Supplementary materials for this article are available at 10.1007/s13253-009-0017-7.

Improving Estimates of Abundance by Aggregating Sparse Capture-Recapture Data

Andrea R. LITT and Robert J. STEIDL

Inferences about abundance often are based on unadjusted counts of individuals observed, in part, because of the large amount of data required to generate reliable estimates of abundance. Where capture-recapture data are sparse, aggregating data across multiple sample elements by pooling species, locations, and sampling periods increases the information available for modeling detection probability, a necessary step for estimating abundance reliably. The process of aggregating sample elements involves balancing trade-offs related to the number of aggregated elements; although larger aggregates increase the amount of information available for estimation, they often require more complex models. We describe a heuristic approach for aggregating data for studies with multiple sample elements, use simulated data to evaluate the efficacy of aggregation, and illustrate the approach using data from a field study. Aggregating data systematically improved reliability of model selection and increased accuracy of abundance estimates while still providing estimates of abundance for each original sample unit, an important benefit necessary to maintain the design and sampling structure of a study. Within the framework of capture-recapture sampling, aggregating data improves estimates of abundance and increases the reliability of subsequent inferences made from sparse data. Additional tables and datasets may be found in the online supplements.

Key Words: Abundance estimation; Data aggregation; Mark-recapture; Program CAPTURE; Program MARK; Population parameters.

1. INTRODUCTION

Many ecological studies seek to make inferences about changes in population size over space, across time, or in response to experimental manipulations, and often base these inferences on counts of organisms that have not been adjusted for imperfect and varying detection probability. During a survey, many factors make it unlikely that all individuals

Andrea R. Litt (✉) is Post-Doctoral Research Associate, School of Natural Resources, University of Arizona, 325 Biological Sciences East, Tucson, AZ 85721, USA and is now Assistant Professor, Caesar Kleberg Wildlife Research Institute, Texas A&M University-Kingsville, 700 University Blvd., MSC 218, Kingsville, TX 78363, USA (E-mail: *andrea.litt@tamuk.edu*). Robert J. Steidl is Associate Professor, School of Natural Resources, University of Arizona, 325 Biological Sciences East, Tucson, AZ 85721, USA (E-mail: *steidl@ag.arizona.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 15, Number 2, Pages 228–247
DOI: 10.1007/s13253-009-0017-7

will be detected without error. Consequently, the ability to draw reliable inferences from counts depends on accounting for individuals not observed. The growing literature on estimating population parameters emphasizes the fundamental need to account for imperfect detectability to make inferences reliable (reviews in Seber 1982, 1986, 1992; Schwarz and Seber 1999).

For methods based on capture-recapture sampling, the framework and tools for modeling detection probability have become increasingly powerful and sophisticated. Within this framework, researchers ideally would generate estimates of abundance by modeling variation in detection probability for each species, sample plot, and sampling period in a study. Choosing an appropriate model for detection probability as the basis for generating estimates of abundance, however, requires a large amount of data (Otis et al. 1978; Rosenberg, Overton, and Anthony 1995). Therefore, even when sampling effort is high, these data demands may be impossible to meet when population sizes are naturally small or detectability is low (e.g., McKelvey and Pearson 2001; Bowden et al. 2003; MacKenzie et al. 2005).

When data are insufficient to reliably choose a model for detection probability for each species, sample plot, and sampling period, studies often rely on unadjusted counts or enumeration statistics, such as the number of unique individuals captured or the total number of captures, to draw inferences about relative differences in abundance over space or time (McKelvey and Pearson 2001). This approach has been vigorously criticized because it assumes detection probabilities are equal among groups being compared, an assumption that is likely to be met only in rare circumstances (Nichols 1992; MacKenzie and Kendall 2002). Further, variation in detection probability among species makes interspecific and community-scale comparisons based on unadjusted counts unreliable (Nichols 1986).

At least three methods have been used commonly to overcome the high data requirements for estimation procedures. One method is to choose a single model for detection probability that is then applied to all sample units (e.g., Rosenberg, Swindle, and Anthony 2003). A second solution is to use traditional hypothesis tests (Skalski, Robson, and Simmons 1983) or equivalence tests (MacKenzie and Kendall 2002) to assess the assumption of equal detection probability to justify the use of unadjusted counts. A third method is to estimate detection probability for a spatial or temporal subset of sample units where data are sufficient and use these estimates of detection probability to generate estimates of abundance for the remaining sample units (e.g., Lynam et al. 2009). With sparse data, however, there may be little information with which to rigorously evaluate the reliability of any of these alternatives. Consequently, metrics with lower data requirements than abundance estimation, such as occupancy or species richness, have increased in popularity (MacKenzie et al. 2005). If study objectives dictate inferences based on abundance, however, these metrics may not be suitable alternatives.

When abundance is the parameter of interest and data are sparse, an additional solution is to aggregate or pool data to increase the information available to generate estimates of abundance that have been adjusted for detection probability. Initial approaches to aggregation focused on simplifying the sampling structure within which the data were collected by pooling data across capture occasions, sites, or sampling periods in ways such that elements of the original sample structure were lost (e.g., Hargrove and Borland 1994). Contempo-

rary approaches to aggregation, such as we describe here, result in no loss of information or structure and use all available data as part of the aggregation process (Boyce et al. 2001; Bowden et al. 2003; MacKenzie et al. 2005; White 2005; Conn et al. 2006).

Although aggregating data from multiple sample units is a common practice by experts in population analysis and some basic information is available on the process (Burnham and Anderson 2002; MacKenzie et al. 2005; White 2005; Conn et al. 2006), our goal is to encourage increased use of data aggregation in practice by providing a clear, synthetic description of the process and making these methods accessible to a wider range of ecologists. Therefore, we describe a heuristic approach that uses biological and empirical information to guide the aggregation process for studies based on capture-recapture sampling. We develop a general framework for aggregation, use simulated data to examine its efficacy, and illustrate the approach with field data.

