

# Real and quasi-experiments in capture-recapture studies

**CARL JAMES SCHWARZ**, *Dept of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada*

**ABSTRACT** *The three key elements of experimental design are randomization, replication, and variance identification and control. Capture-recapture experiments usually pay sufficient attention to the first two elements, but often do not pay sufficient attention to sources of variation. These include blocking factors and different sizes of experimental units. By casting capture-recapture studies in an experimental design framework, the various roles of these sources of variation become clear and the sources that are pooled when these experiments are analysed using existing software is also clear. This formulation also shows that care must be taken with pseudo-replication and different sized experimental units.*

## 1 Introduction

The use of capture-recapture methods does not end with the estimation of survival, abundance, or density. Questions concerning natural or human-induced changes in these parameters often require quantification, often via hypothesis testing or estimation of the difference among various group-specific parameters.

Field studies can be broadly classified into three categories:

- (1) observational studies where randomization is restricted solely to selecting animals from the population of interest and no manipulation of experimental conditions is performed, e.g. a comparison of male and female survival rates. Observational studies can produce inference of various types based on *a priori* hypotheses or based on hypotheses developed after the fact. Although the strength of inference is limited in either case by the lack of control over the system, the former approach is stronger than the latter. These will not be discussed in this paper.

*Correspondence:* Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6. E-mail: cschwarz@stat.sfu.ca

- (2) Environmental impact or assessment studies where site specific comparisons between a non-random assignment of impact and possible randomly selected control sites are required, e.g. the impact of an oil spill upon survival rates.
- (3) Manipulative studies where the experimenter randomly assigns population units to treatment and control groups, e.g. a comparison of survival rates among groups subject to different doses of lead.

Fisher (1935) outlined three principles that are essential for good experimental design: replication, randomization, and variance-reduction (blocking). Surprisingly, only the first two have received much attention in the capture-recapture literature and the 'state of the art' in using experimental design in capture-recapture studies is surprisingly primitive.

Randomization is the key to representativeness. A random selection of animals that are tagged is what ensures that estimates derived from the experiment can be extrapolated to the population as a whole. This is often glossed over in many CJS experiments as tagged animals are often obtained using a convenience sample from the population without much thought to randomly selecting them from the entire population at risk. However, proper random sampling will be crucial if abundance or population growth is to be estimated, as a key assumption is that animals are captured at random from the entire population of interest.

Replication, or sample size, controls precision. Through simulation studies it is relatively easy to determine the necessary sample size to obtain a specified precision in the estimates. The danger in ecological experiments is pseudo-replication (Hurlbert, 1984) where the observational unit (e.g. bird) differs from the experimental unit (habitat type) and inference can be limited.

Blocking, or variance reduction and control, is infrequently seen in capture-recapture studies. In this context, 'variability' refers to the variation in survival and capture probabilities from all sources, not just from the binomial response of an animal to a probabilistic survival event.

Much of classical experimental design deals with issues of variance decomposition, reduction, and control—it is somewhat surprising that this has not permeated the capture-recapture field. Many studies still continue to pool over various sub-groups not formally part of the analysis framework, e.g. pooling over batches of releases—yet recognize that this may not be fully successful as the variance inflation factor is often used to adjust the final estimates for over or underdispersion.

Finally, classical experimental design also carefully examines issues as to different sizes of experimental units; the distinction between experimental and observational units; and fixed versus random effects. These distinctions are absent from most capture-recapture studies. Most studies assume there is but a single size of experimental unit—the animal—and treat the experimental and observation unit as identical. Random effects are starting to appear, but only in restrictive ways as variation over time. Clearly, further research into this topic needed.

In the remainder of this paper, I will examine assessment and manipulative experiments. I will focus upon the analysis of survival rates—similar comments also apply to the analysis of capture probabilities. Although my final conclusions may seem pessimistic, researchers should not shy away from formal experimentation. The experience in medical studies has shown that planned, randomized experiments are the key to inferring causation (Hill, 1971). Observational studies have limited inferential ability in the face of multi-factor, often highly interrelated study variables.

## 2 Some general comments

### 2.1 Terminology

The literature in experimental design is fairly standardized and I will use it throughout this paper.

A factor is a condition that is either under the control of the experimenter (e.g. dose) or is simply observed (e.g. gender). Each factor has two or more values called levels. For example, the factor gender has two levels, M and F. The particular set of levels that is attached to an observational unit is called a treatment, e.g. if there are two factors (dose and gender), each at two levels (H versus L dose; M versus F), then the treatments that may occur in the experiment are the combinations of dose and gender, e.g. HF, HM, LF, LM.

All experiments take place by assigning factor levels to experimental units. Measurements are taken on the observational units. A common error is the failure to distinguish between the experimental unit and the observational unit. This can lead to sub-sampling designs (multiple observations are taken from the same experimental unit) and pseudo-replication (Hurlbert, 1984) where ‘replicates’ taken in a study are simply sub-samples from the true larger experimental unit.

