

CHAPTER 3

USING RESAMPLING TO TEST THE RELIABILITY OF SURVEY DESIGNS: A CASE STUDY WITH THE THREATENED MARBLED MURRELET.

ABSTRACT

Marbled Murrelets (*Brachyramphus marmoratus*) are threatened seabirds of the Pacific Northwest that typically nest in coastal old-growth forests. In an effort to determine nesting distribution and status of this species, a survey protocol has been developed that is based on detecting individuals as they commute to and from nest sites. However, high levels of variability in daily detection counts have raised concern over using these data to seek temporal or spatial differences in daily detections. In response, we developed a process termed ‘reliability analysis’ to determine how effectively various survey strategies estimated measures of daily mean and standard deviation or detection counts of murrelets within a breeding season. We used an intensive field-based survey effort (50 - 65 survey days / breeding season) to estimate measures of central tendency and variance of daily Marbled Murrelet detections. We then used computer-aided resampling techniques to determine the reliability of 12 survey strategies of differing intensity (4 - 14 survey days / breeding season) and scheduling (i.e., date restricted versus random) to simultaneously estimate measures of central tendency and variability for numbers of daily detections during a single breeding season. We extrapolated reliability results to a wider range of possible murrelet detection data by producing statistically-generated detection data from a distribution form similar to the field data. Results indicate that it would be difficult to obtain reliable estimates of murrelet detections with sampling efforts up to 14 days/season. However, it appears that estimates of mean and standard deviation for daily detections during a breeding season may be reliably estimated to within + 50% with similar or less effort. Furthermore, survey strategies without date restrictions were never

less reliable than date-restricted survey strategies indicating that temporal variability was inconsistent among sites and years. The power of survey strategies to detect annual declines in detections of 25% and 50% were very low and moderate, respectively, except when variability was quite low (annual CV for daily detections < 45%). Higher levels of variability (CV > 75%) appeared to decrease power substantially.

INTRODUCTION

Background

Although large-scale population estimates of Marbled Murrelets (*Brachyramphus marmoratus*) are derived from marine surveys, there is currently no technique available to estimate numbers of nesting Marbled Murrelets at inland forest sites. However, observer-based surveys at inland forests are conducted to determine distribution and probable nesting status of murrelets. The methods for these surveys are all guided by an established protocol (Ralph et al. 1994) that requires a minimum of four surveys/season/survey station for two years to obtain a 95% probability of detecting birds if they are present. The protocol also requires inland surveys to record and tally audio and aural detections and behaviors (e.g., circling above or below the canopy during flight; see Chapter 2) of Marbled Murrelets each survey day. Daily and seasonal counts of detections, along with records of observed behavior, serve as an index to activity and intensity of habitat use and are thought to positively reflect habitat quality and nesting effort in the area around the survey station.

Observer-based surveys have been conducted for up to 10 years in some locations and managers and biologists are considering or have already begun using, both formally and informally, daily detection data to search for temporal trends in numbers of detections within stands or to compare habitat quality among stands. However, due to logistical

difficulties inherent in detecting approaching and departing murrelets during dawn and pre-dawn hours, the protocol has never recommended that daily detection data be used to estimate or monitor local populations in this manner. Furthermore, daily detections are known to exhibit a high degree of temporal and spatial variability. To date, only two studies have been designed to specifically examine the extent and potential causes of this variability (Rodway et al. 1993, Jodice, this volume, Chapter 2) and no studies have yet to analyze the potential implications of variability in daily on long-term monitoring efforts.

The goal of this study was to examine the effect of temporal variability of Marbled Murrelet detections on the efficacy of various survey strategies to estimate daily detections within years and detect changes in daily detections among years at forest stands. We accomplished this by recording detection data at multiple stands on a near-daily basis throughout the nesting season and then using these data to test the reliability of a series of less intensive survey strategies.

Statistical considerations

Ecological studies are increasingly using power analyses to aid in study design and interpretation of results (e.g., Hatfield et al. 1996, Hayes and Steidl 1997, Taylor and Gerrodette 1993, Zielinski and Stauffer 1996). At a more basic level, however, one must consider the reliability of the sampling design and resultant data. For example, ecological studies typically assume that the data being collected are reliable and therefore provide an adequate representation of the population or phenomena under consideration. For our purposes, we define reliability as the probability that an estimate of the metric of interest, derived from the data set under consideration, is within some predefined range of acceptability.

All else being equal, reliability of data tends to increase with sample size. However, issues other than sample size may affect reliability. For example, temporal variability has

been shown to be an important aspect of activity data for many species (e.g., Hayes 1997, Hatch and Hatch 1989). An increase in sample size alone may not necessarily improve reliability of data that are subject to temporal variation; surveys may instead need to be stratified by date to account for temporal variability. The unit of measurement or the shape of the statistical distribution also may affect reliability of data (Cohen 1988). Ecologists rarely have an opportunity to test the assumption of data reliability, although doing so would provide useful information for designing sampling regimes, setting effect sizes and acquiring variance estimates for power analyses. Also, such an assessment would provide solid base-line data for long-term monitoring studies. In an effort to test the assumption of reliability, we developed a ‘reliability analysis.’

This reliability analysis may be thought of as a two-step process to determine the effectiveness of sampling strategies to estimate population value(s). The first step involves the near-exact estimation of parameters through very intensive sampling efforts (i.e., greater efforts than typical sampling protocols). The second step uses one of many computer-intensive sampling methods (e.g., resampling, Monte Carlo simulations, bootstrapping) to determine how effectively less intensive or temporally stratified sampling strategies estimate a population parameter. For example, a reliability analysis may determine that the sampling strategy under consideration typically provided data that estimated a population value to within + 20%. Considered in this context, reliability analyses provide a useful tool for determining effective survey designs for Marbled Murrelets, specifically, and for other wildlife species in general. Results of reliability analyses also highlight the importance of collecting baseline data prior to initiating long-term monitoring studies and of exploring the extent of temporal variability in count data.

Objectives

Our objectives were to: (1) apply sufficient survey effort in the field to obtain de-

pendable estimates of measures of central tendency and variance of daily Marbled Murrelet detections within years and trends of daily detections among years; (2) use computer-aided resampling techniques to determine the reliability of survey strategies with differing intensity and scheduling to simultaneously estimate measures of central tendency and variability for detections during a single breeding season; (3) determine the power of these same survey strategies to detect trends in detections over time; and (4) extrapolate reliability and power analyses to a wider range of possible murrelet detection data than provided by our field data by producing statistically-generated detection data that were similar in statistical nature to the field data. These analyses provide an assessment of the feasibility of obtaining reliable estimates of counts of daily detections and subsequently using this metric for determining annual trends in murrelet activity.

METHODS

Study Sites

Seven survey stations are located in five Douglas-fir (*Pseudotsuga menziesii*) old-growth forest stands in the central Oregon Coast Range (Fig. 2.1): Valley of the Giants Meadow (VGM), Valley of the Giants Upper Plateau 1 (VGUP1) and 2 (VGUP2), Spencer Creek Main Fork (SCMF), Spencer Creek Upper Fork (SCUF), 2x4 east (E2x4), and 2x4 west (W2x4). VGUP1 and VGUP2 are within the same stand and ca. 150m apart; E2x4 and W2x4 are within the same stand and are ca. 300m apart. Location, elevation, and general descriptions of each survey stand appear in Methods, Chapter 2.

Data Collection

Surveys for Marbled Murrelets were conducted on a near-daily basis (50 - 64 days/station/year) between 1 May and 4 August (95 possible survey days hereafter referred to

as the breeding season), in 1994 (VGM, VGUP1, SCMF, SCUF), 1996 (VGM, VGUP1, VGUP2), and 1997 (VGM, VGUP1, SCMF, E2x4, W2x4), resulting in 12 site*year combinations. Daily survey data were collected following procedures outlined in Methods, Chapter 2, and, except for number of survey days, generally followed guidelines established in the Marbled Murrelet survey protocol (Ralph et al. 1994).

Reliability of Survey Strategies

We used resampling techniques to evaluate the reliability of 12 survey strategies of which nine were stratified temporally and three were not (Table 3.1; surveys without temporal stratification component hereafter called ‘completely random’). Survey strategies included the existing protocol (Ralph et al. 1994), methodologies considered to be logistically feasible based upon sampling effort and scheduling, and methodologies that considered breeding phenology. Each survey strategy was evaluated for each site*year combination. Daily detection data (i.e., counts of daily detections and date) from the 12 site*year surveys constituted the population (hereafter called observed data) from which resampled surveys (hereafter called samples) were drawn. Within the constraints set by the 12 survey methodologies (i.e., temporally stratified or completely random), samples of daily detections were randomly selected without replacement for each survey strategy (i.e., each survey day in a sample was unique). Furthermore, no two samples contained an identical set of days.