2. OVERVIEW OF DATA AGGREGATION

Data aggregation involves assembling data from multiple sample units or "elements" into a single dataset to increase the information available for selecting an appropriate model for detection probability as the basis for estimating abundance. Elements to consider for aggregation will vary by study, but might include data collected from the same sample unit over time, from multiple sample units over space, from multiple species (MacKenzie et al. 2005; White 2005), or even data from different studies, especially for rare species. Aggregating data from multiple sample elements assumes that one model can be used effectively to describe the different processes driving variation in detection probability for all elements in the aggregate. In studies where data arise from a complex set of sample elements, the decision as to how best to aggregate data for estimation involves considering trade-offs related to the size of the aggregate. In general, larger aggregates are more likely to combine sample elements that vary with respect to the processes that drive detection probability (e.g., heterogeneity, behavior, time). More complex models and larger datasets (i.e., more individuals) are needed to describe multiple detection processes and to represent the more complex sampling structure of the elements combined (e.g., species, seasons, plots) in larger aggregates. In contrast, in smaller aggregates sample elements are likely to be more homogeneous with respect to detection processes and have simpler sampling structures, requiring simpler models for detection probability and, correspondingly, less data. Therefore, the process of aggregating data should seek to balance the benefits of increased information available in larger aggregates with the increased complexity resulting from combining sample elements with disparate detection processes and more complex sampling structures. For studies with many sample elements where several potential aggregates are possible, we suggest that the process of aggregation should begin by considering biological information as the basis for refining the set of potential aggregates, using available data to explore the complexity of processes driving detection probability in potential aggregates, and using model-selection procedures to choose among candidate models for detectability.

2.1. CONSIDER BIOLOGICAL INFORMATION

Decisions about which sample elements to aggregate should begin by considering biological information about the species, environmental variation, and sampling structure of the study to help narrow the range of possible aggregates (Alldredge et al. 2007) and reduce the complexity and number of candidate models (Figure 1). Biological information to consider should include life-history attributes of a species that could affect the processes driving variation in detection probability, such as whether the species is known to respond behaviorally to trapping or whether the processes might be expected to vary seasonally or during periods of reproductive activity. For example, if one species responds behaviorally to trapping and another species displays temporal variation in detection probability, a simple model may not accurately describe variation in detection probability for both species. Aggregating elements with many different detection processes may bias estimates, the magnitude of which will depend in part on the robustness of the estimator.

To illustrate this issue with a simple example, we created two datasets, each representing a population with true abundance of 100 sampled in a single survey with five sampling occasions. In one set, detection probability changed behaviorally in response to trapping (probability of initial capture $= 0.2$, probability of recapture $= 0.6$, $\{p(\cdot), c(\cdot)\}$, Table 1) and in the other set it changed temporally (probability of capture on day $1 = 0.2$ increasing by 0.1 each subsequent day, $\{p(t) = c(t)\}$). We aggregated data from these two heterogeneous sample elements and when we estimated abundance using a model with temporal variation in detection probability ($\{p(t) = c(t)\}$), the estimate for the element with be-

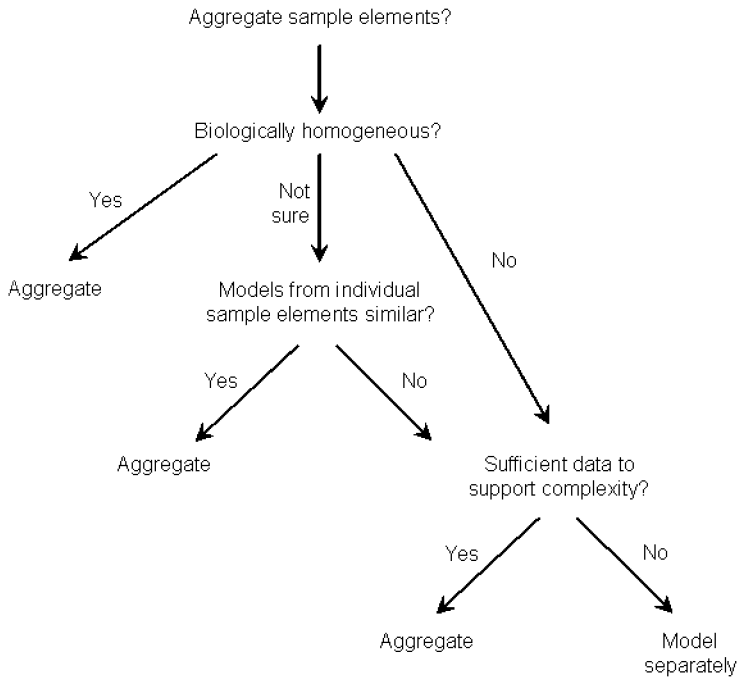

Figure 1. Decision tree for the process of aggregating data from multiple sampling elements.

Table 1. Notation used for general models of detection probability.

Processes driving variation in detection probability	Otis et al. (1978) notation	Expanded notation
Constant (null)	M_o	$\{p(\cdot) = c(\cdot)\}$
Behavioral response	M_b	$\{p(\cdot), c(\cdot)\}$
Heterogeneity	M_h	$\{p_a(\cdot) = c_a(\cdot), p_b(\cdot) = c_b(\cdot), \pi\}$
Temporal	M_h	$\{p(t) = c(t)\}$
Behavioral response, heterogeneity	M_{bh}	$\{p_a(\cdot), c_a(\cdot), p_b(\cdot), c_b(\cdot), \pi\}$
Temporal, heterogeneity	M_{th}	$\{p_a(t) = c_a(t), p_b(t) = c_b(t), \pi\}$
Temporal, behavioral response	M_{tb}	$\{p(t), c(t)\}$
Temporal, behavioral, heterogeneity	M_{tbh}	$\{p_a(t), c_a(t), p_b(t), c_b(t), \pi\}$

NOTE: Expanded notation describes model parameters: p = probability of capture, c = probability of recapture, and π = probability of belonging to a subgroup of animals (mixture) that has homogeneous detection probability. Probability of capture and recapture may be constant $(\cdot)$ or may vary over time (t), be based on a behavioral response to trapping $(b,\ p \neq c)$, or vary among heterogeneous mixtures (h, two mixtures denoted by $_a$ and $_b$). In subsequent tables, if recapture parameters (c) are not specified, then $p = c$.

havioral variation was negatively biased by 33%, whereas the estimate for the element with temporal variation was positively biased by 5%. The estimator for temporal variation in detection probability is not robust to variation due to behavioral responses to trapping (Otis et al. 1978), demonstrating the potential pitfalls of aggregating across elements where detection probability is heterogeneous and that these elements might be better modeled independently. Although a more complex model that included both temporal and behavioral variation could be used to model detection probability for these aggregated elements ($\{p(t), c(t)\}$), modeling complex aggregates requires considerably larger datasets. As such, considering biological characteristics of sample elements can reduce complexity and potential biases associated with aggregating disparate elements.

