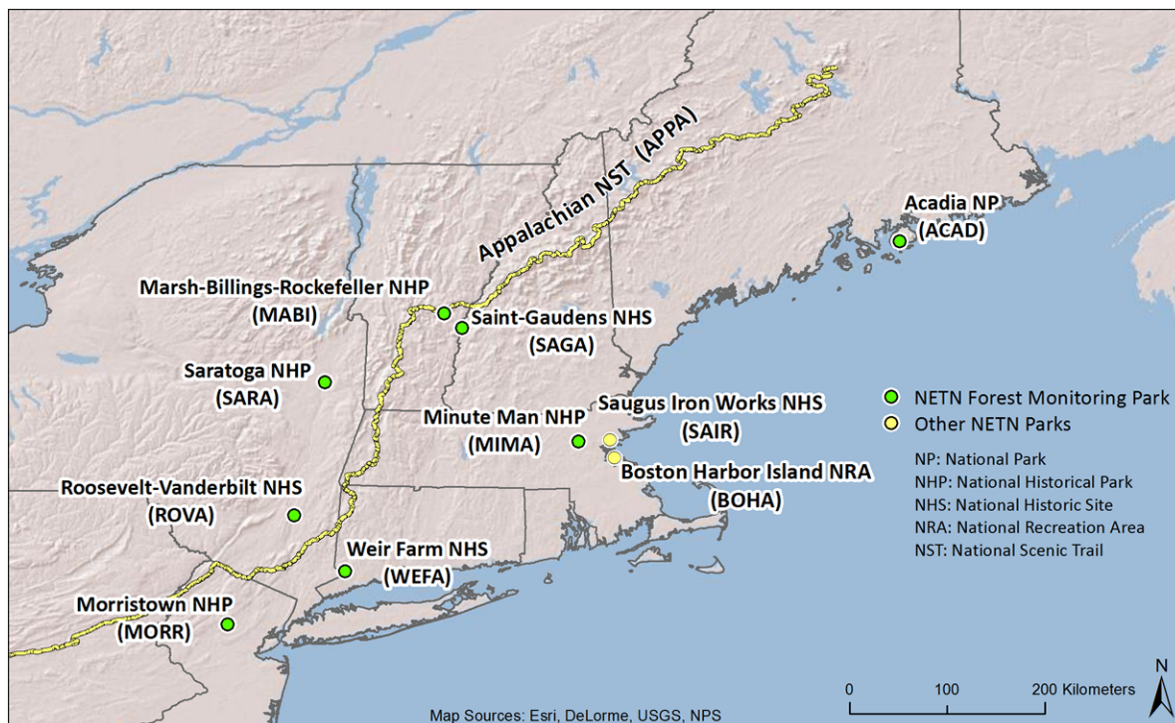


Fig. 1. Map of the Northeast Temperate Inventory and Monitoring Network.

(NPS) in eight national park units in the northeastern U.S. since 2006 (Tierney et al. 2013). NETN adapted the monitoring protocol from the U.S. Forest Service Forest Inventory and Analysis (FIA) Program to focus more on ecological rather than silvicultural metrics. Our methods are also compatible with the five other NPS Inventory and Monitoring (I&M) networks currently monitoring forest condition in eastern U.S. national parks. NETN includes national park units in New Jersey, New York, and New England (Fig. 1). Forests in these parks are primarily composed of northern hardwoods with soils derived from glacial sediment, except at Acadia National Park (ACAD) where the dominant forest type is spruce-fir.

The overall goal of the NETN long-term forest monitoring program is to monitor status and trends in the structure, function, and composition of NETN forested ecosystems in order to inform management decisions affecting those systems (Tierney et al. 2013). We interpret forest condition to park managers using an Ecological Integrity Scorecard that we developed specifically for the parks in NETN. The Ecological Integrity Score-

card rates condition for a suite of metrics that address forest structure, composition and function (Tierney et al. 2009). While the Ecological Integrity Scorecard is a useful tool for evaluating status of forest health in NETN parks, tracking change in key forest metrics over time is also an important function of our monitoring program. We therefore must be capable of detecting change in key metrics of forest condition with sufficient statistical power. While a Type I error rate of 5% and Type II error rate of 20% are the conventional error rates applied in most ecological literature, this approach weighs the cost of making a Type I error much higher than the cost of making a Type II error (Di Stefano 2003, Legg and Nagy 2006). For our forest monitoring program, the cost of missing an actual trend (Type II error) is potentially very high, and it must be balanced with the cost of falsely detecting a trend (Type I error). Therefore, we define sufficient statistical power for the NETN forest monitoring program as having (at a minimum) 80% power to detect a 40% change (effect size) over a 12-year period (i.e., three survey cycles) while controlling Type I (alpha) at 10% and Type II (beta) error at 20%.

Table 1. Northeast Temperate Network forest monitoring panel design. Each panel represents one year of sampling and lists the number of plots sampled per panel in each park. The total number of plots is the number of plots monitored in each park, and is the sample size generated by each bootstrap simulation. Plots in italics comprised the initial dataset that was bootstrap sampled for this analysis.

Park	Cycle 1				Cycle 2				Total no. plots
	2006 Panel 1	2007 Panel 2	2008 Panel 3	2009 Panel 4	2010 Panel 1	2011 Panel 2	2012 Panel 3	2013 Panel 4	
ACAD	44	44	44	44	44	44	44	44	176
MABI	12		12		12		12		24
MIMA	10		10		10		10		20
MORR		14		14		14		14	28
ROVA		20		20		20		20	40
SAGA	10		10		10		10		20
SARA	16		16		16		16		32
WEFA		5		5		5		5	10

Several power analyses were performed by other eastern NPS I&M networks during the protocol development stage for NETN forest monitoring (Comiskey et al. 2009, Schmit et al. 2009). While these analyses helped inform our methods, plot size, and plot allocation in each park, they were based on pilot data that were collected from one survey cycle. Without incorporating repeated measures of the same plots (i.e., temporal variability), results from these studies most likely underestimated the power to detect trends and were considered preliminary. In addition, the very nature of long-term monitoring, which involves repeated measurements from fixed locations, generates datasets that violate one or more of the underlying assumptions of simple linear regression (e.g., independence and homogeneity). We developed our simulation tool to overcome these problems and help us evaluate our ongoing monitoring program.

METHODS

Sampling methods

We sample permanent forest plots at four-year intervals using a rotating panel design, with one survey cycle consisting of four panels (Table 1). We sample one of the four panels each year, with each panel including one-fourth of the ACAD plots and half of the plots for a group of the remaining parks. We sample each of the historical parks or historic sites in alternate years. For example, in 2006 and in 2010, we sampled the first 44 plots in ACAD and the first 12 plots in Marsh-Billings-Rockefeller National Historical

Park (MABI). In 2007 and 2011, we sampled the second 44 plots in ACAD and the first 14 plots in Morristown National Historical Park (MORR). We randomly located plots within each park using generalized random tessellation stratified (GRTS) sampling (Stevens and Olsen 2004).

The permanent plot design is illustrated in Fig. 2. In the full, fixed-area ($15 \times 15 \text{ m}^2$ at ACAD; $20 \times 20 \text{ m}^2$ at the other parks) square plots, we collected basic information describing the site (e.g., slope, aspect, physiographic class), qualitatively assessed stand structure and disturbances, and photographed six standard scenes of the plot to facilitate interpretation of other data collected in the plot (Tierney et al. 2013). In the full square plot, we tagged and measured each tree $\geq 10 \text{ cm}$ diameter-at-breast-height (DBH). For each tree, we identified the species, measured the DBH, and assessed its status (live/dead), crown position (e.g., dominant, co-dominant, intermediate, etc.) and condition. Within three 2-m radius circular microplots per plot, we measured tree regeneration by tallying tree seedlings by species and height class, and measuring DBH of saplings ($\geq 1 \text{ cm}$ and $< 10 \text{ cm}$ DBH). For coarse woody debris (CWD), we used line intersect sampling along three 15-m transects originating at plot center. To monitor understory composition, we estimated percent cover by species of all vascular plants within eight 1-m^2 quadrats. We assessed forest floor condition by visually inspecting for evidence of microtopography, trampling, and earthworms. To monitor forest soils, we collected soil samples from a location adjacent to the plot and had the soils analyzed chemically. Tierney et

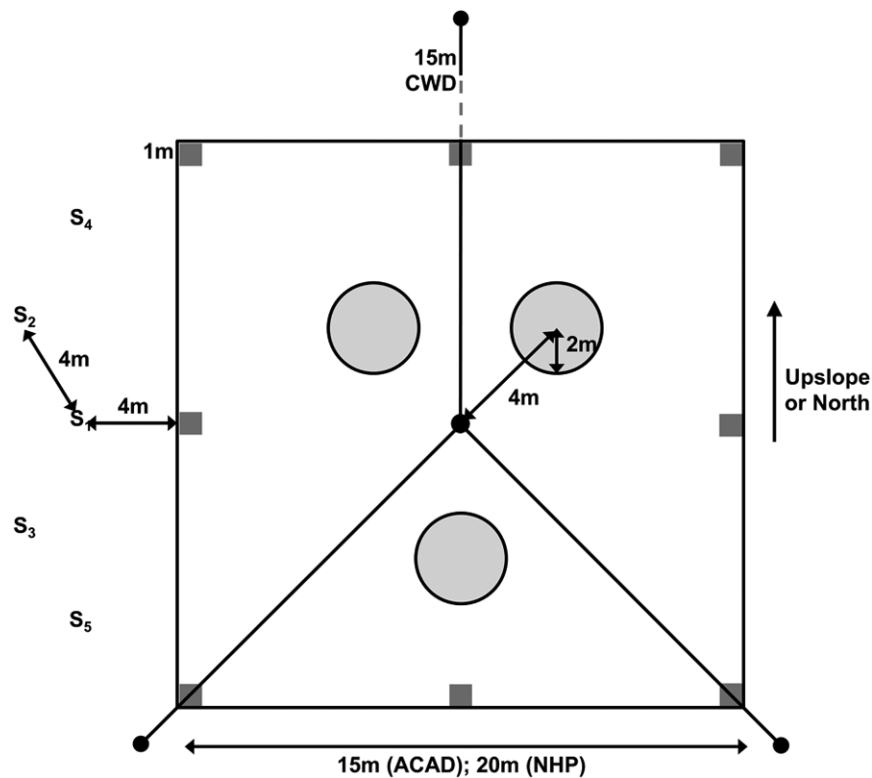


Fig. 2. Plot layout showing square tree plot with three nested 2-m radius regeneration microplots, eight 1-m² vegetation quadrats, and three 15-m coarse woody debris (CWD) transects. S_x is location of soil sample.

al. (2013) contains complete details of the sampling design and methods.

Data analysis

We performed power analyses on key forest metrics related to ground layer abundance and composition, tree regeneration, species richness, CWD, snag abundance and live tree basal area (Table 2). We derived average percent cover by

averaging percent cover class midpoints across eight quadrats in each plot. While cover classes were not evenly distributed (e.g., 1–2% versus 57–75%), estimation of mean plant cover using similar cover classes has been shown to perform equally to continuous percent cover estimation (Damgaard 2014). Table 3 lists groups that were included in percent cover and percent frequency analyses. The invasive species list contained over

Table 2. Description of metrics included in power analyses. Regeneration stocking index was developed by McWilliams et al. (2002). Abbreviations: CWD, coarse woody debris; DBH, diameter at breast height.