For each sample, the mean and SD of number of detections per day were calculated and were compared to the mean and SD of the observed data set under consideration. The reliability of a survey strategy was then defined as the proportion of samples whose estimates of the mean and SD fell within predefined limits of the observed mean and SD.

Three such limits (hereafter called accuracy windows) were used to assess reliability; mean and SD of samples each within +10, 20, and 50% of observed mean and SD. These

Table 3.1. Marbled Murrelet survey strategies evaluated for estimating mean and variance of daily detections. For each survey strategy and site-by-year combination, days were randomly selected without replacement from observed data which was collected on a near-daily basis at seven survey stations in the Oregon Coast Range, 1 May – 5 August 1994, 1995, 1997. Each set of survey days within a survey strategy and site-by-year combination also was unique.

No. survey days	Temporally stratified (TS) or completely random (CR)	Sampling methods (all days randomly selected)	Survey strategy acronym
4	CR	Selected from entire season	CR4
4	TS	1 day from May; 1 day from June; 1 day between 21 June and 21 July; 1 day between 10 July and 4 Aug. At least 6 but no more than 30 days between surveys.	P4 ¹
4	TS	Selected from May	MY4
4	TS	Selected from June	JN4
4	TS	Selected from July	JY4
7	CR	Selected from entire season	CR7
7	TS	1 day selected from each 2 week period	BIWK
8	TS	Selected from May	MY8
8	TS	Selected from June	JN8
8	TS	Selected from July	JY8
14	CR	Selected from entire season	CR14
14	TS	1 day selected from each week	WEEK

¹ An approximation of the current Marbled Murrelet survey protocol (Ralph et al. 1994).

accuracy windows provide a hierarchy of reliability criteria (e.g., Fig. 3.1). The inner-most window represents a ‘best-case’ scenario where estimates were highly accurate; the middle window represents estimates that were moderately accurate; the outermost window represents minimally acceptable standards where accuracy was low but data still provided a useful ‘ball park’ estimate of the metrics.

Reliability within each accuracy window (also referred to as the reliability index) was compared among survey strategies. Survey strategies with higher reliability indices were considered to be more effective at estimating the observed statistics. Additionally, for each accuracy window and survey strategy, all samples were assigned to 1 of 9 error categories based on whether observed means and SDs fell below, within, or above the limits of an accuracy window (e.g., Table 3.2). The proportion of samples within each of the nine categories was calculated for each survey strategy and accuracy window and used to determine the direction and magnitude of error in sample means and SDs.

Reliability analyses were extended beyond the scope of the field data by generating data from statistical distributions that represented a range of detection means and variances likely to be recorded during a season of murrelet surveys (Table 3.3). We generated 1000 sets of 4, 7, and 14 gamma variates (i.e., survey strategy CR4, CR7, and CR14, respectively) for each cell in the GDM and assessed reliability as described above. We used these three survey strategies with generated data because we found little difference in reliability between temporally stratified and completely random survey strategies (see Results).

The gamma distribution was chosen for data generation because it is very flexible (Evans 1993), tends to represent count data well, fit 11 of our 12 field survey sets well (Kolmogorov Smirnov $P > 0.3$ for 11 of 12 cases), and also fit two similarly-sized murrelet detection data sets from British Columbia well (Kolmogorov Smirnov $P > 0.8$; Rodway et al. 1993). Gamma variates were generated (SAS procedure RANGAM; SAS Institute, Inc., 1985) for each cell in the generated data matrix (GDM) using shape and

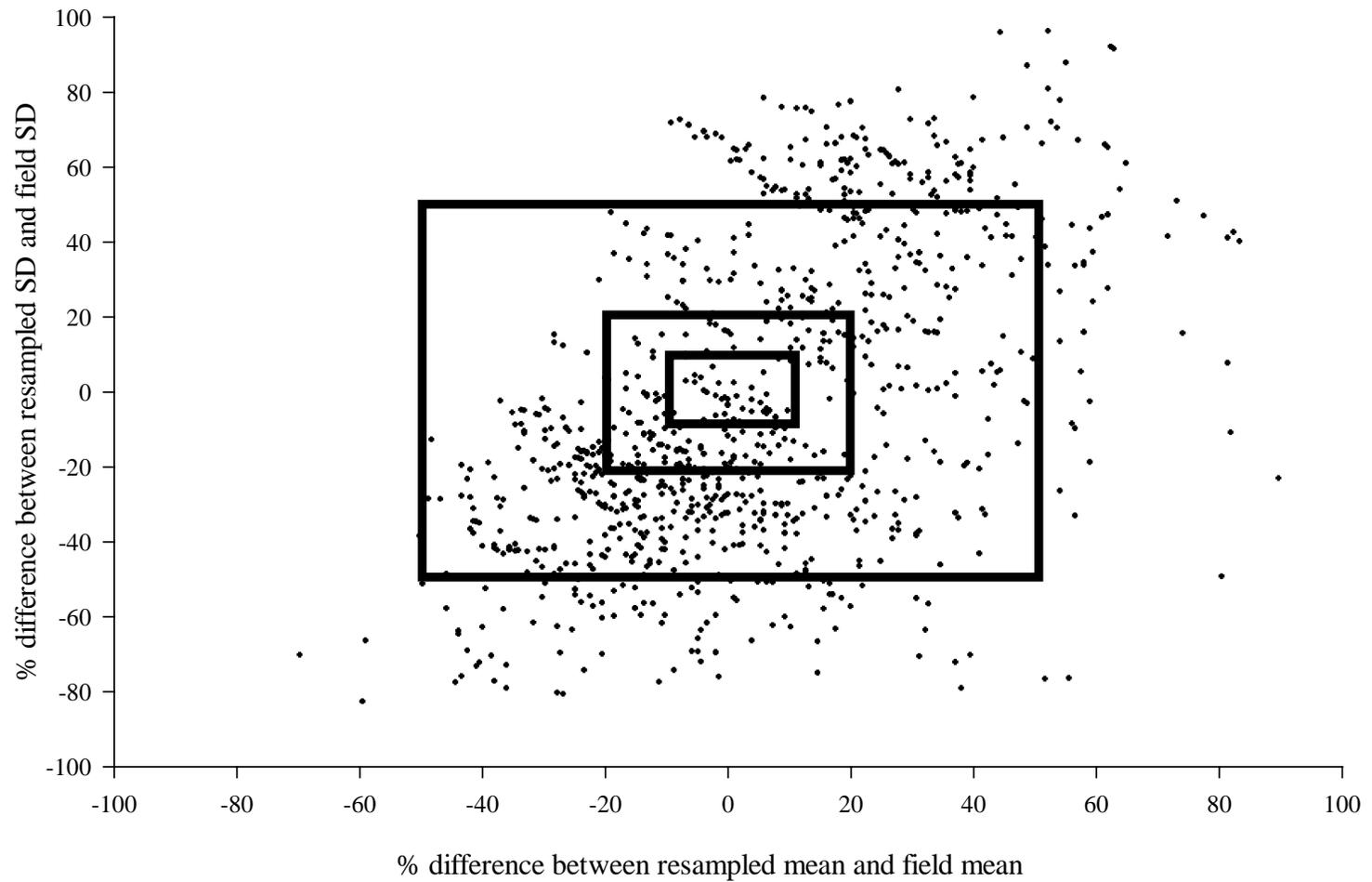


Figure 3.1. Differences between resampled and field estimates of means and SDs of daily detections. Data derived from 1000 resampled surveys using survey strategy P4 with source data from the E2x4 site, 1997.

Table 3.2. Categories of reliability from survey strategy P4 (see Table 3.1), accuracy window $\pm 20\%$, site E2x4, 1997. The value in each cell is the proportion of the 1000 resampled surveys (i.e., samples) that met that cells definition. The value in the cell 'mean reliable, SD reliable' is the reliability index.

	Mean low	Mean reliable	Mean high	Sum
SD low	12.3	22.7	4.8	39.8
SD reliable	6.3	18.0	6.6	30.9
SD high	<u>0.1</u>	<u>13.6</u>	<u>15.6</u>	29.3
Sum	18.7	54.3	27.0	

Table 3.3. Mean and SD values used for generating gamma variates to evaluate survey strategies for Marbled Murrelets. Cells with 'O' and 'BC' had field data from our study in Oregon and a similar study in British Columbia (Rodway et al. 1993), respectively, that had means and SDs similar to the corresponding row and column. Cells with 'G' had data generated from gamma distributions..