2.2. Model Detection Probability

When the number of elements to aggregate and the number of candidate models remain large after biological information has been considered, fitting simple "general" models that describe the basic processes driving detection probability (Table 1) to individual sample elements that are data rich can help to further reduce the number of potential aggregates and candidate models (Table 1, Figure 1). If similar models emerge for individual elements, then it seems reasonable to aggregate these elements. For example, if similar models emerge regardless of season, aggregation across seasons should be effective and introduce little complexity to the aggregate. In contrast, if many different models emerge for elements sampled in different seasons, modeling these potentially disparate elements within a single aggregate will require a more complex model and larger datasets, suggesting that aggregating across seasons may be less effective than creating separate aggregated datasets for each season to reduce model complexity. Ideally, information used to assess the appropriateness of different aggregates would be gleaned from the current study; however, previous studies might also help to guide the process. This step should be considered

exploratory, as these smaller datasets may not be sufficient to identify reliably complex variation in detection processes.

After establishing aggregates, the set of candidate models that will be used to model detection probability for the aggregated data are developed. Starting with simple general models that describe the basic processes driving detection probability (Table 1), classification variables or covariates that represent the structure of the sample elements combined in the aggregate are incorporated into models. These "flexible" models can be created by denoting each aggregated element as a unique "group" in the input dataset containing capture histories, building design matrices to identify the structure of the individual sample elements, and incorporating additive or multiplicative terms to constrain variables and effectively create a wide range of related flexible models. As such, values for detection probability can be allowed to vary for individual elements or sets of elements in the aggregate. The process of starting with a general model then adding parameters to increase model flexibility for capture-recapture data is analogous to selecting a key function then incorporating a series expansion within the context of distance sampling (Buckland et al. 1993); that is, adding parameters increases the ability of the model to describe variation inherent in the data.

The number and complexity of flexible models can increase quickly as the number of elements aggregated and the number of classification variables increase, which can easily require hundreds of parameters. For example, fitting a model with both temporal variation and heterogeneity in detection probability for data with five capture occasions requires as many as 11 parameters for just one sample element ($\{p_a(t) = c_a(t),\ p_b(t) = c_b(t), \pi\}$). For an aggregate that includes data for one species from five plots sampled in three seasons, the fully multiplicative time and heterogeneity model for these 15 sample elements requires 165 parameters ($\{p_a(t*season*plot) = c_a(t*season*plot),\ p_b(t*season*plot) = c_b(t*season*plot), \pi\}$). Thus, using biological information to refine the aggregate and candidate model set is necessary to ensure that the number of models, the number of model parameters, and the size of design matrices remain manageable. In addition, detecting very complex patterns of variation in detection processes require rich datasets (Anderson, Burnham, and White 1994). Once the set of candidate flexible models has been refined and fit, support for these models can be assessed within an information-theoretic framework (Figure 1).

2.3. GENERATE ESTIMATES

Estimates of abundance could be generated based on the flexible model with the most support (e.g., smallest AIC_c) or averaged across competing models to account for uncertainty in model selection (Burnham and Anderson 2002). Because aggregated elements are identified uniquely as groups within the input dataset, unique estimates of abundance can be generated for each sample element (Figure 1) as data are pooled only for the purposes of improving estimates. These element-specific estimates can be used within a hypothesis-testing framework to address questions about treatment effects because the experimental units and design structure for randomized experiments has been retained and traditional

estimates of experimental error can be used as a basis for inference. Although other approaches also could be used to answer these questions, a frequentist approach offers advantages within the context of randomized, replicated designs (Burnham and Anderson 2002).

3. EFFICACY OF AGGREGATING

3.1. SIMULATION METHODS

We used simulations to explore consequences of aggregating data for two challenges associated with sparse data: choosing an appropriate model for detection probability and generating accurate estimates. Simulated survey data were based on five trapping occasions, a duration common for small mammal studies (McKelvey and Pearson 2001). We created aggregates that included three sample elements that could represent multiple surveys of the same sample unit or three sample units surveyed simultaneously, for example.

We manipulated three factors in simulations: (1) the true flexible model used to generate the data, (2) overall detection probability, and (3) true abundance (N) of the entire aggregate. We modeled three types of variation in detection probability to represent a range of detection processes: heterogeneity, behavioral response, and temporal variation. For simulated data with heterogeneity, we established two groups of equal size ($\pi = 0.5$), each with different detection probabilities (Appendix 1). For simulated data with a behavioral response, we set recapture probability to be higher than the probability of initial capture, a trap-happy response (Appendix 2). For simulated data with temporal variation, we set detection probability to be lowest on the first occasion, slightly higher and constant on the second through fourth sampling occasions, and highest on the fifth occasion (Appendix 3), a scenario that might reflect studies where trapping begins without a prebaiting period. We explored different flexible models to generate data where detection probability among elements in the aggregate was (1) constant (e.g., $\{p(t) = c(t)\}$), (2) varied additively (e.g., $\{p(t + group) = c(t + group)\}$), and (3) varied multiplicatively (e.g., $\{p(t * group) = c(t * group)\}$). We examined two levels of detection probability (low and high, Appendices 1–3) and six values of true abundance for the entire aggregate (60, 150, 300, 600, 1200, or 1500 individuals), with true abundance varying among the three elements (each element was 1/3, 1/6, 1/2 of the total true abundance) in the aggregate. The resulting aggregate size (number of individuals captured) was a function of true abundance of the aggregate and detection probability. We considered all levels of factors in all combinations, yielding 84 sets of simulations, and used the identity link function to establish parameter values for generating models.