Metric	Definition	Units
Average percent cover	Average quadrat percent cover of guild or indicator	%
Quadrat frequency	Percent quadrat frequency of guild or indicator	%
Native richness	No. native vascular species in each plot	no. species/plot
Invasive richness	No. indicator invasive species in each plot	no. species/plot
CWD volume	Volume of CWD ≥10 cm diameter and ≥1 m long	m ³ /ha
Snag abundance	Density of snags ≥10 cm DBH	stems/ha
Stocking index	Weighted measure of tree seedling density by height class	n/a
Seedling density	Density of tree seedlings that are ≥15 cm tall and <1 cm DBH	stems/ha
Sapling density	Density of saplings that are 1–9.9 cm DBH	stems/ha
Live tree basal area	Calculated for all trees ≥10 cm DBH	m ² /ha

Table 3. Description of species guilds examined in average percent cover and quadrat percent frequency power analyses. Tierney et al. (2013) lists the species in the invasive, deer-preferred, and deer-avoided species groups.

Group	Definition
Native herbs	Native herbaceous species
Ferns	Native species of ferns and fern allies
Graminoids	Native species in Cyperaceae, Juncaceae, and Poaceae families
Invasives	Priority invasive exotic plant species
Deer-preferred species	Indicator list of herbaceous species that are preferably browsed by deer
Deer-avoided species	Indicator list of herbaceous species that are unpalatable or toxic to deer, or are resistant to deer browse.

two dozen exotic plant species that are highly invasive in northeastern forest, woodland and successional habitats, and that can dominate and/or persist under shaded conditions (Tierney et al. 2013). The deer-preferred species list included 15 common, highly visible herbaceous species that are preferred browse for deer, and/or have been shown or predicted to be intolerant of deer browsing due to life history traits (Augustine and Jordan 1998, Augustine and deCalesta 2003, Tierney et al. 2013). The deer-avoided indicator list was comprised of three herbaceous species and one genus (*Carex*) that are considered unpalatable to deer or resistant to deer browse, and that have been shown to increase in abundance under heavy deer browse pressure (Augustine and Jordan 1998, Horsley et al. 2003, Tierney et al. 2013). Species richness was the total number of vascular species observed in each plot. The status of each species (native or exotic) was determined using NatureServe Explorer (NatureServe 2012).

To calculate CWD volume, expressed in m^3/ha , we used Huber's formula, which estimates CWD volume using the diameter of each CWD piece measured at the point of intersection with the transect line (Marshall et al. 2000). We then corrected CWD volume estimates for slope (Van Wagner 1982). Snag abundance was the number of snags within a plot converted to a per hectare basis to account for different plot sizes among parks. The stocking index was developed by McWilliams et al. (2002), and assigned a weight to each seedling based on its height class, with higher weights for larger height classes. We calculated the stocking index as the sum of the weights across all height classes, and averaged this sum over three 2-m radius microplots per plot. The seedling density metric was the sum of all tree seedlings that were greater than 15 cm

tall, and less than 1 cm diameter at breast height (DBH), and was averaged across three microplots and converted to a per hectare basis. The sapling density metric was the number of tree species that were 1.0–9.9 cm DBH, averaged over three microplots and converted to a per hectare basis.

We performed bootstrap power analysis simulations using mixed effect models on key forest metrics of NETN forest health data that represented half of two survey cycles (panels 1 and 2 from cycle 1 and cycle 2) and eight park units. We did not use complete survey cycles (four panels) because we began this analysis mid-way through the second cycle of sampling, and only had resample data from the first two panels of plots. Because we used GRTS sampling to randomly locate plots, we could analyze a subset of the full collection of plots, and as long as the plots were consecutively ordered by GRTS priority number, the subset of plots was spatially balanced and representative of the target population (Stevens and Olsen 2004). A bootstrapped dataset is a plausible dataset produced by randomly selecting data with replacement from a pilot or sample dataset until a desired sample size is reached (Manly 2007), which in this case is the total number of plots currently established in each park. For each metric, our simulation generated 250 bootstrapped datasets for each effect size, model, and number of survey cycles in order to produce a distribution of potential results for a range of effect sizes. We chose mixed effects models for our simulations because this type of model allows estimation of certain “fixed” or “treatment” effects (e.g., park unit or trend over time) while accounting for “random” effects, such as randomly located plots (Zuur et al. 2009). We used the mixed effects model functions available in the nlme package in R for our analyses (R Development Core Team 2011,

Pinheiro et al. 2012). Our sampling strategy is a design-based spatial sample with a model-based temporal sample, which Brus and de Groot (2012) term a “hybrid sampling approach”. Our simulation and modeling process assumed that sample units were spatially independent and did not incorporate design-based spatial variance because the focus was on temporal trend detection rather than identification of spatial patterns. If the temporal trends in nearby sites are correlated, our approach would likely underestimate the true power that could be obtained by incorporating spatial design-based variance. Because we treated time as a fixed (model-based) factor, our simulation tool is not appropriate for monitoring approaches that are fully design-based in time (such as complex temporal revisit patterns). We also included park as a fixed effect in the mixed effects models, and used the `varIdent` function to model different variances for each park.

Prior to bootstrapping, our R code performed a user-specified data transformation to prevent simulations from using impossible values (e.g., negative counts, proportions below zero or above one, etc.). Metrics that could not be negative (i.e., live tree basal area, snags per hectare) were log transformed. Metrics that ranged between 0 and 1 (e.g., percent cover data) were transformed using a modified logit following Warton and Hui (2011). The modified logit transformation added a small fixed constant (ϵ) to the logit numerator and denominator to solve the issue with sample proportions that equal 0 and 1, which otherwise resulted in undefined values. The constant ϵ was defined as the smallest non-zero number in the dataset, or the smallest difference between 1 and the largest number less than 1, whichever was smaller. In a few cases, bootstrapped datasets had a tendency to result in all zeros for a park, which resulted in no variance and ended the simulation (e.g., average cover of invasives in ACAD). To resolve this issue, we “jittered” the original data by randomly adding small non-zero values to 10% of the plots in the park with the problem.

Bootstrapped datasets were generated by randomly selecting values from the first sampling cycle by park (first two panels in cycle one of actual NETN data) to generate initial data values for the desired total number of plots in

each park. In this case, the number of plots generated in each bootstrap was the total number of plots NETN currently monitors in each park. Data for a resurvey of these plots were simulated by applying an effect (between -50% and 50% , see below) to the initial data values, and then adding simulated sampling variation. The simulated sampling variation was generated from a normal distribution with a mean of zero and a park-specific standard deviation equal to the standard deviation of the differences in metric values between the first and second sampling cycle of each plot (based on the actual data). This approach will overestimate sampling variation in the presence of a temporal trend or annual variation in the metric, as it assumes that all variation between the two cycles is due to sampling variation. Nevertheless, it provided a conservative estimate of sampling variation suitable for small datasets. Additional occasions were simulated by applying an additional effect to the initial data values, and again adding simulated sampling variation.

Statistical power was determined by repeatedly generating datasets with a specified percent change between sampling cycles in the value of each metric, and then testing whether the mixed-effects model could detect the change (modeled as a linear trend). For the purposes of this analysis, we used a Type II error rate of 20% (i.e., 80% power), and used both the conventional Type I error rate of 5%, and a 10% Type I error rate to see how results compared. Our modeling approach assumed variables were continuous with normal error distribution. While some metrics, such as seedling density and species richness, may more appropriately be modeled as count data (e.g., Poisson distribution), the `nlme` package we used for this simulation tool does not have a generalized mixed effects model function. We chose the `nlme` package because of the variance structure modeling features (e.g., `varIdent`). Our simulation tool could easily be modified to run generalized mixed effects models, such as those offered by the `glmer` function in the `lme4` package (Bates et al. 2014), but this is beyond the scope of this study.

We ran simulations for percent changes ranging from a 50% decline to a 50% increase in 5% increments. When simulating multiple cycles of data, the R code allowed the level of change to

occur between each cycle (cycle-to-cycle change) or across all cycles (total change). For this analysis, the percent change was the total change. That is, for three cycles of data, changes were limited to $\pm 25\%$ between each cycle. There is an option in our simulation tool to apply a constant effect size across all cycles (e.g., applying a 50% increase to every simulated cycle), but the results of those analyses are beyond the scope of this study. The percent change was first applied to all parks to test the additive model with no interaction, and then for each park (with no change applied to the other parks) to test the model with an interaction effect. The additive model assessed the power to detect a uniform change across all parks (best-case scenario), while the interaction model assessed the power to detect change at only one park (worst-case scenario). Note that in the situation where there is no effect for any park (0% change), the analysis does not assess power. In this case, the percentage of times when a trend is detected should be approximately equal to the Type I error rate (alpha). Following the notation approach used in Sullivan et al. (1999) and Raudenbush (1993) to describe hierarchical models, the model statement for our additive model was

$$\text{Level 1: } y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_j^2)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + v_{0j}, v_{0j} \sim N(0, \tau_0^2)$$

$$\beta_{1j} = \gamma_{10} + v_{1j}, v_{1j} \sim N(0, \tau_1^2)$$

where i = plot (1–350); j = park (ACAD, ..., WEFA); y_{ij} = the metric value at the i th plot within the j th park; β_{0j} = the intercept for the j th park; β_{1j} = the slope associated with cycle (time; 1, ..., 4) for the j th park; x_{1ij} = covariate (cycle) for the i th plot within the j th park; ε_{ij} = error term for level 1 model, normally distributed with a mean of 0 and park-specific variance σ_j^2 ; γ_{00} = overall mean intercept for all parks; γ_{10} = overall mean slope (due to cycle) for all parks; v_{0j} = random error term for level 2 model of level 1 intercept, normally distributed with a mean of 0 and variance τ_0^2 ; v_{1j} = random error term for level 2 model of level 1 slope (due to cycle), normally distributed with a mean of 0 and variance τ_1^2 ; and the model statement for our interactive model was

$$\text{Level 1: } y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_j^2)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + v_{0j}, v_{0j} \sim N(0, \tau_0^2)$$

$$\beta_{1j} = \gamma_{10} + v_{1j}, v_{1j} \sim N(0, \tau_1^2)$$

where β_{2j} = the slope associated with cycle $\times$ park for the j th park; x_{2ij} = covariate (cycle $\times$ park) for the i th plot within the j th park and all other terms are defined as in the additive model.

In both model statements, the metric value was the raw or transformed (logit or log) data, park was a site containing multiple sampling units (plots), and cycle was the index for the sampling cycle.

For each set of simulations for a given metric, additive and interaction mixed effects models were fit based on two, three and four survey cycles, and confidence intervals were calculated for each bootstrapped population and multiple levels of alpha (0.01, 0.05, 0.10, and 0.20) using the parameter estimate for time (i.e., change between cycles) and its standard error. Confidence intervals that did not contain 0 were considered significant, and the results of each significance test were stored by effect size, park, and simulation. Power was calculated as the proportion of simulations where a trend was detected (i.e., proportion of confidence intervals that did not contain 0), and was calculated for each effect size across the network and within each park. This confidence interval approach assumed that the error distributions of the model were normal, and based on examination of residual plots from the first model in each simulation, this was a valid assumption for our data. We included code in our simulation tool that plots the residuals against fitted values and residuals by park and cycle for the additive and interactive models that were derived from the first of the 250 simulations with an effect size of 0 to allow users to inspect the error distributions. If the residual plots show patterns or violations of normality, this confidence interval approach may not be an appropriate method to assess whether a trend was detected. Using the current number of plots established in each park, we ran a total of 141,750 simulations on each metric: 250 bootstrapped populations at each of 21 effect sizes, two to four cycles of data, and nine scenarios

(network-wide change or change at one specific park). The R code that we used for the simulations in this article, along with an example dataset, are available as a Supplement.