SD (multiple of mean)	Mean detections / day					
	<u>10</u>	<u>30</u>	<u>50</u>	<u>70</u>	<u>90</u>	
0.25 * mean	G	G	G	G	G	
0.35 * mean	G	G	G	G	G	
0.45 * mean	G	O	BC	G	G	
0.55 * mean	G	G	G	G	G	
0.65 * mean	G	O	O	G	G	
0.75 * mean	O	G	BC	G	G	
0.85* mean	O	O	G	G	G	
0.95* mean	G	G	G	G	G	
1.05* mean	G	G	G	G	G	
1.15* mean	O	O	G	G	G	

scale parameters, where $\text{shape} = (\text{mean}/\text{sd})^2$ and $\text{scale} = \text{s}^2/\text{mean}$ (N.B., not all statistical software use the same equation to generate shape and scale; see Evans et al. 1993 for other equations). Similarity in statistical distributions between generated and field data were verified in two ways. First we compared frequency distributions of field and generated data that shared similar but not identical means and variances (i.e., cells in the GDM where field data were located). Visual observations indicated these frequency distributions appeared similar in all cases; two such distributions are displayed (Fig. 3.2). Second, we compared the results of reliability analyses from field and generated data that shared similar mean and SD parameters. For a given cell in the GDM, we assumed that any given survey strategy should produce similar reliability indices whether data were generated from random gamma variates or collected in the field. No significant differences were detected between the reliability of CR4, CR7, or CR14 survey strategies with field or generated data in accuracy window 10 (paired-t = 0.73_{20} , P = 0.47), 20 (paired-t = 1.17_{20} , P = 0.25), or 50 (paired-t = 0.77_{20} , P = 0.457). Therefore, generated data from the gamma distributions appeared to match field data well enough to proceed with analyses.

Power Analysis

We calculated the power of selected survey strategies to detect negative trends in daily detections during 2, 3, or 5 year periods. Analyses used data from SCMF (1994 and 1997), VGM (1994, 1996, and 1997) and VGUP (1994, 1996, and 1997), and generated data (3 and 5 years). For each of the field sites, we used all of the observed data to calculate the slope of daily detections regressed upon year (i.e., observed trends). One thousand samples for each of three survey strategies (CR4, CR7, CR14; Table 3.1) were generated for each field site and year. For each site and survey combination, sequential samples from each year were combined (e.g., sample 1 of year 1 and 2 combined, sample 2 of year 1 and 2 combined, etc.), and the slope of daily detections regressed on year was

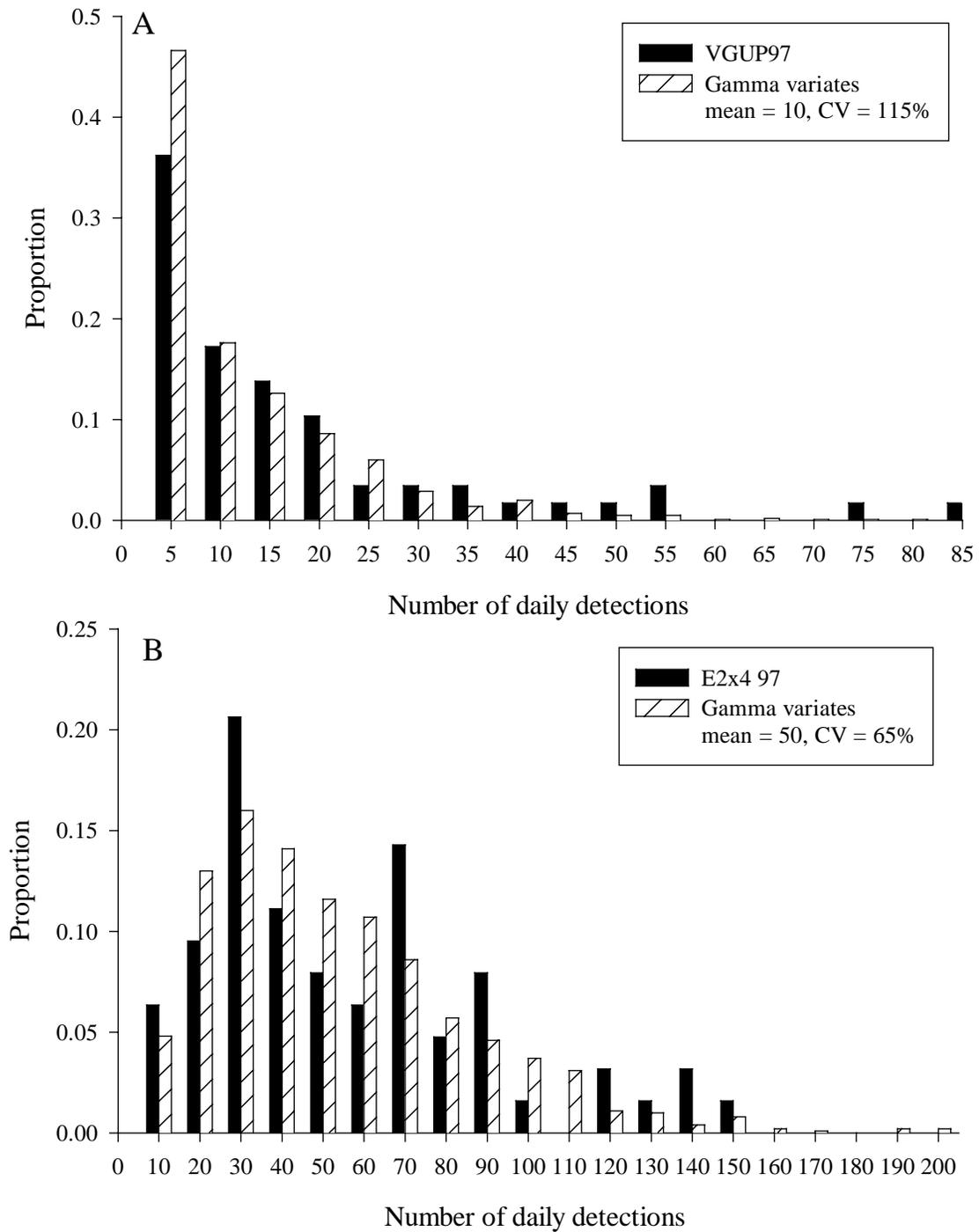


Figure 3.2. Comparison of frequency distributions between field data and data generated from a gamma distribution. For each plot, field and generated data fall within the same cell of the generated data matrix and therefore share similar means and SDs (see Table 3.3). X-axis labels are upper values of intervals. (A) Generated and observed data each having low mean and high SD; (B) generated and observed data each having moderate SD and high mean.

derived. The percent difference between the slope of each of the 1000 sample aggregates and the corresponding slope from the observed data were calculated. Power of a survey strategy for a site was calculated as the proportion of samples whose slopes were within +10, 20, 30, 40, or 50% of the observed slope. Data were log transformed prior to regression analyses to linearize trends.

Using data generated from gamma distributions, we calculated the power of CR4, CR7, and CR14 survey strategies to detect 25% and 50% annual declines in the number of detections for 3 and 5 year periods. We generated 1000 samples of gamma variates (i.e., 4, 7, or 14) with a starting mean of 50 and CV of 45% and 85%. We then reduced the mean by either 25% per year or 50% per year over 3 and 5 year periods (holding the CV constant) and generated 1000 samples of gamma variates for each year, mean, and CV combination. Power was then calculated as described above (also see Table 3.4).

RESULTS

Survey Data

We conducted 681 Marbled Murrelet surveys, averaging 56.7 survey days/survey station/breeding season. At least one detection was recorded on 616 mornings, although 9 of 12 site*year combinations had at least one day without any detections. We recorded 16,105 detections, ca. 141,000 keer calls, and ca. 31,000 minutes of activity. About 29% of detections were visual, 58% audio and 13% audio-visual, although these proportions varied within and among sites and years (Table 2.2). We observed 12,244 birds during all surveys. A more complete description of survey data appears in Chapter 2.

Mean counts of detections/day at all stations and in all years varied between 7 and 51. There was significant variability in means within and among stations and years (Table 3.5). The mean number of daily detections varied by month at most sites during

Table 3.4. Range of annual declines represented by the percent difference between resampled slopes and observed slopes used in determination of power. Analyses were conducted by regressing mean detections day⁻¹ upon year and using generated gamma variates to represent multiple years of surveys.

% difference between resampled slope and observed slope	25% annual decline		50 % annual decline	
	Lower limit	Upper limit	Lower limit	Upper limit
± 10 %	27.12	22.81	53.35	46.41
± 20 %	29.19	20.55	56.48	42.57
± 30 %	31.19	18.23	59.39	38.45
± 40 %	33.14	15.85	62.11	34.03
± 50 %	35.04	13.39	64.65	29.29

Table 3.5. Summary statistics for Marbled Murrelet detection data obtained during observer-based audio-visual surveys at seven survey stations in the Oregon Coast Range, 1 May - 4 August, 1994, 1996, 1997.