A simple example to illustrate the difference between the experimental and observational unit is an experiment done on fish held in tanks. The experiment is to investigate the effect of various chemicals upon the growth of fish over time. Fish are individually marked, weighed, randomly assigned to tanks (ten per tank), and then chemicals are added to the tanks (two tanks each for the chemical and the control group). At the completion of the experiment, the individual fish are again weighed, and the increase in weight recorded for each fish. It is tempting to believe that the experiment has 20 replicates for each treatment (ten fish in each of two tanks), but in fact, the treatment (the chemical or control) was applied to the *tank* not to individual fish, and there are only two replicates. The analysis of variance table for this experiment would look like:

Source	Df	Error term
Treatment	1	tank(treatment)
tank(treatment)	3	fish(tank $\times$ treatment)
fish(tank $\times$ treatment)	36	

and the test for a treatment effect would involve the ratio of  $ms(\text{treatment})$  to  $ms(\text{tank}(\text{treatment}))$ . The individual fish are observational units—treating the individual fish as the experimental unit will typically lead to an increase in Type I error rates as the variation among tanks is usually much greater than among fish.

### 2.2 Structure

Any experimental or assessment design has three embedded, independent, structures—the treatment structure (i.e. what combinations of treatment levels are present in the experiment); the experimental unit structure (i.e. what are the various experimental and observational units); the randomization structure (i.e. how are factors assigned to experimental units). Many textbooks on experimental and assessment designs often fail to distinguish among these three structures by providing a cookbook approach that wraps all three aspects under one name.

By far the most common experimental design situation involves the factorial treatment structure where every treatment combination occurs in the experiment.

In the case of many factors, this may lead to an inordinate number of treatments, so fractional factorial designs (where only certain treatment combinations appear) are used. Most experiments also profess to conduct complete randomization of factors to experimental units. Closer reading of these papers often shows that, in fact, this did not occur and often systematic or haphazard allocation schemes are used. For the purposes of this paper, I will assume that complete randomization occurred.

### 2.3 Treatment of time effects—repeated measures?

Most capture-recapture experiments treat ‘time’ as a factor with several levels. By definition, capture-recapture studies are longitudinal studies, i.e. the same animal is followed (imperfectly) over time. In the traditional experimental design literature, these are often referred to as repeated measure studies and there are various ways to treat this ‘repeated factor’. The simplest is to simply treat time as any other factor and construct an analysis of variance table accordingly. However, unless the covariance structure of the repeated measure takes a special form (e.g. compound symmetric), the F-tests for the repeated factors do not follow the nominal F-distributions. Consequently, either corrections to the standard F-tests (Greenhouse-Geisser or Huynh-Feldt) are performed or multivariate tests are conducted.

It is ‘unusual’ to think of the capture history of an animal as representing repeated measurements on the same animal. In the standard likelihood theory (Lebreton *et al.* 1992), the likelihood factors into pieces representing the probability of next capture conditional upon each release of the animal. These are often treated as independent events, even though a particular animal could be ‘reused’ several times in the likelihood function. In addition, the fate of the animal is unknown after its last capture—but this is analogous to ‘missing values’ in an experimental design framework.

Consequently, in the remainder of the paper, I will view the capture history of a animal as representing repeated measurement on the survival of the animal to make it easier to cast capture-recapture experiments into an experimental design framework.

### 2.4 Fixed versus random effects

When considering factors in an experiment, a distinction is made between fixed and random effects. A fixed effect is one where the levels present in the experiment are the only ones of interest; inference will only be about these levels; and if the experiment were to be repeated the same levels would be re-used. Examples of fixed effects are gender, dose of chemical, type of band applied, etc. A random factor has levels in the experiment that were chosen, at random, from a larger set of levels; inference is to this larger set of levels; and if the experiment were to be repeated, a different set of levels would typically be used. Examples of random effects are often location, sites, etc.

Inference for these two types of factors is quite different. For fixed effects, interest typically lies in how the levels affect the mean response; for random effects, interest lies in the overall mean response and the size of the variation of the level effects for that factor. For example, for fixed effects, the size of the difference between the two means is estimated while for random effects, the variance of the effects (variance component) is estimated.

### 3 Assessment studies

In assessment studies, site specific comparisons between a non-random assignment of impact and possible randomly selected control sites are required, e.g. the impact of an oil spill upon survival rates.

#### 3.1 Intervention (monitoring) designs

The simplest design is often used in monitoring studies. It consists of simply monitoring the parameter of interest over time (survival or population size) and examining whether the level has changed from before to after an intervention. The design is often an ‘unplanned’ experiment in that the intervention may be accidental (e.g. oil spill), but it is often used in environmental impact studies where the intervention is planned, e.g. the construction of a new power plant. In a classical experimental design analysis, estimates would be obtained at every time point before and after the study and a ‘t-test’ would be used to compare the population parameters before and after the intervention. This ‘simple’ analysis ignores any covariances among the estimates that may be obtained from a multi-year longitudinal study and typically will underestimate the variability of the estimates that will result in a lower  $p$ -value that would be warranted and an increased Type I error rate. Time series intervention analyses are more complex, model-based approaches to try and account for this time covariance.