For each combination of factors, we used the Huggins closed-capture simulation platform in Program MARK (version 5.1, White and Burnham 1999) to generate 1000 datasets. To assess the consequences of aggregation on model fitting and estimation, for each dataset we generated an estimate of abundance for each element based on four types of models of detection probability: (1) the generating model used to create the data (e.g., $\{p(t + group) = c(t + group)\}$), (2) other flexible models based on the same general model

(e.g., $\{p(t) = c(t)\}$, $\{p(t * group) = c(t * group)\}$), (3) flexible models with constant detection probability, as they are likely to emerge with sparse data and because they are relatively robust to temporal variation in detection probability (Otis et al. 1978) ($\{p(\cdot) = c(\cdot)\}$, $\{p(\cdot + group) = c(\cdot + group)\}$), and (4) other flexible models incorrectly specifying variation in detection probability (e.g., based on general models with heterogeneity and behavior for generating models with temporal variation). We fit a total of nine candidate models (Table 2) to each set of simulated data and used the logit link function for estimation.

For each dataset, we generated a list of competing models (defined as $\Delta\text{AIC}_c \leq 2$) and estimates of abundance for each of the three aggregated elements. To determine how aggregation affected selection of an appropriate model, we computed the percentage of times the true generating model was among the list of competing models for all datasets. To determine how aggregation affected bias of estimates, we computed the average absolute value of percent relative bias (PRB) of abundance estimates for each element in the aggregate for all competing models. To determine how aggregation affected precision of estimates, we computed the interquartile range (IQR) for PRB.

3.2. SIMULATION RESULTS AND DISCUSSION

As true abundance of the aggregate and detection probability increased, the frequency with which the generating model was chosen increased (Table 2). When generating models included heterogeneity (e.g., $\{p_a(\cdot), p_b(\cdot), \pi\}$), other models were selected more often than the true model unless aggregate size was large (Table 2). Further, if the generating model was complex (e.g., included multiplicative effects), both true abundance and detection probability had to be large before the generating model was selected consistently (Table 2), especially when the generating model included heterogeneity in detection probability. When generating models included heterogeneity and capture probabilities were low, generating models were never selected most often, regardless of the aggregate size. Instead, simpler models with constant detection probabilities ($\{p(\cdot) = c(\cdot)\}$, $\{p(\cdot + group) = c(\cdot + group)\}$) were selected most frequently (Table 2).

As expected, estimates of abundance were relatively consistent and unbiased when estimated with the correct model; as aggregate size increased, precision increased and bias decreased (Table 3). Even when estimates were generated with competing models other than the true model, estimates were usually consistent and reasonably unbiased (Table 3). The exception was when both true abundance and detection probability were low and a competing model incorrectly included behavioral variation, when estimates were more variable and had higher bias (Table 3). When generating models included heterogeneity, estimates from the true model had lower precision and higher bias than estimates from competing models (Table 3). In almost all circumstances, however, aggregating data improved accuracy of estimates by improving selection of an appropriate model, especially when detection probabilities were high. Even when based on competing models rather than the generating model, estimates usually provided acceptable accuracy, which is important given that for real data the "true" model is unknown.

Note that we only evaluated aggregates that included sample elements subject to the same general detection process; the efficacy of aggregating sample elements with different

Table 2. Percentage of times where each candidate model was among the list of competing models ($\Delta AIC_c \leq 2$) based on 1000 simulated datasets. We explored the results under seven different generating models as well as low and high values for detection probability. Results from each generating model are identified in bold.

Detection probability	N	$\{p_a(\cdot), p_b(\cdot), \pi\}$	$\{p_a(\cdot * group), p_b(\cdot * group), \pi\}$	$\{p(\cdot), c(\cdot)\}$	$\{p(\cdot + group), c(\cdot + group)\}$	$\{p(t)\}$	$\{p(t + group)\}$	$\{p(t * group)\}$	$\{p(\cdot)\}$	$\{p(\cdot + group)\}$
Low	60	**6**	1	31	3	6	2	1	38	14
	150	**9**	1	30	4	7	2	1	36	13
	600	**16**	3	27	3	6	2	1	33	11
	1500	**24**	4	23	3	5	2	1	29	10
High	60	**14**	2	26	4	6	2	0	32	13
	150	**25**	2	24	3	4	1	0	29	11
	600	**65**	6	9	2	3	1	0	10	4
	1500	**91**	9	0	0	0	0	0	0	0
Low	60	6	**2**	27	3	6	3	1	35	17
	150	7	**1**	26	5	5	3	1	32	21
	600	9	**6**	12	9	3	7	1	17	36
	1500	6	**15**	1	14	1	9	1	3	51
High	60	13	**3**	20	6	5	3	1	27	23
	150	17	**6**	11	9	3	6	1	16	33
	600	9	**44**	1	9	0	6	1	1	31
	1500	0	**97**	0	1	0	1	0	0	1
Low	60	4	1	**41**	3	8	2	1	31	11
	150	3	1	**51**	7	11	4	1	16	6
	600	0	0	**73**	14	8	3	1	0	0
	1500	0	0	**78**	19	2	1	0	0	0

Table 2. (Continued.)

Detection probability	N	$\{p_a(\cdot), p_b(\cdot), \pi\}$	$\{p_a(\cdot * group), p_b(\cdot * group), \pi\}$	$\{p(\cdot), c(\cdot)\}$	$\{p(\cdot + group), c(\cdot + group)\}$	$\{p(t)\}$	$\{p(t + group)\}$	$\{p(t * group)\}$	$\{p(\cdot)\}$	$\{p(\cdot + group)\}$
High	60	2	1	**65**	9	9	3	1	8	3
	150	0	0	**76**	15	7	2	0	0	0
	600	0	0	**83**	17	0	0	0	0	0
	1500	0	0	**38**	17	0	0	0	0	0
Low	60	4	1	29	**5**	5	7	2	23	25
	150	2	2	26	**20**	4	15	1	8	23
	600	0	1	1	**75**	0	21	1	0	2
	1500	0	0	0	**94**	0	5	1	0	0
High	60	2	1	41	**28**	4	11	2	3	9
	150	0	0	25	**64**	1	9	1	0	1
	600	0	0	0	**100**	0	0	0	0	0
	1500	0	0	0	**100**	0	0	0	0	0
Low	60	2	1	34	1	**23**	7	2	22	8
	150	0	0	29	3	**44**	15	2	6	2
	600	0	0	2	0	**68**	27	2	0	0
	1500	0	0	0	0	**72**	26	2	0	0
High	60	1	0	28	3	**42**	14	2	7	2
	150	0	0	8	1	**66**	22	3	0	0
	600	0	0	0	0	**70**	27	3	0	0
	1500	0	0	0	0	**70**	27	3	0	0

Table 2. (Continued.)