Survey station	Year	No. survey days	Mean detections/day	CV	Minimum no. detections	Maximum no. detections
SCMF	1994	64	32.62	1.310	0	198
SCMF	1997	61	10.56	1.525	0	83
SCUF	1994	58	16.22	1.341	0	112
VGM	1994	55	27.31	0.684	0	79
VGM	1996	50	7.66	1.134	0	38
VGM	1997	58	15.29	0.874	1	56
VGUP	1994	56	36.14	0.493	1	88
VGUP	1996	51	14.09	0.721	0	39
VGUP	1997	56	14.69	1.216	0	85
VGUP2	1996	51	16.25	0.830	0	51
E 2X4	1997	62	51.29	0.695	2	147
W 2X4	1997	59	34.26	0.872	0	125

most years. Separate ANOVAs were conducted for each of the 12 site*year combinations and 7 of these documented significant relationships between month and mean daily detections. However, the strength and temporal pattern of the relationship between these two variables was inconsistent within sites among years and among sites within years (Table 3.6). Temporal variation in counts of daily detections also was high and it was not uncommon to observe near-minimum and near-maximum counts of daily detections at a site within the same week (Fig. 2.3). CVs for daily detections varied three-fold among all stations and years, and also varied within stations among years. CVs were not consistently lower during any particular time period of the breeding season (Table 2.3, Fig. 3.3).

Reliability of Survey Strategies for Observed Data

Most of the 12 survey strategies we used for resampling did not provide reliable estimates of observed means and SDs (Fig. 3.4). The percentage of samples meeting the strictest reliability criteria (i.e., accuracy window + 10%) was typically < 20% for all survey strategies (Fig. 3.4a). This percentage increased for the + 20% accuracy window, but was still generally low (Fig. 3.4b). Within this accuracy window, CR14 and WEEK surpassed 70% reliability in some cases, but none of the survey strategies reliably estimated the observed means or SDs > 50% of the time when averaged among sites and years. Within accuracy window + 50% (Fig. 3.4c), all survey strategies resulted in > 70% reliability for at least 1 site in 1 year. CR7 and BIWK surpassed 70% reliability on average, and CR14 and WEEK surpassed 70% reliability for all sites in all years. In general, resampled surveys tended to under- or over-estimate both the observed mean and SD in accuracy window + 10% and underestimate the SD in accuracy windows + 20 and 50% (Table 3.7).

Table 3.6. Variability in counts of daily Marbled Murrelet detections by month for 12 site-by-year combinations in the Oregon Coast Range. For significant models, months sharing any identical letters have daily detection means that are not significantly different ($P > 0.05$; Tukey -Kramer post-hoc analyses). Data were log transformed for analysis but raw values are shown.

Site & year	ANOVA <i>F</i>	<i>P</i>	May	June	July
SCMF '94	9.74	<0.001	7.7 (a)	39.0 (b)	59.3 (b)
SCMF '97	0.87	0.426	10.9	10.2	11.7
SCUF '94	29.84	<0.001	4.1 (a)	10.1 (b)	37.2 (c)
VGM '94	5.46	0.007	28.1 (a)	18.8 (b)	29.6 (a)
VGM '96	2.78	0.073	8.4	10.6	5.7
VGM '97	7.45	0.001	9.4 (a)	11.9 (a)	25.3 (b)
VGUP '94	0.11	0.899	37.9	35.1	39.8
VGUP '96	0.50	0.611	15.1	15.0	14.2
VGUP '97	4.04	0.023	8.0 (ab)	13.5 (a)	24.9 (b)
VGUP2 '96	2.02	0.144	22.5	18.0	14.0
E2x4 '97	13.22	<0.001	48.5 (a)	32.4 (b)	81.3 (a)
W2x4 '97	13.00	<0.001	29.8 (a)	16.8 (b)	57.8 (a)

Figure 3.3. Coefficients of variation of daily Marbled Murrelet detections from 11 inland survey stations in the Oregon Coast Range, 1 May – 4 August 1994, 1996, 1997.

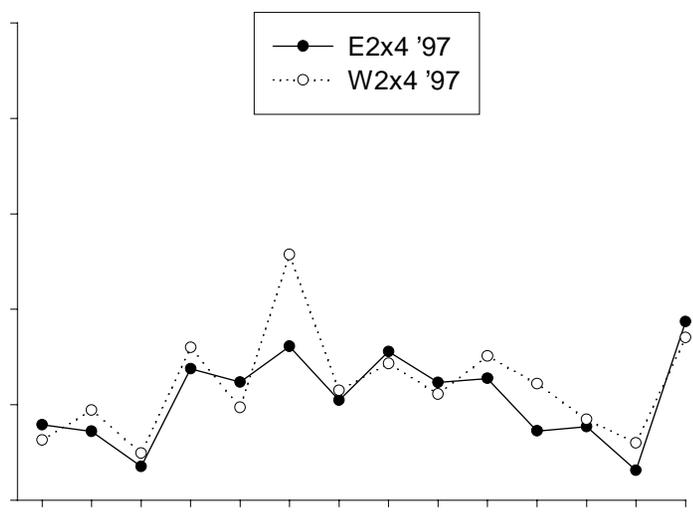
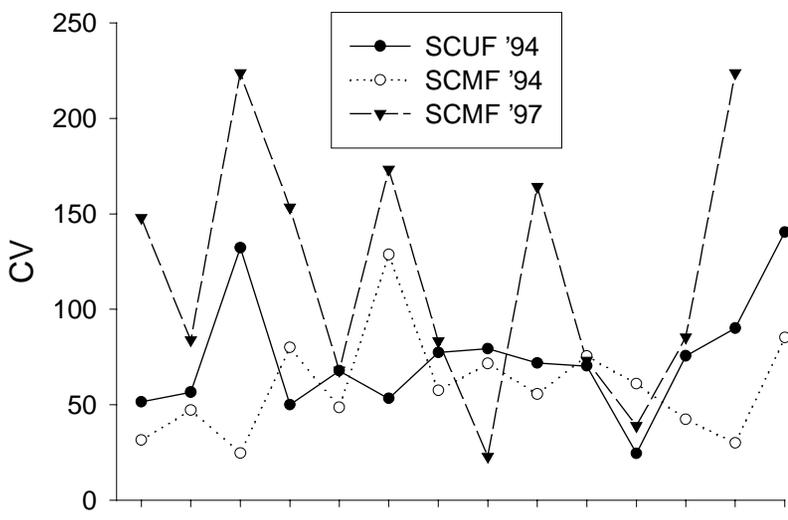
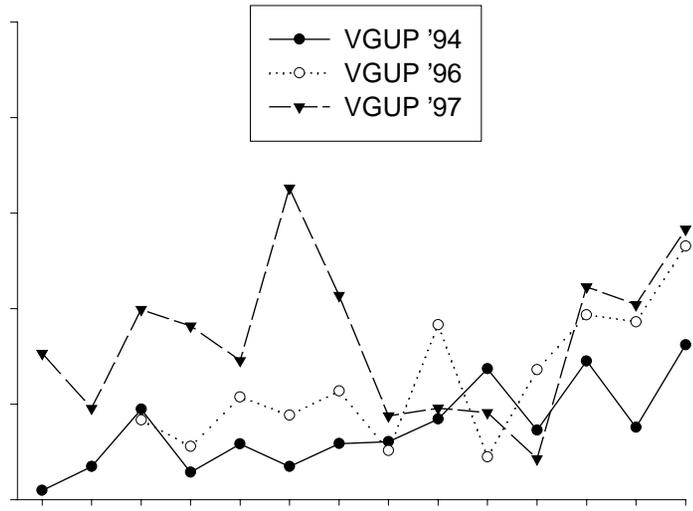
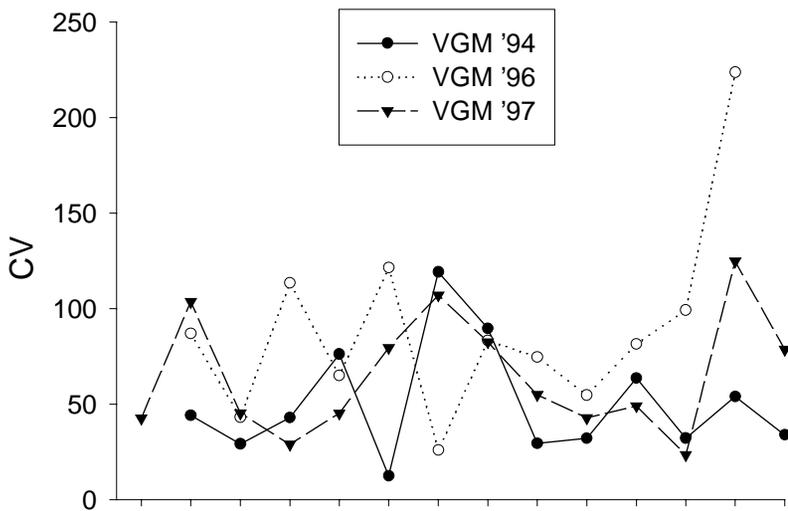


Figure 3.3.

Week (1 May = Week 1)

Week (1 May = Week 1)

Figure 3.4. Reliability indices of 12 different sampling strategies (Table 3.1) for 12 site*year combinations in (A) accuracy window $\pm 10\%$, (B) accuracy window $\pm 20\%$, and (C) accuracy window $\pm 50\%$. Reliability = the proportion of 1000 resampled surveys satisfying the accuracy window criteria.