In the capture-recapture context, the modelling is done ‘internally’, i.e. a linear model is developed that typically includes models survival before and after the intervention, and standard likelihood ratio tests, or the AIC method is used to determine if a simpler model without intervention effects is tenable. These linear models can be quite complicated and they can also be used to account for existing trends over time, where a relevant test is to examine if the slope has changed between before and after the intervention—a simple example of the analysis of covariance.

An example of this type of study is found in Piper *et al.* (1999) where the intervention consists of supplemental feeding of the Cape Griffon vulture from 1984 onwards. The authors fit a series of Cormack-Jolly-Seber (CJS) models that included the intervention and found that survival increased across the intervention. Aebischer (1995) investigated the association between changes in hunting regulations and subsequent annual survival of pigeons and doves. Lebreton *et al.* (1992) also illustrate the analysis of unplanned interventions (floods) upon the survival of European dippers using similar methods.

As Green (1979) pointed out, the fundamental flaw with these designs is that any changes in the parameters are totally confounded with temporal effects, i.e. the change in survival may have occurred regardless of any intervention (planned or unplanned). Another potential problem is that capture-recapture methods cannot distinguish between experiments where the animals involved are a different set before and after the intervention, or they are the same set of animals simply followed over time. Classical experimental design carefully distinguishes between the independent sample and the paired sample experiments but the likelihood theory for CJS models makes no distinction. For example in the Piper *et al.* (1999) study, only the survival of first year birds was of interest, which essentially made this an unpaired experiment and the comparison is of different cohorts of birds. Aebischer’s (1995) study involved following cohorts over long periods of time,

which essentially made this a paired experiment. The disadvantage of a paired experiment, is that survival after the intervention is of course conditional upon survival prior to the intervention. Perhaps, just by chance, heterogeneity in survival rates among birds leads to a higher observed mortality before the intervention took place. Such an increase in survival would be consistent with the heterogeneous survival models of Burnham & Rexsted (1993).

### 3.2 BACI designs

To resolve the confounding of the intervention and temporal effects, a class of experimental designs has evolved that go under the rubric of 'Before-After-Control-Impact (BACI)' designs (Green, 1979; Underwood, 1993, 1994; Stewart-Oaten & Murdoch, 1986).

In these designs, studies are carried out at both control and impacted sites, and these studies commence before the planned intervention and continue after the impact. In the simplest design, a single control and single impact site are monitored over time. It is now the change in the differences between the mean response at the two sites between the before and after points (i.e. an interaction between site and time effects) that provides evidence for an impact. This comparison can again be easily performed by using a linear model approach to tagging studies.

There are two 'major' problems with the simple BACI design. First, in the simplest design, measurements are taken at each site at a single time point before and after the intervention. There may be multiple measurements at this single time point, which allows estimates of the precision of that particular measurement to be derived. However, natural variation over time in the parameters at each site may, by chance, lead to cases where the observed difference is certainly within normal bounds. The design is unable to detect this as, without any replication over time, no estimate of variation over time is possible. This is resolved by taking a time series of measurements at both sites both before and after the impact and looking at the difference in the average response before and after the experiment. As capture-recapture studies are usually longitudinal, this will be a natural aspect of the design.

Second, this design is unable to extrapolate any 'effect' beyond that which occurred between two specific sites. Again, no estimate of site-to-site variation is obtainable and this is an example of pseudo-replication—the animals in the study are the observational units and not the experimental units. The enhanced-BACI design was developed in response. As it is unlikely that the intervention site can be replicated, the usual procedure is to use replicated control sites with the implicit assumption that the variation among replicated impacted sites is the same as among control sites.

The analysis of an enhanced-BACI design is fairly straightforward in standard experimental designs. The analysis of variance table would take the form:

Source	'Error term'
Treatment	site(treatment)
site(treatment)	
Time	Residual
Time × treatment	Residual
Residual = time × bird(treatment)	

which to a classical experimentalist, looks like a ‘split-plot-in-time’ design. Alternatively, the difference in response before and after the intervention is found for each site. This reduces the data to a set of single numbers—one of which belongs to the impacted site and the others to the control site—and a two-independent-samples test (e.g. a *t*-test) is performed. Both analyses are equivalent, except the latter analysis ‘averages’ out any time and interaction tests.

In contrast, the CJS models make no distinction between the experimental unit (the site) and the observational unit (the bird). Both the effects of time and of treatment are tested using the same residual deviance. This will be examined in more detail later in this paper.

#### 4 Some common experimental designs

As noted earlier, experimental designs are among the most convincing ways of inferring causation mechanisms as the separate effects of designed factors can be extracted from the combined effects of factors in observational studies. In this section, I will review several of the common experimental designs that have been used in the capture-recapture literature.

