

## OPTIMIZING THE ALLOCATION OF COUNT DAYS IN A MIGRATION MONITORING PROGRAM

LEN THOMAS, GEOFFREY R. GEUPEL, NADAV NUR, AND GRANT BALLARD

*Abstract.* Many migration monitoring stations cannot operate on every day of the migration period. In this paper, we used migration count data from two stations (Point Reyes Bird Observatory fall migration and Long Point Bird Observatory spring migration) to examine the relationship between the proportion of count days (frequency of sampling) and the statistical power to detect long-term population trends. We found that power to detect trends at a single station declined at an accelerating rate as the frequency of sampling decreased. Stations that operate on one or two days per week are unlikely to detect changes in abundance for most species that would be well monitored at higher sampling frequencies. The effect of missing counts can be mitigated to some extent by the choice of sampling design (method of allocating count days over the migration period). We compared a number of candidate designs and found that systematic sampling was the most accurate, although stratified random sampling may be preferred in situations where little is known about the pattern of migration. Designs that clump count days together, such as sampling only at weekends, should be avoided because adjacent count days tend to duplicate the same information.

*Key Words:* avian migration monitoring, population trends, power analysis, sampling frequency, survey design.

One of the principal objectives of songbird migration monitoring is to determine whether the abundance of birds arriving at a monitoring station has changed over time. To achieve this, birds are surveyed at the station on as many days as possible during the migration season. A number of survey techniques may be employed (including mist netting), depending on the characteristics of the location. Regardless of the survey method used, the daily counts are converted into annual indices of abundance, and population trends are estimated from the annual indices by regression. The use of count data to monitor migration in this way has been reviewed elsewhere (Dunn and Hussell 1995, Dunn in press).

At most of the larger migration monitoring stations in North America (e.g., Long Point, Point Reyes, and Manomet bird observatories), counts take place on essentially every day of the migration season. However, many smaller stations are constrained by funding or by the availability of volunteers and cannot operate every day. Gaps in the daily counts result, which introduce additional variability into the annual index estimates, and in turn reduce the ability of the station to detect population trends. This additional variability is called “sampling variance” and is a function of the “frequency of sampling” (the proportion of days on which counts take place) and the “sampling design” (method of allocating count days over the season). In this paper, we address two questions: (1) To what degree can gaps in the count data reduce our ability to detect long-term trends?

and (2) Can the effect of gaps be minimized by the choice of an appropriate sampling design?

To answer these questions, we used data from two stations where monitoring occurs continually throughout the migration season: Point Reyes and Long Point bird observatories. By analyzing the pattern of counts in these “complete” data sets, we could estimate the sampling variance that would arise from different frequencies of sampling and sampling designs. The two observatories differ in the environment of the stations, the methods of data collection, and the species seen. In addition, we used fall data from Point Reyes and spring data from Long Point. We reasoned that, by using very different datasets, any similarities in the results between stations would be of more general applicability to other migration monitoring stations in North America. This does mean, however, that we cannot interpret any differences in results between the data sets as being due to differences between stations, since they could also be due to differences between season.

The ability of a monitoring program to detect a given trend can be measured using the concept of “statistical power.” Statistical power is the probability of getting a significant result in a statistical hypothesis test, given that there is an effect (i.e., trend) of specified magnitude (Cohen 1988, Gerrodette 1987, Nemas 1991, Steidl and Thomas 2001). In the context of avian population monitoring, the Monitoring Working Group of Partners in Flight have proposed that a successful monitoring

program is one that has a 90% chance of detecting a 50% decline in a species' population over a 25 year period (Butcher et al. 1993:199). Thus, the importance of gaps in the count data can be assessed by quantifying the impact of sampling frequency on the statistical power to detect a population change of this magnitude. To do this, we estimated statistical power to detect a 50% population change over 25 year for sampling frequencies ranging from one to seven days per week. We assumed that count days are selected at random, that the statistical test used to detect trends is a linear regression of annual indices against time, where annual indices are the mean of the log-transformed daily counts, and that the test was statistically significant when  $P \leq 0.05$ .

At monitoring stations that do not operate every day, there is often some flexibility in the way that the count days can be allocated through the season. A number of familiar sampling designs are discussed in standard textbooks on sampling (e.g., Cochran 1977), such as simple random, stratified random, and systematic. These designs vary in the ease with which they can be implemented, and in the sampling variance of the resulting annual indices. To quantify the differences in sampling variance that could be expected for migration monitoring, we computed the "design effect" of a number of candidate sampling designs, over a range of sampling frequencies. The design effect is the ratio of (1) the sampling variance obtained from the candidate sampling design divided by (2) the sampling variance obtained from simple random sampling at the same sampling frequency (reference in Cochran 1977:85). Design effects of  $<1$  indicate an improvement in precision over random sampling, and the design with the lowest design effect should be preferred by those designing monitoring programs, all other factors being equal.

This paper is aimed at those designing a migration monitoring program at a single station. We do not consider the trade-off between the frequency of sampling at multiple stations versus the number of stations that can be sampled. The solution to this problem will depend in part upon the variability between stations, which is not well known for migration monitoring. A treatment of the topic in the context of extensive surveys such as the Breeding Bird Survey is given by Link et al. (1994).

## METHODS

### DATA USED

Data from Point Reyes were collected at the Palomarin Field Station in coastal California using constant-effort

mist netting (see DeSante and Geupel 1987 for details of the field methods). We used fall migration data collected between August 18 and November 26 (101 days) in the years 1980 to 1992 (13 years). The field protocol calls for 20 fixed nets to be placed for 6 h each day during the migration period, making 120 net-h in total. However, inclement weather and other eventualities sometimes prevented the protocol from being followed: the mean percent of days when no nets were set was 4.7/year, and the mean net-hours for the remaining days was 110.7. Our analyses required a complete data set, so we substituted for the missing value on days when no nets were run the mean of the counts from the previous two days and the next two days. We standardized all daily counts to the total new birds banded per 120 net-h.

Long Point is a peninsula on Lake Erie, Ontario. There, a combination of standardized transect counts, unstandardized netting, and casual observations were used to produce a daily estimated total of each bird species present at the monitoring station (see Dunn et al. *this volume a*, for a full description of the method of data collection). In this paper, we used spring migration data from the station at the tip of the Point, collected between 1961 and 1993. Of these data, we excluded years with any missing counts, leaving 16 years: 1963, 1967, 1971–1972, 1975, 1978, 1980, 1982–1985, 1987, 1989–1990, 1992–1993. The seasonal timing of data collection varied between years, with a mean starting date of April 10 and ending date of June 13, giving a mean of 65 daily counts per year.

Having derived daily counts from each station, we treated both datasets identically. At each station, we chose migration periods separately for each species using a procedure similar to that of Hussell et al. (1992), as follows. First, we excluded counts in the first or last two weeks of the data collection periods that were separated by more than four days from any other count in any other year. Then we selected start and end dates so as to encompass the middle 98% of days on which the species was recorded. To simplify the comparison of sample designs (see below), we further truncated the data such that all migration periods at the same station began on the same day of the week, and were an integer number of weeks in length.

After the selection of migration periods, species with a mean daily count over all years of less than 1.0 were excluded from the analysis. This criterion was necessary because the methods we used to estimate trends are known to be unreliable for species with low abundances (see DISCUSSION). The counts were then log-transformed (after the adding 0.5 to all counts, to prevent taking the log of a zero count). Annual indices were calculated as the mean of the logged counts.

To better understand our results, we performed a number of descriptive analyses of the log-transformed count data for each species at each station. We studied the frequency distribution of the counts, and the distribution of counts through the migration period. We also plotted "correlograms," which show the correlation between counts taken in the same season against the number of days the counts are spaced apart. Correlation between counts taken

in the same season is known as “autocorrelation.” The degree of autocorrelation has important implications for the optimal choice of sampling design, as is discussed below.

All of the analyses described here were performed using SAS for OS/2 version 6.10 (SAS Institute 1993).