In cases where parks had poor trend detection due to high metric variability, we tested whether post-stratifying the data (e.g., by land use history) improved power. The stratified datasets included data from all plots in the respective park that had been sampled through 2013 (two complete cycles of data), to increase the pool of data per group for the bootstrap. Similarly, we tested whether adding the second panel of plots in Weir Farm National Historic Site (WEFA), which increased the number of plots from 5 to 10, improved the power. In MORR, we stratified by Mature and Successional stand types. Mature stands were defined as second growth stands that had no history of cultivation and that originated from areas cleared by General Washington's forces around 1788 to 1789. The Successional category represented younger stands that originated on abandoned field or pasture around 1920 or later. Soils in successional stands were likely plowed, and forests were mostly even-aged (Shaw and Patterson 2006). In MABI we stratified by Natural and Plantation stand types. Natural stands were second growth forests that were primarily composed of naturally regenerated hardwood species. Plantations consisted of forests that were planted with conifer species, and had been managed more intensively for timber by the park. In Saratoga National Historical Park (SARA), we stratified forest plots based on land use history. The Field category represented areas that were open field in 1927 when the park was established; the soils in these areas were likely plowed and some soil amendment (e.g., manure, liming) may have occurred. The Forest category included sites that were forest or woodlot in 1927 and have been continuously forested since 1927.

RESULTS

Average percent cover

Minimum detectable percent change at 80% power and $\alpha = 0.10$ ranged from 5% to 10% for all species groups and across two, three, and four survey cycles in NETN and ACAD (Fig. 3). These results suggested more than adequate

sample sizes to detect network-wide changes in average percent cover of important plant groups, as well as changes occurring only in ACAD. Minimum detectable percent change was also 10% or better for NETN and ACAD at the $\alpha = 0.05$ level, but remaining parks were more likely to meet the 40% trend detection target at $\alpha = 0.10$ (Figs. 3 and 4). Average percent cover of deer-preferred species performed well for all parks, with all parks meeting the target of detecting 40% change over three cycles at 80% power and $\alpha = 0.10$. Trends in average percent cover of deer-avoided species were also detectable at 80% power and $\alpha = 0.10$, but minimum detectable percent change tended to be 5–10% higher than for deer-preferred species, and WEFA did not meet the trend detection target. Trend detection in fern and graminoid cover was consistently better than for native herbaceous and invasive cover (Fig. 3).

While most parks met the 40% change target for each plant group at 80% power and $\alpha = 0.10$, exceptions were invasive cover in Minute Man National Historical Park (MIMA) and MORR, deer-avoided species in WEFA, and native herb cover in SARA. Poor trend detection of invasive cover in MIMA and MORR was likely the result of high variability of invasive cover in both parks, and small sample size may have been an issue in MIMA (Table 4, Fig. 5). In MORR, invasive cover was significantly higher in successional forests compared with mature stands (Miller et al. 2012). Post-stratifying invasive cover in MORR by successional stage improved trend detection for the mature group, but not for the successional group, which failed to detect even a 50% change over four cycles (Fig. 6). Native herb cover in SARA followed a similar pattern as MIMA and MORR with invasive cover (Fig. 5). Post-stratifying SARA by land use history improved trend detection for the forest group, but both the field and forest groups failed to meet the trend detection target (Fig. 6).

Small sample size and high metric variability were likely the reasons for the inability to detect the target level of change in cover of deer-avoided species in WEFA (Fig. 3). At the time this power analysis was initiated, only five plots of data were available for the simulation, which was a fairly small pool for the bootstrap to pull from. After running the simulation with 10 plots

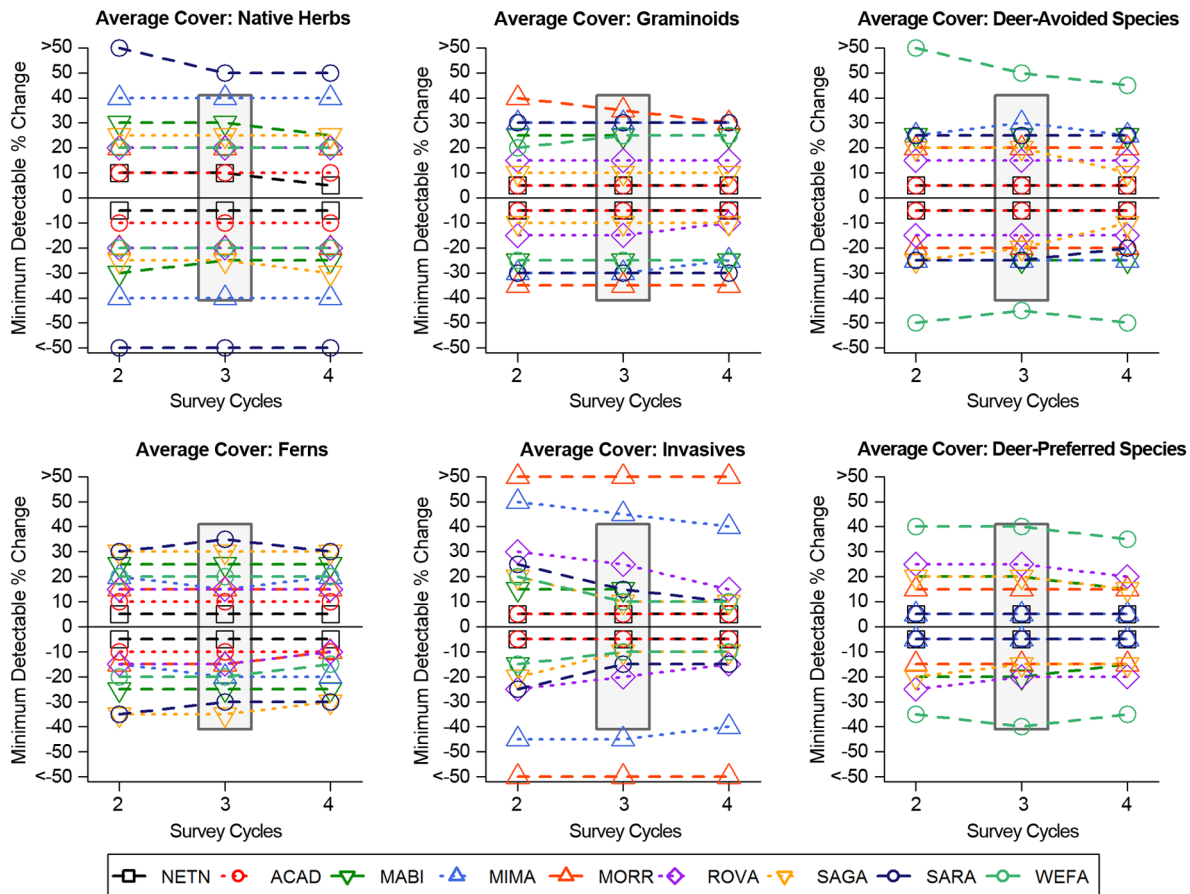


Fig. 3. Results of power simulations on average percent cover by plant group and for each park using $\alpha = 0.10$. For each park, results are displayed for the minimum detectable percent decline and increase in effect size across 2, 3 and 4 survey cycles. The gray box in the center of each plot represents the power target of the monitoring program (80% power to detect a 40% change after three cycles of data collection), and parks inside the box have met the target.

of data (i.e., both panels), WEFA's results improved by 10% or more and met the trend detection target (Fig. 6).

Quadrat frequency

Minimum detectable percent change ranged from 5% to 15% at 80% power and $\alpha = 0.10$ for all species groups in NETN and ACAD, suggesting more than adequate sample sizes to detect changes in quadrat percent frequency of important plant groups in ACAD and NETN (Fig. 7). In contrast, every plant group failed to meet the target of detecting a 40% change at 80% power and $\alpha = 0.10$ in at least one park, suggesting that average percent cover was a better approach for analyzing changes in under-

story composition.

Species richness

The native species richness metric performed well for NETN overall and all parks; we were able to detect small (5–10%) changes in native species diversity at both α levels, across two to four survey cycles, while maintaining 80% power (Figs. 8 and 9). In contrast, trend detection for invasive species richness varied across parks, and several parks did not meet the target of detecting a 40% change at $\alpha = 0.10$ (Fig. 8). Trend detection was greatest in MIMA, MORR, Roosevelt-Vanderbilt National Historic Sites (ROVA), and SARA, where invasive species were relatively widespread (Fig. 10). Trend detection

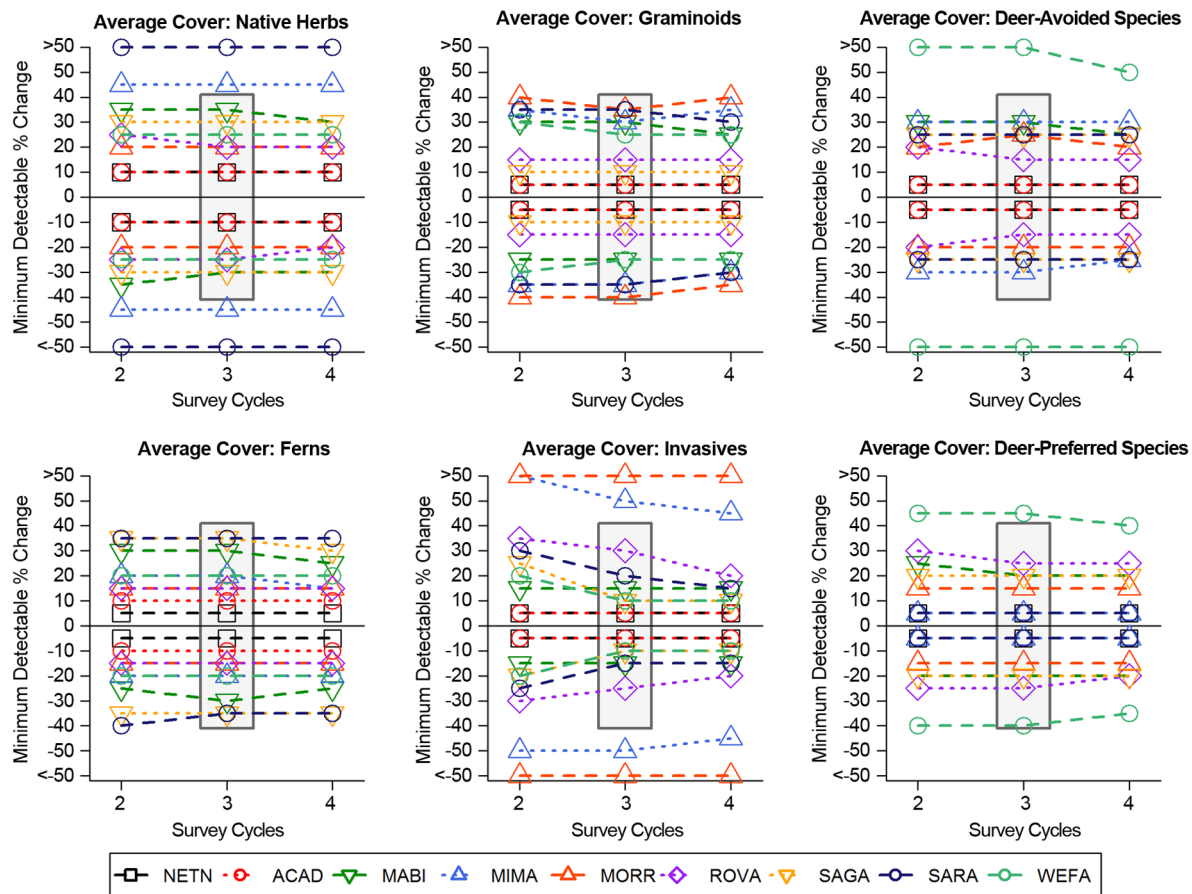


Fig. 4. Results of power simulations on average percent cover by plant group and for each park using alpha = 0.05. See Fig. 3 caption for additional details.

of invasive richness improved in most parks between two and four cycles. This pattern was particularly pronounced in MORR, where the minimum detectable percent increase improved by 15% between two and four cycles. Parks that were relatively uninvaded (i.e., ACAD, MABI and Saint-Gaudens National Historic Site [SAGA]), and which have a high proportion of 0's in the NETN dataset, lacked power to detect even a 50% change in invasive species richness (Table 5). A 50% increase in invasive richness in these parks, which averaged less than 1 invasive species per plot, was still a small value, and not of great concern. Poor trend detection with invasive richness in WEFA was more of a concern, as it was the result of small sample size, high metric variability and also potentially due to high between cycle variability, which was treated as sampling variation in the simulation

(Table 5).