##### 4.1 Factorial treatment structure; single size of e.u.; complete randomization

This is the simplest possible experimental design for a capture-recapture study. In this study, treatments are randomly assigned to animals (the experimental unit) at the time of initial capture and the subsequent survival rates of the treatment groups are compared over time. In most cases, there is a single factor, but this description is completely general and also applies to multiple factor designs, i.e. in a two factor design with *a* and *b* levels respectively, then all *ab* treatment combinations are completely randomized to animals before release.

There are many examples of this type of studies. For example, Samuel *et al.* (1999) investigated the effect of inoculations against avian cholera upon the survival rate of Lesser Snow Geese. As geese were captured, they were randomly assigned to either the treatment group (given an inoculation) or the control group (no inoculation), banded, and released. Resightings took place over a number of years.

As before, it is helpful to cast this experiment into the traditional ANOVA framework where the various sources of variation are extracted and the expected mean-squares are used to determine how certain hypotheses are to be tested.

A classical experimentalist would recognize that there are two sources of variation in the individual survival rates over time—that of the individual bird assigned to each treatment group, and the variation over time within each bird. The model under a classical experimental design approach is:

$$\phi_{ijk} = \mu + \tau_i + b_{ij} + t_k + \tau t_{ik} + e_{ijk}$$

where  $b_{ij}$  and  $e_{ijk}$  are the (random) bird and residual random variation.

In most CJS analyses, the analyst would write a model for survival as

$$\phi_{ijk}^* = \mu + \tau_i + t_k + \tau t_{ik} + e_{ijk}^*$$

The discrepancy between the two approaches is the explicit recognition of a ‘bird’ effect over and above any residual variation.

The 'standard' ANOVA table for the classical experimental design would be structured as:

Source	'Error term'
Treatment	Bird(treatment)
Bird(treatment)	
Time	Residual
Time $\times$ treatment	Residual
Residual = time $\times$ bird(treatment)	

Here, it is explicit that the test for treatment relies upon a different error term from the test for time and time  $\times$  treatment interactions.

When using MARK or POPAN or SURGE to analyse this experiment, all sources of variation are combined in one residual deviance. Going back to the ANOVA table, a single 'error term' would be appropriate for all model tests if, in fact, there was no heterogeneity among birds within each treatment group. If the heterogeneity among birds can be 'explained' by other factors such as gender, then these can also be entered into the model. Nevertheless, an explicit assumption of all capture-recapture models is homogeneity among birds. The effects of ignoring this source of variation can also be speculated based on what happens in traditional experimental designs— if the bird-to-bird variation is larger than the within-bird variation, then the pooled error term will be too small for the tests of treatment effects (resulting in an increased Type I error rate and too small standard errors) and too large for the test of time and interaction effects (resulting in a decrease in the power to detect such effects and too large standard errors for the time and interaction effects).

The effects of heterogeneity in survival and capture rates has been extensively studied. For example, Pollock & Raveling (1982) and Nichols *et al.* (1982) investigated the effects of heterogeneity on estimates of survival in bird banding models and found that bias induced by heterogeneity was small relative to the standard errors, and that the standard errors were underestimated. The latter point is most crucial as this leads to inflated Type I errors. Burnham & Rexstad (1993) detected heterogeneity in many waterfowl datasets they examined and developed models to allow for heterogeneity in individual survival rates. However, they conclude that 'heterogeneity of survival probabilities is a biologically important issue, but presents great difficulties in estimation. Successful estimation requires long-term data sets of long-lived species with large number of marked individuals.' Barker (1992) studied the effects of heterogeneity upon coverage of confidence intervals and found that, for most studies with not severe heterogeneity and less than 500 birds marked, the actual confidence interval coverage was close to nominal levels for individual estimates, but could be poor for estimates of the mean survival. He concludes 'The problem of introduction of heterogeneity through data pooling also needs to be considered. . . . Such pooling increases the likelihood of heterogeneity, and because of the increased sample size, exacerbates the effect of any heterogeneity that is present.'

Some of the heterogeneity may be caused by a failure of birds to operate independently. For example, some species of birds form pair bonds and the fate of these animals is closely tied. As an extreme, if a pair of birds had complete dependency, the actual experimental unit would be the pair rather than the individual bird. Schmutz *et al.* (1995) examined the effects of dependency in survival rates caused, for example, by pair bonding. Schmutz *et al.* (1995) used random lots (selecting one from each member of the pair) and bootstrapping to estimate the empirical variance and compared it with the model-based variances (including any inflation factor from

overdispersion). They found that the VIF from the standard models in Release accounted for less than 10% of the increase in variation caused by pair bonding. Samuel *et al.* (1999) recognized the potential deleterious effects of pair bonding in Lesser Snow Geese and only tagged males or females within a group of releases.