IMPORTANCE OF SAMPLING FREQUENCY

For each species, we calculated the power to detect a 50% change in annual indices over 25 year under the null hypothesis of no change as

$$\text{power} = 1 - \Phi\left(t_{1-\alpha/2,df,\delta}\right) + \Phi\left(-t_{1-\alpha/2,df,\delta}\right)$$

where  $t_{x,df,\delta}$  denotes the  $x$ th quantile of the noncentral t-distribution, given  $df$  degrees of freedom and the noncentrality parameter  $\delta$ , and  $\Phi(t_{x,df,\delta})$  is the cumulative distribution function of the appropriate noncentral t-distribution, evaluated at  $x$  (Nemac 1991). The degrees of freedom were 23 throughout (number of years over which we hope to detect the 50% change - 2), and  $\alpha$  was set at 0.05. The noncentrality parameter was calculated as:

$$\delta = \frac{(\beta_1 - 0)}{\sqrt{(S_{reg}^2 + S_{samp}^2)/SS_{yr}}}$$

where  $\beta_1$  is the slope of the log-linear regression line under the alternate hypothesis (50% change in population size over 25 years is a slope of 0.277 on the log scale),  $SS_{yr}$  is the sum of squares of the year variable (which, with 25 years of continuous data, is 1,300),  $S_{reg}^2$  is the variance of the annual indices that is unexplained by the regression line when there are no gaps in the daily count data, and  $S_{samp}^2$  is the additional variance due to missing counts (i.e., the sampling variance). We estimated  $S_{reg}^2$  for each species at each station as the mean square error from the linear regression of annual indices against year. We calculated statistical power at seven levels of  $S_{reg}^2$ , assuming sampling frequencies of one to seven days per week and a simple

random sampling design. At a sampling frequency of seven days per week (i.e., no gaps in the count data),  $S_{samp}^2 = 0$ . At the other six levels of sample frequency, was calculated for each species as the mean sampling variance over all years of count data. The calculation of the sampling variance for each year, assuming a simple random design, is outlined in the next section.

COMPARISON OF SAMPLING DESIGNS

At each station, we calculated the sampling variance for the sampling designs for each species and year and over a range of sampling frequencies. We then calculated the design effect of each sampling design for each species at each sampling frequency as (1) the mean sampling variance over years for the sampling design divided by (2) the mean sampling variance for random sampling at the same sampling frequency. The sampling designs are explained below, and are shown diagrammatically in Figure 1. The sampling frequencies used and the method of comparing sampling designs are described at the end of this section.

In “simple random sampling,” our baseline design, the number of count days is fixed, but their location in the season is chosen at random. The variance of the annual index in a single year was calculated as

$$V_{rand} = \frac{S^2}{n} \frac{(N-n)}{N}$$

(Cochran 1977, formula 2.8), where  $S^2$  is the variance of the log-transformed daily counts,  $N$  is the number of days in the migration period and  $n$  is the number of days on which counts took place. Here,  $n = \text{frequency of sampling per week} \times N/7$ .

In “stratified random sampling,” the season is divided into strata and the sample consists of count days randomly selected from within each stratum. For simplicity, we chose to use equal-size strata of one week. We used two different methods to allocate sample days within strata: “proportional

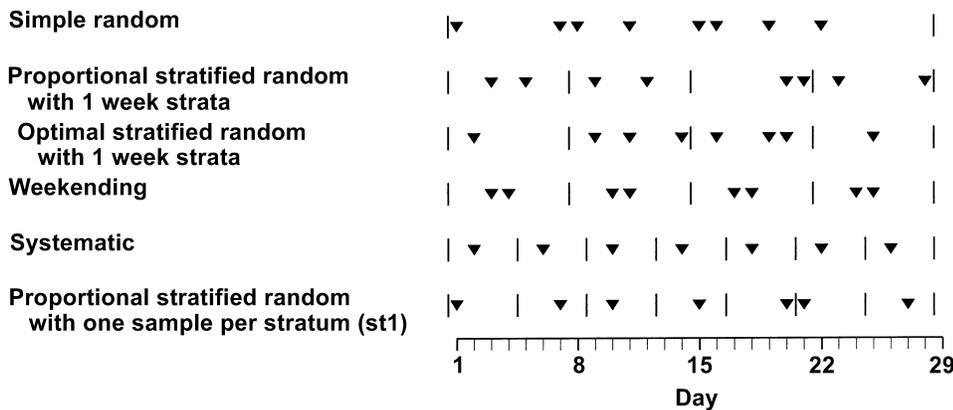


FIGURE 1. A schematic representation of possible sample allocations under six different sampling designs. Vertical lines represent the sampling strata for each design.

allocation" and "optimal allocation." In proportional allocation, the proportion of count days is the same in all strata. Because strata were all the same size, the number of sample days was the same in each stratum. In optimal allocation, strata in which the daily counts are more variable are given a greater frequency of sampling than strata with relatively homogeneous counts. With equal-size strata, the overall sample variance is minimized when the count days are allocated as follows:

$$n_h = n \frac{S_h}{\sum_{h=1}^H S_h}$$

(from Cochran 1977, formula 5.26), where  $n$  is the total number of count days,  $n_h$  is the number of count days in stratum  $h$ , and  $S_h$  is the standard deviation of the log-transformed daily counts in stratum  $h$  ( $h = 1 \dots H$ ). Because the timing of migration varied between species and years, no one allocation of count days could be optimal for all species in all years. We thus constructed a compromise allocation scheme at each monitoring station such that the number of count days in each stratum was proportional to the mean of the optimal allocation for that stratum over all years and species. When the compromise allocation of count days was non-integer, we used the nearest integer value, unless that value was zero, in which case we used one, or was greater than seven, in which case we used seven. For both stratified random designs, the sampling variance was calculated as

$$V_{strat} = \frac{1}{H^2} \sum_{h=1}^H \frac{S_h^2}{n_h} (N_h - n_h)$$

(Cochran 1977, formula 5.6), where  $H$  is the number of strata (number of weeks),  $N_h$  is the stratum size,  $n_h$  is the number of count days, and  $S_h^2$  is the variance of the log-transformed daily counts in stratum  $h$ .

In the "weekending" design, counting is concentrated on two consecutive days each week. No formulae are available to determine the sampling variance of the annual indices under such a design, so we calculated the variance empirically. At a sample frequency of two days per week, there were seven possible sample allocations for each year, each subsample consisting of day  $i$ ,  $i+1$ ,  $i+7$ ,  $i+8$ ,  $i+14$ ,  $i+15$ ,... where  $i = 1 \dots 7$ . We calculated the annual index from each subsample, and used the variance of these seven indices as an estimate of the sampling variance. At higher sample frequencies, additional counts were randomly located during each week, and we used 50 subsamples at each level of  $i$  to calculate the sampling variance.

In "systematic sampling," the sample consists of a fixed number of count days spaced at regular intervals throughout the season. The sampling variance is

$$V_{sys} = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2$$

(Cochran 1977, formula 8.1), where  $S^2$  and  $N$  are defined as with random sampling,  $k$  is the interval of the count days (e.g., when sampling on alternate days,  $k = 2$ ),  $n$  is the

number of count days and  $S_{wsy}^2$  is the mean within-sample variance. Here,

$$n = N / k$$

and

$$S_{wsy}^2 = \frac{1}{k(n-1)} \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{i\bullet})^2$$

where  $y_{ij}$  is the  $j$ th log-transformed count in sample  $i$ , and  $\bar{y}_{i\bullet}$  is the mean of the log-transformed counts in sample  $i$ . To simplify the calculations, we removed the last few counts when the migration period was not an integer multiple of the count interval ( $k$ ). In these cases, we used the same data to calculate the sampling variance for random sampling when determining the design effect.

Systematic sampling is often compared with "proportional stratified random sampling with one sample per stratum" (*stl*) because the two designs differ only in the allocation of samples within strata (Fig. 1). In *stl*, the stratum size is equal to  $k$  and one sample is drawn from each of the  $n$  strata. We calculated the variance for the *stl* design using the same data as for systematic sampling and the formula given above for stratified random sampling.