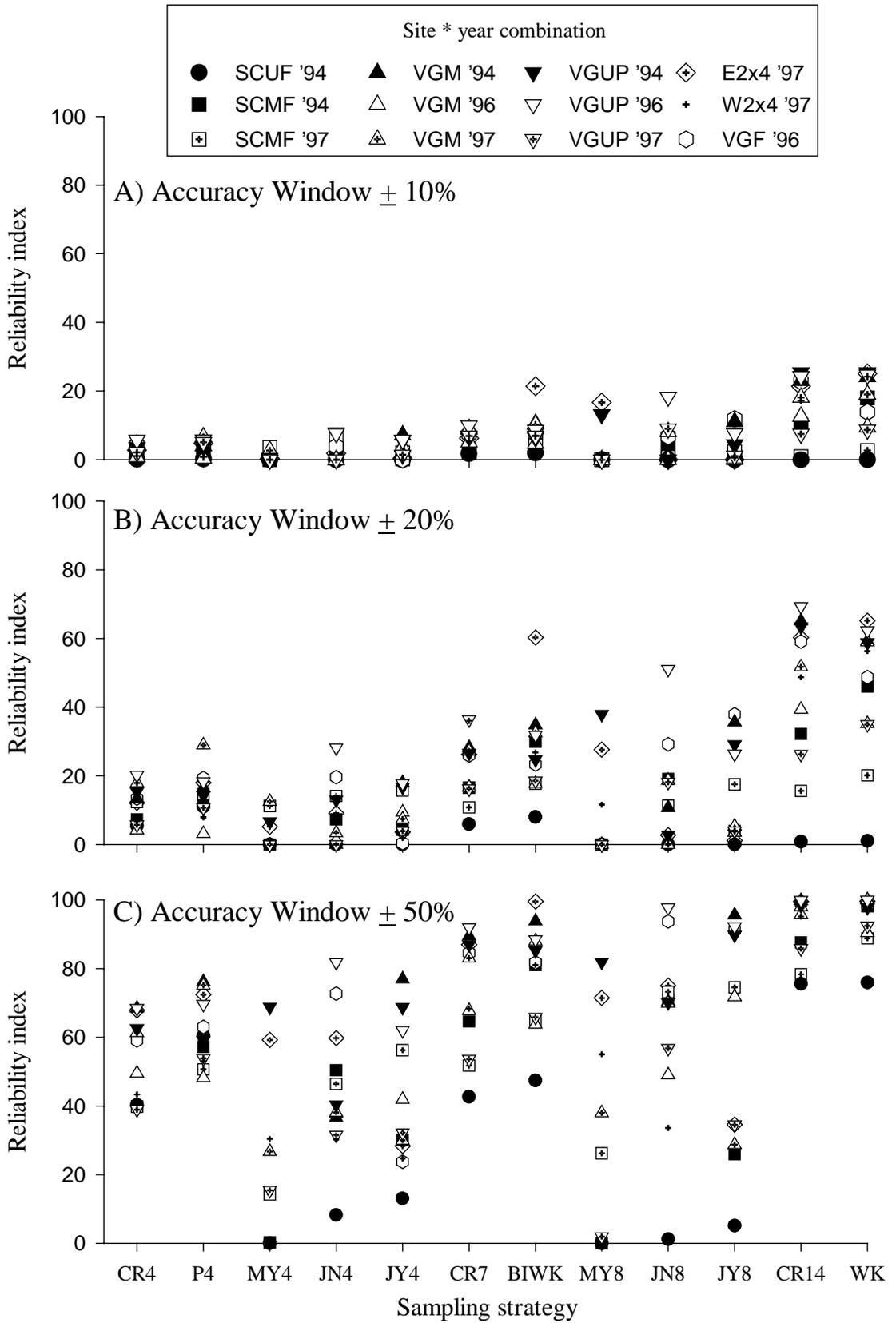


Figure 3.4.

Table 3.7. Mean rank (± 1 s.e.) of the proportion of samples in each of the nine error categories of the reliability matrix, by accuracy window. Mean rank is based on values of all survey strategies combined within each accuracy window. Order of category labels (e.g., low/low) is mean/sd. 1 = highest rank (i.e., most common), rel. = reliable.

Accuracy window	low/low	rel./low	high/low	low/rel.	rel./rel.	high/rel.	low/high	rel./high	high/high
$\pm 10\%$	1.42 (0.79)	3.17 (1.11)	5.17 (1.90)	7.17 (0.93)	5.83 (2.08)	4.50 (1.57)	9.00 (0)	5.58 (0.99)	2.42 (1.24)
$\pm 20\%$	2.21 (1.52)	2.83 (1.53)	7.42 (1.08)	6.92 (1.38)	2.92 (1.56)	4.33 (1.61)	8.92 (0.29)	5.50 (1.31)	3.54 (1.53)
$\pm 50\%$	4.04 (1.79)	2.42 (0.79)	7.92 (0.29)	5.92 (1.73)	1.17 (0.58)	3.79 (1.34)	8.58 (0.51)	4.75 (1.42)	5.42 (0.99)

Reliability of Temporally Stratified versus Completely Random Survey Strategies

The reliability of temporally stratified versus completely random surveys with identical or similar effort (i.e., 4 days, 7 - 8 days, 14 days) varied with the accuracy window being considered (Table 3.8). For accuracy window + 10% there were no differences in reliability for survey strategies of similar effort. Within accuracy window + 20% and 50%, however, there were significant differences in reliability among surveys with 4 days and among survey with 7 - 8 days. In each case, single month efforts (e.g., 4 or 8 days in May) were less reliable than either completely random surveys or stratified surveys conducted throughout the breeding season (i.e., P4, BIWK, WEEK). Completely random survey strategies were never less reliable than any temporally stratified survey strategies. Therefore, only completely random survey strategies were used in reliability analyses with generated data and power analyses with field and generated data.

Reliability of Survey Strategies for Generated Data

Reliability indices for generated data were similar to those for observed data. Reliability was low for most survey strategies for most accuracy windows (Fig. 3.5). CR4 resulted in low reliability (Fig. 3.5a) which exceeded 70% only when the SD was $< 0.65 * \text{mean}$, and the accuracy window was + 50%. Although CR7 surpassed 70% reliability in accuracy window + 50% for all but the highest SDs, it never exceeded 50% reliability for accuracy windows + 10 or + 20% (Fig. 3.5b). Reliability in CR14 was moderate to high in accuracy windows + 50 and + 20% when SDs were high and $< 0.55 * \text{mean}$, respectively (Fig. 3.5c). CR14 never produced reliable estimates in accuracy window + 10% even with low SDs. Additionally, consistent differences or patterns in reliability were not apparent among means within or among SD values (Fig. 3.5), indicating the mean had little effect on reliability.

Table 3.8. Results of paired t-tests (survey strategies with $n = 14$ survey days) and ANOVAs (all other survey strategies) testing whether the proportion of reliable samples varied between survey strategies of similar effort that were temporally stratified versus completely random. Survey strategies tested and their associated survey effort appear in Table 1. Post-hoc comparisons made with Tukey-Kramer test and significant results are $P \leq 0.05$.

Survey effort (days)	Accuracy window	df	<i>F</i> statistic	<i>t</i> statistic	<i>P</i>
4	$\pm 10\%$	4,51	1.82	-	0.139
4	$\pm 20\%$	4,51	3.10	-	0.023 ¹
4	$\pm 50\%$	4,51	5.38	-	0.001 ²
7 - 8	$\pm 10\%$	4,51	1.32	-	0.273
7 - 8	$\pm 20\%$	4,51	2.75	-	0.038 ³
7 - 8	$\pm 50\%$	4,51	4.32	-	0.004 ⁴
14	$\pm 10\%$	11	-	0.79	0.443
14	$\pm 20\%$	11	-	0.57	0.581
14	$\pm 50\%$	11	-	1.62	0.134

¹ P4 > MY4.

² P4 & CR4 > MY4.

³ no pairwise differences at $P = 0.05$.

⁴ BIWK & CR7 > MY8.

Figure 3.5. Reliability indices for survey strategies (A) CR4, (B) CR7, and (C) CR14 in three accuracy windows (± 10 , 20, and 50%) with data from generated gamma variates (see Table 3.3). Each series of data points for each SD value represents mean daily detections of 10, 30, 50, 70, and 90, respectively.