Rather than trying to model heterogeneity among birds directly, empirical methods can be used to avoid model-based standard errors by dividing the releases into random lots, e.g. by the final digit of the tag number. Each lot is analysed separately to obtain a set of independent estimates of each effect. The final estimate is a (weighted) average of the individual estimates, and the (weighted) standard error among random lots can be used to estimate empirically the standard error of the estimates. In the sampling literature, this is quite commonly used and is known as interpenetrating sub-samples. Burnham *et al.* (1987, Sections 4.2.3 and 4.3) illustrate its use on a fish release experiment.

#### 4.2 Factorial treatment structure; single size of e.u.; restricted randomization via blocking

Often, heterogeneity in the animals may be related to some outside factor that is not really an experimental factor, e.g. areas of release; time of release etc. In these cases, the experimental units (animals) may be grouped into more homogeneous sets (called blocks). Within each block, complete randomization of animals to treatments is done much as in the previous section. The simplest examples of these types of block designs are often called 'paired-release experiments', where each 'block' has only two treatment levels.

An example of this type of design is the study by Deuel (1985) of the effect of lead upon the survival of pintails. Birds were caught in several areas of California; banded; and sequentially assigned to the control or dosed group. Data from subsequent hunting seasons were used to estimate the survival rates. A subset of the data (limited to five areas and males only) is presented in Burnham *et al.* (1987, Section 7.3).

This is an example of Generalized Randomized Block design (Gates, 1995). It differs from a standard randomized block design in that there are multiple experimental units (birds) assigned to treatments within each block. A classical experimentalist would write a model for the variation of survival rates among areas, doses, birds, and over time as:

$$\phi_{ijkl} = \mu + a_i + d_j + ad_{ij} + b_{ijk} + t_l + at_{il} + dt_{jl} + adt_{ijl} + e_{ijkl}$$

where, at a minimum,  $ad_{ij}$ ,  $b_{ijk}$  and  $e_{ijkl}$  and are random effects. (Note that because this is a block design, an implicit assumption is no interaction between blocks and treatments; consequently, the  $ad_{ij}$  term represents experimental error.) There would be some debate about the role of areas—are they fixed or random effects.

The ANOVA table would break out these sources of variation as follows

Source	Error term
Block = Area	
Dose	dose $\times$ area
Dose $\times$ area = experimental error	bird(dose $\times$ area)
Bird(dose $\times$ block)	
Time	Residual
Time $\times$ area	Residual
Time $\times$ dose	Residual
Residual	

Again notice that there are several different error terms. In the usual CJS modelling approaches, all of these error terms are pooled and, unless they are very small, can lead to test statistics and standard errors that may not fully reflect the various sources of variability.

In the case of a treatment structure consisting of only two treatments (so-called ‘paired-release experiments’), the effects of block experimental error can be explicitly included and quantified. Burnham *et al.* (1987, Section 7.3) analysed a subset of the Deuel (1985) data, by first finding separate estimates of the treatment effect in each block. These individual estimates will then be free of block effects, but include in their empirical variation, the first experimental error variation. The overall effect can then be found as a (weighted) average of the individual estimates from each block. An empirical variance estimate can be found from the observed variation of the estimates over blocks. Furthermore, they also demonstrate how to estimate this error term variation using a method-of-moments approach similar to what is done in an ANOVA decomposition.

Unfortunately, there is no simple method than can be applied to cases with more than two treatment levels per block unless one constructs all pairwise contrasts. How the overall Type I error rate would be controlled is not clear at all in this case.

Other examples of blocks are colonies or release groups (cohorts) and are often not explicitly recognized. For example, Piper *et al.* (1999) studied the effect of supplementary feeding on the survival rates of vultures in two colonies 120 km apart. These would seem to be natural blocks, but the analysis in that paper did not include any colony effect. They also state ‘While we have no direct evidence that the nestlings ringed in a particular year are prone to suffer higher or lower mortality than those ringed in other years, we tested for this factor in our models.’ Unfortunately, if a cohort is to be treated as a block, then it cannot be ‘tested’ (refer to many of the articles in experimental design on the question of whether ‘block effects can be tested’). Similarly, Samuel *et al.* (1999) released Lesser Snow Geese in groups but pooled the data over all groups when fitting the final models.

#### 4.3 Factorial treatment structure; multiple e.u.; complete randomization

Designs with multiple experimental units of different sizes are among the most difficult to recognize in classical experimental design and the ones most often analysed incorrectly. The most common unrecognized design is the split-plot design.

In these designs, factor levels are assigned to different sized experimental units. Consequently, any analysis of these experiments must recognize the different sizes of experimental units, and replication of the different sizes units is required in order to perform valid statistical tests for factor effects.

An example of this type of study is found in Boudjemadi *et al.* (1999). They designed an ingenious experiment to investigate the effects of habitat type (grassland or wood clearance) and connectivity of habitat fragments (connected or not connected) upon survival and a number of variables. Within a single patch of each habitat, four experimental enclosures were constructed, of which two were randomly assigned to each connectivity level. Lizards were introduced and followed over three sampling periods using mark-recapture methods.