Detection probability	N	$\{p_a(\cdot), p_b(\cdot), \pi\}$	$\{p_a(\cdot * group), p_b(\cdot * group), \pi\}$	$\{p(\cdot), c(\cdot)\}$	$\{p(\cdot + group), c(\cdot + group)\}$	$\{p(t)\}$	$\{p(t + group)\}$	$\{p(t * group)\}$	$\{p(\cdot)\}$	$\{p(\cdot + group)\}$
Low	60	2	1	24	2	16	**21**	4	13	18
	150	0	1	10	4	12	**59**	7	1	6
	600	0	0	0	0	0	**78**	22	0	0
	1500	0	0	0	0	0	**60**	40	0	0
High	60	1	1	17	9	23	**36**	3	4	8
	150	0	0	2	5	10	**77**	6	0	0
	600	0	0	0	0	0	**91**	10	0	0
	1500	0	0	0	0	0	**88**	12	0	0
Low	60	2	0	34	2	20	8	**2**	23	10
	150	1	0	23	6	34	22	**4**	6	3
	600	0	0	1	2	32	46	**19**	0	0
	1500	0	0	0	0	5	34	**61**	0	0
High	60	1	1	22	7	30	21	**5**	8	6
	150	0	0	6	7	29	43	**14**	0	0
	600	0	0	0	0	1	29	**70**	0	0
	1500	0	0	0	0	0	1	**99**	0	0

NOTE: Detection probabilities are provided in Appendices 1–3. We show complete results for N = 60, 150, 300, 600, 1200, and 1500 in Appendix 4.

Table 3. Bias (average absolute value of percent relative bias) and precision (interquartile range for percent relative bias) for estimates from the true generating model and from all competing, but incorrect, models for simulated datasets.

True model	Detection probability	N	Bias (%)		Precision (%)	
			True model	Competing models	True model	Competing models
$\{p_a(\cdot), p_b(\cdot), \pi\}$	Low	60	17	15	22	20
		150	26	10	18	13
		600	15	5	10	6
		1500	12	5	7	4
	High	60	15	6	10	9
		150	9	4	8	5
		600	5	4	4	3
		1500	2	14	3	5
$\{p_a(\cdot * group), p_b(\cdot * group), \pi\}$	Low	60	28	16	26	22
		150	20	13	24	16
		600	15	8	9	9
		1500	12	6	7	6
	High	60	18	7	10	10
		150	11	5	8	7
		600	9	3	5	4
		1500	7	3	4	3
$\{p(\cdot), c(\cdot)\}$	Low	60	40	22	44	19
		150	17	24	25	14
		600	7	21	12	20
		1500	5	13	8	20
	High	60	10	16	15	14
		150	6	14	9	13
		600	3	5	5	8
		1500	2	3	3	5
$\{p(\cdot + group), c(\cdot + group)\}$	Low	60	58	30	52	33
		150	42	24	39	27
		600	23	22	19	23
		1500	15	22	13	25
	High	60	23	12	21	17
		150	12	9	12	16
		600	5	12	6	10
		1500	3	*	4	*
$\{p(t * group)\}$	Low	60	13	34	19	33
		150	8	24	11	24
		600	4	6	6	7
		1500	2	3	4	4
	High	60	4	8	5	8
		150	3	4	4	5
		600	1	1	2	2
		1500	1	1	1	1
$\{p(t)\}$	Low	60	10	41	17	43
		150	7	41	11	48
		600	3	8	5	9
		1500	2	3	3	5

Table 3. (Continued.)

True model	Detection probability	N	Bias (%)		Precision (%)	
			True model	Competing models	True model	Competing models
$\{p(t)\}$	High	60	4	10	6	11
		150	2	6	4	7
		600	1	1	2	2
		1500	1	1	1	1
$\{p(t + group)\}$	Low	60	18	42	19	38
		150	9	39	12	37
		600	4	5	6	6
		1500	3	3	4	4
	High	60	4	14	4	10
		150	3	11	4	7
		600	1	1	2	2
		1500	1	1	1	1
$\{p(t * group)\}$	Low	60	13	34	19	33
		150	8	24	11	24
		600	4	6	6	7
		1500	2	3	4	4
	High	60	4	8	5	8
		150	3	4	4	5
		600	1	1	2	2
		1500	1	1	1	1

NOTE: Detection probabilities are provided in Appendices 1–3. We show complete results for $N = 60$, 150, 300, 600, 1200, and 1500 in Appendix 5. $*$ = no competing models.

detection processes will vary, in part, with complexity of the aggregate, amount of data available, and robustness of individual estimators. Therefore, during the design phase of a study, all efforts should be made to increase capture success (e.g., prebaiting, using a sufficient number of traps with appropriate spacing), as accuracy of model selection and the resulting estimates increase appreciably as detection probability increases.

4. CASE STUDY

4.1. FIELD METHODS

To illustrate the process of data aggregation, we explored data collected to quantify how abundance of small mammal populations varied in response to differences in dominance of nonnative grass cover and prescribed fire. Data were a subset from a larger study (Litt 2007) collected between spring 2000 and winter 2002 in grasslands of southern Arizona on 27 study plots established in areas with three levels of nonnative grass: (1) dominated by nonnative grass (nonnative), (2) dominated by native grass (native), and (3) a mixture of native and nonnative grasses (mixed), with nine plots established at each level. Plots were randomly assigned to one of three fire treatments: (1) no fire, (2) fire in spring 2001, or (3) fire in summer 2001.

We trapped small mammals for five consecutive nights during sampling periods in spring, summer, and winter each year. We used an 8×8 grid of Sherman live traps on each plot and marked individuals uniquely. Because richness of small mammals in this grassland community is high (24 species) and trapping grids relatively small, we rarely had sufficient data to reliably choose models to estimate abundance by species, plot, and sampling period, even for the most common species. Because we wanted to generate plot-specific estimates of abundance to examine treatment effects from this replicated experiment, we aggregated data to facilitate choosing models to generate estimates of abundance.