CWD volume

Minimum detectable percent change of CWD volume at 80% power and alpha = 0.10 varied across network parks (Fig. 8). NETN and ACAD were able to detect changes of $\pm 10\%$, and MORR, ROVA, SAGA and SARA met the 40% detection target over three survey cycles. In contrast, MABI, MIMA and WEFA did not reach the 40% target. The low power in MIMA and WEFA can be explained by a combination of low sample size, high standard deviation relative to the mean value, and many plots with no CWD (Table 5, Fig. 10). The poor trend detection in MABI was primarily due to low sample size and high metric variability, including a higher between cycle variability than other parks with fewer than 30 simulated plots. Post-stratifying

Table 4. Table of summary statistics for average percent cover by plant group collected in the first survey cycle in each park. Mean is the average value, including 0. SD is the standard deviation of the metric by park for cycle 1. The % zeros column represents the proportion of plots in the dataset where the metric value is 0 for each park. SD_{BC} is the standard deviation of metric differences between cycles.

Park	Mean %	SD	% zeros	SD_{BC}
Native herbs				
ACAD	3.00	4.66	11.39	3.08
MABI	9.44	12.57	8.33	2.83
MIMA	8.17	9.95	0.00	6.14
MORR	9.15	13.92	0.00	3.61
ROVA	10.50	11.01	5.00	6.16
SAGA	3.83	4.98	0.00	2.69
SARA	28.58	15.09	0.00	12.64
WEFA	10.26	6.91	0.00	2.85
Graminoids				
ACAD	0.58	2.04	51.90	1.96
MABI	0.63	1.05	8.33	1.09
MIMA	0.90	1.54	30.00	3.15
MORR	10.71	13.22	0.00	11.02
ROVA	1.72	4.73	40.00	1.97
SAGA	0.14	0.33	70.00	0.76
SARA	7.43	9.05	6.67	12.54
WEFA	2.91	3.25	0.00	1.43
Ferns				
ACAD	1.92	3.78	54.43	3.07
MABI	7.53	10.70	16.67	3.67
MIMA	2.60	3.98	40.00	3.40
MORR	1.24	3.78	64.29	0.29
ROVA	3.47	9.54	40.00	5.25
SAGA	3.19	4.98	0.00	2.76
SARA	13.46	15.30	6.67	10.04
WEFA	6.09	5.86	20.00	2.79
Invasives				
ACAD	<0.01	<0.01	89.87	<0.01
MABI	0.04	0.13	91.67	0.13
MIMA	6.10	6.68	0.00	8.10
MORR	24.30	30.17	14.29	9.92
ROVA	4.70	7.54	20.00	5.06
SAGA	1.35	4.25	90.00	2.77
SARA	3.34	4.10	6.67	8.39
WEFA	0.09	0.12	40.00	0.19
Deer-avoided species				
ACAD	0.22	1.01	73.42	2.01
MABI	1.43	2.83	8.33	3.08
MIMA	0.22	0.35	50.00	1.47
MORR	4.81	7.75	7.14	6.08
ROVA	3.34	6.79	45.00	5.01
SAGA	1.38	3.55	30.00	3.77
SARA	2.57	3.34	0.00	12.43
WEFA	8.46	7.21	0.00	3.86
Deer-preferred species				
ACAD	0.04	0.19	91.14	0.15
MABI	0.15	0.32	66.67	0.35
MIMA	<0.01	<0.01	90.00	<0.01
MORR	0.08	0.18	64.29	0.22
ROVA	1.78	3.58	20.00	1.21
SAGA	0.04	0.08	50.00	0.10
SARA	<0.01	<0.01	86.67	<0.01
WEFA	0.19	0.42	60.00	0.42

MABI by forest type improved trend detection for plantations, but weakened trend detection for natural stands (Fig. 6). In this case, results for WEFA did not improve after running the simulation on 10 plots (i.e., both panels), and WEFA did not meet the trend detection target (Fig. 6).

Snag abundance

We were able to detect a 40% or smaller change in snag abundance at 80% power and $\alpha = 0.10$ over three survey cycles in only half of the parks in NETN (Fig. 8). MABI, MIMA, MORR and WEFA were the parks that missed this target (Fig. 8). At $\alpha = 0.05$, ROVA also missed the trend detection target (Fig. 9). MABI, MIMA, MORR, and WEFA had fewer than 30 simulated plots, had relatively high standard deviations in relation to the low mean value for this metric, and had a high proportion of plots with 0 values (Table 5). This resulted in low power both because the magnitude of a 40% change was relatively small, and because the high standard deviation meant that a large change had to occur to be significant. SAGA, in contrast, had a similar number of simulated plots, but a higher mean snag abundance and power for this park was much higher.

We reran the simulations with 10 plots of actual data for WEFA and post-stratified MABI and MORR by land use history. Post-stratifying MABI by natural stands and plantations resulted in slightly improved power, but only for natural stands (Fig. 6). In MORR, trend detection was actually worse for the mature group than for the full collection of plots, and did not improve for the successional group (Fig. 6). Despite doubling the number of plots of data used by the bootstrap, power also remained poor for WEFA (Fig. 6).

Tree regeneration

At 80% power and $\alpha = 0.10$, we failed to detect a 40% change over three survey cycles (12 years), for all three regeneration metrics (stocking index, seedling density, and sapling density) at MABI, MIMA, SAGA, and SARA (Fig. 8). In MORR, we failed to meet the 40% change target for seedling density. Parks with poor trend detection in the regeneration metrics often had low simulated sample size, a relatively high

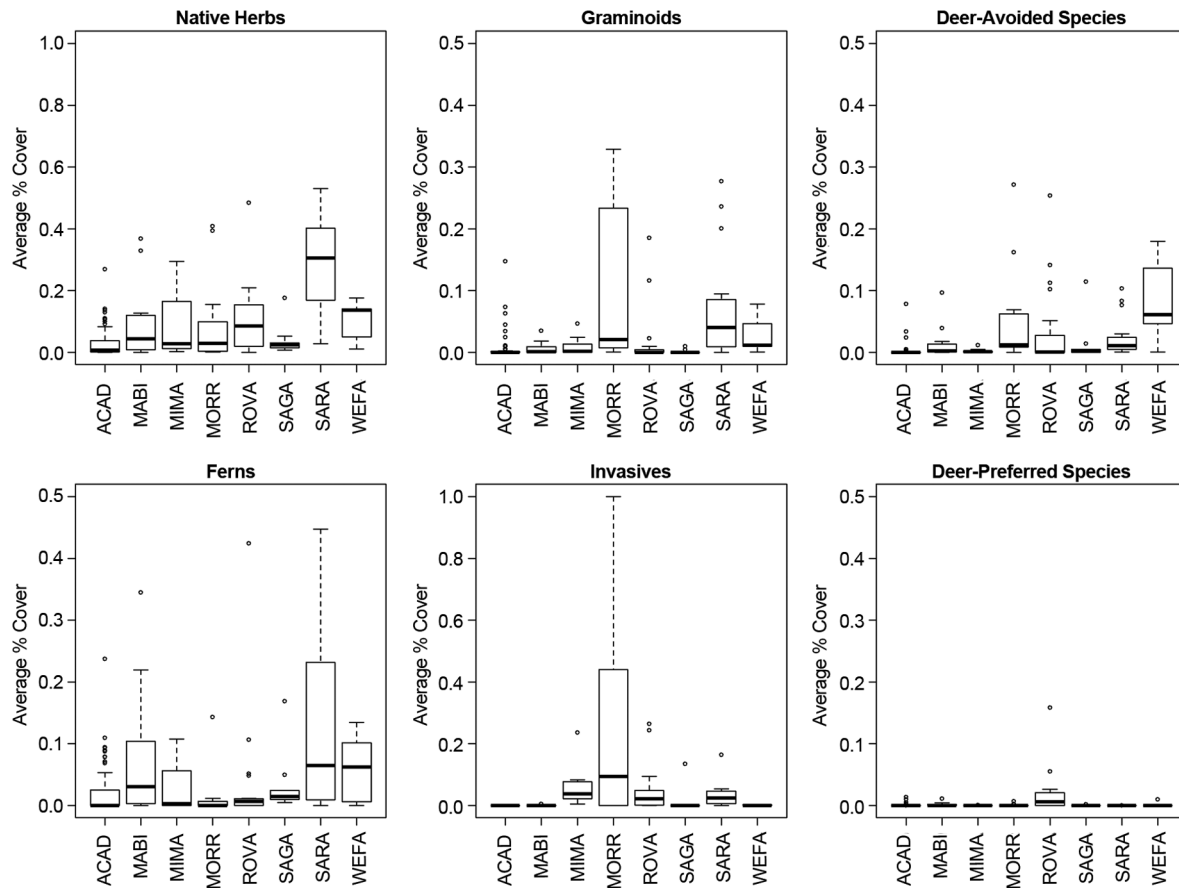


Fig. 5. Boxplots of average percent cover data from the first survey cycle by guild for each park.

standard deviation relative to the mean value, a high proportion of 0 values, and high between cycle variability (Table 5, Fig. 10). SAGA also had consistently high variability across the three regeneration metrics, and also had one of the smaller simulated sample sizes ($n = 20$). ACAD and ROVA met the 40% detection target for all three regeneration metrics at both alpha levels and across two, three, and four survey cycles. At $\alpha = 0.10$, WEFA also met the 40% detection target for all three regeneration metrics. WEFA was the exception to the poor trend detection in parks with low simulated sample sizes. For stocking index, WEFA had a relatively low standard deviation relative to the mean metric value and a fairly high mean value. For the other two regeneration metrics, WEFA had low between cycle variability when compared to the parks with low power.

Post-stratifying MABI, MORR and SARA by

land use history resulted in slight improvements in trend detection, but often for only one of the two groups per park (Fig. 11). In MABI, the minimum detectable percent change improved slightly in natural stands for both the stocking index and seedling density, but MABI still did not meet the trend detection target. Trend detection in MORR was worse for both groups, compared with the full collection of plots (Fig. 11). In SARA, the minimum detectable percent change in seedling density greatly improved for the forest group, but trend detection remained poor for the field group and the stocking index overall.