This is an example of a split-plot design. Habitat type operates on large sections of land; connectivity operates on the enclosures within each habitat type; multiple individuals are observed over time within each enclosure. The ANOVA table is:

Source	df	Error term
Habitat type	1	sections(habitat)
sections(habitat)	0	Note: there were <i>no</i> replicates of sections of each habitat!
Connectivity	1	enclosure(connectivity $\times$ sections $\times$ habitat)
connectivity $\times$ habitat	1	enclosure(connectivity $\times$ sections $\times$ habitat)
connectivity $\times$ sections(habitat)	0	
enclosure(connectivity $\times$ sections $\times$ habitat)	4	
Time	2	residual
time $\times$ habitat	2	residual
time $\times$ sections(habitat)	0	
time $\times$ connectivity	2	residual
time $\times$ connectivity $\times$ habitat	2	residual
Residual	8	

Note that this experiment first suffers from pseudo-replication as there were no true replicates of the habitat type—only a single section of each type was used. Inferences are limited to differences between these two particular sections—their effects are totally confounded with that of habitat type. Connectivity operates on the enclosure level and so a test of connectivity effects must be performed relative to the variation among enclosures. Finally, time operates on the individual lizard. There are two sizes of experimental units (sections and enclosures) and lizards are observation units within the lowest experimental unit.

In their paper, the authors do a pooled analysis using SURGE with effects for habitat, connectivity and time but only a single source of variation. As in past examples, this residual variation will be a pooling of the three error variances from above and may not be appropriate for testing particular factors.

Similarly, Horak & Lebreton (1998) examined the effects of habitat (urban or rural), sex and time upon the survival of adult Great Tits. In classical experimental design literature, this is again a split-plot design with habitat operating on the main plot experimental units (replicates of habitat types), gender operating on the subplot units (individual birds), and time being a repeated measure on each bird. As in the above example, there were no true replicates of habitat type (i.e. the experiment pseudo-replicated this factor). Various models were fit in SURGE with considerable underdispersion present (but, as noted by the authors, this may have been an artefact of the sparse data).

These problems can also occur in 'natural experiments'. For example, Franklin *et al.* (1999) describe a meta-analysis of survival rates of the northern spotted owl. There were 15 study areas in the Pacific Northwest that were divided into four broad ecological provinces, or land ownership categories. Here, the experimental unit for ecological province is the study area—it is likely that local conditions within each study area influence the survival rates of the individual birds within the study area. The observational unit within each study area is the bird.

The classical ANOVA table for this experiment for a simple model that included

province effects, local study area effects, and simple time effects upon survival would look like:

Source	Error term
Province	area(province)
area(province)	bird(area $\times$ province)
Time	residual
time $\times$ province	residual
time $\times$ area(province)	
residual	

Tests for province effects should use the variation of the survival rates among the area as an error term. Franklin *et al.* (1999) first fit individual CJS models to each study area but conducted a meta-analysis by combining all data and fitting CJS models incorporating ecological provinces or ownership status as a simple factor in the linear model but treating all birds as the individual experimental units. They found that one model that included ownership effects was close to, but not, the best model in the AIC hierarchy. They also found moderate overdispersion—this may be caused by the pooling of the variation from the various sources.

#### 4.4 An Incomplete block design

Burnham *et al.* (1987) reports on an experiment by Bellrose (1959) involving the effects of different lead dose upon the survival of mallards. There were a total of four dose levels (control, 1, 2 or 4 pellets). In each year (a block), a 'paired-release' experiment was performed that compared the effect of a dose of lead and a control group that received no lead. This is an example of an incomplete-block experiment as not all treatments occurred in every block. Burnham *et al.* (1987) analysed this experiment by estimating the difference in survival for each block (i.e. the contrast between the control group and the experimental group). This is equivalent to an inter-block analysis. However, the analysis of incomplete block designs can also recover intra-block information. Both estimates are then combined using appropriate weights. A combined inter- and intra-block analysis can be fit by fitting a more complex model. As in previous examples, this model has no counterpart in the models fit by SURGE, POPAN, or MARK.

#### 4.5 Random effect models

Another area of divergence between classical experimental design and capture-recapture studies is the distinction between fixed and random effects. Burnham (2001) has recently developed a model where the time effects upon survival are to be treated as random effects around a long term mean, but there has been little work where other factors in the model could be treated as random effects. For example, Hastings & Wardtesta (1998) performed a capture-recapture study on seal populations and were interested in the differential effects of birth colonies, which change location from year to year. Here, birth colony would naturally be treated as a random effect. In some cases, cohort effects, i.e. animals all born in a single year, could also be considered random effect. In the study of Franklin *et al.* (1999), there were several study sites within each land classification, but these should likely be considered as a 'sample' taken from all possible sites under each

classification—again a random effect. Furthermore, Burnham & Anderson (1998, Section 6.7.6) indicate that there are serious problems in using AIC in model selection with random coefficient models.