4.2. CASE STUDY RESULTS AND DISCUSSION

4.2.1. Considering Biological Information

We considered potential aggregates based on data pooled across species (24 total species), sampling seasons (winter, spring, summer), sampling years (2000–2002), vegetation composition (nonnative, mixed, native), and fire treatments (no fire, spring fire, summer fire). A single aggregated dataset that combined all of these elements would include 3,888 different sample elements. Consequently, we first considered biological and empirical information as a basis to restrict candidate aggregates and refine the set of candidate models (Figure 1).

The 24 species in 12 genera that we captured offered multiple potential aggregates. We restricted potential aggregates to species within the same genus, assuming that variation in detection probability was more likely to be driven by the same processes for congeners than noncongeners. This reduced the number of sample elements in the aggregate for the genus *Perognathus*, for example, with two species, to 324 elements. To evaluate whether variation in detection probability seemed to be driven by similar processes, we first considered variation in morphology and behavior. For the two species of *Perognathus*, *P. flavus* (silky pocket mouse) and *P. hispidus* (hispid pocket mouse), *P. flavus* averaged 7.5 g total mass (SE = 0.05, $n = 1297$) and 56.9 mm body length (SE = 0.12), whereas *P. hispidus* averaged 33.3 g (SE = 0.28, $n = 1418$) and 96.8 mm (SE = 0.33). Individual *P. flavus* were captured on fewer occasions per sampling period (mean = 1.6 occasions, SE = 0.4) than *P. hispidus* (mean = 2.6, SE = 0.4). *P. hispidus* also enters torpor during winter, emerging only on particularly warm days, a pattern we did not observe with *P. flavus*. Although detection processes for landbirds surveyed with point-count methods can be similar (Alldredge et al. 2007), detection processes for small mammals can vary widely among species (Hammond and Anthony 2006). This difference suggests that aggregating data across small mammal species will usually require fitting complex models. Given that these biological differences between species were likely to affect patterns of detection, we did not aggregate data across species in the same genus to simplify models for aggregates. Instead, we considered a separate aggregated dataset for each species, which reduced the number of sample elements in each aggregate to 162.

We also considered torpor in *P. hispidus* as a reason not to aggregate across sampling seasons as major seasonal differences in activity could result in different processes driving variation in detection probability. For example, detectability for *P. hispidus* may vary daily

with variation in temperature during winter, where a temporal model of detection probability might be most appropriate. This model might be less appropriate in other seasons when these animals are consistently active, more detectable, and somewhat trap-happy. We chose to apply the same aggregation strategy for all species; therefore, if a level of aggregation was not supported for some species or circumstances, we rejected that level for all species. As such, we considered a separate aggregated dataset for each species and each season, reducing the number of elements in each aggregate to 54.

We could not envision a biological reason why the process driving variation in detection probability would differ among years for a given season or vary with the amount of nonnative grass, therefore, for each species we aggregated data over years and over levels of vegetation composition. Because we were interested in comparing changes in abundance in response to prescribed fire treatments, we also considered aggregating over fire season. By including plots that received different fire treatments in the same aggregate, we avoided confounding potential biases due to model choice with any treatment effects. Therefore, we created aggregated datasets for each species and season that included multiple sampling years, all levels of vegetation composition, and all fire treatments.

4.2.2. Modeling Detection Probability

Although 54 aggregated elements and the resulting candidate models seemed relatively reasonable, we explored support for this candidate aggregate empirically, using the richest datasets from potential elements—plots or sets of plots that received the same treatment and were sampled at the same time. For *P. flavus*, we considered eight datasets, each with 45–101 captured individuals (Table 4). These datasets represented all sampling years (5 datasets for 2000, 2 for 2001, and 2 for 2002), all categories of vegetation composition (5 datasets for native, 2 for mixed, and 1 for nonnative), but only unburned areas. We used these data to explore eight general models for detection probability (Table 1) and gauged consistency in the set of competing models ($\Delta AIC_c \leq 2$) for datasets from different years, vegetation composition categories, or fire treatments (Table 4). A general model with temporal variation was among competing models for 7 of 8 datasets and was the model with

Table 4. Competing general models ($\Delta AIC_c \leq 2$) and number of individuals captured (M_{t+1}) for individual datasets for *Perognathus flavus*.

Data set	Season	Year	Vegetation	M_{t+1}	Competing models
1	Spring	2000	Mixed	45	$\{p(\cdot), c(\cdot)\}, \{p(\cdot)\}, \{p(t)\}$
2	Winter	2002	Native	45	$\{p(t)\}$
3	Summer	2000	Nonnative	46	$\{p(\cdot)\}, \{p_a(\cdot), p_b(\cdot), \pi\}, \{p(\cdot), c(\cdot)\}$
4	Spring	2001	Native	48	$\{p(\cdot)\}, \{p(t)\}$
5	Summer	2001	Native	61	$\{p(t)\}$
6	Summer	2000	Native	87	$\{p_a(\cdot), p_b(\cdot), \pi\}, \{p(t)\}$
7	Spring	2000	Native	97	$\{p(t)\}$
8	Summer	2000	Mixed	101	$\{p(t)\}, \{p_a(t), p_b(t), \pi\}$

NOTE: Datasets included sets of plots sampled at the same time and that received the same treatment. Competing general models are listed in order of increasing AIC_c value.

the lowest AIC_c value for 4 of 8 datasets (Table 4). For years, a model with temporal variation was among competing models for all three years, suggesting that variation in detection probability might be driven by similar processes. For vegetation, temporal variation in detection probability was clearly evident in two of three categories and suggestive for the third (within 2.5 of the smallest AIC_c) (Table 4, dataset 3), again suggesting that a single process would likely be reasonable to describe variation in detection probability, regardless of vegetation composition. As such, the empirical information we considered seemed to support the level of aggregation suggested based on biological information.