We compared the sampling designs in two groups: (1) proportional stratified, optimal stratified, and weekending, and (2) systematic and *stl*. In the first group, we calculated design effects at all integer sampling frequencies from two to six days per week (two days per week being the minimum for the weekending design). In the second group, we calculated design effects at all integer count intervals from  $k = 2$  (sampling every other day) to 7 (once per week), which gave sampling frequencies of 3.5, 2.33, 1.75, 1.4, 1.17, and 1 day per week. For both groups, we performed the comparison using a repeated-measures analysis of variance, with species as the subject and sampling design and sampling frequency as the within-subject factors. Because design effect is a ratio measurement, all comparisons took place on the log scale, and were then back-transformed to the arithmetic scale for the presentation of results.

## RESULTS

A total of 38 species was observed at the Point Reyes station in fall, and 81 at Long Point in spring. Of these, six species occurred with a mean count of 1.0 bird/120 net-h or greater during the fall migration period at Point Reyes and 46 species had daily counts of 1.0 or greater during spring migration at Long Point. For these more common species, there was considerable variation in the distribution of the log-transformed daily counts (Fig. 2). Most of the species with low mean counts had very skewed distributions, with a majority of zero counts and a few days when many birds were seen. Species with greater mean counts tended to have less skewed distributions. Some differences were also evident between stations (Fig. 3). Species at Long Point tended

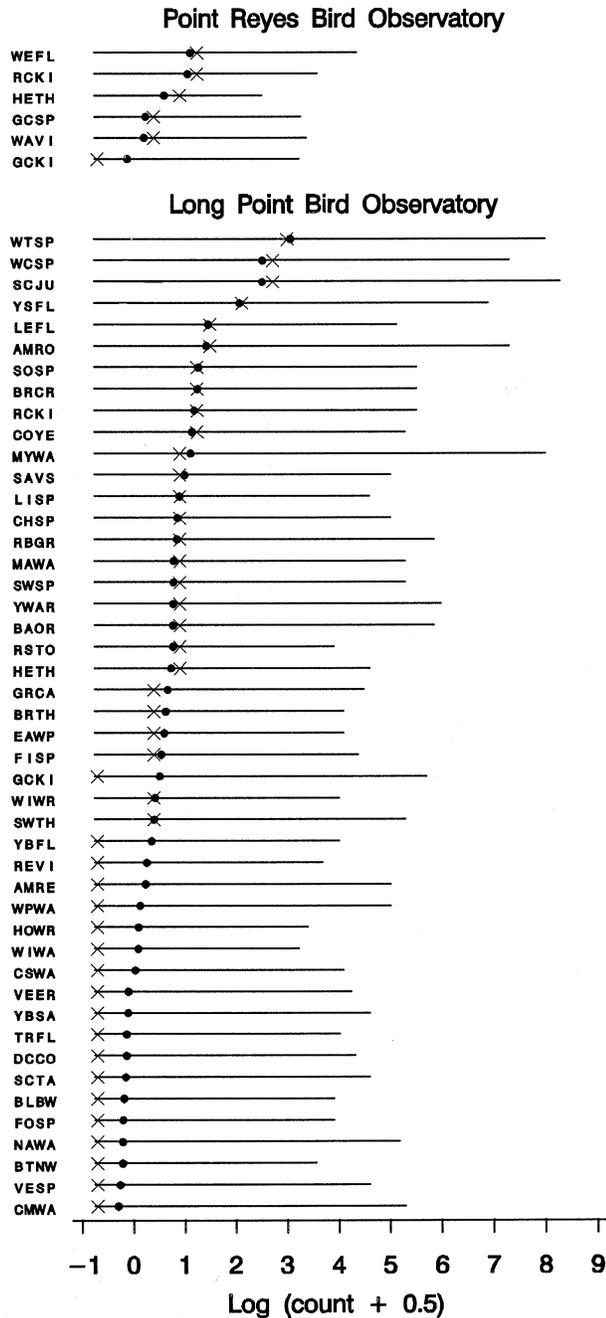


FIGURE 2. The mean (black dot), median (X), and range (horizontal line) of log-transformed daily counts for 52 species at two migration stations (see key to species codes in Appendix A).

to show greater variability in log-transformed counts (larger standard deviation), and have more skewed distributions than at Point Reyes.

The timing of migration varied between species, but the pattern of counts was similar for most

species at both stations, showing a rise from low counts to a period of high counts and then a drop-off (Fig. 4). There were, however, distinct differences in the pattern of autocorrelation between the stations (Fig. 5). Species at Point Reyes tended to show an

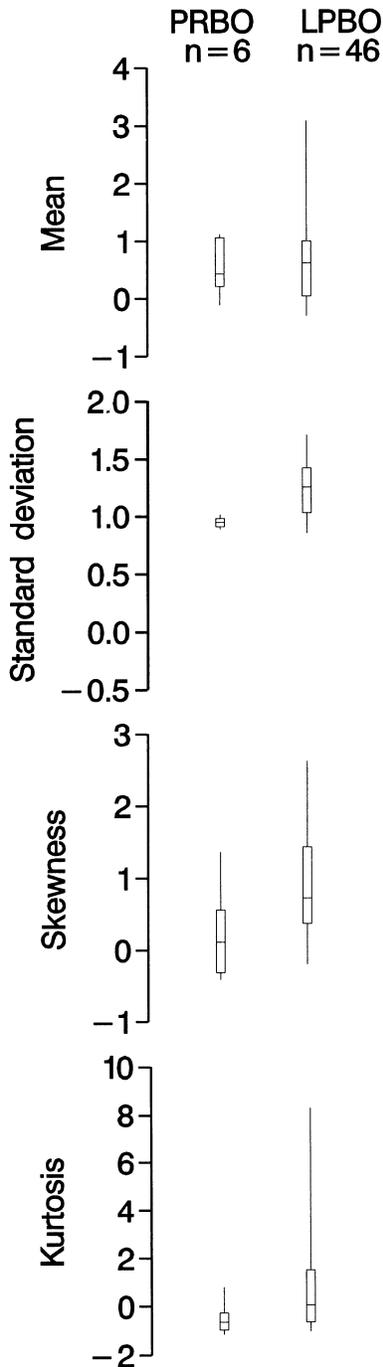


FIGURE 3. Boxplots showing the distribution of four statistics that describe the log-transformed daily counts for six species at Point Reyes Birds Observatory (PRBO) and 46 species at Long Point Bird Observatory (LPBO). The vertical lines show the median, the ends of the box show the inter-quartile range, and the horizontal lines show the range.

approximately linear decline in correlation between counts with increasing separation between count days. A few species at Long Point showed the same pattern, but the majority showed a sharp drop in autocorrelation so that the median correlation between counts spaced three days apart was close to zero. As noted previously, these differences could be due either to differences between stations or between seasons.

#### IMPORTANCE OF SAMPLING FREQUENCY

Statistical power declined at an accelerating rate with decreasing frequency of sampling for all species (Fig. 6). At Point Reyes, median power declined from 0.78 at a sample frequency of seven days per week to 0.59 at a sample frequency of one day per week. At Long Point, median power declined from 0.60 to 0.28 over the same range of sample frequencies.

The overall rate of decline differed between species, with some showing little effect of decreasing sampling frequency (e.g., Golden-crowned Kinglet [scientific names and key to four-letter codes in the Appendix] at Point Reyes; Slate-colored Junco at Long Point) and others being strongly affected (e.g., Hermit Thrush at Point Reyes; Chestnut-sided Warbler at Long Point). Species little affected were generally those with low power even at high sampling frequencies, while those strongly affected tended to have good power at high sampling frequencies (Spearman's rank correlation between rate of decline in power and power at a sampling frequency of seven days per week:  $r_s = 0.89$ ,  $N = 6$ ,  $P = 0.02$  at Point Reyes Bird Observatory;  $r_s = 0.26$ ,  $N = 46$ ,  $P < 0.001$  at Long Point Bird Observatory).