## 5 Discussion

Conceptualizing capture-recapture experiments in terms of an experimental design has the distinct advantage of forcing the experimenter to pay special attention to the question of experimental units; treatment structure; and randomization structure. In the past, this was difficult to do because data were typically collected and displayed as summary statistics—however, with the recent emphasis on the individual capture histories, these experiments can be more closely matched with standard designs. Skalski & Robson (1992) illustrate many of these points when comparing abundance at specific points in time.

It appears that the above examples can be broken into two relatively distinct cases—single size experimental units and multiple sizes of experimental units.

In the case of a single sized experimental unit—the animal itself—the key issue appears to be the effect of pooling various sources of heterogeneity in survival rates—some recognized (e.g. bird-to-bird variation) and some unrecognized (block-to-block) variation. This often leads to overdispersion in the counts of the observed capture-histories. This leads to poor performance of unmodified AIC-based model selection methods (Anderson *et al.* 1994). Fortunately, Anderson *et al.* (1994) also show that simple corrections to AIC based on the estimation of the overdispersion factor seem to perform well. Alternatively, a relatively simple method involving interpenetrating subsampling (random groups of birds) can be used to estimate empirically the true variance. However, as Anderson *et al.* (1994) point out, the use of a single overdispersion factor is only an approximation—it may be that this overcorrects for estimates of time effects (within bird effects) and undercorrects for estimates of treatment effects (among bird effects). This needs to be investigated further.

Multiple sized experimental unit studies are problematic. Unfortunately, I do not believe that these problems can be corrected with a simple multiplicative factor—the problem is that there is no provision within the current modelling software for different sizes of experimental units. As a stop-gap measure, individual estimates of survival can be found for each larger sized experimental unit, extracted from the software packages, and analysed further using the ANOVA package. Although the lower level estimates (e.g. individual yearly survival estimates) are not independent within each larger experimental unit, estimates across units will be independent. Fortunately, simple effects of factors applied to the larger units only rely upon the ‘average’ of lower level effects (regardless of the lower level covariance structure) and hence should give approximate tests until more suitable software can be developed. This stopgap measure is inefficient and does not deal with nuisance parameters such as capture-rates that might or might not vary among the experimental units. A general approach would be one where the separate effects of primary (e.g. survival) and nuisance (e.g. capture) parameters and the estimation of effects under additive, interactive, and other model structures would be preferable.

It is now almost 20 years since Hurlbert (1984) published his paper on pseudo-replication in field studies. Unfortunately, this still occurs in many studies. I suspect it is simply difficult to recognize the difference between the experimental and

observational unit, particularly when most of mark-recapture methods are so tightly tied to the individual animal. As such, it is very simple to fail to recognize pseudo-replication in the larger experimental units.

Lastly, further work needs to be done on incorporating random effects into models for other than the temporal component. Coull & Agresti (1999) propose a mixed logit model that is very similar to the normal theory mixed effect models of experimental design that could be modified to incorporate random blocks. Burnham & Anderson (1998, Section 6.7.6) indicate that shrinkage estimators may be a pragmatic way to fit random coefficient models. Under this approach, a model is fit treating all random coefficients as fixed effects; these estimates are shrunk towards zero; and the likelihood is re-evaluated at these shrunken estimates. Alternatively, a full Bayesian approach could be developed generalizing the work of Brooks *et al.* (2000), Chavez-Demoulin (1999), Dupuis (1995), and Vounatsou & Smith (1995).

### Acknowledgements

This work was funded by a Research Grant from the Natural Science and Engineering Research Council (NSERC) of Canada.

### REFERENCES

- AEBISCHER, N. J. (1995) Investigating the effects of hunting on the survival of British pigeons and doves by analysis of ringing recoveries, *Journal of Applied Statistics*, 22, pp. 923-934.
- ANDERSON, D. R., BURNHAM, K. P. & WHITE, G. C. (1994) AIC model selection in overdispersed capture-recapture data, *Ecology*, 75, pp. 1780-1793.
- BARKER, R. J. (1992) Effect of heterogeneous survival on bird-banding confidence interval coverage rates, *Journal of Wildlife Management*, 56, pp. 111-116.
- BELLROSE, F. C. (1959) Lead poisoning as a mortality factor in waterfowl populations, *Illinois Natural History Survey Bulletin*, 27, pp. 235-288.
- BOUDJEMADI, K., LECOMTE, J. & CLOBERT, J. (1999) Influence of connectivity on demography and dispersal in two contrasting habitats: an experimental approach, *Journal of Animal Ecology*, 68, pp. 1207-1224.
- BROOKS, S. P., CATCHPOLE, E. A. & BARRY, S. C. (2000) On the Bayesian analysis of ring-recovery data, *Biometrics*, 56, pp. 951-960.
- BURNHAM, K. P. (2001) Random effect models in ringing and capture-recapture data, *Journal of Agricultural, Biological, and Environmental Statistics*, in press.
- BURNHAM, K. P. & REXSTAD, E. A. (1993) Modeling heterogeneity in survival rates of banded waterfowl, *Biometrics*, 49, pp. 1194-1208.
- BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C. & POLLOCK, K. H. (1987) Design and analysis methods for fish survival experiments based on release-recapture, *American Fisheries Society Monograph*, 5.
- BURNHAM, K. P. & ANDERSON, D. A. (1998) *Model selection and inference—a practical information-theoretic approach*. (New York, Springer).
- CHAVEZ-DEMOULIN, V. (1999) Bayesian inference for small-sample capture-recapture data, *Biometrics*, 55, pp. 727-731.
- COULL, B. A. & AGRESTI, A. (1999) The use of mixed logit models to reflect heterogeneity in capture-recapture studies, *Biometrics*, 55, pp. 294-301.
- DEUEL, B. (1985) Experimental lead poisoning of northern pintails in California, *California Fish and Game*, 71, pp. 125-128.
- DUPUIS, J. A. (1995) Bayesian estimation of movement and survival probabilities from capture-recapture data, *Biometrika*, 82, pp. 761-772.
- FISHER, R. A. (1935) *The Design of Experiments*. (Edinburgh, Oliver and Boyd).
- FRANKLIN, A. B., BURNHAM, K. P., WHITE, G. C., ANTHONY, R. J., FORSMAN, E. D., SCHWARZ, C. J., NICHOLS, J. D. & HINES, J. E. (1999) Range-wide status and trends in northern spotted owl