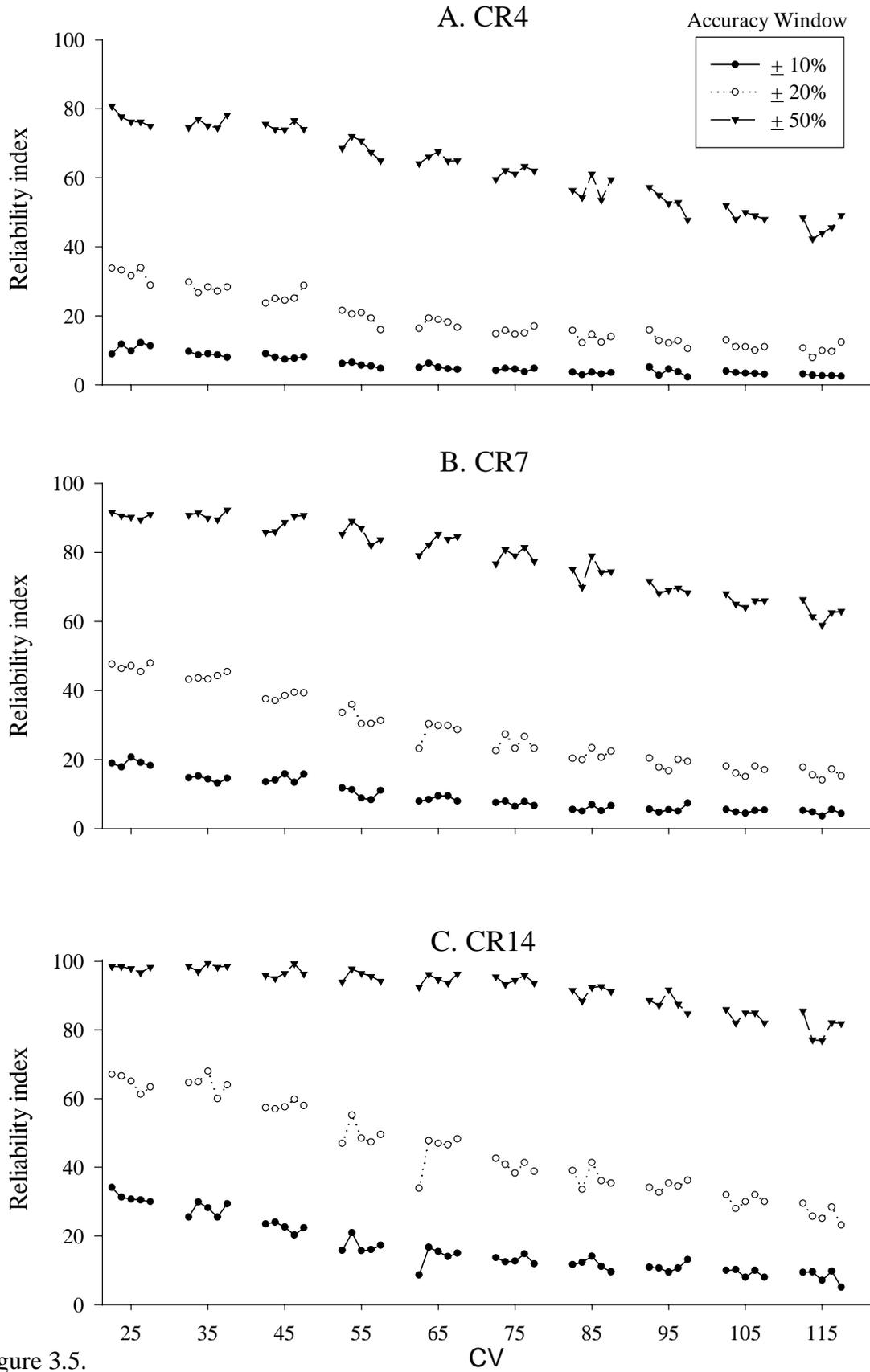


Figure 3.5.

Power of Survey Strategies to Detect Annual Trends in Detections

Significant negative trends in annual mean numbers of daily detections were observed at each site (range -28% per year to -71% per year; Fig. 3.6). While all survey strategies (i.e., CR4, 7, and 14) correctly identified the direction of the slope in > 85% of resampled regressions, none of the survey strategies consistently displayed reasonable power (e.g., >70%) to estimate the value of the observed slope of detection trends (Fig. 3.7). For example, at SCMF where the annual decline in detections was steepest and the variability in the numbers of daily detections highest, power was < 70% even for CR14 (Fig. 3.7a). Survey results were similar at VGM (Fig. 3.7b). However, at VGUP, where the annual decline was about 57% per year, CR7 surpassed 70% power when the percent difference between the observed and sample slope was + 50%. Similarly, CR14 surpassed 70% power when the percent difference between the observed and sample slope was + 30% (Fig. 3.7c).

Power of regressions with generated data was greater than power with observed data (Fig. 3.8); however, CVs of daily detections were held constant in all years for these analyses. With a low CV (i.e., 45%), the power to detect a 50% decline/year was adequate with three survey years and 4 samples / year (e.g., within + 30% of observed slope, power > 70%; Fig. 3.8a). The power to detect a less severe decline of 25% per year exceeded 70% only when 14 surveys were conducted and the percent difference between the observed slope and sample slope was > + 40%. Increasing the CV to 85% reduced power sufficiently enough so that detecting annual declines of 25% per year was improbable. Power to detect a steeper decline of 50% per year with a CV of 85% was sufficient only when 14 surveys were conducted and the percent difference between the observed slope and sample slope was > + 40% (Fig. 3.8c). Increasing the number of survey years to 5 in this same scenario improved the power to detect the decline even for survey strategy CR4 (Fig. 3.8d). However, with the higher CV, the power to detect the 25%

Figure 3.6. Results of daily detections regressed upon year from three survey stations in the Oregon Coast Range, 1 May – 4 August 1994, 1996, and 1997. Each regression equation was significant at $P < 0.05$. One data point from SCMF, 1994, was not shown in order to maintain clarity and consistent scales among plots. Value of that point = 198 detections.

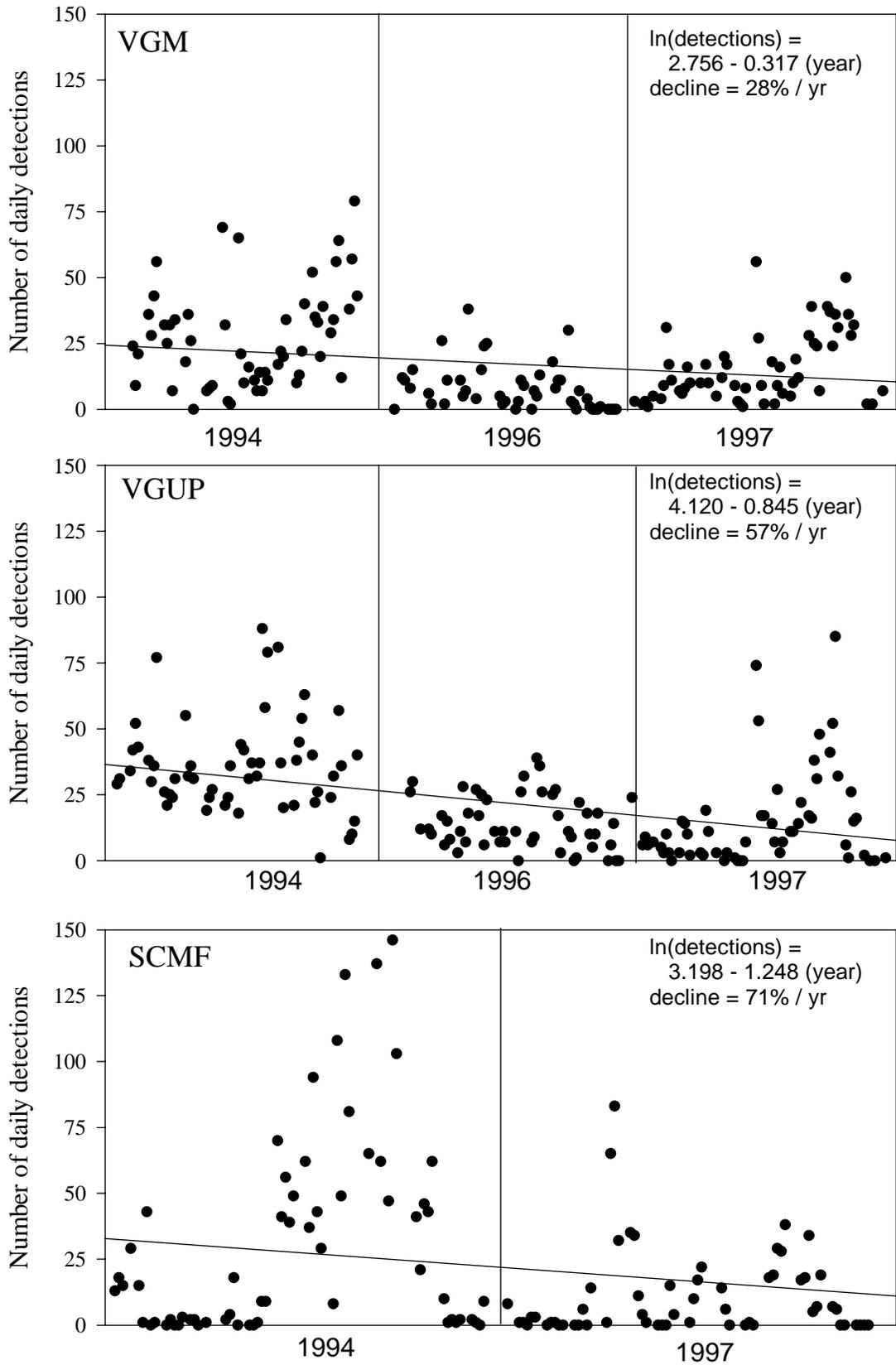


Figure 3.6.