#### COMPARISON OF SAMPLING DESIGNS

There were consistent differences in design effect (and therefore in sampling variance) among sampling designs. These differences were statistically significant for both groups of sampling designs at Long Point, but not at Point Reyes (Table 1), although the sample size was low at the latter station (six species). Averaged over all sampling frequencies, optimal stratified sampling had the lowest design effect in group 1, and systematic was the lowest in group 2 at Long Point (Table 2). Although direct comparisons between all sampling designs are not strictly appropriate, since the two groups were measured at different sampling frequencies, it is clear that weekend sampling performed very poorly (i.e., had the highest design effect), and the systematic design had the best (i.e., had the lowest design effect; Table 2). Average design effects were

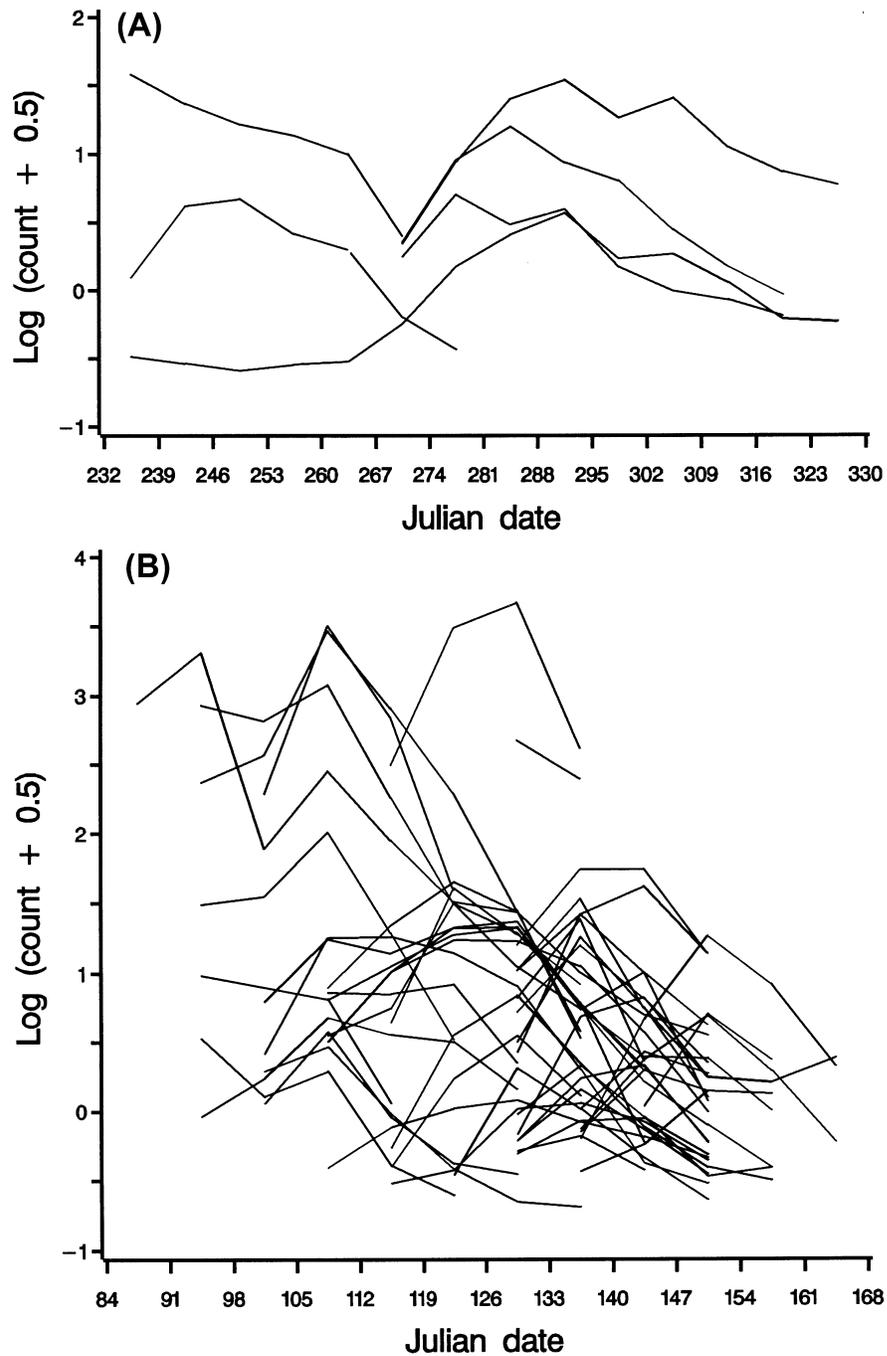


FIGURE 4. Change in daily counts over time at two migration monitoring stations. Lines show weekly mean (pooled across years) of the log-transformed counts during the migration period for each of six species at Point Reyes Bird Observatory (fall migration) and 46 species at Long Point Bird Observatory (spring migration).

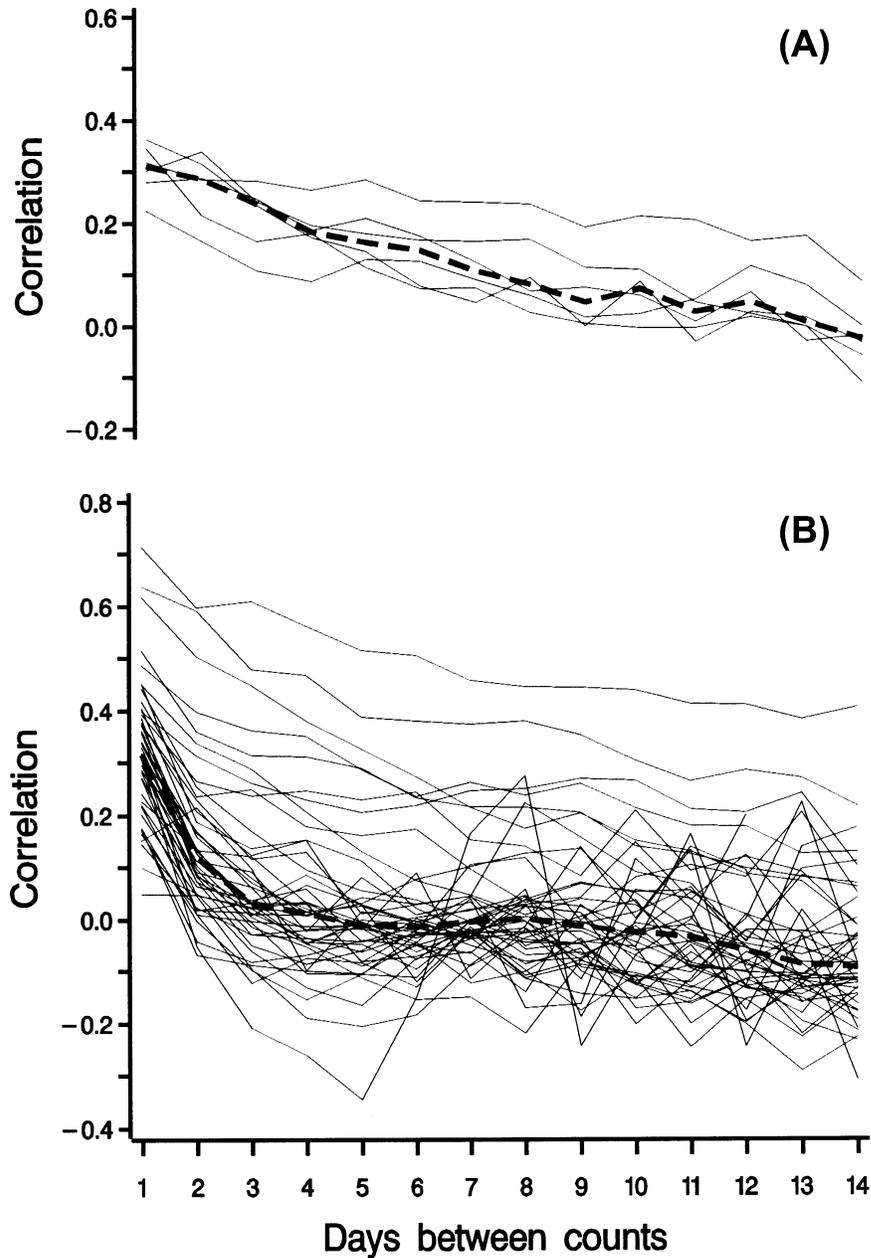


FIGURE 5. Autocorrelation between daily counts at two migration monitoring stations. Thin solid lines show the mean correlation over years between log-transformed counts for each of six species at Point Reyes Bird Observatory and 46 species at Long Point Bird Observatory. The thick dashed line shows the median of the species correlations.

similar among sampling designs at Point Reyes, and were not significantly different (Table 2).