Live tree basal area

The live tree basal area metric performed very well, with all parks detecting changes at or below 25% while maintaining 80% power at both alpha levels, and across two, three, and four survey

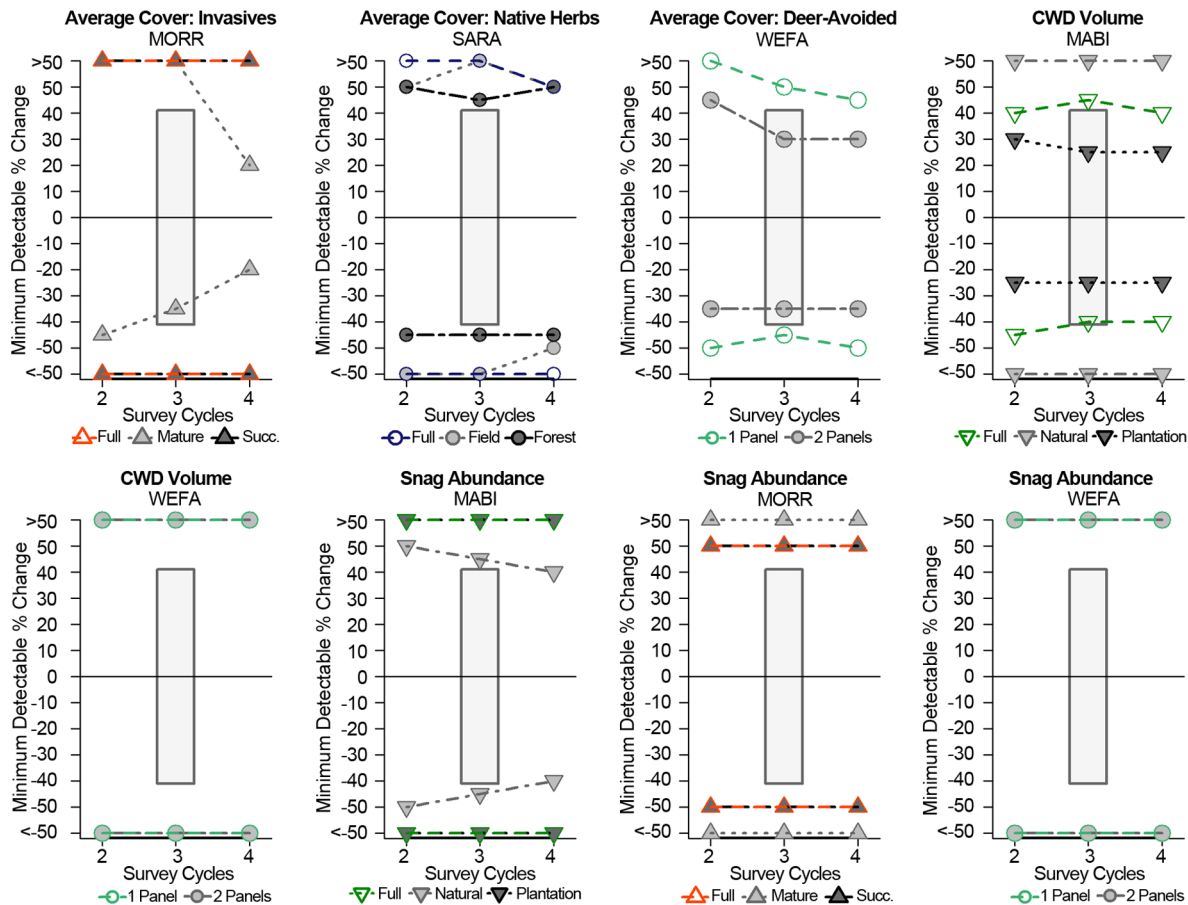


Fig. 6. Results of power simulations on metrics that were post-stratified within a park. Succ., successional. See Fig. 3 caption for additional details.

cycles (Figs. 8 and 9). This suggested that even small changes in live tree basal area will likely be detected in all NETN parks, making this a reliable metric for examining patterns in forest stand dynamics.

DISCUSSION

Power analyses indicated that, for most key forest metrics, our monitoring program met the target goal of detecting a 40% change in a metric over three survey cycles with 80% power and $\alpha = 0.10$. In general, we also were able to detect trends using the standard 0.05 alpha level. However, there were a number of cases where a park missed the 40% target with $\alpha = 0.05$, and met the target with $\alpha = 0.10$, and minimum detectable percent change also tended

to be better with the higher alpha level. We think that for many of the key forest metrics having a relatively high probability of falsely detecting a trend (i.e., 10% rather than 5%) is warranted given the risk of missing an important ecological change. Therefore, we believe the results from this power analysis justify using a default Type I error rate of 0.10 for future trend analyses.

The power to detect trends was very good for NETN overall and ACAD, with a minimum detectable change of $\pm 10\%$ or better for nearly every metric at both alpha levels. This was expected, as ACAD had the highest number of forest plots ($n = 176$) in the network, and the NETN-wide dataset was based on a sample size of 350 plots. ROVA had the second highest sample size ($n = 40$ plots) in NETN, and all but a few percent frequency metrics met the trend

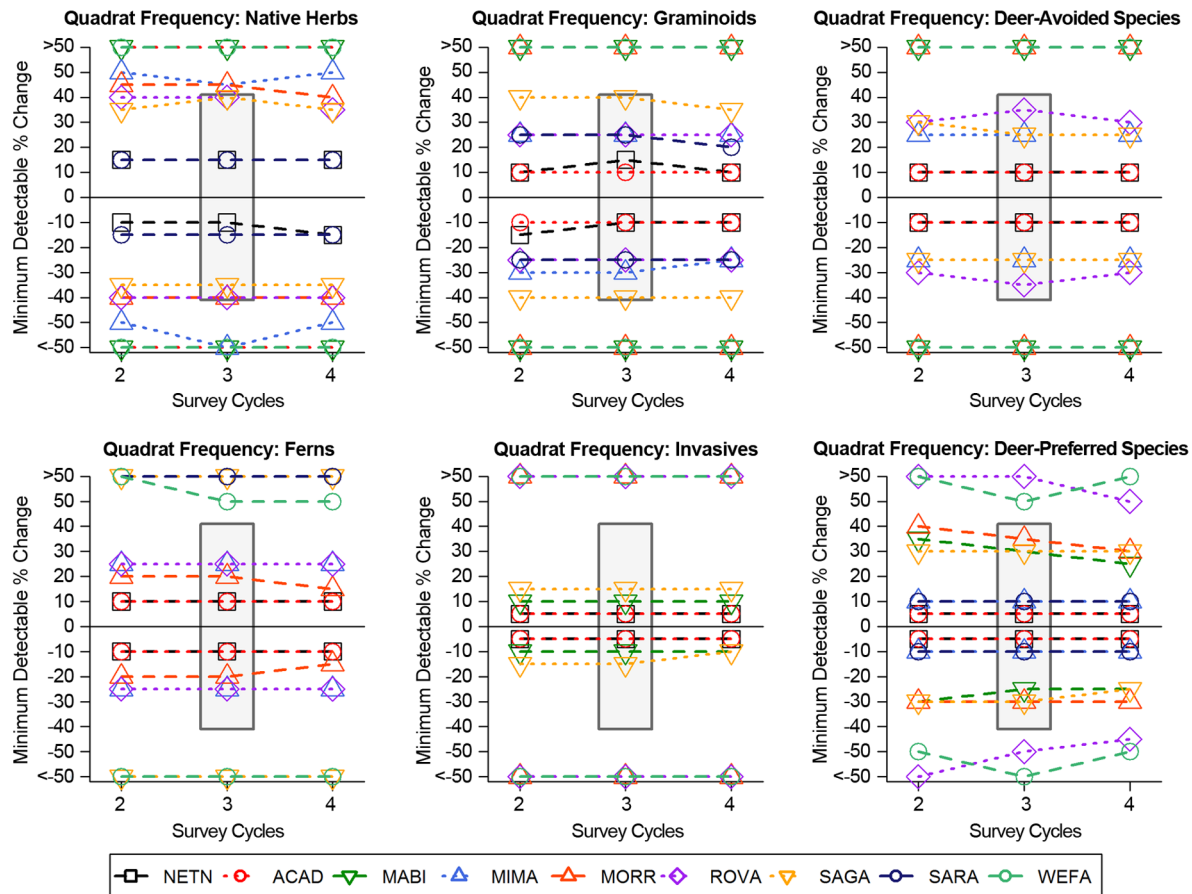


Fig. 7. Results of power simulations on quadrat frequency by plant group and for each park using $\alpha = 0.10$. See Fig. 3 caption for additional details.

detection target at $\alpha = 0.10$ in ROVA. The power for remaining parks (with 32 or fewer plots) varied, and metrics containing a high proportion of zeros and low park-specific mean values were the most common cause for poor change detection (e.g., invasive species richness in ACAD, MABI and SAGA; CWD Volume in MIMA and WEFA; and snag abundance in MIMA, MORR and WEFA). Poor trend detection in this case is not a big concern, because a 40% change in a metric with low values is generally not biologically meaningful.

In contrast, several cases of poor change detection were the result of high metric variability within the initial plot data or between sampling cycles (e.g., average percent cover of invasive species in MIMA and MORR, average percent cover of native herbs in SARA, and CWD volume in MABI). This is of greater concern,

because only large changes in a metric can be detected, even if metric values are high to begin with. These problems with poor detection could be remediated by increasing the sample size. We also recommend carefully reviewing the sampling methods to determine whether improvements could be made to whether sampling variability. For example, spatial and temporal CWD volume variability could be reduced by using permanent markers to delineate the transects more clearly or by considering a different method of tallying CWD (e.g., the point relascope method; Brissette et al. 2003).

Metric variability can also be reduced by stratifying the data into reasonable groups based on factors like habitat type or land use and that reduce the within-group variance as much as possible. We tested whether post-stratifying improved power for MABI, MORR and SARA,

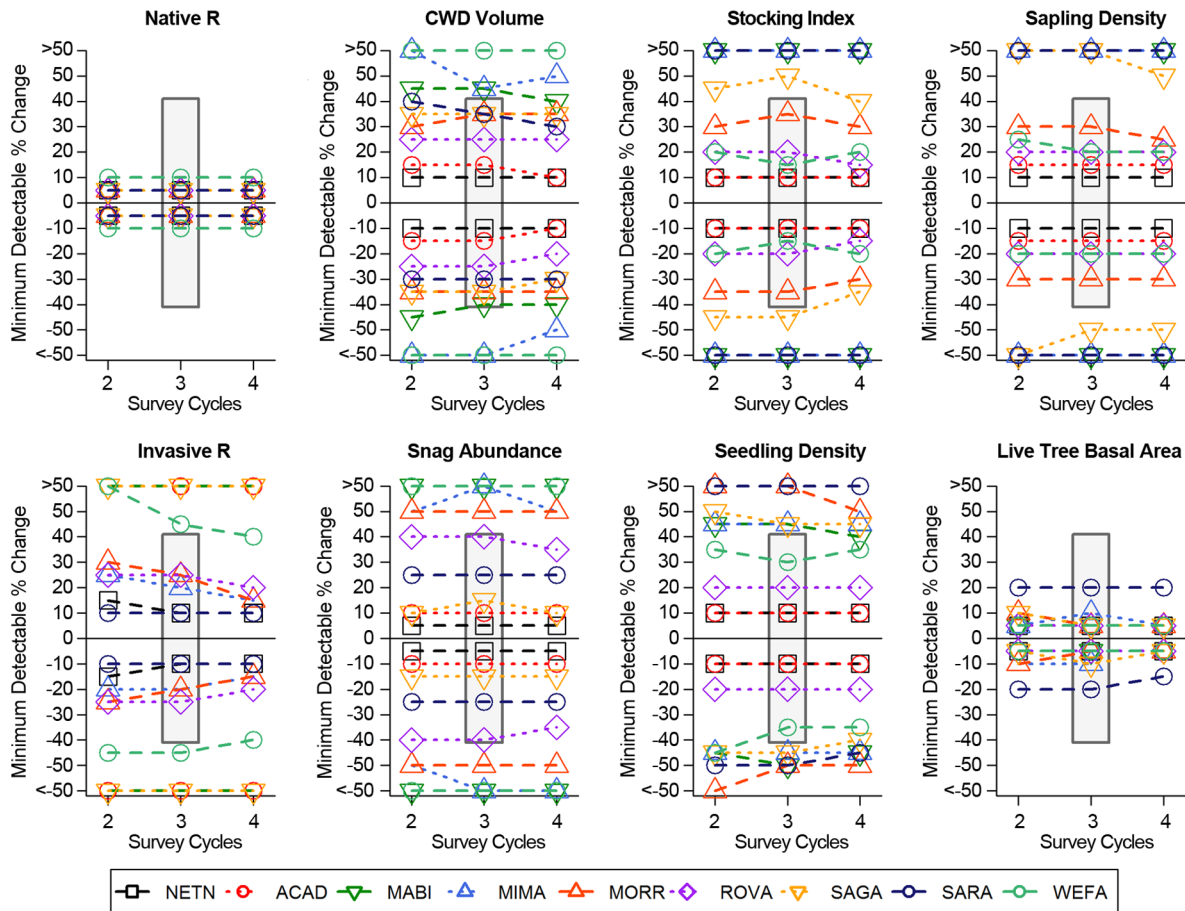


Fig. 8. Results of power simulations on key forest metrics for each park using $\alpha = 0.10$. R stands for richness in the Native R and Invasive R plots. CWD, coarse woody debris. See Fig. 3 caption for additional details.