Ultimately, we created three aggregated datasets for each species, one each for winter, spring, and summer sampling seasons, and used these to generate estimates of abundance, which we illustrate for *P. flavus* during summer. Each aggregate was comprised of 27 individual plots sampled in each of two years, resulting in 54 aggregated elements, each element identified as a unique group in the input dataset. We considered seven general models (first seven models in Table 1) and constructed associated flexible models that incorporated both additive and multiplicative terms to represent aggregated elements (sampling year, vegetation composition, application of fire, and fire season; examples provided in Appendix 6). We created design matrices in Program MARK that incorporated classification variables to identify aggregated elements and built models by constraining specific variables in the design matrix (Appendix 6).

4.2.3. Generating Estimates

After fitting all candidate models, there were several competing models, all of which were based on two general models, one with heterogeneity and temporal variation and the other with temporal variation only (Table 5). All competing models indicated that detection probability varied with differences in vegetation composition. Because there was support for several flexible models, we generated model-averaged estimates of abundance to provide estimates of abundance for the original sample elements in the aggregate (Table 6).

Table 5. Flexible models (shown where AIC_c weight ≥ 0.01) used to generate model-averaged estimates of abundance for the aggregate of plots sampled in summer for *Perognathus flavus* ($M_{t+1} = 319$).

Flexible model	ΔAIC_c	AIC_c weight	No. parameters
$\{p_a(t + veg + yr), p_b(t + veg + yr), \pi\}$	0.00	0.37	17
$\{p_a(t + veg), p_b(t + veg), \pi\}$	0.82	0.25	15
$\{p(t * veg)\}$	1.58	0.17	15
$\{p_a(t + veg + yr + burn), p_b(t + veg + yr + burn), \pi\}$	3.07	0.08	19
$\{p_a(t + veg + burn), p_b(t + veg + burn), \pi\}$	3.19	0.08	17
$\{p_a(t + veg + burn + fireseas), p_b(t + veg + burn + fireseas), \pi\}$	5.45	0.02	19
$\{p_a(t + veg + yr + burn + fireseas), p_b(t + veg + yr + burn + fireseas), \pi\}$	5.80	0.02	21
$\{p(t * yr)\}$	8.28	0.01	10

Table 6. Total number of individuals captured and model-averaged estimates of abundance for aggregated data for *Perognathus flavus*. We show a subset of 10 of 54 plots in the aggregate of plots sampled in summer.

		Aggregated data	
Plot	M_{t+1}	$\hat{N}$	SE
1	23	32.6	4.3
2	1	1.4	0.8
3	17	24.1	3.6
4	5	9.1	2.9
5	4	7.2	2.6
6	5	9.1	2.9
7	0	0.0	0.0
8	5	8.7	2.8
9	10	17.4	4.4
10	13	18.4	3.1

5. DISCUSSION

Ecological studies of vertebrates based on capture-recapture approaches often fail to generate sufficient data to estimate abundance reliably at the level of individual sample units, for all sampling periods, and for all species of interest (McKelvey and Pearson 2001). In our case study, for example, sampling effort was 51,840 trap nights that resulted in the capture of nearly 5,600 individual small mammals, and the larger study (Litt 2007) included approximately 210,000 trap nights and over 11,000 individuals. Nonetheless, for many species, reliably selecting a model to estimate abundance at the level of the individual plot was impossible. Even for common species, data occasionally were sparse for some plots and sampling periods. Aggregating data increased the ability to account for variation in detection probability, allowing for more reliable estimates of abundance and subsequent inferences compared to unadjusted counts. Because the quality of results from aggregating depends on how well the model used for estimation captures the various processes driving variation in detection probability within the aggregate, using biological and empirical information to evaluate the potential variation in an aggregate is an essential step in the aggregation process.

Larger pools of information provide more precise estimates of detection probability and abundance (Burnham and Anderson 2002; White 2005) if the model used for estimation describes well the range of processes driving variation in detection probability in the aggregate. With sparse data, the "true" underlying detection processes may not be represented among competing models, however, reasonable estimates can still be generated (Anderson et al. 1994). We found that estimates generated from competing but misspecified models generally were comparable to those from the true model, indicating an increased degree of robustness gained through data aggregation (Table 3). Although estimates from aggregated datasets may not be completely unbiased, they likely will be less biased than results based on unadjusted counts (White 2005). Further, model-averaged estimates account for uncertainty in the model-selection process when data do not clearly support a single model (Burnham and Anderson 2002). Capabilities in estimation software, such as

Program MARK, allow more realistic, flexible, and complex models to be built, improving considerably on previous alternatives, especially with sparse data.

Hierarchical Bayesian models provide an alternative approach to inform the aggregation process we described. This framework offers the ability to accommodate the same types of complexity we considered when evaluating models for aggregates, but it can also incorporate random effects and explicit structures for addressing parameter and model uncertainty (Congdon 2003; Clark 2005; Clark and LaDeau 2006). The approach involves building a model for pooled data by identifying and modeling hierarchical relationships among aggregated elements without the need to identify precisely the underlying processes or factors that might be influencing detection probability. When it is reasonable to assume similar relationships among different elements, such as for multiple sites sampled over time or space, information can be shared to improve estimation procedures (e.g., Congdon 2003; Kéry and Royle 2008) much like the approach we describe. Resources for applying Bayesian tools in ecology have been increasing rapidly (Clark 2005; Clark and Gelfand 2006; McCarthy 2007; Kéry and Royle 2008).

Data aggregation provides a promising alternative for capture-recapture studies with sparse data and is almost certainly a better strategy than relying on unadjusted counts or a single estimator to make comparisons and draw inferences. Increasing sample sizes by aggregating improves the ability to model variation in detection probability, ultimately reducing bias and increasing precision of parameter estimates regardless of the sampling framework. With more information, a model for estimation can be selected that describes variation in detection probability that is reliably grounded in data and provides estimates with higher precision (MacKenzie et al. 2005). Increased reliability of species-specific estimates also provides a better foundation for interspecific or community-wide comparisons that are inadvisable with unadjusted count data (Nichols 1986). Further, data are aggregated only to choose among models for detection probability, as unique estimates of abundance are generated for each element in the aggregate, which offers the advantage of retaining individual experimental units for analysis of replicated experiments. Because of these advantages, data aggregation can improve the reliability of ecological inferences in a wide variety of sampling circumstances.