There were also statistically significant effects of sampling frequency on design effect, and interactions between sampling frequency and sampling design

for both groups of sampling designs at Long Point (Table 1). The interactions are shown in Figure 7 (lower panel). For the group 1 designs, weekending performed very poorly at a sampling frequency of two days per week, and was the only sampling design to

TABLE 1. RESULTS OF REPEATED-MEASURES ANOVA TEST OF THE NULL HYPOTHESIS THAT DESIGN EFFECT WAS INDEPENDENT OF SAMPLING DESIGN AND SAMPLING FREQUENCY AT TWO MIGRATION MONITORING STATIONS.

Factor	Point Reyes			Long Point		
	df	F	P	df	F	P
Group 1:						
Sampling design	2, 10	1.37	0.30	2, 90	112.61	<0.001
Sampling frequency	4, 20	1.15	0.36	4, 180	112.04	<0.001
Sampling design × sampling frequency	8, 40	1.52	0.18	8, 360	72.95	<0.001
Group 2:						
Sampling design	1, 5	2.24	0.19	1, 45	44.92	<0.001
Sampling frequency	5, 25	0.70	0.63	5, 225	58.25	<0.001
Sampling design × sampling frequency	5, 25	0.43	0.83	5, 225	4.76	<0.001

Notes: Design effect measures the sampling variance of a sampling design relative to that of random sampling. Response variable was the log-transformed design effect for 6 species at Point Reyes Bird Observatory and 46 species at Long Point Bird Observatory. Group 1 designs (proportional stratified random with 1 week strata, optimal stratified random with 1 week strata, weekending) were tested at sampling frequencies of 2, 3, 4, 5 and 6 days per week. Group 2 designs (systematic, proportional stratified random with 1 sample per stratum) were tested at sampling frequencies of 1, 1.17, 1.4, 1.75, 2.33, 3.5 days per week. The sampling designs are defined in the text.

have a higher sampling variance than simple random sampling (i.e., design effect >1.0). At higher sampling frequencies weekending improved, becoming similar to the proportional stratified random design. Optimal stratified random sampling was similar to the proportional design at low sample frequencies, but improved as the frequency of sampling increased. For the group 2 designs, both systematic sampling and *st1* improved with increasing frequency of sampling, but in an erratic manner such that they were quite similar at a sampling frequency of 1.75 days per week ( $k = 4$ ), but not similar at other frequencies.

At Point Reyes (Fig. 7, upper panel), the weekending design showed the same pattern of convergence upon the proportional stratified design with increasing sampling frequency, but, unlike Long Point, the optimal design and proportional designs were very similar at all frequencies of sampling. The *st1* and systematic designs showed no consistent patterns. For clarity, error bars were not shown on Figure 7, but they were very wide for all designs at

Point Reyes, due to the low number of species in the analysis.

## DISCUSSION

Some readers will be disappointed to see that, even when there are no gaps in the count data, the median power to detect a population change of 50% over 25 years was 0.78 for the fall migration data from Point Reyes and 0.60 for the spring migration data from Long Point. This falls short of the goal suggested by the Monitoring Working Group of Partners in Flight (90% chance of detecting a decline of this magnitude; Butcher et al. 1993). However, a number of caveats should be made regarding our power analysis, and we begin the discussion by pointing these out. These limitations affect the level of power assigned to each species when there are no gaps in the count data; they do not greatly alter our main results regarding the relationship between statistical power and sampling frequency, which we

TABLE 2. GEOMETRIC MEAN DESIGN EFFECTS FOR FIVE SAMPLING DESIGNS AT TWO STATIONS. SMALLER DESIGN EFFECTS INDICATE SMALLER SAMPLING VARIANCE (RELATIVE TO RANDOM SAMPLING).

Design	PRBO	LPBO
Group 1:		
Proportional stratified random with 1 week strata	0.69 A	0.77 B
Optimal stratified random with 1 week strata	0.69 A	0.75 C
Weekending	0.72 A	0.87 A
Group 2:		
Systematic	0.67 B	0.62 E
Proportional stratified random with 1 sample per stratum ( <i>st1</i> )	0.66 B	0.70 D

Notes: Means were calculated over six species at Point Reyes Bird Observatory (PRBO) and 46 species at Long Point Bird Observatory (LPBO), and over five sampling frequencies for Group 1 and six sampling frequencies for Group 2. Values in a column with the same letters were not significantly different in paired comparison of means from a repeated-measures ANOVA for each group with log-transformed design effect as the response variable and design and sampling frequency as factors (Bonferroni *t*-tests with experimentwise  $\alpha = 0.05$ ).

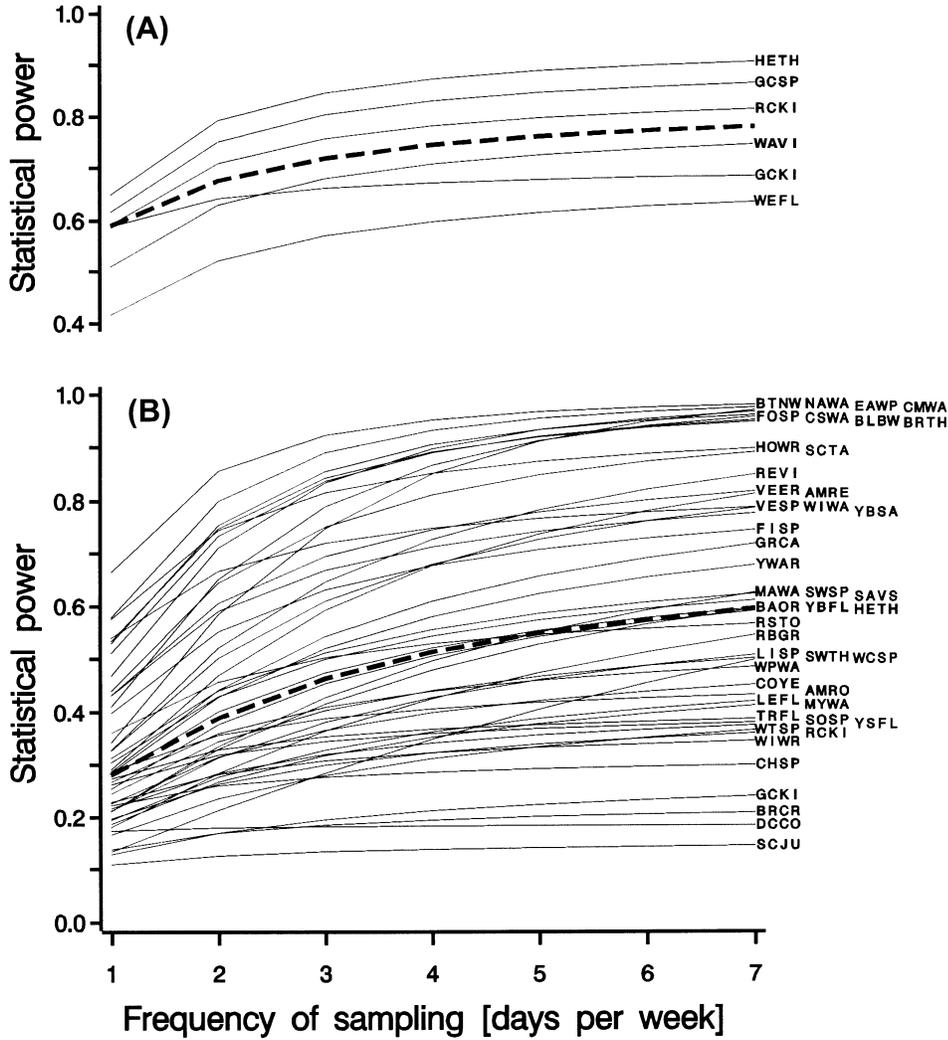


FIGURE 6. Statistical power to detect a 50% change in counts over 25 year and over a range of sampling frequencies at two migration monitoring stations, assuming random allocation of count days. Thin solid lines show the power for each species (species abbreviations at right of plot; see Appendix A for key). The thick solid line shows the median power at each sampling frequency.

discuss next. We finish by discussing the comparison of sampling designs.