and results varied. In most cases, post-stratifying improved trend detection for one of the two groups in a park, but only resulted in a group meeting the trend detection target in two cases (CWD volume in MABI and seedling density in SARA). There were also several instances where post-stratifying resulted in worse trend detection for a group than with the full collection of plots. These results indicated that post-stratification, at least for the groups we tested, may not actually improve trend detection. This is contrary to Johnson et al. (2008), which suggested that post-stratifying plots based on land use history, forest type, or management unit may improve power to detect change (Johnson et al. 2008). We believe that post-stratification will only be beneficial when the reduction in within-group variability gained by stratification outweighs the

detrimental effects of a reduced sample size.

Trend detection in WEFA did not vary greatly between simulations with a dataset containing five plots of data, and a dataset with 10 plots. In both datasets, WEFA did not meet the trend detection target for CWD volume and snag abundance. Trend detection did improve for average cover of deer-avoided species. Despite the lack of marked improvement in power, using as large a sample as possible for the bootstrap is recommended since a larger sample provides a better approximation of the true population distribution. If the data used are not a representative sample of the population, the bootstrap simulation will be biased and will produce inaccurate power estimates. Although we used only five plots in the initial simulations for WEFA, we recommend using at least 10 plots of

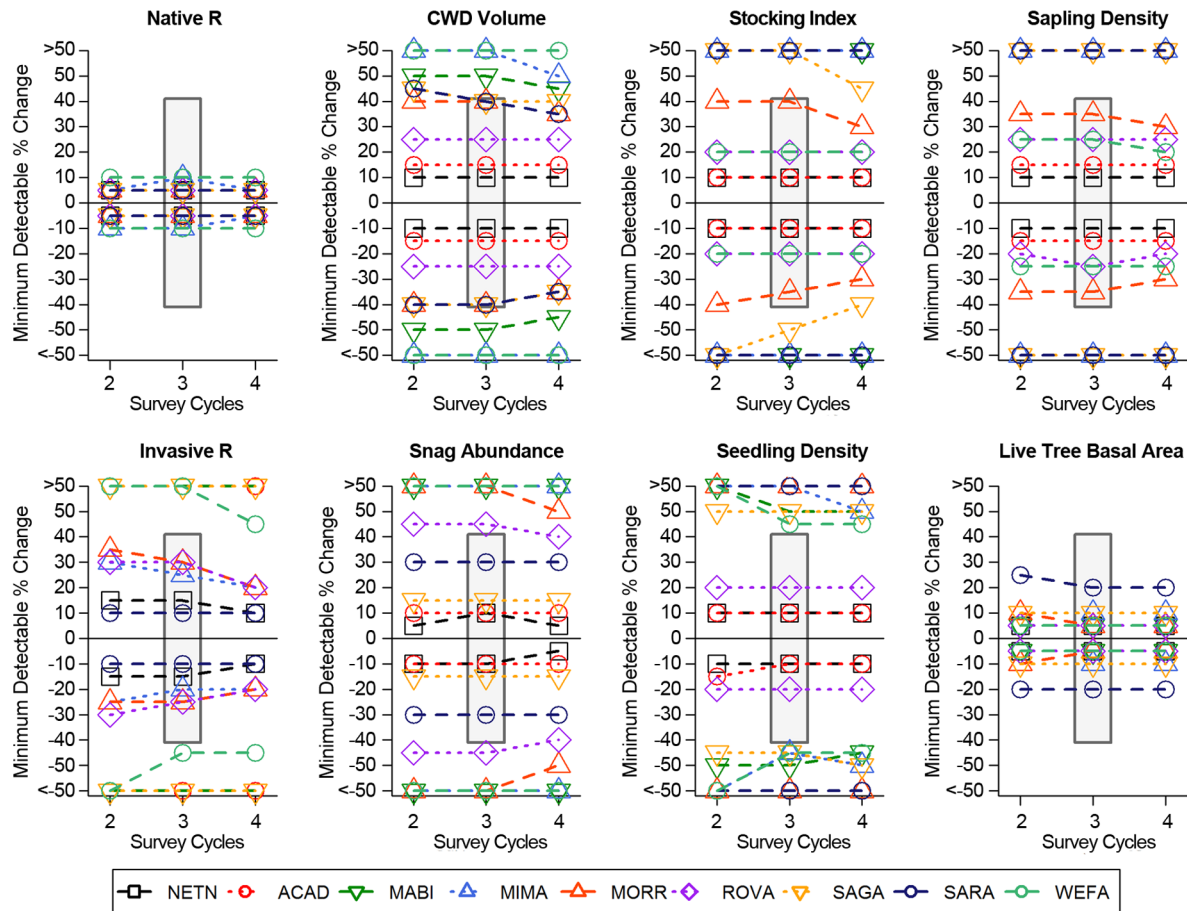


Fig. 9. Results of power simulations on key forest metrics for each park using $\alpha = 0.05$. R stands for richness in the Native R and Invasive R plots. CWD, coarse woody debris. See Fig. 3 caption for additional details.

data per site, and perhaps more in the case of highly variable metrics.

Trend detection results were consistently better for average percent cover of plant guilds than quadrat percent frequency. These results were contrary to Johnson et al. (2008), where quadrat frequency was favored over quadrat percent cover for assessing trends in understory composition. The consistently poor performance of quadrat frequency across all plant groups was likely due to the small number of quadrats that we sampled in each plot ($n = 8$), versus the 30 quadrats sampled by the hybrid plot design proposed by Johnson et al. Our power results for quadrat frequency were also potentially affected by our simulation approach, because the simulation was unable to take the discrete nature of the plot frequency data into account. Frequency data

were converted to proportions, and the simulation generated continuous proportion data even though true values were limited to 0, 0.125, 0.25, and so on. The extent to which our simulation approach biased the quadrat frequency results is unclear and warrants additional investigation. Nevertheless, the results from this analysis suggested that our sampling effort was adequate to detect important changes in understory composition using percent cover. This is valuable information, as the amount of time and level of expertise required to identify species in each quadrat is quite high, and adding more quadrats to a plot would come at a considerable cost to NETN.

While most parks met the 40% trend detection target at 80% power and $\alpha = 0.10$ in every plant group for average cover, some plant groups

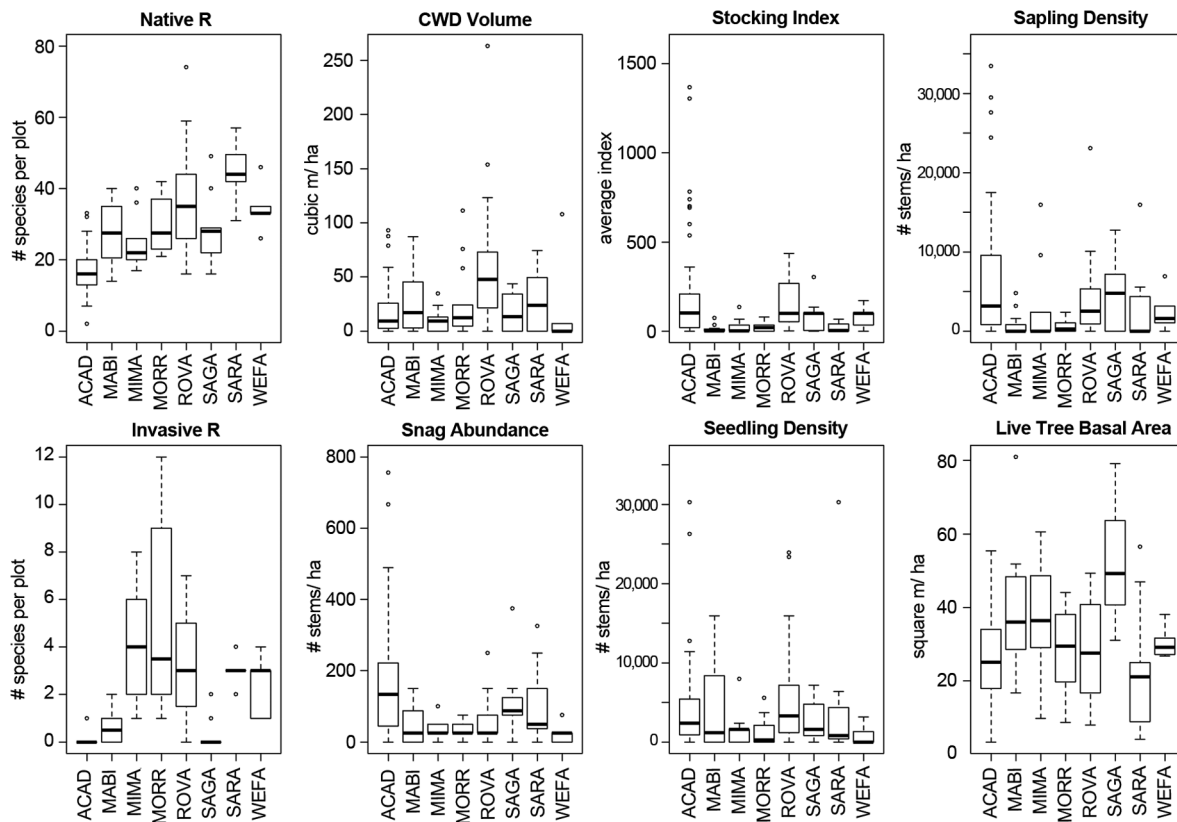


Fig. 10. Boxplots of metric data collected in the first survey cycle for each park.

had consistently better trend detection than others. The fern and graminoid groups were consistently strong metrics for trend detection, and performed better than the native herbaceous and invasive species group. Ferns and graminoids often respond favorably to moderate deer browse impacts and/or earthworm invasions while herbaceous species show a decline (De La Cretaz and Kelty 2002, Wiegmann and Waller 2006). Therefore, changes in the understory may be identified by a significant increase in fern or graminoid cover before a decrease in native herbaceous cover is detected. Invasive species also tend to respond favorably to deer browse and earthworm invasion, but results indicated that changes in fern and graminoid cover may be detected before change in invasive species cover. The deer browse indicator groups also performed well, and will be useful metrics to assess impacts or recovery from deer overabundance.