SUPPLEMENTAL MATERIAL

Datasets: Datasets used in this article are available as supplemental material online. (13253_2009_17_MOESM1_ESM.pdf)

ACKNOWLEDGEMENTS

The Department of Defense Legacy Resource Management Program and the BIO5 Institute for Collaborative Bioresearch at the University of Arizona funded our work. William J. Matter, Guy R. McPherson, Carl J. Schwarz, and two anonymous reviewers provided helpful comments on earlier drafts.

[Received January 2008. Revised October 2008. Published Online January 2010.]

REFERENCES

Alldredge, M. W., Pollock, K. H., Simons, T. R., and Shriner, S. A. (2007), "Multiple-Species Analysis of Point Count Data: A More Parsimonious Modelling Framework," *Journal of Applied Ecology*, 44, 281–290.

Anderson, D. R., Burnham, K. P., and White, G. C. (1994), "AIC Model Selection in Overdispersed Capture-Recapture Data," *Ecology*, 75, 1780–1793.

Bowden, D. C., White, G. C., Franklin, A. B., and Ganey, J. L. (2003), "Estimating Population Size With Correlated Sampling Unit Estimates," *Journal of Wildlife Management*, 67, 1–10.

Boyce, M. S., MacKenzie, D. I., Manly, B. F. J., Haroldson, M. A., and Moody, D. (2001), "Negative Binomial Models for Abundance Estimation of Multiple Closed Populations," *Journal of Wildlife Management*, 65, 498–509.

Buckland, S. T., Anderson, D. R., Burnham, K. P., and Laake, J. L. (1993), *Distance Sampling: Estimating Abundance of Biological Populations*, New York: Chapman & Hall.

Burnham, K. P., and Anderson, D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer.

Clark, J. S. (2005), "Why Environmental Scientists Are Becoming Bayesians," *Ecology Letters*, 8, 2–14.

J. S. Clark, and A. E. Gelfand (eds.) (2006), *Hierarchical Modeling for the Environmental Sciences*, Oxford: Oxford University Press.

Clark, J. S., and LaDeau, S. (2006), "Synthesizing Ecological Experiments and Observational Data With Hierarchical Bayes," in *Hierarchical Modeling for the Environmental Sciences*, eds. J. S. Clark, and A. E. Gelfand, Oxford: Oxford University Press, pp. 41–58.

Congdon, P. (2003), *Applied Bayesian Modelling*, West Sussex: Wiley.

Conn, P. B., Arthur, A. D., Bailey, L. L., and Singleton, G. R. (2006), "Estimating the Abundance of Mouse Populations of Known Size: Promises and Pitfalls of New Methods," *Ecological Applications*, 16, 829–837.

Hammond, E. L., and Anthony, R. G. (2006), "Mark-Recapture Estimates of Population Parameters for Selected Species of Small Mammals," *Journal of Mammalogy*, 87, 618–627.

Hargrove, J. W., and Borland, C. H. (1994), "Pooled Population Parameter Estimates From Mark-Recapture Data," *Biometrics*, 50, 1129–1141.

Kéry, M., and Royle, J. A. (2008), "Hierarchical Bayes Estimation of Species Richness and Occupancy in Spatially Replicated Surveys," *Journal of Applied Ecology*, 45, 589–598.

Litt, A. R. (2007), "Effects of Experimental Fire and Nonnative Grass Invasion on Small Mammals and Insects," unpublished Ph.D. dissertation, University of Arizona, School of Natural Resources.

Lynam, A. J., Rabinowitz, A., Myint, T., Maung, M., Latt, K. T., and Po, A. H. T. (2009), "Estimating Abundance With Sparse Data: Tigers in Northern Myanmar," *Population Ecology*, 51, 115–121.

MacKenzie, D. I., and Kendall, W. L. (2002), "How Should Detection Probability Be Incorporated Into Estimates of Relative Abundance?" *Ecology*, 83, 2327–2393.

MacKenzie, D. I., Nichols, J. D., Sutton, N., Kawanishi, K., and Bailey, L. L. (2005), "Improving Inferences in Population Studies of Rare Species That Are Detected Imperfectly," *Ecology*, 86, 1101–1113.

McCarthy, M. A. (2007), *Bayesian Methods for Ecology*, Cambridge: Cambridge University Press.

McKelvey, K. S., and Pearson, D. E. (2001), "Population Estimation With Sparse Data: The Role of Estimators versus Indices Revisited," *Canadian Journal of Zoology*, 79, 1754–1765.

Nichols, J. D. (1986), "On the Use of Enumeration Estimators for Interspecific Comparisons, With Comments on a 'Trappability' Estimator," *Journal of Mammalogy*, 67, 590–593.

——— (1992), "Capture-Recapture Models: Using Marked Animals to Study Population Dynamics," *BioScience*, 42, 94–102.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978), "Statistical Inference From Capture Data on Closed Animal Populations," *Wildlife Monographs*, 62, 1–135.

Rosenberg, D. K., Overton, W. S., and Anthony, R. G. (1995), "Estimation of Animal Abundance When Capture Probabilities Are Low and Heterogeneous," *Journal of Wildlife Management*, 59, 252–261.

Rosenberg, D. K., Swindle, K. A., and Anthony, R. G. (2003), "Influence of Prey Abundance on Northern Spotted Owl Reproductive Success in Western Oregon," *Canadian Journal of Zoology*, 81, 1715–1725.

Schwarz, C. J., and Seber, G. A. F. (1999), "Estimating Animal Abundance: Review III," *Statistical Science*, 14, 427–456.

Seber, G. A. F. (1982), *Estimation of Animal Abundance and Related Parameters*, New York: Macmillan.

——— (1986), "A Review of Estimating Animal Abundance," *Biometrics*, 42, 267–292.

——— (1992), "A Review of Estimating Animal Abundance II," *International Statistical Review*, 60, 129–166.

Skalski, J. R., Robson, D. S., and Simmons, M. A. (1983), "Comparative Census Procedures Using Single Mark-Recapture Methods," *Ecology*, 64, 752–760.

White, G. C. (2005), "Correcting Wildlife Counts Using Detection Probabilities," *Wildlife Research*, 32, 211–216.

White, G. C., and Burnham, K. P. (1999), "Program MARK: Survival Estimation From Populations of Marked Animals," *Bird Study*, 46 (Suppl.), 120–138.