LIMITATIONS OF THE POWER ANALYSIS

Our study was not designed as a rigorous evaluation of the ability of the two monitoring stations to detect population trends. To treat all species at both stations in the same manner, and to make the analysis tractable, we made a number of simplifying assumptions. Thus, a number of caveats should be made regarding our results.

First, we assumed that population trends could be estimated using a linear regression of annual indices against time on the log scale. However, for many species the indices do not conform very closely to the assumptions of a linear regression model. A number of other statistical models of trend could be used (such as non-parametric models or empirical smoothing, with different error models; Gerrodette 1987, Thomas and Martin 1996, Thomas 1997), which would almost certainly produce different estimates of power. In addition, we calculated annual indices as the mean of the logged daily counts. However, much of the

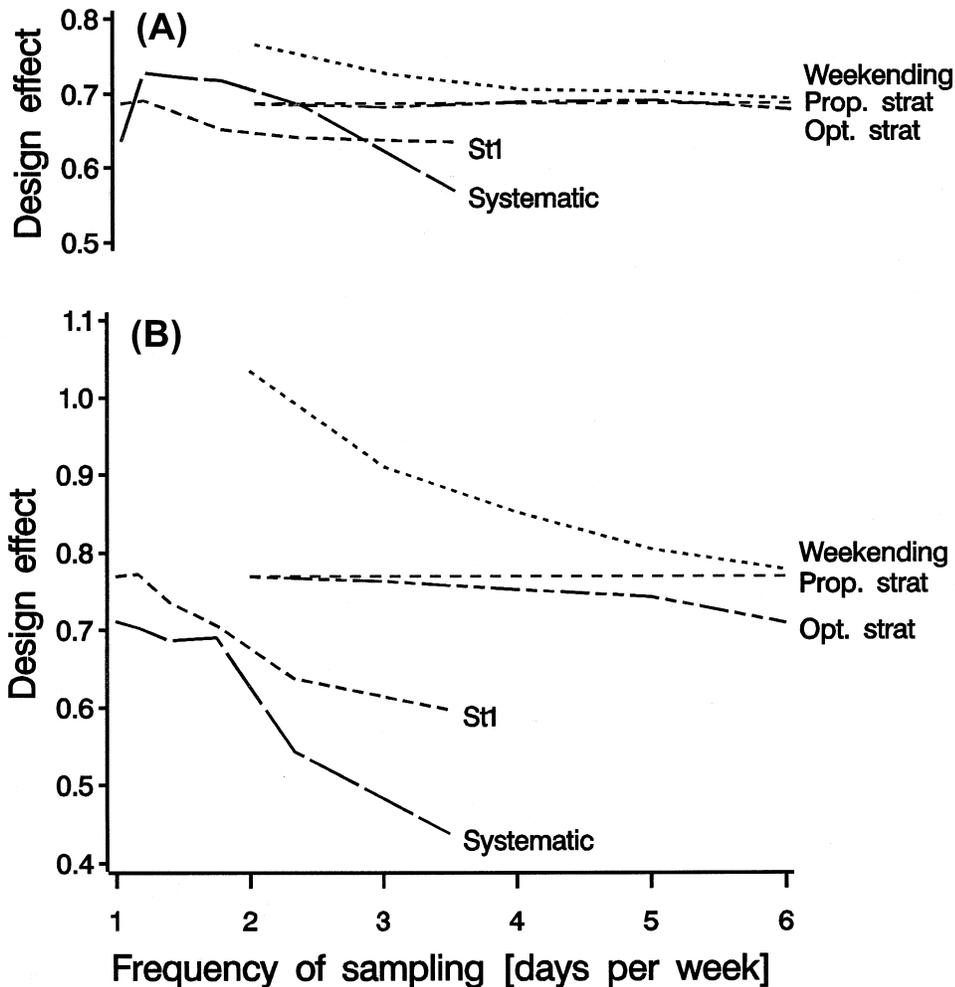


FIGURE 7. Design effect of five sampling designs over a range of sampling frequencies at two migration monitoring stations. Each line is the geometric mean design effect for the sampling design over six species at Point Reyes Bird Observatory and 46 species at Long Point Bird Observatory. Smaller design effects indicate smaller sampling variance (relative to random sampling). See Appendix A for key to species codes.

day-to-day variation in counts may be attributed to environmental factors such as wind direction and phase of the moon. Multiple regression approaches can be used to correct for these factors (e.g., Hussell et al. 1992, Pyle et al. 1994), which can lead to less variability in the trend estimates (Pyle et al. 1994) and increased power to detect trends.

The second caveat regards the selection of data. Our method of selecting a migration period for each species was designed to exclude data collected during the period when the species was not migrating, and also to produce migration periods that were an integer number of weeks in length. Our criterion of

excluding species with mean daily counts of less than 1.0 was designed to minimize the bias associated with adding a constant to the daily counts before log transformation (see below). All such criteria are arbitrary in nature, and differences among analysts will undoubtedly lead to different data being selected and thus to different estimates of power.

Third, we estimated power at only one magnitude of trend (50% over 25 years). Statistical power is dependent upon the size of trend, number of years of monitoring and significance criterion (“ $\alpha$ -level”). Thus, even species with low power using our criterion will show a statistically significant trend if the