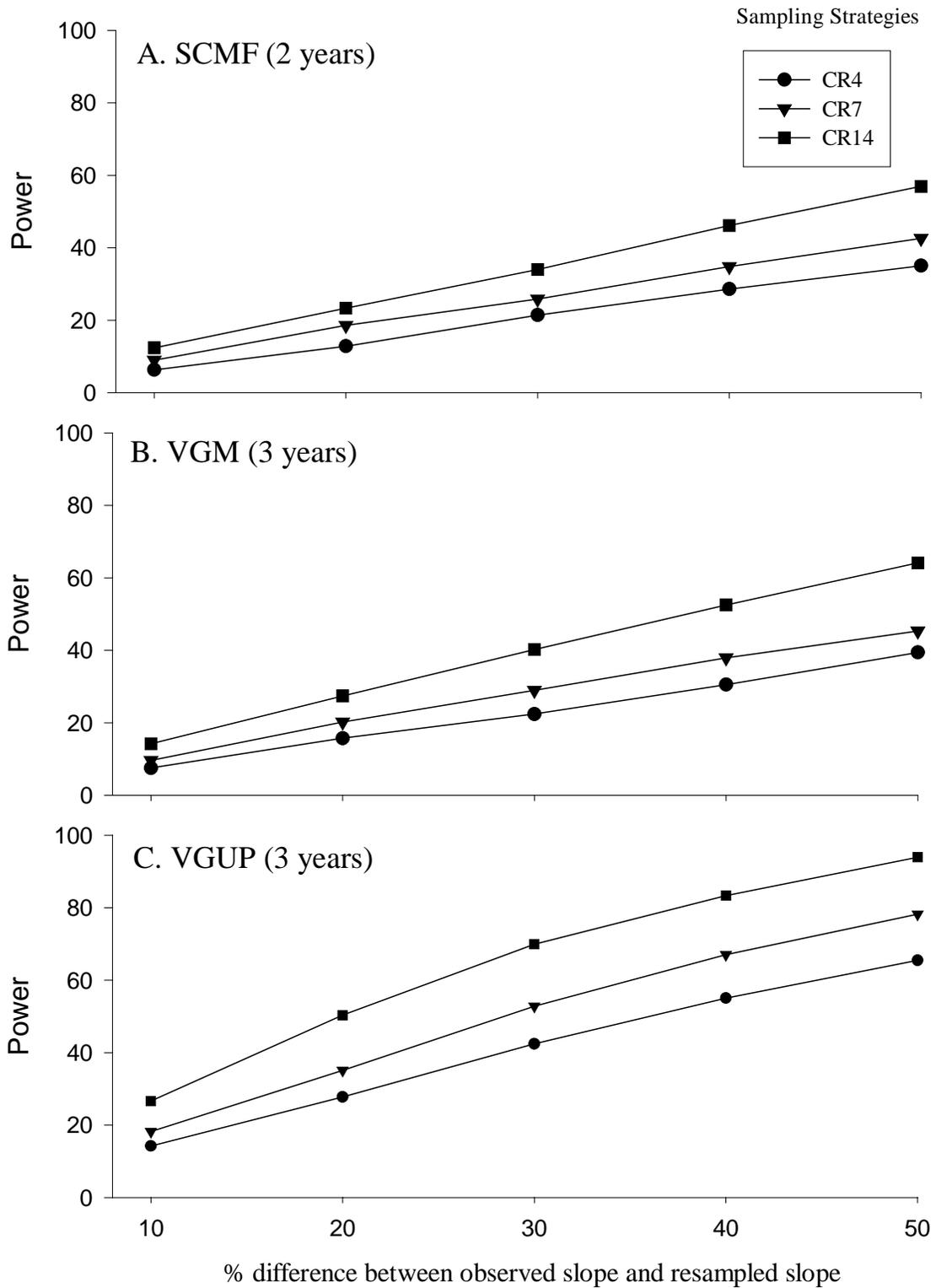


Figure 3.7. Power of resampled regressions to estimate known slopes from regressions of daily detections upon year (see Fig. 3.6) at three sites in the Oregon Coast Range: (A) SCMF 1994 & 1997, (B) VGM 1994, 1996, & 1997, and (C) VGUP 1994, 1996, and 1997.

Figure 3.8. Power of three strategies (CR4, CR7, and CR14; see Table 3.1) to estimate annual declines of 25% and 50% in daily detections using data generated from a gamma distribution (starting mean (i.e., year 1) = 50, CV = 0.45 and 0.85). Power = the proportion of the 1000 resampled surveys where the sample slope of detections regressed upon year was within $\pm 10, 20, 30, 40,$ or 50% of the observed slope which was built into the generated data sets for each year. (A) CV = 45%, years of survey = 3; (B) CV = 45%. Years of survey = 5; (C) CV = 85%, years of survey = 3, and; (D) CV = 85%, years of survey = 5.

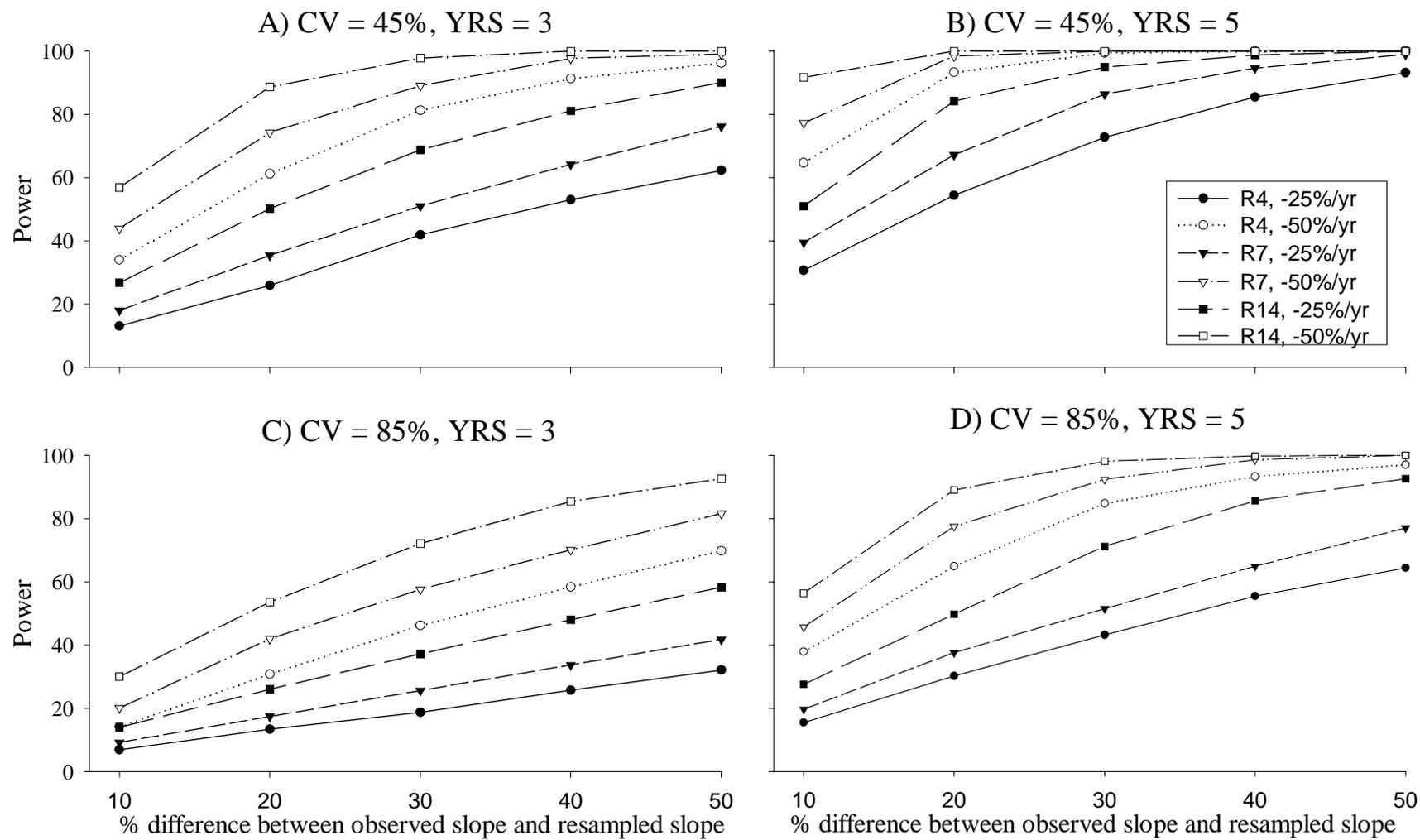


Figure 3.8.

decline/year was never adequate. Lastly, a low CV and 5 years of survey allowed even moderate survey efforts to produce sufficient power to detect the decline of 25% per year (Fig. 3.8b).