The power to detect change in tree regeneration varied considerably between parks and

metrics, and only NETN overall, ACAD, ROVA and WEFA met the 40% trend detection target over three cycles with $\alpha = 0.10$ and 80% power. With the two largest sample sizes (176 for ACAD, and 40 for ROVA) and relatively high regeneration rates, the results in ACAD and ROVA were to be expected. In contrast, WEFA had the smallest sample size (10 plots), and it was a bit surprising that the regeneration metrics performed so well in this park. A fairly high mean metric value and low standard deviation relative to the mean (for the stocking index) and relatively low spatial and between cycle variability (for the other metrics; Table 5) were possibly the reasons for good trend detection in WEFA. MORR had extremely low rates of regeneration due to high deer browse pressure (Shaw and Patterson 2006), and for the two metrics where the park met the trend detection target the park had very low between cycle variability. Trend detection was weaker across all three regeneration metrics in the remaining parks with low

Table 5. Table of summary statistics for each metric by park and collected in the first survey cycle. Mean is the average value, including 0. SD is the standard deviation of the metric by park for cycle 1. The % zeros column represents the proportion percent of plots in the dataset where the metric value is 0 for each park. SD_{BC} is the standard deviation of metric differences between cycles of the actual data.

Park	Mean	SD	% zeros	SD _{BC}
Native richness				
ACAD	16.66	5.73	0.00	2.93
MABI	27.42	8.63	0.00	4.37
MIMA	24.80	7.51	0.00	5.69
MORR	29.57	7.82	0.00	4.63
ROVA	37.05	14.53	0.00	5.90
SAGA	28.60	9.61	0.00	2.11
SARA	44.80	6.88	0.00	8.86
WEFA	34.60	7.23	0.00	9.04
Invasive richness				
ACAD	0.06	0.25	93.67	0.28
MABI	0.58	0.67	50.00	0.75
MIMA	4.20	2.30	0.00	1.17
MORR	5.36	3.99	0.00	1.54
ROVA	3.05	1.90	5.00	1.63
SAGA	0.30	0.67	80.00	0.71
SARA	3.00	0.65	0.00	0.59
WEFA	2.40	1.34	0.00	1.52
Live tree basal area (m ² /ha)				
ACAD	26.06	12.27	0.00	5.04
MABI	39.68	16.47	0.00	6.35
MIMA	37.92	15.36	0.00	5.94
MORR	28.70	10.71	0.00	3.18
ROVA	28.36	14.18	0.00	3.36
SAGA	52.36	15.92	0.00	10.16
SARA	22.54	16.57	0.00	4.71
WEFA	30.34	4.61	0.00	1.56
Stocking index				
ACAD	177.18	258.83	10.12	145.92
MABI	10.86	22.07	41.67	38.63
MIMA	24.17	44.33	20.00	55.80
MORR	25.81	29.47	28.57	16.58
ROVA	154.07	132.71	0.00	94.81
SAGA	92.22	94.32	11.11	47.49
SARA	22.29	25.78	33.33	35.60
WEFA	81.47	66.92	20.00	45.65
Seedling density (stems/ha)				
ACAD	4291.09	5181.34	17.72	3559.31
MABI	4378.98	5746.34	33.33	5018.28
MIMA	1751.59	2337.87	30.00	2792.03
MORR	1194.27	1787.14	42.86	1088.10
ROVA	6064.23	7183.57	5.00	6883.22
SAGA	2742.39	2761.23	22.22	1558.83
SARA	3980.89	7630.74	26.67	3988.05
WEFA	902.34	1399.30	60.00	581.45
Sapling density (stems/ha)				
ACAD	6016.69	7538.70	24.68	5810.81
MABI	796.18	1592.36	75.00	1391.16
MIMA	2786.62	5506.50	70.00	4016.71
MORR	587.66	730.00	50.00	479.67
ROVA	4325.90	5230.54	5.00	3875.17
SAGA	4954.00	4587.15	33.33	3019.47
SARA	2547.77	4257.88	53.33	4414.59
WEFA	2547.77	2690.83	20.00	2333.34
CWD volume (m ³ /ha)				
ACAD	17.27	21.25	24.05	14.07
MABI	26.19	28.94	16.67	17.91
MIMA	10.77	11.20	30.00	5.59

Table 5. Continued.

Park	Mean	SD	% zeros	SD _{BC}
MORR	25.86	32.78	7.14	15.52
ROVA	60.37	61.21	5.00	36.07
SAGA	17.58	16.75	30.00	11.78
SARA	27.65	28.94	33.33	20.18
WEFA	22.98	47.51	60.00	5.93
Snag abundance (stems/ha)				
ACAD	175.53	169.45	13.92	76.21
MABI	45.83	54.18	41.67	27.09
MIMA	42.50	33.44	10.00	44.41
MORR	33.93	27.05	21.43	30.57
ROVA	53.75	61.38	20.00	30.86
SAGA	117.50	99.34	10.00	48.88
SARA	103.33	100.39	6.67	66.46
WEFA	25.00	30.62	40.00	41.83

sample sizes and moderate regeneration abundance and variability (MABI, MIMA, SAGA and SARA).

Minimum detectable percent change tended to be 5–10% better for the stocking index than seedling and sapling density metrics. The stocking index was developed and used by the U.S. Forest Service to assess regeneration in PA forests (McWilliams et al. 2002), and has been used by multiple NPS Inventory and Monitoring Networks to assess condition of regeneration (Perles et al. 2010, Comiskey and Wakamiya 2012, Miller et al. 2012), and has been proposed as a metric for parks undergoing deer management to assess regeneration response (National Park Service 2008, National Park Service 2009, National Park Service 2011). Results from this analysis suggested that the stocking index is a decent metric for assessing change in regeneration abundance in parks with low rates of regeneration (e.g., parks with high deer browse pressure). For parks with moderate regeneration abundance and variability, current metrics may not be sufficient for detecting trends in regeneration. Post-stratifying improved power in a few cases, but we may need to explore other metrics to track regeneration in these parks.

The results from this analysis are informative beyond NETN, as there are a number of federal and state agencies that are implementing or developing forest monitoring programs in the eastern U.S. This includes five other NPS I&M networks that are implementing similar protocols in over 50 park units from Minnesota to Virginia. The I&M forest monitoring protocols are adapted from the U.S. Forest Service's Forest Inventory

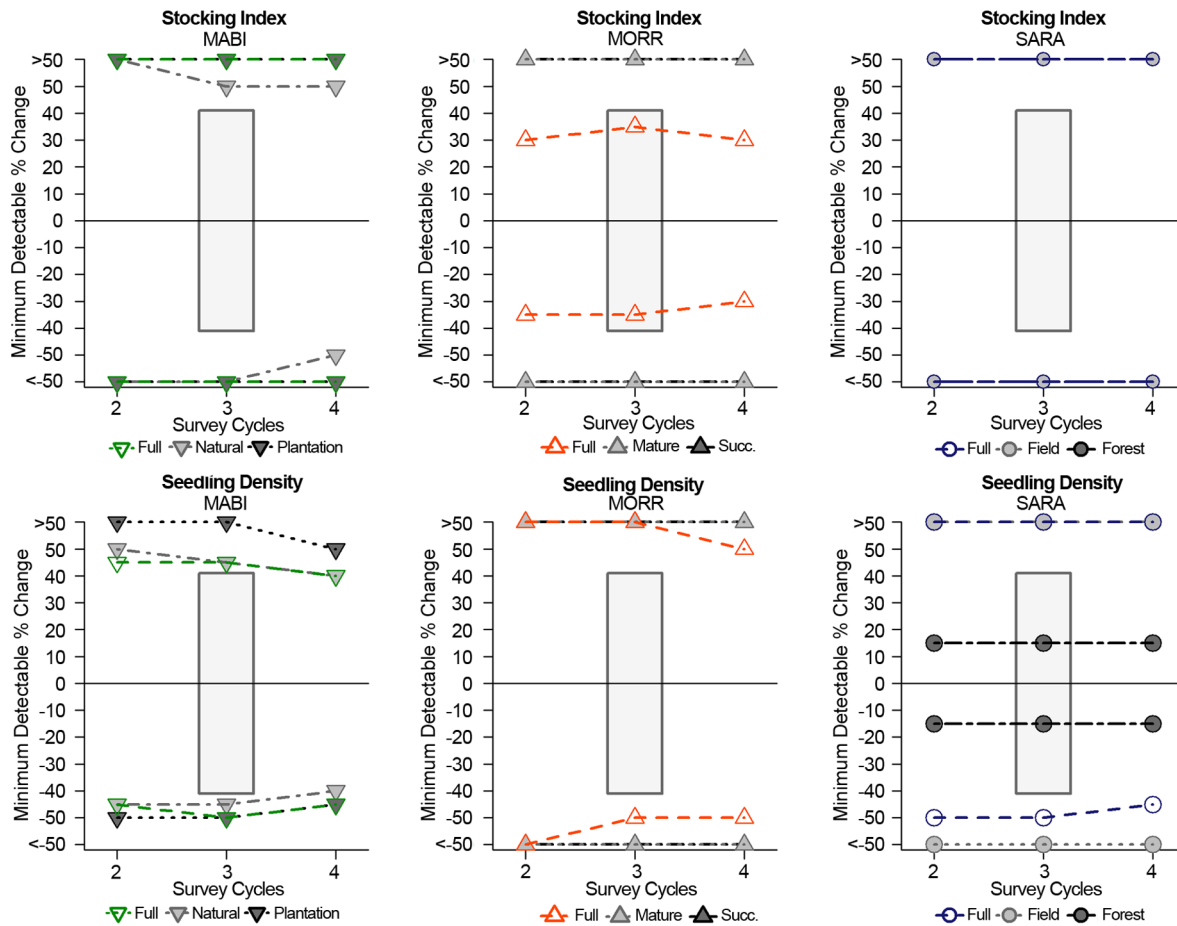


Fig. 11. Results of power simulations on regeneration metrics that were post-stratified within a park. Succ., successional. See Fig. 3 caption for additional details.

and Analysis Program, which covers the entire U.S., and includes timber surveys dating back to the 1930s. Multiple state agencies in our region also have forest monitoring programs, including the Maine Natural Areas Program in Maine (Cutko 2005), and a forest monitoring cooperative in Vermont (Vermont Monitoring Cooperative 2009).

The power analysis simulation we developed represents one of the few tools available that assesses the power to detect change over time using mixed effects models, and that appropriately models datasets that are organized as random plots within multiple sites, with repeated measures at fixed intervals and unequal variance among sites. The analyses we presented here applied effect size as a percent change across all cycles (e.g., a 40% change between

three cycles was a 13.3% change in each cycle). However we included the option in our simulation tool to treat the change as “per cycle” (e.g., always a 40% change). A 40% change per cycle will likely result in much higher power and marked improvement in power over time, but testing this is beyond the scope of this study. The simulation assumes that the differences in metric measurements between plot visits and within each site are normally distributed, but alternative distributions can easily be used (e.g., the glmer function in lme4) and the simulation otherwise makes no assumptions beyond those required by the mixed effects model. Commonly used power analysis tools, such as MONITOR and TRENDS, are based on least-squares regression or analysis of variance, and therefore assume equal variances and independence (Gerrodette 1993, Hatch

2003, Gray and Burlew 2007, Gibbs and Ene 2010). While these programs are valuable and easy to use, the results have limited application when datasets violate the underlying assumptions. By incorporating fixed (e.g., site and time) and random (plot) factors, the mixed effects model appropriately handles repeated measurements (Wang and Goonewardene 2004, Zuur et al. 2009). Our approach also uses the `varIdent` function in the mixed effects model, which allows each group (e.g., park) to have its own variance structure. Finally, we designed the simulation in R to require a minimal amount of user input to run (e.g., dataset file name, metric name, number of cycles, number of simulations to run, and transformation type), and to be applicable to a range of data types (e.g., proportional and count data), for any number of sampling areas (e.g., parks) and sampling units (e.g., plots), and a range of repeated, fixed-interval surveys.