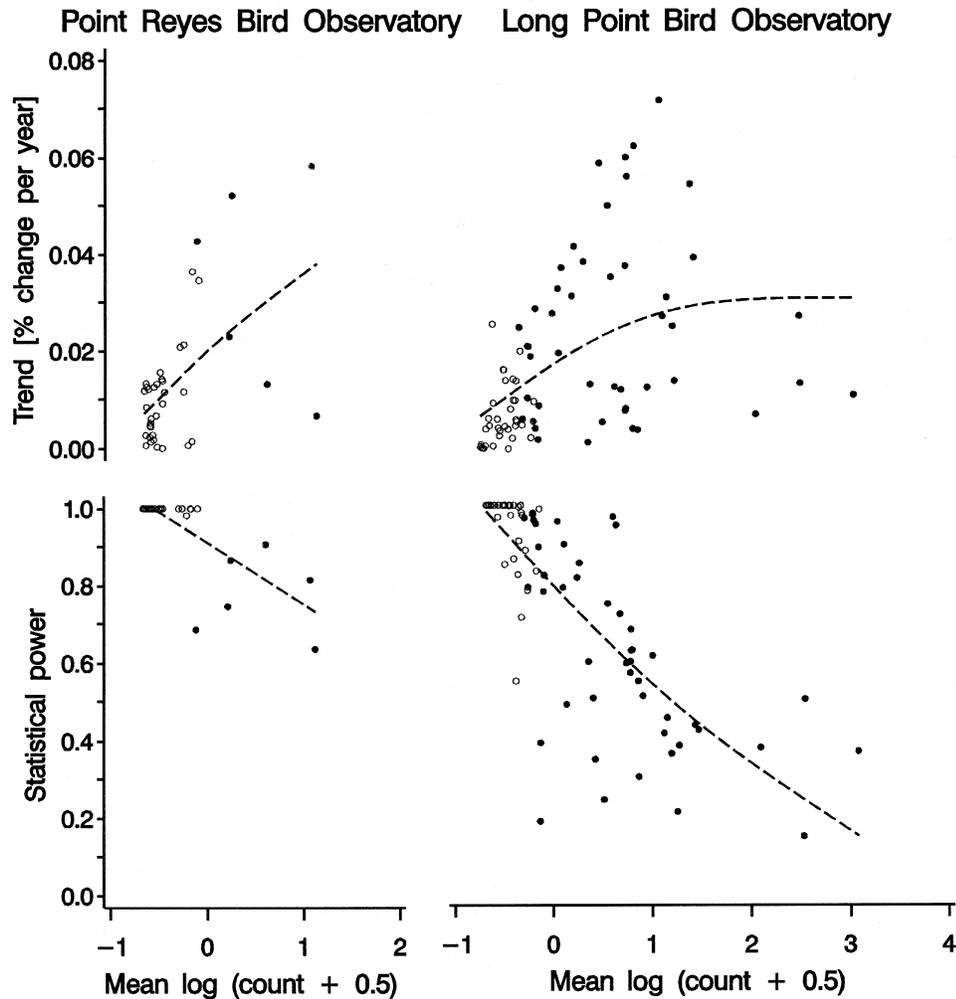


FIGURE 8. Relationship between absolute estimate of trend and statistical power ( $y$ -axis) and mean log-transformed count ( $x$ -axis) at two migration monitoring stations. Species that were excluded from our analysis (i.e., with a mean count of less than 1.0) are plotted with an open circle; those that were included are plotted with closed circles. A cubic spline (tension parameter 0.75) has been added to illustrate the pattern.

rate of change is steep enough, if there are enough years of data, or if the significance criterion is raised to greater than the customary level of 0.05. In addition, our analysis used two-tailed tests for population change, while the Partners in Flight criteria only refer to population declines. If we were not interested in detecting population increases then one-tailed tests could be used, which would result in higher statistical power (Gerrodette 1987, Cohen 1988, Steidl and Thomas 2001).

Lastly, we should emphasize that statistical power is a measure of the precision of an estimate,

assuming that the estimator is unbiased. In other words, we assumed that the annual indices, on average, reflect the true patterns of population change in the species they measure. There are two reasons why this may not be the case. Firstly, the estimate of trend will be biased if the proportion of the population that is counted varies with population size (see Sauer and Link *this volume*). The proportion of birds counted is often called the detectability, and cannot be measured directly from count data. Dunn and Hussell (1995) review the factors that may lead to differences in detectability in migration monitoring, and Sauer and

Link (*this volume*) show the statistical consequences. The second possible source of bias in the trend is the statistical procedure used to derive the estimate. In this case we added a constant before log transforming the counts. This is known to bias trend estimates (Link and Sauer 1994, Thomas 1997), causing underestimates of the size of the trend and also decreasing the variance of the estimate. The bias is greater at low abundances, where the constant is large relative to the counts, and so tends to swamp out any natural variability. Thus low abundance species will tend to have artificially high estimates of statistical power (because these estimates are based on the variance of the trend estimate), when in fact the bias towards no trend means that there is little chance of detecting a change in population size if one occurs. In our data, including all species observed at the two stations, there was a clear correlation between the size of species' trend estimates and their mean count, and between statistical power and mean count (Fig. 8). To control this bias, species with low counts are usually excluded from trend analyses that involve log-transformed data (e.g., linear route regression analysis of the Breeding Bird Survey; Geissler and Sauer 1990), as we did in our analyses by excluding species with a mean daily count of less than 1.0. This reduced the correlation between trend estimate and mean count, but the relationship between statistical power and mean count was still quite strong at Long Point (closed circles in Fig. 8). It is thus possible that our estimates of power for low-abundance species were overly optimistic.

In conclusion, due to the limitations outlined above, we caution readers not to place too much emphasis on the actual levels of power assigned to individual species, especially those with low mean counts. Further research is required before we can evaluate the importance of these limitations for assessing the ability of the two stations to detect trends. We do feel, however, that our results regarding the variation in statistical power with sampling frequency are qualitatively robust to these limitations. Because sampling variance increases at an increasing rate with decreasing frequency of sampling, missing count days will always be more important when the overall frequency of sampling is low. We discuss this further in the next section.

#### IMPORTANCE OF SAMPLING FREQUENCY

Gaps in the daily counts introduce sampling variance into the annual indices, which increases the unexplained variability about the trend line, and thus decreases the statistical power. Our results

show that power declines at an accelerating rate with decreasing sampling frequency (Fig. 6), due to the accelerating rate of increase in sampling variance. Hence, a few missing count days have little effect on the power to detect trends for these species, but the effect of additional missing days becomes greater the more there are. Species most affected tended to be those with highest power, that is, those that are the best monitored.

We are reluctant to make general recommendations about a minimum frequency of sampling that should be used, because much depends upon the individual circumstances of each station. However it is plain from our results that frequencies of one or two days per week will likely lead to annual indices for most species that are too imprecise to be able to detect large population trends if they occur. This is supported by the recommendation in Hussell and Ralph (1998) that sampling take place on at least 75% of days within a species' migration window. In addition, if analysis methods that incorporate weather variables are to be used then a large sample of days is required to detect consistent effects (Francis and Hussell 1998).

#### COMPARISON OF SAMPLING DESIGNS

Many of the differences between sampling designs are explained by two features of the data: the strong seasonality in counts for each species (Fig. 4), and the autocorrelation between counts taken on days that are close to one another (Fig. 5). Seasonality favors sampling designs that lead to counts being taken throughout the migration period; these will tend to consistently capture seasonal variation in counts and, because of this consistency, will have low sampling variance. All of the designs we compared limited the number of count days per week (or per  $k$  in the case of systematic sampling and *stl*), and thus had lower sampling variances on average than simple random sampling (i.e., average design effects < 1.0). The autocorrelation between counts taken on adjacent days was relatively high (median  $\cong 0.3$ ), and in most species decreased with increasing distance apart of count days (Fig. 5). Thus in the weekend design, with two count days per week, the count data collected on the second day of each "weekend" contained similar information to that already collected on the first day, making it less efficient than the other designs. At higher sample frequencies, additional random days were sampled during the week, and allocation of count days became similar to that of proportional stratified random sampling.

Comparing the systematic and *stl* designs,

Cochran (1977:219–221) has shown that systematic sampling will necessarily have a lower design effect than *stl* if the shape of the correlogram is concave upwards. Many species in the Long Point spring migration data set exhibited correlograms that approximated this pattern, especially at four days between counts and less (Fig. 5). Species in the Point Reyes fall data set did not tend to show the same pattern, and systematic sampling did not appear to be better than *stl*, although the small sample size prevents us from making any strong inferences about differences between the designs at this station.

Overall, our results indicate that systematic sampling should be preferred over the other designs if sampling variance is the sole criterion. Systematic sampling also has the advantage of being easy to implement. However, three drawbacks of the design should be mentioned (see Cochran 1977 and references therein for details). The first is that there is no reliable way of calculating the sampling variance from the sample. Treating the data as if it came from a random sample will almost always result in an overestimate of the true variance. This is not a problem in the current application because the variance of the annual indices does not need to be calculated to estimate the variance of the trend estimate. Secondly, systematic sampling is very imprecise if the counts show a linear trend within the season. This could be a problem for migratory species with breeding populations at the station, because the abundance of birds will tend to show a monotonic increase in the spring and decrease in the fall. However, many migration stations use only the number of new birds captured in mist nets as the daily count (e.g., Point Reyes), or attempt to exclude the resident population from analysis using other techniques (see Dunn and Hussell 1995). We saw little evidence of linear trends in counts at either station in this study (Fig. 4). Where linear trends are suspected, the problem of imprecision may be avoided by making simple adjustments to the formula for calculating annual indices (“end corrections,” Cochran 1977). Thirdly, systematic sampling can be very imprecise if the counts show regular periodic variation within the season. This is a potential problem at some stations, such as Long Point, where bird abundance at the station is thought to be related to the regular passage of weather fronts. However, even without adjusting the counts for environmental variables, we found systematic sampling to have a lower design effect than stratified random at Long Point. We thus suspect that imprecision due to periodic variation is unlikely to be a major concern.