DISCUSSION

Implications for Monitoring Marbled Murrelets

Our analyses suggest that it is difficult to reliably estimate the mean and the variance of daily Marbled Murrelet detections at forested sites during a breeding season to within + 10 - 20%. Furthermore, it also appears that estimates of the mean and SD within + 50% are unlikely to be obtained except when survey effort is high (i.e., CR14) or SDs are low (i.e., < ca. 60%). Survey efforts of 4 - 7 days may prove reliable for wide accuracy windows if SDs are low (i.e., < 50%), although variability in daily murrelet detections tends to be quite high (Rodway et al. 1993, Jodice this document). Survey effort with < 7 days of effort will likely provide misleading data for detecting temporal differences in detections.

Temporal variability in daily Marbled Murrelet detections was higher in this study than previously reported for this species (Rodway et al. 1993, Chapter 2). High levels of daily variability of our data sets contributed strongly to the unreliable estimates provided by most of our survey strategies. This is supported by the inverse relationship observed between SD and reliability with generated data. Furthermore, the temporal variability in observed detections within years was the likely reason we did not observe a distinct difference in reliability between most temporally stratified and completely random survey strategies. In fact, preliminary explorations of 10 other temporally stratified and 2 other completely random survey strategies, each with 5 - 15 days of survey effort, yielded similar, unreliable results (Jodice unpublished data). Temporal patterns of variability in

detections also were not consistent within sites among years and this likely explains the differences in reliability of survey strategies within sites among years (e.g., P4, VGM '96 & '97, AW 20; Fig. 3.4b). The survey strategies we tested did not adequately account for the high levels of daily or annual variability in detections and therefore do not consistently provide reliable estimates of detection means and variances.

Our results suggest that interpretation of survey results may be improved by examining the type, magnitude, and direction of errors observed from resampled surveys (i.e., reliability matrix, Table 3.7). We observed that the direction and type of error encountered with survey strategies varied with the range of reliability chosen. Unreliable surveys tended to either under – or over- estimate both the mean and SD for accuracy window + 10% and tended to underestimate the SD for accuracy windows + 20 and 50%. Furthermore, reliability of survey strategies improved with sample size, indicating that sampling effort was too low in most cases. Typically when sample sizes are low we expect variability to be high, i.e., surveys strategies should overestimate variance. This is contrary to what we observed in many cases and what is assumed based on typical sampling results. Therefore, the errors defined in the reliability matrix can be used to formalize data regarding the direction and magnitude of the error and thus allow biologists to not rely merely on assumptions. Accordingly, sample sizes may be increased to provide more reliable estimates of mean and variance or, variance estimates intended for use in analyses (e.g., power) may be adjusted in the appropriate direction based on the data in the error matrix.

The accuracy window concept we employed is similar to 'effect' in power analysis, which is often referred to as the minimum detectable response that will be considered biologically significant (Steidl et al. 1997). Based on the similarity between these two concepts, it appears unlikely that the survey efforts we examined would both detect an effect of the same or lesser magnitude and simultaneously provide reliable estimates of the mean and variance. However, most resampled surveys using either field or generated data did reliably estimate detection means and variances to within + 50%. While this

level of reliability appears moderately useful at best, it does suggest that survey efforts of 4 - 14 days per breeding season could detect effects on the order of 50% at most sites in most years. This should allow biologists to at least detect changes in numbers of detections/day of catastrophic or extreme proportions (Zielinski and Stauffer 1996).

Given that single-year results indicated it was unlikely that most survey strategies could estimate means and variance measures to much better than $\pm 50\%$ most of the time, it was not surprising that regression analyses with multiple years of detection data were not very powerful when attempting to estimate similar magnitudes of annual declines in detections. Although most regression analyses from resampled surveys with field data did correctly identify the direction of the slope, the power of these resampled surveys to reliably estimate observed slopes tended to be low, even when criteria for estimating negative slopes were very relaxed ($\pm 50\%$) and sample sizes were relatively high (i.e., CR14). It is clear that, at the field sites, the magnitude and variability in annual CVs was an important factor in power determination; greatest power was achieved at VGUP where the average annual CV was least and, conversely, lowest power was achieved at SCMF where the average annual CV was greatest. This power ranking among sites occurred despite the fact that the known slopes being estimated were highest at SCMF and lowest at VGUP.

Power of regressions using generated data tended to be higher than those using field data. This likely occurred because the CV remained constant among years for regressions with generated data. As among-year variability increases it becomes more difficult to detect any significant trend (Sokal and Rolf 1987). However, these results do provide guidelines to consider for detecting annual trends in detection data. For example, an increase in survey effort from 3 to 5 years with either a low or moderate CV appears to sufficiently increase power so that substantially fewer surveys within each year are required. However, it is unreasonable to expect that declines of 25% per year could be reliably detected in all but the least variable conditions and with significant effort. Con-

versely, annual declines of 50% per year should be detectable under most of the conditions we simulated with generated data, keeping in mind these included a constant CV among years.

Implications for Use with Count Data

The concept of a reliability analysis as described herein is not necessarily new, although examples in the published literature are few (Schwagmeyer and Mock 1997). Reliability analyses may be viewed as an extension of pilot studies. For example, preliminary data may be used to test the effectiveness of various sampling strategies and subsequently justify the selected sampling strategy, a rarely documented decision (Beier and Cunningham 1996). Such analyses also may provide reliable estimates of variance for the metric of interest that can then be used in retro- and prospective power analyses (Steidl et al. 1997, Ribic and Ganio 1996). Additionally, reliability analyses have the advantage of being applicable to any metric. While statistics like the standard error of the mean provide, in essence, a measure of reliability for that metric, such statistics are not available for most metrics.

Another advantage of reliability analysis is its relative simplicity. We designed our approach in close collaboration with the Bureau of Land Management and in so doing attempted to maintain a data collection and analysis process that was easily repeatable. We also reasoned that, for a technique such as this to be useful, it should be easily understood by managers and ecologists that might not possess extensive statistical training. Our reliability analysis asks the basic question ‘how well do my data estimate what is really going on?’ and answers that question by simply stating the chance of obtaining a good or reliable estimate within some predefined range of acceptability. This is the major reasoning behind the use of accuracy windows, a concept we believe most managers will easily grasp and apply. This may be contrasted with ‘effect size’, the equivalent concept

in power analysis that requires an understanding and knowledge of the sample and population SD.

Lastly, we believe the extension of our reliability analysis to the generated data set broadens the scope of inference for our results. Our generated data represent any count data that fit a gamma distribution and are either free from the influences of temporal variation or cannot be surveyed in a manner to account for temporal variation (i.e., random sampling in time). As the gamma distribution tends to fit count data well, it appears that estimates of reliability and statements about sampling effort may be applicable to a wide range of studies that employ count data. Furthermore, it would not be difficult to apply this style of analysis to count data from statistical distributions other than the gamma, given that they provided a good fit to the data of interest (Beier and Cunningham 1996).

Management Recommendations

Based on the results of our analyses we suggest that the use of Marbled Murrelet detection data for quantitative analyses be limited and considered at great length prior to initializing research or management efforts. It appears that, given the range of data we tested, it would be difficult to obtain reliable estimates of murrelet detections with the observer-based method described and sampling efforts up to 14 days/season. However, it does appear that detection data during a breeding season may be reliably estimated to within + 50% with similar or less effort. Additionally, it may be feasible to detect declines of 50% per year over 3 - 5 years with substantial effort despite not reliably knowing the actual detection values.

We suggest that similar, long-term data sets be collected from other portions of the species range to document the degree of annual variability. Further analyses with RADAR will likely provide insight to quantifying and understanding patterns in variability as

well. Continued efforts to use radio-telemetry to monitor movements and gather behavioral data at inland forest stands also should be a priority.

Furthermore, although our analyses focused on temporal issues such as the ability to detect annual trends in detections, the techniques could have been used just as effectively with spatial issues given that variability due to observer and site-specific environmental differences could be accounted for. Therefore, using detection data to compare habitat quality among stands is prone to the same issues of reliability as using detection data to seek annual trends in activity over time within stands.

Our analyses do not suggest or make any changes to the current Marbled Murrelet survey protocol (Ralph et al., 1994). The main objective of the protocol is to determine presence and probable nesting status at inland forest sites and we did not examine the reliability or power of the protocol to accomplish that task. We do, however, caution managers and biologists against using detection data to seek temporal or spatial differences in Marbled Murrelet activity patterns without fully considering the implications of temporal variability.