CONCLUSION

The results from our power analysis indicate that for most key forest metrics, our monitoring program meets our target goal of detecting a 40% change in a metric over three survey cycles with 80% power and $\alpha = 0.10$. These results suggest that our sampling methods are sound, and there are only a few areas that require improvement. After examining the causes of poor trend detection, we identified which metrics had low power due to a high proportion of zeros and low park-specific mean values, and which metrics had low power due to high metric variability. High metric variability is of most concern to us, because only large changes in a metric can be detected, even if metric values are high to begin with. We will attempt to improve power for metrics with high variability by augmenting our sample design with other existing data sources in our parks, such as incorporating similar data collected in more than 150 silvicultural inventory plots in MABI. We are also carefully reviewing the sampling methods, such as for CWD, to determine whether improvements could be made to lower sampling variability. We will evaluate potential modifications to our monitoring protocol or new metric calculations using this simulation tool to ensure

any changes will improve power over original methods. We also plan to run similar power analyses for many of our other long-term monitoring protocols (e.g., breeding landbirds and freshwater wetlands) to ensure that our methods are effective for detecting long-term changes over time, and we encourage other monitoring programs to do the same.

ACKNOWLEDGMENTS

We are grateful for the efforts of numerous individuals who have contributed to NETN's forest monitoring program. We are indebted to Geri Tierney for developing the forest monitoring protocol, and for her continued support. Adam Kozlowski's data management support is greatly appreciated and crucial to this program. We are grateful to Jesse Wheeler, our forest crew leader, and all of technicians who have sampled our plots over the last eight years. Many thanks also to our park resource managers and staff for their feedback and continued support. We thank Jim Comiskey, John Karish, and Stephanie Perles for their helpful suggestions on early versions of this manuscript, Tom Philippi for assistance with the R code, and John Paul Schmit for reviewing the final manuscript. Finally, we thank two anonymous reviewers for their valuable feedback.

LITERATURE CITED

- Augustine, D. J., and D. deCalesta. 2003. Defining deer overabundance and threats to forest communities: From individual plants to landscape structure. *Ecoscience* 10:472–486.
- Augustine, D. J., and P. A. Jordan. 1998. Predictors of white-tailed deer grazing intensity in fragmented deciduous forests. *Journal of Wildlife Management* 62(3):1076–1085.
- Bates, D., M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, and H. Singmann. 2014. *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-5. <http://cran.r-project.org/web/packages/lme4/>
- Brissette, J. C., M. J. Ducey, and J. H. Gove. 2003. A field test of point relascope sampling of down coarse woody material in managed stands in the Acadian forest. *Journal of the Torrey Botanical Society* 130(2):79–88.
- Brus, D. J., and J. J. de Gruijter. 2012. A hybrid design-based and model-based sampling approach to estimate the temporal trend of spatial means. *Geoderma* 173-174:241–248.
- Cohen, J. E. 1988. *Statistical power analysis for the behavioral sciences*. Second edition. Lawrence-Erlbaum, Hillsdale, New Jersey, USA.

- Comiskey, J. A., J. P. Schmit, and G. L. Tierney. 2009. Mid-Atlantic Network forest vegetation monitoring protocol. Natural Resource Report NPS/MIDN/NRR—2009/119. National Park Service, Fort Collins, Colorado, USA.
- Comiskey, J. A., and S. M. Wakamiya. 2012. Forest vegetation monitoring: Mid-Atlantic Network 2011 summary report. Natural Resource Data Series NPS/MIDN/NRDS—2012/324. National Park Service, Fort Collins, Colorado, USA.
- Cutko, A. 2005. Ecological Reserves monitoring: project update, June 2005. Maine Natural Areas Program, Augusta, Maine, USA.
- Damgaard, C. 2014. Estimating mean plant cover from different types of cover data: a coherent statistical framework. *Ecosphere* 5(2):20.
- De La Cretaz, A. L., and M. J. Kelty. 2002. Development of tree regeneration in fern-dominated forest understories after reduction of deer browsing. *Restoration Ecology* 10 (2):416–426.
- Di Stefano, J. 2001. Power analysis and sustainable forest management. *Forest Ecology and Management* 154:141–153.
- Di Stefano, J. 2003. How much power is enough? Against the development of an arbitrary convention of statistical power calculations. *Functional Ecology* 17:707–709.
- Foster, J. R. 2001. Statistical power in forest monitoring. *Forest Ecology and Management* 151:211–222.
- Gerrodette, T. 1993. Program TRENDS: user's guide. Southwest Fisheries Science Center, La Jolla, California, USA. http://swfsc.noaa.gov/uploadedFiles/Divisions/PRD/Programs/ETP_Cetacean_Assessment/trends_guide.pdf
- Gibbs, J. P., and E. Ene. 2010. Program Monitor: estimating the statistical power of ecological monitoring programs. Version 11.0.0. <http://www.esf.edu/efb/gibbs/monitor>
- Gray, B. R., and M. M. Burlew. 2007. Estimating trend precision and power to detect trends across grouped count data. *Ecology* 88:2364–2372.
- Hatch, S. A. 2003. Statistical power for detecting trends with applications to seabird monitoring. *Biological Conservation* 111:317–329.
- Horsley, S. B., S. L. Stout, and D. S. deCalesta. 2003. White-tailed deer impact on the vegetation dynamics of a northern hardwood forest. *Ecological Applications* 13:98–118.
- Johnson, S. E., E. L. Mudrak, E. A. Beever, S. Sanders, and D. M. Waller. 2008. Comparing power among three sampling methods for monitoring forest vegetation. *Canadian Journal of Forest Research* 38:143–156.
- Legg, C. J., and L. Nagy. 2006. Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management* 78:194–199.
- Manly, B. F. J. 2007. Randomization, bootstrap and Monte Carlo methods in biology. Third edition. CRC Press, Boca Raton, Florida, USA.
- Marshall, P. L., G. Davis, and V. M. LeMay. 2000. Using line intersect sampling for coarse woody debris. Forest Research Technical Report TR-003. British Columbia Ministry of Forests, Vancouver Forest Region, Nanaimo, British Columbia, Canada.
- McWilliams, W. H., et al. 2002. Measuring tree seedlings and associated understory vegetation in Pennsylvania's forests. Pages 21–26 in R. E. McRoberts, G. A. Reams, P. C. Van Deusen, W. H. McWilliams, and C. J. Cieszewski, editors. Proceedings of the Fourth Annual Forest Inventory and Analysis Symposium. General Technical Report NC-252. U.S. Forest Service, North Central Research Station, St. Paul, Minnesota, USA.
- Miller, K. M., B. R. Mitchell, and J. S. Wheeler. 2012. Forest health monitoring in the Northeast Temperate Network: 2011 summary report. Natural Resource Technical Report NPS/NETN/NRTR—2012/604. National Park Service, Fort Collins, Colorado, USA.
- Morrison, L. W. 2007. Assessing the reliability of ecological monitoring data: power analysis and alternative approaches. *Natural Areas Journal* 27:83–91.
- National Park Service. 2008. Final white-tailed deer management plan/environmental impact statement. Catocin Mountain Park, Thurmont, Maryland, USA.
- National Park Service. 2009. Final white-tailed deer management plan/environmental impact statement. Valley Forge National Historical Park, King of Prussia, Pennsylvania, USA.
- National Park Service. 2011. Final white-tailed deer management plan/environmental impact statement. Rock Creek Park, Washington, D.C., USA.
- NatureServe. 2012. NatureServe Explorer: An online encyclopedia of life. Version 7.1. NatureServe, Arlington, Virginia. <http://www.natureserve.org/explorer>
- Perles, S. J., K. K. Callahan, and M. R. Marshall. 2010. Condition of vegetation communities in Allegheny Portage Railroad National Historic Site and Johnstown Flood National Memorial: Eastern Rivers and Mountains Network summary report 2007–2009. Natural Resource Data Series NPS/ERMN/NRDS—2010/034. National Park Service, Fort Collins, Colorado, USA.
- Pinheiro, J., D. Bates, S. DebRoy, and D. Sarkar. and the R Development Core Team. 2012. nlme: Linear and nonlinear mixed effects models. R package version 3.1-103. <http://cran.r-project.org/web/packages/nlme/>
- Raudenbush, S. W. 1993. Hierarchical linear models and experimental design. Pages 459–496 in L. K. Edwards, editor. Applied analysis of variance in

- behavioral science. Marcel Dekker, New York City, New York, USA.
- R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Schmit, J. P., G. Sanders, M. Lehman, and T. Paradis. 2009. National Capital Region Network long-term forest monitoring protocol. Version 2.0. Natural Resource Report NPS/NCRN/NRR—2009/113. National Park Service, Fort Collins, Colorado, USA.
- Shaw, S., and W. A. Patterson III. 2006. Strategies for managing the forest at Morristown National Historical Park. Technical Report NPS/NER/NRTR—2006/040. National Park Service, Boston, Massachusetts, USA.
- Stevens, D. L., Jr., and A. R. Olsen. 2004. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* 99:262–278.
- Sullivan, L. M., K. A. Dukes, and E. Losina. 1999. Tutorial in biostatistics and introduction to hierarchical linear modelling. *Statistics in Medicine* 18:855–888.
- Tierney, G. L., D. Faber-Langendoen, B. R. Mitchell, W. G. Shriver, and J. P. Gibbs. 2009. Monitoring and evaluating the ecological integrity of forest ecosystems. *Frontiers in Ecology and the Environment* 7.
- Tierney, G. L., B. R. Mitchell, K. M. Miller, J. A. Comiskey, A. J. Kozlowski, and D. Faber-Langendoen. 2013. Northeast Temperate Network long-term forest monitoring protocol: 2013 revision. Natural Resource Report NPS/NETN/NRR—2013/639. National Park Service, Fort Collins, Colorado, USA.
- Urquhart, N. S., S. G. Paulsen, and D. P. Larsen. 1998. Monitoring for policy-relevant regional trends over time. *Ecological Applications* 8:246–257.
- Van Wagner, C. E. 1982. Practical aspects of the line intersect method. Petawawa National Forestry Institute Information Report PI-X-12. Canadian Forestry Service, Chalk River, Ontario, Canada.
- Vermont Monitoring Cooperative. 2009. Vermont's changing forests: Key findings on the health of forested ecosystems from the Vermont Monitoring Cooperative. University of Vermont, South Burlington, Vermont, USA. <http://www.uvm.edu/vmc/documents/synthesisReport.pdf>
- Wang, Z., and L. A. Goonewardene. 2004. The use of MIXED models in the analysis of animal experiments with repeated measures. *Canadian Journal of Animal Science* 84:1–11.
- Warton, D. I., and F. K. Hui. 2011. The arcsine is asinine: the analysis of proportions in ecology. *Ecology* 92:3–10.
- Wiegmann, S. M., and D. M. Waller. 2006. Fifty years of change in northern upland forest understories: Identity and traits of “winners” and “loser” plant species. *Biological Conservation* 129:109–123.
- Zar, J. H. 1999. *Biostatistical analysis*. Fourth edition. Prentice Hall, Upper Saddle River, New Jersey, USA.
- Zuur, A., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Springer, New York, New York, USA.

SUPPLEMENTAL MATERIAL

SUPPLEMENT

R scripts and an example dataset for conducting the power analysis simulations ([Ecological Archives C005-009-S1](#)).