Despite the higher expected design effects, stratified random designs have the advantage that it is

always possible to derive an unbiased estimate of the sampling variance from the sample. Because of this, stratified random designs are often preferred when little or nothing is known about the distribution of the data being sampled (such as the possibility of periodic variation or linear trends in abundance within the migration period). In this study, optimal allocation performed similarly to proportional allocation at low frequencies of sampling, where the constraint of at least one sample per stratum made the two designs very similar. Even at higher frequencies, the optimal design was only slightly better. It appears that the optimal allocation for individual species were different enough to prevent the compromise optimum allocation from providing much overall benefit. In addition, it should also be noted that a real-life implementation of the optimal allocation formula would not have the true within-stratum variances to work with, but only estimates based on previous years’ sampling. We conclude that optimal allocation schemes are probably not worth the extra effort involved in their implementation. If a stratified random design is chosen, perhaps because little is known about the region being sampled, then we recommend a proportional scheme over an attempt at optimization.

Sampling only on two adjacent days (weekending) produced the highest sampling variance of all the designs we compared. We recommend that if constraints are such that sampling can only occur on two days per week, then sampling one day in the weekend and once in the middle of the week should be encouraged.

## CONCLUSIONS

1. The frequency of daily sampling will likely have an important effect on the ability of a migration monitoring station to detect trends for some species, regardless of the statistical method used to calculate annual indices and trends.

2. The effect of missing count days is not great when the overall frequency of sampling is high, but increases with decreasing sampling frequency. Species that are well monitored (i.e., high statistical power) are more strongly affected than species that are not well monitored. Single stations that operate on 1–2 days per week are unlikely to be able to detect large changes in the abundance of species that would be well monitored at higher frequencies of sampling.

3. Of the designs we compared, systematic sampling (i.e., counting at regular intervals) performed the best for the Long Point spring migration data,

yielding the lowest sampling variance over a range of sampling frequencies. We had too few species to tell whether the systematic design was any better than the others for the Point Reyes fall data, but it did not appear to be very much different. There are a number of problems associated with systematic sampling, but these are unlikely to be important in the context of migration monitoring. Also, a major advantage of this design is that it is easy to implement, because the sampling days are regular and predictable (e.g., every second day).

4. Stratified random sampling (i.e., dividing the season into one week strata and counting on random days within each week) yielded the next lowest sampling variance for the Long Point spring migration data. Proportional stratified random sampling (i.e., the same number of counts in each week) may be preferred over systematic sampling under conditions where an unbiased estimate of the variance of the annual indices is required. We also evaluated an optimal allocation scheme, which allocated more sampling effort to weeks in which the abundance of birds was more variable. This performed slightly

better than proportional allocation, but would be difficult to implement in practice and so is not recommended.

5. Sampling only at weekends produced the largest estimates of variance of the designs we compared. If the station can only be open two days a week, we recommend counting once at the weekend and once during the middle of the week.

#### ACKNOWLEDGMENTS

We are indebted to the field staff and volunteers at the Palomarin Field Station of Point Reyes Bird Observatory and at Long Point Bird Observatory for their work in collecting the data that we used in this analysis. We thank M. S. W. Bradstreet for permission to use the data from Long Point. Reviews by E. H. Dunn, J. M. Hagan, III, K. Martin, and C. J. Ralph helped to improve earlier versions of the manuscript. This research was supported through grants from the Canadian Wildlife Service (National Wildlife Research Centre) and the Canadian Commonwealth Scholarship and Fellowship Program to L. Thomas and the National Sciences and Education Research Council of Canada to K. Martin.

#### APPENDIX. KEY TO SPECIES CODES IN FIGURES

Code	Species)
DCCO	Double-crested Cormorant ( <i>Phalacrocorax auritus</i> )
YBSA	Yellow-bellied Sapsucker ( <i>Sphyrapicus varius</i> )
YSFL	Northern Flicker ( <i>Colaptes auratus</i> )
EAWP	Eastern Wood-Pewee ( <i>Contopus virens</i> )
YBFL	Yellow-bellied Flycatcher ( <i>Empidonax flaviventris</i> )
TRFL	Alder/Willow Flycatcher ( <i>E. alnorum/E. traillii</i> )
LEFL	Least Flycatcher ( <i>E. minimus</i> )
WEFL	Pacific-slope Flycatcher ( <i>E. difficilis</i> )
WAVI	Warbling Vireo ( <i>Vireo gilvus</i> )
REVI	Red-eyed Vireo ( <i>V. olivaceus</i> )
BRCR	Brown Creeper ( <i>Certhia americana</i> )
HOWR	House Wren ( <i>Troglodytes aedon</i> )
WIWR	Winter Wren ( <i>T. troglodytes</i> )
GCKI	Golden-crowned Kinglet ( <i>Regulus satrapa</i> )
RCKI	Ruby-crowned Kinglet ( <i>R. calendula</i> )
VEER	Veery ( <i>Catharus fuscescens</i> )
SWTH	Swainson's Thrush ( <i>C. ustulatus</i> )
HETH	Hermit Thrush ( <i>C. guttatus</i> )
AMRO	American Robin ( <i>Turdus migratorius</i> )
GRCA	Gray Catbird ( <i>Dumetella carolinensis</i> )
BRTH	Brown Thrasher ( <i>Toxostoma rufum</i> )
NAWA	Nashville Warbler ( <i>Vermivora ruficapilla</i> )
YWAR	Yellow Warbler ( <i>Dendroica petechia</i> )
CSWA	Chestnut-sided Warbler ( <i>D. pensylvanica</i> )
MAWA	Magnolia Warbler ( <i>D. magnolia</i> )
CMWA	Cape May Warbler ( <i>D. tigrina</i> )
MYWA	Yellow-rumped Warbler ( <i>D. coronata</i> )
BTNW	Black-throated Green Warbler ( <i>D. virens</i> )
BLBW	Blackburnian Warbler ( <i>D. fusca</i> )

#### APPENDIX. CONTINUED

Code	Species
WPWA	Palm Warbler ( <i>D. palmarum</i> )
AMRE	American Redstart ( <i>Setophaga ruticilla</i> )
COYE	Common Yellowthroat ( <i>Geothlypis trichas</i> )
WIWA	Wilson's Warbler ( <i>Wilsonia pusilla</i> )
SCTA	Scarlet Tanager ( <i>Piranga olivacea</i> )
RSTO	Eastern Towhee ( <i>Pipilo erythrophthalmus</i> )
CHSP	Chipping Sparrow ( <i>Spizella passerina</i> )
FISP	Field Sparrow ( <i>S. pusilla</i> )
VESP	Vesper Sparrow ( <i>Pooecetes gramineus</i> )
SAVS	Savannah Sparrow ( <i>Passerculus sandwichensis</i> )
FOSP	Fox Sparrow ( <i>Passerella iliaca</i> )
SOSP	Song Sparrow ( <i>Melospiza melodia</i> )
LISP	Lincoln's Sparrow ( <i>M. lincolni</i> )
SWSP	Swamp Sparrow ( <i>M. georgiana</i> )
WTSP	White-throated Sparrow ( <i>Zonotrichia albicollis</i> )
WCSP	White-crowned Sparrow ( <i>Z. leucophrys</i> )
GCSP	Golden-crowned Sparrow ( <i>Z. atricapilla</i> )
SCJU	Dark-eyed Junco ( <i>Junco hyemalis</i> )
RBGR	Rose-breasted Grosbeak ( <i>Pheucticus ludovicianus</i> )
BAOR	Baltimore Oriole ( <i>Icterus galbula</i> )