- populations. Prepared for Bureau of Land Management, US Fish and Wildlife Service, and US Forest Service, 71pp.
- GATES, C. E. (1995) What really is experimental error in block designs, *American Statistician*, 49, pp. 362-363.
- GREEN, R. H. (1979) *Sampling Design and Statistical Methods for Environmental Biologists* (New York, Wiley).
- HASTINGS, K. K. & WARDTESTA, J. (1998) Maternal and birth colony effects on the survival of Weddell seal offspring from McMurdo Sound, Antarctica, *Journal of Animal Ecology*, 67, pp. 722-740.
- HILL, A. B. (1971) *Principles of Medical Statistics*, 9th edn (New York, Oxford University Press).
- HORAK, P. & LEBRETON, J.-D. (1998) Survival of adult Great Tits (*Parus major*) in relation to sex and habitat: a comparison of urban and rural populations, *Ibis*, 140, pp. 205-209.
- HURLBERT, S. H. (1984) Pseudoreplication and the design of ecological field experiments, *Ecological Monographs*, 54, pp. 187-211.
- LEBRETON, J.-D., BURNHAM, K. P., CLOBERT, J. & ANDERSON, D. R. (1992) Modelling survival and testing biological hypotheses using marked animals. A unified approach with case studies, *Ecological Monographs*, 62, pp. 67-118.
- NICHOLS, J. D., STOKES, S. L., HINES, J. E., & CONROY, M. E. (1982) Additional comments on the assumptions of homogeneous survival rates in modern bird banding estimation models, *Journal of Wildlife Management*, 46, pp. 953-960.
- PIPER, S. E., BOSHOF, A. F. & SCOTT, H. A. (1999) Modelling survival rates in the Cape Griffon (*Gyps coprotheres*) with emphasis on the effect of supplementary feeding, *Bird Study*, 46 (suppl), pp. S230-S238.
- POLLOCK, K. H. & RAVELING, D. G. (1982) Assumptions of modern band-recovery models with emphasis on heterogeneous survival rates, *Journal of Wildlife Management*, 46, pp. 88-98.
- SKALSKI, J. R. & ROBSON, D. S. (1992) *Techniques for Wildlife Investigations* (New York, Academic Press).
- SAMUEL, M. D., TAKEKAWA, J. Y., BARANYUK, V. V. & ORTHMEYER, D. L. (1999) Effects of avian cholera on survival of Lesser Snow Geese (*Anser caerulescens*): an experimental approach, *Bird Study*, 46 (suppl), pp. S239-S247.
- SCHMUTZ, J. A., WARD, D. H., SEDINGER, J. S. & REXSTAD, E. A. (1995) Survival estimation and the effects of dependency among animals, *Journal of Applied Statistics*, 22, pp. 673-681.
- STEWART-OATEN, A. & MURDOCH, W. M. (1986) Environmental impact assessment: 'pseudoreplication' in time? *Ecology*, 67, pp. 929-940.
- UNDERWOOD, A. J. (1993) Things environmental scientists (and statisticians) need to know to receive and (and give) better statistical advice. In: D. J. FLETCHER & B. F. J. MANLY (Eds) *Statistics in Ecology and Environmental Monitoring* (Dunedin, New Zealand, Otago Conference Series), 2, pp. 33-61
- UNDERWOOD, A. J. (1994) On beyond BACI: sampling designs that might reliably detect environmental disturbances, *Ecological Applications*, 4, pp. 3-15.
- VOUNATSOU, P. & SMITH, A. F. M. (1995) Bayesian analysis of ring-recovery data via Markov chain Monte Carlo simulation, *Biometrics*, 51, pp. 687-708.